

Pragmatic Reasoning and the Evolution of Adjectival Monotonicity: an Experimental Approach

Jaap Kruijt, 5515084

Supervisors: Rick Nouwen, Yoad Winter (Utrecht University)
& Marieke Schouwstra (University of Edinburgh)



Utrecht University

Final Thesis Project
Master Artificial Intelligence
Utrecht University
31 January 2020

Abstract

How do humans learn to categorise concepts so quickly? Finding out which factors influence human learning can help bring AI and human learning closer together. Research has suggested the biases for simplicity and informativeness play an important role in how categories are learned. This in turn influences the evolution of these categories over generations, which leads to more efficient and learnable categories. However, previous studies have also shown the importance of pragmatic reasoning, in particular leading to monotonic categorisations of scalar adjectives. Monotonic scalar adjectives are widespread in natural language. Therefore, pragmatic reasoning should be implemented in AI, to make AI reasoning more like human reasoning and aid human-AI communication. In this thesis, I aim to test the importance of pragmatic reasoning in human category learning. In a (human) behavioural Iterated Learning experiment, I explore which factors and biases are most important in learning categories, and how they influence the evolution of categorisations. The results show a tendency towards monotonic categorisations of scalar adjectives, but a bias for informativeness was not properly induced in the experiment and participants behaved in unexpected ways. The results are statistically inconclusive, and cannot be used to back up my statement that pragmatic reasoning should be implemented in AI. However, there are some interesting findings which ask to be explored in more detail.

1 Introduction

1.1 Learning categorisations

Humans are capable of quickly categorising objects and concepts (Ashby & Maddox, 2005). This helps in making the world around them more understandable and less fuzzy. For instance, humans often do not perceive differences between sounds and concepts gradually, but rather direct: there is a strong boundary where humans stop perceiving an object as orange, and instead perceive it as red. Similarly, there is a boundary between objects perceived as ‘small’ and ‘big’, although this depends on context. Humans are able to quickly learn these categories and the boundaries between them (Eimas, Siqueland, Jusczyk & Vigorito, 1971). How humans learn to categorise concepts so quickly is an interesting question in Artificial Intelligence research. Specifically, what are the most important biases for learning a categorisation? These biases influence the ways in which humans learn. Because language is shaped by the continuous process of humans learning and using it, these biases also influence the evolution of language itself. Artificial Intelligence agents are generally good at creating their own biases, for instance through deep learning. However, in this case, we generally have little influence on what these biases are. If we want to create Artificial General Intelligence agents who reason as humans do, and whose language is shaped like human language, their biases should be based on the learning biases in humans.

In the study of language evolution, two important learning biases in human language can be distinguished: *simplicity* and *informativeness* (Regier, Kemp & Kay, 2015). Naturally, a language that is simpler is easier to learn. Fewer words or less complicated

phrases require less time and effort to remember. However, if a language becomes too simple, it loses its informativeness because words have too many meanings associated to them or because a phrase is too short to accurately convey a complex meaning. Experiments and models of language evolution have shown that language evolves due to the interplay of these opposite pressures for simplicity and informativeness. When both pressures are in place, structured languages emerge (Kirby, Cornish & Smith 2008; Kirby, Tamariz, Cornish & Smith 2014). However, structure is not the only way in which languages become informative and learnable. Another way in which this is achieved is through *pragmatic reasoning*: humans are able to use their knowledge about the world and about the context of the conversation to infer speaker meanings that are not mentioned explicitly. Grice (1975) proposed this behaviour arises due to the assumption that speaker utterances are as informative and relevant to the conversation as possible. If something is left unmentioned or said in a roundabout way, the speaker must therefore additionally want to convey something else, which is not said explicitly. This type of inference is called an *implicature*. The semantics of such an utterance is *underspecified*, and pragmatic reasoning is used to fill in the gaps. Letting a hearer infer something implicit instead of mentioning it explicitly costs less energy, and so language becomes more efficient and learnable without losing informativeness.

One instance in which implicatures arise, is in scalar adjectives such as ‘big’ and ‘strong’. ‘Big’, ‘huge’ and ‘small’ are all adjectives describing the size of an object. Though there is no overlap in meaning between the adjectives ‘small’ and ‘big’, there is overlap in the meanings of ‘big’ and ‘huge’, with ‘big’ completely covering the meaning of ‘huge’. Although an object that is ‘big’ need not necessarily be ‘huge’, an object that is ‘huge’ is in fact always ‘big’. In contrast, an object cannot at the same time be ‘big’ and ‘small’ or ‘huge’ and ‘small’. Scalar adjectives like ‘big’, ‘small’ and ‘huge’ are all known as *monotonic*:

Definition 1. Monotonicity

An adjective P is monotonic iff $x \prec_1 y$ then $P(x) \prec_2 P(y)$ (with \prec_1 and \prec_2 being relevant relations between x and y and $P(x)$ and $P(y)$, respectively)

In words, this means that applying P to arguments x and y , which are part of a certain structure \prec_1 , will lead to a similar structure \prec_2 in the objects which emerge from this application. For scalar adjectives in particular this means that an adjective is monotonic if from adjective $P(x)$ it follows directly that adjective $P(y)$ holds whenever $x \prec y$ (with \prec being an ordering between x and y , for instance an ordering in size). Concretely, if object A is bigger than object B (ordering $B \prec_1 A$) and object B is ‘big’ (applying B to adjective P meaning ‘big’, i.e. $P(B)$), then it immediately follows that object A is also ‘big’ (ordering $P(B) \prec_2 P(A)$). There is no point where an object stops being ‘big’ when growing in size, which means that the adjective does not have an *upper bound*. It does have a *lower bound*, where the object stops being ‘big’ and becomes ‘small’ when decreasing in size, although this lower bound is arbitrary and very vague. Almost all scalar adjectives in natural languages are monotonic: the same reasoning above can be repeated with ‘big’ replaced with ‘small’ or ‘huge’. There

does not exist an adjective for size ‘nuge’ meaning *big but not huge*¹. This adjective would have both an upper *and* a lower bound, and would be called *non-monotonic*. Monotonicity is not only found in scalar adjectives in language, but can also be found in, for instance, generalised quantifiers such as ‘some’ and ‘all’. These quantifiers overlap in meaning, and ‘some’ could in theory be used to describe ‘all’. This monotonicity is so widespread in natural language that some research suggests it can be considered a *semantic universal*: a property that all human languages share (Steinert-Threlkeld & Szymanik, 2019). Interestingly, then, humans choose not to make a strong distinction in meaning between ‘big’ and ‘huge’, and place no strong boundary between the two meanings. Instead, the semantic difference between ‘big’ and ‘huge’ is underspecified.

One may object that we use ‘big’ and ‘huge’ as if they have distinct meanings. That is, an object that is ‘big’ is generally regarded to be smaller in size than an object that is ‘huge’. One might be able to say: ‘this house is not *big*, it is *huge*’. However, this use of the adjectives ‘big’ and ‘huge’ is not hard-wired into their meaning. Rather, the way in which we use these adjectives is a result of pragmatic reasoning. Since it would be more informative to say that an object is ‘huge’, if a speaker instead says that the object is ‘big’, listeners infer that the object is not ‘huge’. This is a kind of implicature which is known as *scalar implicature* (Grice, 1975). Even though it might seem like ‘big’ has both a lower and an upper bound, this is in fact not the case. In its literal meaning, ‘big’ can mean everything from a certain size and larger. In its non-literal, pragmatically induced meaning, ‘big’ stops being an adequate adjective to describe very large objects at a certain point, and ‘huge’ becomes the preferred adjective.

Even though this is an adequate description of the way humans categorise scalar adjectives, this is no direct evidence that humans indeed categorise in such a way. After all, it would be more informative to have a literal meaning that includes both an upper and lower bound for ‘big’. Researching human behaviour in producing and learning categorisations is difficult, since the pragmatic reasoning humans use in everyday speech is also employed by human participants in an experiment. This pragmatic reasoning complicates interpretation of the results in behavioural experiments where we want to find out whether pragmatic reasoning leads to monotonic categorisations. Pragmatic reasoning can be reduced in behavioural experiments by increasing cognitive load on participants (de Neys & Schaeken, 2007) or by using certain types of questions or instructions (Bott & Noveck, 2004) to reveal a more literal reading. This gives us insights into whether and how pragmatic reasoning influences the interpretation of monotonic adjectives. However, it does not provide insight into why these monotonic adjectives are more abundant than non-monotonic adjectives, and whether pragmatic reasoning is the cause of this abundance of monotonic adjectives. Therefore, a good way to investigate why scalar adjectives are monotonic is by looking at how they evolved. How could language have evolved in such a way that the literal meanings of scalar adjectives such as ‘big’ and ‘huge’ overlap? Research in language evolution suggests that this happened

¹There do exist a few non-monotonic adjectives in natural language. For example, ‘tepid’ means ‘warm but not hot’, or ‘somewhere between cold and warm’. However, adjectives such as these are very rare.

precisely because humans are capable of reasoning pragmatically, in combination with competing cognitive biases towards simplicity and informativeness. This research will be described in the following section.

1.2 Evidence from model simulations

Language evolution can be studied experimentally, using human participants or computer models to simulate language evolution as it would have taken place over the course of a hundred thousand years. Various models have shown that adjectival monotonicity evolved in languages due to an ability to reason pragmatically. Brochhagen, Franke & van Rooij (2016) devised a computational Bayesian learning model based on game theory which simulates the interaction between a speaker and a listener learning a language over generations. In this model, agents could choose from a set of 6 languages, each consisting of two meanings. These meanings could have either an upper bound or no upper bound. In this way, some of the languages modelled categorisations consisting of non-monotonic scalar adjective categories, while others modelled monotonic scalar adjective categorisations. Their model shows that languages that do not encode an upper bound on scalar adjectives are preferred by learners in case there is a bias for simplicity. However, languages made up of these monotonic scalar adjectives are only preferred by the model when the speaker and listener use pragmatic reasoning. Without pragmatic reasoning, learners prefer languages that encode a clear upper bound on scalar adjectives.

Carcassi, Schouwstra & Kirby (2019) designed three computational Iterated Learning models that show how biases for simplicity, informativeness and pragmatic reasoning influence the evolution of monotonicity in scalar adjectives. Similar to Brochhagen *et al.* (2016), two agents interact to convey a message. However, in the models by Carcassi *et al.* (2019), the languages consisted of categorisations of three meanings, corresponding to three degrees of a certain scale (such as size). In their models, the biases for simplicity and informativeness and pragmatic reasoning were treated as three separate parameters which were included or excluded in each model, leading to different results for each of the three models. The three models will be explained in more detail in Section 2.1 below. In contrast to the model results by Brochhagen *et al.* (2016) which suggested that a bias for simplicity and pragmatic reasoning are sufficient to create a preference for monotonic categories, the models by Carcassi *et al.* (2019) showed that *both* the biases for simplicity and informativeness and pragmatic reasoning are necessary for monotonic categories to evolve. These three factors all need to be present: without pragmatic reasoning, model agents reasoned literally and thus preferred non-monotonic categories due to their added informativeness, because they were not able to distinguish between two categories with overlapping meanings.

1.3 Importance for Artificial Intelligence

These findings have important implications for AI models. Pragmatic reasoning and pragmatic logics are valuable to commonsense reasoning (Bell, 1991; Bell, 1999). Commonsense reasoning is a type of reasoning which uses context from world knowledge and

is based on the assumption that a change in one factor does not mean the world around it changes together with that factor, but instead stays constant. Without commonsense reasoning, an AI agent might have false beliefs about the world. A few authors argue for more commonsense and pragmatic reasoning in AI (Bell, 1999; Davis & Marcus, 2015), but it seems AI research in general pays little attention to commonsense and pragmatic reasoning, and the research in this area has stalled in recent years. However, I argue that pragmatic reasoning is essential to AI research, as without it, AI agents cannot reason the way humans do and will therefore have false beliefs about the world. The findings by Carcassi *et al.* (2019) provide an example of how an AI agent could have such false beliefs. They show that when agents interpret language literally, nonmonotonic categories, which include both an upper and a lower bound in their literal meaning, are more informative to such an agent. An AI agent learning a language without supervision might therefore favour nonmonotonic categories and develop a language which is unlike human language. Monotonic categories can be used in a wider range of circumstances and contexts than non-monotonic categories, due to the underspecified semantics of monotonic categories. An AI agent developing a categorisation on its own might arrive at a categorisation which is similar to that of human languages, but it could have arrived there through a different interpretation of the meanings of the categories. For instance, instead of monotonic categories, the categorisation could consist of non-monotonic categories. Therefore, when a human and an AI agent converse about the same category, but the human uses the monotonic form and the agent uses the non-monotonic form, this will lead to misinterpretations of each others' utterances. Specifically, since the monotonic form of a category has a less specific and broader meaning than the non-monotonic form, the human is better capable of understanding the agent than the other way round. Having AI agents reason the way humans do helps bridge the gap between AI and humans, aids communication between AI and humans and helps bring us closer to Artificial General Intelligence which accurately simulates human behaviour. Simulating human-like reasoning in AI means incorporating pragmatic reasoning, which is a key part of human reasoning. The preference for monotonic categorisations discussed by Carcassi *et al.* (2019) is just one example of how humans act pragmatically when using language. There are many other uses of pragmatic reasoning in human language, which all make language more efficient. Thus, to make language in AI more human-like, we must incorporate pragmatic reasoning.

Sections 1.1 and 1.2 above show there is linguistic evidence for the existence of monotonic categories. Concepts related to monotonicity in generalised quantifiers and adjectives are considered universals (Von Stechow & Matthewson, 2008). In the past years, many experiments have been performed which explore the importance of monotonicity in the human interpretations of utterances. Chemla, Buccola & Dautriche (2019) found that a more general version of monotonicity, *connectedness*, could be found in natural languages and helped bring the interpretations of content and logical words closer together. Steinert-Threlkeld & Szymanik (2019) developed a computational learning model which showed how semantic universals arise and which suggests that monotonicity is a semantic universal. There is thus a broad range of research which shows that mono-

tonicity exists, but we know less about how this phenomenon arose. Specifically, this research does not address what role pragmatic reasoning, together with biases towards simplicity and informativeness, could have played in the emergence of monotonicity in natural language. However, the model simulations from Carcassi *et al.* (2019) suggest these factors are the key to monotonicity in language. Therefore, it is useful to explore how these model simulations relate to similar simulations with human participants. Do humans indeed categorise in the way predicted by Carcassi *et al.*'s (2019) models? If so, is such a categorisation preferred because of competing biases for simplicity and informativity and the ability to reason pragmatically? We should find answers to these questions before implementing pragmatic reasoning in AI. Therefore, in this thesis, I aim to determine what biases are most important in the categorisation of scalar adjectives in humans, and how these biases could lead to monotonic scalar adjectives.

2 Theoretical background

2.1 Experiments by Carcassi *et al.* (2019) and Carr *et al.* (2018)

The experiment by Carcassi *et al.* (2019) consisted of three computational Iterated Learning models. An Iterated Learning model is a probabilistic model which simulates language evolution through repeated learning and use of a language over generations (Kirby & Hurford, 2002). In real-life language evolution and language learning, it is impossible for humans to encounter all constructions and words possible in a language during the time they learn it. This is known as the *bottlenecking effect* (Kirby & Hurford, 2002). To simulate this effect in Iterated Learning experiments, agents or participants are only taught a subset of the complete language. After learning a subset of the language from the previous generation, the agents or participants then reproduce that language. Due to the bottlenecking effect, the language is learned imperfectly, or assumptions are made that may not reflect the language in the previous generation. The language therefore changes over generations, as natural language does. The model or experiment is run for a number of generations, or until the language converges. Iterated learning experiments and models are set up in such a way that the cognitive biases influencing language evolution become apparent. In general, the competing biases for simplicity and informativeness influence learning in such a way that the language becomes simpler and more structured over generations. Iterated learning models and experiments therefore aim to explain language evolution through the biases that are at play during repeated learning and use.

Carcassi *et al.* (2019) showed in a systematic way what biases need to be in place for monotonic scalar adjectives to evolve. They developed three models, each model building on the previous one by adding an extra bias. In all the models, languages encode a set of three degrees on a scale, modelling scalar adjectives. All possible combinations of three meanings are encoded. These meanings may cover a part of the scale, or the entire scale. The complexity of these languages is based on the amount of boundaries that a meaning has. For instance, a monotonic category such as ‘big’, which covers the top

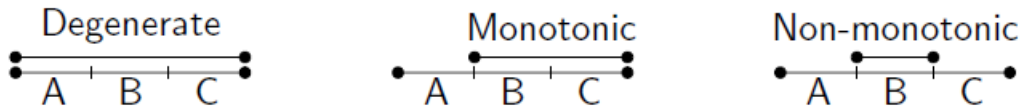


Figure 1: Degenerate, monotonic and non-monotonic meanings for scalar adjectives. ‘A’, ‘B’ and ‘C’ are the three degrees of the scale. Figure by Carcassi *et al.* (2019).

part of the scale, would have 1 boundary. A non-monotonic scalar adjective meaning ‘big but not huge’ which covers the middle section of the scale has 2 boundaries. If an adjective covers the entire scale, it has 0 boundaries. Such an adjective is called *degenerate*. It is not communicatively useful, since it makes no distinctions in meanings. However, a degenerate adjective has the least amount of boundaries, and thus the lowest complexity in Carcassi *et al.*’s (2019) model. The three types of adjectives considered in the model can be found in Figure 1. The evolution of languages is modelled as follows: Each model agent has access to the entire set of possible languages, out of which it chooses one language to use and teach to the next generation. The agent chooses the language which has the highest probability based on the language that their teacher used in the previous generation. The choice of language is also influenced by the agent’s own prior probability distribution over all possible languages, which is based on the language’s complexity. Agents favour less complex (i.e. simpler) languages, and thus these languages are assigned a greater prior probability. This way, the model simulates a bias for simplicity in languages.

In Carcassi *et al.*’s (2019) first model, only this bias for simplicity is modelled. A speaker/teacher sends a signal to a listener/learner together with its meaning. The listener/learner uses this signal to update his own probability distribution over possible languages. This updated probability distribution determines which language the listener will use himself. Languages that have a higher probability given the bias for simplicity will be used more frequently. Through this iterated learning process, one language will outperform the others in frequency of use by the agents. Carcassi *et al.* (2019) found that the bias for simplicity alone does not lead to languages encoding monotonic categories, but rather to degenerate languages. These languages are easy to learn, but they are not usable for communication, and thus do not portray natural languages accurately.

In the second model, Carcassi *et al.* (2019) added a communicative element to the model to include a bias for informativeness². This bias favours languages that are as informative and therefore communicatively useful as possible. Including this second bias did not lead to monotonic categorisations either. Rather, agents favoured languages with non-monotonic categorisations. Even though these are the most complex to learn, their informativeness outweighed the bias for simplicity. Since humans do not use non-monotonic categories, where the literal meaning of a scalar adjective has both an upper

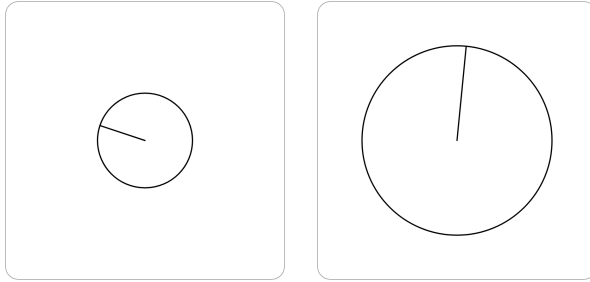


Figure 2: Examples of Shepard circles of varying size and angle.

and a lower bound, this model did not portray natural languages accurately either.

Thus far, the agents had been interpreting the message sent by a speaker literally rather than pragmatically. In such a case, there is no difference in meaning for an adjective such as ‘huge’ and an adjective such as ‘big’. Therefore, in the third model, Carcassi *et al.* (2019) included pragmatic reasoning in a way that is similar to the Rational Speech Act (RSA) model (Frank & Goodman, 2016). Now, because agents could distinguish between meanings such as ‘big’ and ‘huge’ using pragmatic reasoning, the informativeness of a nonmonotonic category no longer outweighed the simplicity of a monotonic category. Agents now favoured languages which encode monotonic categories. The models therefore show that monotonicity in scalar adjectives could have emerged due to the combined pressure for simplicity and informativeness, given that humans reason pragmatically.

Carr, Smith, Culbertson & Kirby (2018) also studied category learning and evolution in humans experimentally. However, the focus of their experiments was not specifically on scalar adjectives, but rather on categories in general. They devised models and performed experiments to test which learning bias is most important in learning categorisations. Similar to Carcassi *et al.*’s (2019) models, they distinguished between simplicity and informativeness as the two main competing biases, but they did not factor in pragmatic reasoning. Carr *et al.*’s (2018) research was twofold: they tested the success of reproducing a certain categorisation and let the categorisation evolve on its own. The first setup thus tested which types of categorisations are more learnable, while the second setup tested how this learnability influences the evolution of categorisations. In the first setup, participants and model agents learned different types of categorisations during a training round and reproduced them in a series of production tasks in the subsequent test round. The words used as category labels were one-syllable consonant-vowel-consonant nonce words such as ‘reb’, ‘wud’ or ‘zix’. The pictures used as meanings (hereafter called the meaning space) were Shepard circles (Shepard, 1964) of varying size and angle (Fig. 2). Some of the categorisations were ordered in an ‘informative’ way, where the categories were as compact as possible and each of the four categories was located in one

²Carcassi *et al.* (2019) use the term *communicativity* rather than *informativeness* for this bias. However, these terms are often used interchangeably in Iterated Learning research. Language is communicatively useful by virtue of its informativeness.

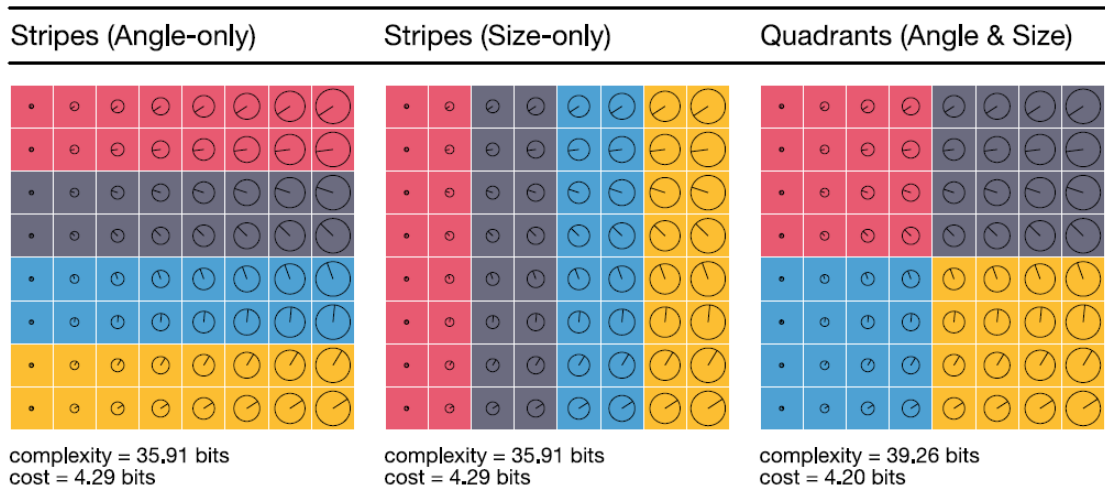


Figure 3: The three types of categorisations distinguished in Carr *et al.*'s (2019) first experiment. The two leftmost categorisations are simple categorisations (the first ordered along angle, the second ordered along size), while the rightmost categorisation is more informative. Images from Carr *et al.* (2019).

of the four corners of the meaning space. Other categorisations were ordered in a more 'simple' way, with the meaning space partitioned in four equal stripes. These simple categorisations could then be ordered horizontally, along the angle property of the Shepard circles, or vertically, along the size property of the Shepard circles. The simplicity and informativeness of each of these types of categorisations was calculated. Figure 3 shows the three types of categorisations used by Carr *et al.* (2019) and their complexity. The experiment and model showed that simpler categorisations were reproduced more successfully than more informative categorisations.

The second setup was an Iterated Learning experiment and model. Like in the previous setup, participants learned a categorisation and reproduced it. However, in this case, the reproduced form of the categorisation was used as input for the next participant. This categorisation constituted a miniature language. This way, the Iterated Learning experiment could be used to gain more insight into the evolution of language, in this case specifically the evolution of categorisations in language. Carr *et al.* (2018) simulated the bottlenecking effect by having participants learn only a subset of the entire meaning space. Because of this bottlenecking effect, participants were unable to reproduce the categorisation completely faithfully. The language therefore changed each generation. This created a chain of participants shaping a language through learning and use, each generation learning the language from the previous generation. The categorisations were initially completely random, with the 4 category labels distributed randomly across the meaning space. Such a categorisation is very inefficient and impossible to learn, and categorisations quickly became more organised over generations as participants tried to make sense of the categorisation. The chains of participants were run until two

subsequent generations used the exact same categorisation, thus converging on that categorisation. In some of the categorisations that were converged on, all meanings collapsed into one, with just one label remaining at the point of convergence. Most of the other chains converged on categorisations which were more similar to the simple categorisations of the first setup, than to the informative categorisations. The Iterated Learning experiments and models thus showed that in learning the categorisation from the previous generation, participants and model agents were biased towards learning simple categorisations. These experiments and models confirmed the findings from the first setup.

Carr *et al.*'s (2018) experiments thus provide empirical evidence for a learning bias towards simplicity, which we could implement in AI models. However, the results did not show a clear case for monotonic categorisations. Many categorisations that were converged on consisted of three categories, with one category covering only the middle of the meaning space. How could this be explained in light of the abundance of monotonic categorisations in human language?

First of all, we cannot know with absolute certainty that the categorisation is not monotonic. As explained in Section 1.1, some adjectives often *look* as though they have an upper bound (such as 'big'). At first sight it is not apparent that this upper bound is actually not present, but rather inferred through pragmatic reasoning. In the same way, the categorisations in Carr *et al.* (2018) do not look monotonic even though they might be to the participants that learned and used them.

Secondly, monotonicity is a phenomenon that is observed in *scalar* adjectives, such as 'big' and 'wide'. Adjectives that are not scalar, such as 'red', cannot be assigned (even arbitrary) upper and lower bounds, and there is no clear hierarchy between 'red' and 'blue', for instance. Therefore, 'red' and 'blue' cannot be described in terms of monotonicity. In Carr *et al.*'s (2018) Iterated Learning experiment, the pictures could be categorised according to their size or to their angle (see Fig. 2 and Fig. 3). While size is a scalar property, angle is not: though angles stand in a certain relation to each other, one cannot assign an ordering where one angle is 'higher' in rank than the other. An angle of 45° is not intrinsically smaller than an angle of 180° , as it may just as well be an angle of 225° . One could compare the angles in the pictures used by Carr *et al.* (2018) to a clock-hand, which does not have an ordering between the hours either: 1 o'clock could come either 'before' (a.m.) or 'after' (p.m.) 12 o'clock. Although both categorisations according to angle and according to size were present in the categorisations that were converged on, there was a clear preference for categorisations according to angle (Carr *et al.*, 2018), and so no conclusion can be drawn on the 'monotonicity' of these categorisations.

Lastly, the models by Carcassi *et al.* (2019) simulated a bias towards informativeness by having two model agents communicate about a statement with each other. Without this communicative element, only a bias towards simplicity is present. Previous Iterated Learning experiments with human participants (Kirby, Cornish & Smith (2008); Kirby, Tamariz, Cornish & Smith (2015)) have also shown that the evolution of a language is only influenced by a pressure for simplicity, unless participants are required to com-

municate with another participant in some way. The presence of this communication partner leads to a pressure for informativeness. In the Iterated Learning experiment by Carr *et al.* (2018), such a communication partner was not included in the design. Carr *et al.* (2018) acknowledge that this could partly explain the preference for simple categorisations over informative ones. Perhaps this could also explain why many categories, at least at first sight, did not evolve to be monotonic. As the experiments by Carcassi *et al.* (2019) showed, the pressures for simplicity and informativeness both need to be present, together with pragmatic reasoning, for monotonic categories to emerge.

The results from Carr *et al.*'s (2018) behavioural experiments thus do not contradict the findings from Carcassi *et al.* (2019), and it is still possible that monotonic categories are preferred when the categories are scalar, but the results from Carr *et al.* (2018) do not provide a conclusive answer on which biases could lead to monotonicity in scalar adjectives. The combined findings from Carr *et al.* (2018) and Carcassi *et al.* (2019) therefore provide a good starting point for further research into the evolution of scalar categories.

2.2 Research objective and hypothesis

In this research, I will investigate the evolution of scalar category systems and the biases that are at play in this evolution. Firstly, I aim to determine *whether* the biases for simplicity and informativeness and the ability to reason pragmatically influence the evolution of scalar category systems. Secondly, if so, I aim to discover *in what way* these biases influence the evolution of these category systems.

Based on the literature discussed in this section and section 1, my hypothesis is as follows: monotonic categories evolve if and only if the competing biases for simplicity and informativeness and the ability to reason pragmatically are present. All three factors have their own influence on the evolution of categories, and when all three are combined, monotonic categories arise in this process of evolution. The pressure for simplicity on its own favours less categories, and this leads to a decrease in informativeness. The pressure for informativeness on its own favours more categories, which leads to a decrease in simplicity. Combined, one of the two drives will always be stronger, and this will lead to either *degenerate* categorisations, which are not communicatively useful, or non-monotonic categorisations. Pragmatic reasoning adds the ability to distinguish between meanings without having to specify these meanings explicitly, and this means monotonic categories can be used in the same way non-monotonic categories can. Therefore, this paves the way for monotonic categories. In other words, pragmatic reasoning is essential for the evolution of monotonic categories, but the pressures for simplicity and informativeness need to be present as well.

3 Method

3.1 Inducing biases for simplicity, informativeness and pragmatic reasoning

I will perform a (human) behavioural Iterated Learning experiment in which they learn and reproduce a categorisation. This categorisation constitutes a miniature ‘language’. The methodology for my experiment will be based on Carr *et al.*'s (2018) Iterated Learning experiment with human participants, which was made up of a training round and a test round consisting of a series of production tasks. Responses for the production tasks are used as input for the language in the next generation. Carr *et al.*'s (2018) experiment provides evidence bias for simplicity in learning categorisations. However, based on the findings by Carcassi *et al.* (2019), an additional bias for informativeness is needed in order for categorisations to evolve towards monotonicity. This could be done by having two participants communicate about a statement, but this is a costly task. Therefore, in my experiment, the bias towards informativeness is introduced in the form of an additional *acceptability judgment task*. Recall that certain types of questions reduce pragmatic reasoning in humans and lead to more literal responses (Bott & Noveck, 2004). In an acceptability judgment task, participants are asked to judge the acceptability of a statement with a yes/no question such as ‘Is this object big?’. Ellis (2004) states that acceptability judgment tasks tap into more explicit knowledge. This leads to responses that are based more on semantic, literal knowledge than on pragmatic reasoning. For instance, participants are more likely to accept ‘big’ as an alternative for ‘huge’ in such a task. In combining a production task relying on pragmatic reasoning and an acceptability judgment task relying more on literal meaning, participants are required to provide more information about their interpretation of the meaning of a category. This way, the acceptability judgment task induces a bias for informativeness. Furthermore, the acceptability judgment task could solve the problem in interpreting whether a category is monotonic or non-monotonic, which I mentioned regarding Carr *et al.*'s (2018) experiment in section 2.1. Because the task shows the semantic meaning of a statement or word rather than the pragmatic meaning, we gain insight into this semantic meaning, which will overlap for two monotonic categories, but not for two non-monotonic categories. As with the production task, the responses for the acceptability judgment task are used as input for the language in the next generation. An acceptability judgment task has not been included in an Iterated Learning experiment before, and this is a novel approach which I am introducing in this experiment.

The main type of task in my experiment is the production task, in which participants reproduce the labels for pictures which they learned during the training round. In these tasks, which do not increase cognitive load for the participants and are not formulated in a way which reduced pragmatic reasoning, I assume that participants reason as they do in natural conversations, that is, pragmatically. Since my experiment is based on Carr *et al.*'s (2018) experiment, which showed a preference for simple categorisations, I assume the pressure for simplicity is also present in my experiment. The only pressure which is induced due to an additional type of task is the pressure for informativeness. This means I can also compare the evolution of categorisations with and without a

pressure for informativeness by leaving out the acceptability judgment task which induces this pressure. Therefore, in my experiment, I introduce two conditions. Participants are assigned to one of two conditions. In one of these conditions, the results for the acceptability judgment task will be used as input for the language in the next generation. In the other condition, these results will simply be noted, but they will not influence what the language looks like in the next generation. This way, in the latter condition the extra information that is gained from acceptability judgment task is lost, as is the information about the overlap in semantic meaning between two monotonic categories.

3.2 Participants

The participants are selected through Amazon Mechanical Turk. All participants are located in the United States and were native speakers of English. Participants are divided into 2 conditions, which are again divided into 4 chains. Each chain consists of 5 generations. A total of 46 participants performed the experiment. 5 participants scored too low on the training items (with a performance accuracy below 90%) and were excluded from the experiment. One participant performed a wrong version of the experiment due to a mistake by me. In total, 6 experiments were redone.

3.3 Stimuli

The stimuli used in this experiment are the Shepard circles that Carr *et al.* (2018) used in their experiments. These Shepard circles have been tested on their saliency for use in experiments for both the angle and the size property (Canini, Griffiths, Vanpaemel & Kalish, 2014). The meaning space consists of a 6×6 grid with the Shepard circles lined up along the x -axis in increasing size. The sizes range from 50 pixels to 175 pixels in radius, incrementing in steps of 25 pixels. The angle of the circles is distributed in order along the y -axis. The angle ranges from 3.0144 radians to 5.2583 radians, increasing in steps of 0.4488 radians. The grid is used to design the categories and track their evolution. Participants do not see this meaning space in its entirety: instead, pictures and their label are selected from this meaning space using a selection procedure described in Section 3.3.2.1. The angle of the shepard circles is used as a second property both to increase the total amount of pictures and to ensure that participants do not see the same picture for each category each time. The entire meaning space is shown in Figure 4.

From Carr *et al.*'s (2018) findings, it seems that a categorisation according to angle is easier to learn than a categorisation according to size. Because such a categorisation cannot be described in terms of monotonicity, this does not give us any insight into the evolution of scalar adjectives towards monotonicity. Although the methodology by Carr *et al.* (2018) is useful for exploring the evolution of categorisations, it needs to be changed in order to keep the focus of the experiment on scalar adjectives. I aim to achieve this by introducing a categorisation at the start of the experiment which nudges the evolution in the direction of categorisations based on the scalar property of size instead of the non-scalar property of angle. In most Iterated Learning experiments, the




































 <i>Pov</i>	 <i>Pov</i>	 <i>Reb, Zix</i>	 <i>Reb, Zix</i>	 <i>Wud, Zix</i>	 <i>Wud, Zix</i>
 <i>Pov</i>	 <i>Pov</i>	 <i>Reb, Zix</i>	 <i>Reb, Zix</i>	 <i>Wud, Zix</i>	 <i>Wud, Zix</i>
 <i>Pov</i>	 <i>Pov</i>	 <i>Reb, Zix</i>	 <i>Reb, Zix</i>	 <i>Wud, Zix</i>	 <i>Wud, Zix</i>
 <i>Pov</i>	 <i>Pov</i>	 <i>Reb, Zix</i>	 <i>Reb, Zix</i>	 <i>Wud, Zix</i>	 <i>Wud, Zix</i>
 <i>Pov</i>	 <i>Pov</i>	 <i>Reb, Zix</i>	 <i>Reb, Zix</i>	 <i>Wud, Zix</i>	 <i>Wud, Zix</i>
 <i>Pov</i>	 <i>Pov</i>	 <i>Reb, Zix</i>	 <i>Reb, Zix</i>	 <i>Wud, Zix</i>	 <i>Wud, Zix</i>

Figure 4: The complete meaning space with an example categorisation consisting of the categories ‘pov’, ‘reb’, ‘wud’ and ‘zix’. Shaded cells indicate pictures which are used in the acceptability judgment task.

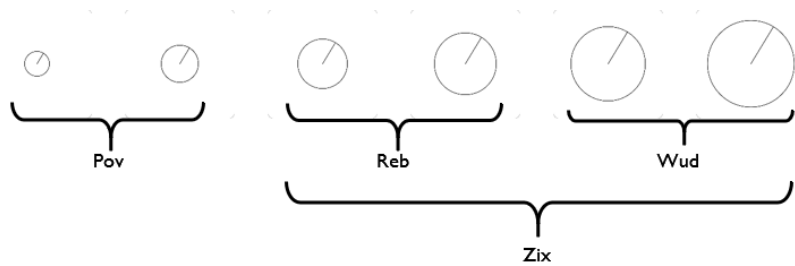


Figure 5: An example categorisation of Shepard circles. ‘Zix’ is the *super category*, ‘reb’ is the *non-monotonic category*. ‘Pov’ and ‘wud’ are both *monotonic ‘base’ categories*.

languages start out randomly, without any structure (e.g. Kirby *et al.* (2014)), as did the categorisations in the experiment by Carr *et al.* (2018). Starting from a categorisation which already contains some structure therefore is a fairly novel approach. One could say that the experiment is started as if a few generations have already passed, so that some structure is already present.

The experiment will start from a categorisation made up of four categories as follows: The meaning space is divided into 3 equal parts according to the size of the Shepard circles (shown along the x -axis). These three ‘base’ categories are all assigned a label. The fourth category is a *super category* that spans the top two base categories. The three base categories all consist of 12 pairs of labels and pictures, and the super category of 24 pairs, giving a total of 60 pairs of labels and pictures. These categories and their places in the meaning space can be found in Figure 4. Two of the base categories and the super category are monotonic, corresponding roughly to ‘small’, ‘huge’ and ‘big’. The middle category, which corresponds roughly to ‘big but not huge’, is non-monotonic. For clarity, the categories from Figure 4 are shown without the entire meaning space in Figure 5. We can view the categorisation of the meaning space in Figure 4 as a language which contains only meanings which are in this meaning space, and we can view the evolution of this categorisation as the evolution of a language.

I use a total of 16 nonce words to use as category labels. The labels are three-letter consonant-vowel-consonant words that have a different first and last letter and are adopted from Carr *et al.* (2018) with their permission. Carr *et al.* (2018) took great care in ensuring the nonce words did not exist in as many languages as possible. The words are grouped together in groups of 4 to be used as category labels in a chain. The groups are as shown in the table below. For each group, the label from the first column is used for the smallest category, the label from the second column is used for the non-monotonic middle category, the label from the third column is used for the largest category and the label from the fourth column is used for the super category.

1	pov	reb	wud	zix
2	gex	juf	vib	wop
3	buv	jef	pid	zox
4	fod	jes	wix	zuv

3.4 Procedure

3.4.1 Training phase

Before the experiment starts, participants receive a written instruction and are presented a practice item of each of the types of tasks that are in the experiment. After this, the experiment begins. Participants first go through a training phase. In this training phase, participants learn a subset of the total set of labels to create a bottlenecking effect. The labels selected for testing are distributed pseudorandomly across the total set of labels using a method by Carr *et al.* (2018). The grid is divided into 9 equal

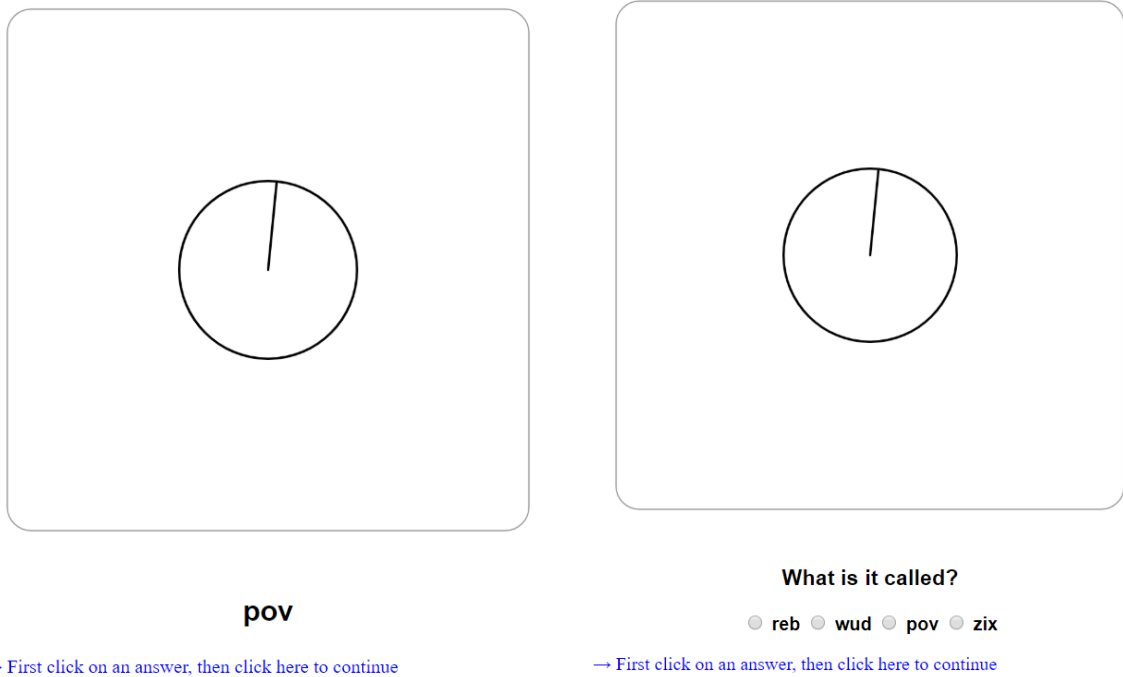


Figure 6: Example of a training item. The left item is shown for 3 seconds, then the right item is shown.

2×2 squares. From each square, two pictures are selected at random. For each picture that is selected, all possible labels are assigned to the picture. This means that a picture that has multiple labels assigned to it (such as the lower category and the super category label) will be presented multiple times, one time for each label. Due to this selection method, participants only encounter approximately half of the labels: because some pictures might have more labels assigned to them than others, the total amount of pictures presented fluctuates somewhat depending on which pictures are selected for training. Participants will see examples of the same category with different angles. First, the picture is shown on its own for half a second and then the label is shown together with the picture for three seconds. After each training item, the participants are tested on how well they pay attention to the training items by having them repeat the label that they just learned (Fig. 6). These responses are used to calculate performance accuracy during the experiment. Participants who reached an accuracy lower than 90% during the training phase were excluded from the experiment. The training phase consists of four training rounds to ensure that participants perform above chance.

3.4.2 Test phase

3.4.2.1 Condition 1: Transfer of acceptability responses

After the training phase, the test phase starts. During the test phase, no feedback is given after each trial so that participants do not change their hypothesis based on the feedback. In the test phase, participants are shown two types of trials. The first and main type of trial is a production trial. In the production trials, participants are asked to name all of the 36 pictures. Each trial is a multiple choice question where participants can choose out of all four labels. First, the picture is shown with the question ‘What is this called?’. The picture is placed in a greyscale frame with a fixed size to improve recognition. After a 1s delay, the four labels appear on the screen.

The second type of trial is an acceptability judgment task. There are 32 acceptability judgment tasks in which the participant is asked whether a given label can be used to describe a given picture. The pictures that are selected for the acceptability judgment task are situated on the diagonal from the bottom left to the top right of the meaning space. Furthermore, two extra pictures are selected from the two rightmost columns, as these are the relevant columns for the experiment (Fig. 4, shown shaded in grey). For each picture, participants are shown four acceptability judgment tasks; one for each label. First, a picture is shown with the question “Is this a ... ?”. The picture is placed in a greyscale frame. After a 1s delay, the yes/no response buttons appear.

The responses for the acceptability judgment task are used together with the responses for the production task as input for the next participant in the chain. First, a label is assigned to a picture based on the production task response for that picture. Next, additional labels are added if a participant accepted them as labels for that picture. For instance, if a participant chose ‘zix’ as the label for a picture during the production task, the picture is assigned the label ‘zix’. If, additionally, the participant accepted ‘wop’ and ‘reb’ as labels for that picture, these are added as well. Labels are only added, not removed. For instance, if a participant chose ‘zix’ as the label for a picture during the production task, but during the acceptability task only accepted ‘wop’ and ‘reb’ as labels, the latter two labels are added to the picture, but the label ‘zix’ is not removed. The additional labels can only be added to at most 8 of the 36 pictures, because for the other pictures no acceptability judgment task is presented.

3.4.2.2 Condition 2: No transfer of acceptability responses

The types of tasks in condition 2 are the same as in condition 1. To the participant, the experiments in each of the two conditions are exactly the same. However, for the second condition, the responses for the acceptability judgment task are not used to assign additional labels to a picture in the next generation. They are only used to gain insight in the participants’ hypothesis.

3.4.3 Transfer to next participant

When a participant has finished the experiment, their responses are used as input for the next participant in the chain. The 6×6 grid is labeled using the responses from

the production task. For condition 1, the responses for the acceptability judgment task are used to assign additional labels to a picture, but not for condition 2. If a category disappears because it was not selected during the production or acceptability judgment task, its label is removed from the production and acceptability judgment task in the next generation.

3.5 Predictions

Based on the findings from Carcassi *et al.* (2019), the following predictions can be made for my experiment. A pressure for simplicity would lead to a decrease in the total number of categories, since it is easier to learn less words. Therefore, if only a pressure for simplicity is present, all meanings will collapse into one, leading to a *degenerate* categorisation. However, a pressure for informativeness favours the four-way categorisation that is already present at the start of the experiment. This categorisation is the most informative and therefore communicatively useful, since it has all meanings a monotonic categorisation would have, but also contains an additional category which can be used to be more specific about a certain size. Finally, pragmatic reasoning renders the non-monotonic category in the four-way categorisation redundant. This is because the same meaning can be achieved through scalar implicature. Altogether, these biases will result in the disappearing of the non-monotonic middle category, while the other three categories are retained.

All of the three categories which are retained are monotonic. For the top and the bottom category, this should be clear: they only have one bound, respectively a lower and an upper bound. The super category is monotonic as well, although this may not become clear immediately. The acceptability judgment task provides the information that we need to show that this category is indeed monotonic in its literal interpretation.

In my experiment, I will also investigate the effect of removing the pressure for informativeness. I hypothesize that if the pressure for informativeness induced by the acceptability judgment task is removed, the non-monotonic category will disappear more quickly, since its informativeness is no longer considered valuable. This will also lead to more degenerate languages, because the pressure for simplicity becomes stronger without a competing pressure for informativeness. The lack of informativeness will also cause monotonicity to be lost, since participants no longer learn a categorisation in which the super category is interpreted as monotonic.

As mentioned in section 3.3, the angle is used as a second property to increase the total amount of pictures in my experiment. However, the focus of this experiment is on the property of size, since angle is not a scalar property. Potential categories of angle cannot be described in terms of monotonicity (as in Carr *et al.* (2018)). However, it may be possible that participants assign some ordering to the different angles, where for instance ‘11 o’clock’ is lower than ‘1 o’clock’. It may be possible that instead of (monotonic) categorisations of size emerging, languages evolve into categorisations of angle which are monotonic according to the interpretation above. However, I do not expect this to happen.

A clear indication of influence from biases for simplicity and informativeness in light

of pragmatic reasoning will back up the model predictions by Carcassi *et al.* (2018). This will in turn confirm the importance of modelling pragmatic reasoning in AI models to more faithfully model human reasoning and to make AI more human-like.

3.6 Calculation of complexity and monotonicity

The languages are measured for three values: complexity, number of non-degenerate categories and monotonicity. The complexity score is used to test for an effect of the pressure for simplicity. The pressure for simplicity alone leads to degenerate languages: languages in which only one category is present or in which all categories have the same meaning which covers the entire meaning space, and are thus not informative. Therefore, the number of non-degenerate categories is used to measure whether languages have turned degenerate or have still retained their informativeness. Finally, the complexity score is used in calculating the monotonicity score. This score indicates whether the categorisation which a language consists of is degenerate, monotonic or non-monotonic. A categorisation is degenerate if all its categories are degenerate. If a categorisation contains only monotonic categories, or only degenerate and monotonic categories but no non-monotonic categories, the categorisation as a whole is monotonic. If a categorisation contains at least one non-monotonic category, the categorisation as a whole is non-monotonic. In my analysis, (non-)monotonicity is thus a property of both an entire categorisation and a single category.

Following Carcassi *et al.* (2019), we use the number of boundaries to define complexity using the following calculation. Recall that monotonic categories have only one boundary, while non-monotonic categories have two (or more) boundaries. Degenerate categories cover the entire meaning space, and therefore have no boundaries. For each of the languages in each generation of my experiment, the number of boundaries for each category is counted, as well as the number of non-degenerate categories. These are counted separately in the direction of the properties of angle and size to explore the possibility that a categorisation might have switched from size to angle over generations. A boundary is defined as a transition from a size or angle where the category label is not used, to a size or angle where it is used (and vice versa). For instance, a monotonic category has one transition between the label being used and the label not being used, and thus has one boundary. Figure 7 shows an example distribution of categories for size for a language in a later stage of its evolution. As can be seen from this graph, ‘pov’ and ‘wud’ are used for sizes 125 pixels and under, and sizes 75 pixels and up, respectively. These two categories are monotonic, since they only have one boundary, ‘Reb’ is used for all sizes, and is therefore degenerate because it has no boundaries. ‘Zix’ is used for sizes 100 pixels and under and sizes 150 pixels and up. This category has two boundaries, and is therefore non-monotonic. Because this categorisation of size has one degenerate category (‘reb’), the total number of non-degenerate categories is 3.

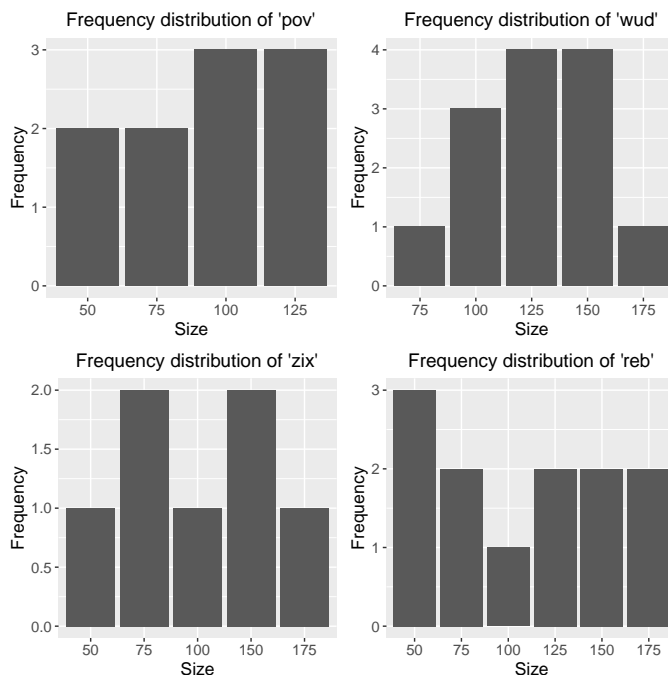


Figure 7: An example distribution showing the frequency distribution of category labels for size. ‘Pov’ and ‘wud’ are monotonic, ‘zix’ is non-monotonic and ‘reb’ is degenerate.

The complexity C_w of a category w is equal to the amount of boundaries it has. The sum of the individual complexities $\sum_w C_w$ is divided by the total number of non-degenerate categories N to obtain the average complexity C_L of a categorisation:

$$C_L = \frac{\sum_w C_w}{N} \quad (1)$$

The average complexity is used to define the monotonicity of a categorisation. If the total number of non-degenerate words is 0, then all categories are degenerate and the categorisation is degenerate as well, and the above equation is not used. If all words that are not degenerate have a complexity of 1 (i.e. have only one boundary), the average complexity of the categorisation will also be 1, and the categorisation will be monotonic. This is in line with the idea that a categorisation where one category is used for the entire meaning space and one category is used for the upper or lower part of that meaning space is still a monotonic categorisation. A complexity greater than 1 means that the categorisation is non-monotonic, because there is at least one non-monotonic category in the language. One non-monotonic category is enough to render the entire categorisation non-monotonic, even if the other categories are all monotonic. Through this calculation of monotonicity, we also obtain the correct classification of the starting categorisation. This categorisation contains one non-monotonic category and

three monotonic categories, and therefore has an average complexity of 1.25. This means that the categorisation is non-monotonic.

For the initial language, the complexity score and number of non-degenerate categories are thus as follows: The complexity of the categorisation of size starts at 1.25. The number of non-degenerate categories for the categorisation of size starts at 4, since the initial language contains 4 categories which are all not degenerate. In the initial language, there is no categorisation of angle. In the direction of angle, the complexity and number of non-degenerate categories are therefore both 0. All four categories in the initial language are degenerate in their meaning of angle. Since degenerate categories have a complexity of 0, the complexity of this ‘categorisation’ of angle is 0.

4 Results

4.1 Qualitative results

The evolution of the languages over 6 generations for all chains can be seen in Table 1, with the last column showing the language after the last generation. For each language in each generation, the size varies from left to right and the angle varies from top to bottom (as in Fig. 4). All languages started out as non-monotonic categorisations of size. Languages reach convergence if, by the last generation, the same language has been used two generations in a row. In my experiment, only one chain has converged on a language after the last generation. All other languages have not yet reached convergence, but do lean towards a certain type of categorisation. In condition 1, all languages lean towards a categorisation of angle. In 3 of these languages, all 4 categories are still in place, while in one language, one category has disappeared. In condition 2, one language converged on a categorisation of size, and one language did not converge on, but leans towards a categorisation of size. In the former, two categories have disappeared and a two-category system has emerged. In the latter, all 4 categories are still in place, although one category is only used for two meanings. In one language in condition 2, two categories are scattered across the meaning space and one category is used for only two pictures, with no clear indication of a categorisation of size nor angle. Finally, one language in condition 2 leans towards a categorisation of angle, with all 4 categories still in place.

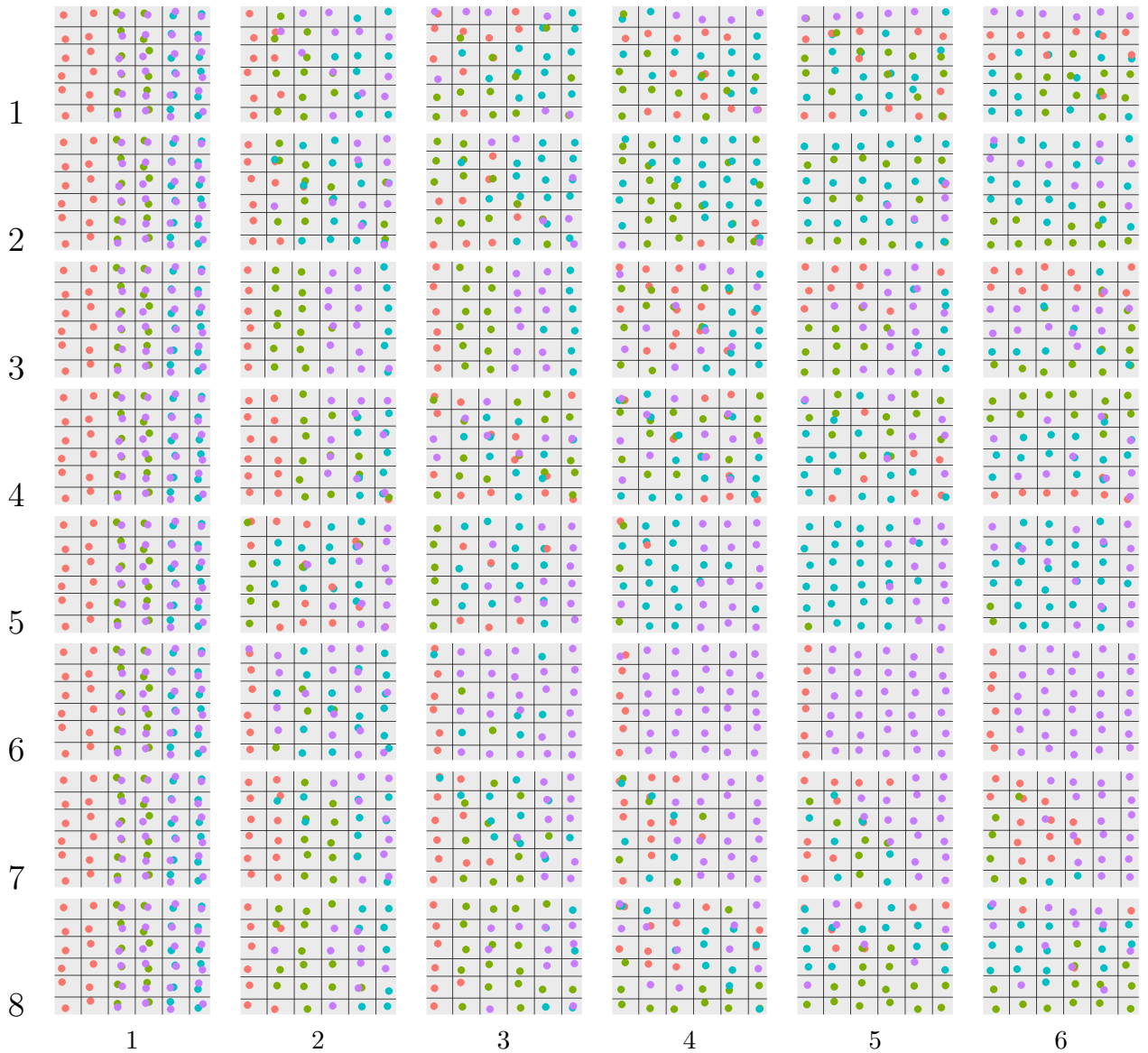


Table 1: Evolution of all 8 languages over 6 generations. The 6 generations are ordered left to right, with the rightmost image being the final language. Each colour corresponds to a category. Languages 1-4 are from condition 1, languages 5-8 are from condition 2. The design of this overview of the languages is based on Carr *et al.* (2018).

4.2 Quantitative results

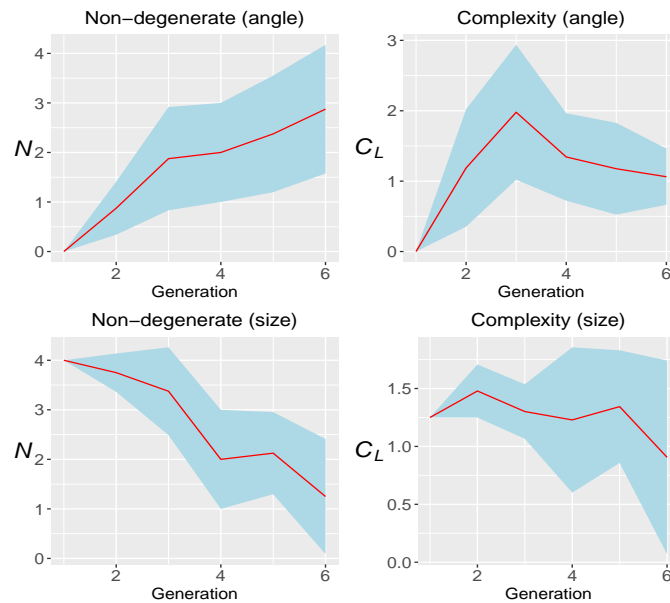


Figure 8: Experimental results for both conditions for average complexity of the size and angle categorisations and total number of non-degenerate words in the size and angle categorisations.

Figure 8 plots the evolution of the average complexity C_L and total number of non-degenerate words N over 6 generations. The red line shows the mean, and the blue shading shows the 95% confidence interval. For each language in each generation, both the categorisations of angle and size are plotted. Figures 9 and 10 plot the same evolution for each condition separately. The plots start at the values of the initial language which are mentioned in section 3.6. I performed a linear regression analysis to test for an effect of generation and condition, and for an interaction effect between generation and condition. For the categorisation of size, both complexity (estimate = $-.28$, $p = .067$) and number of non-degenerate words (estimate = -1.06 , $p < .001$) decrease over generations. However, for complexity, this result is not significant. There is a significant interaction effect of generation and condition for the number of non-degenerate words (estimate = $.34$, $p = .04$). The same interaction effect can be observed for complexity, but this is not significant (estimate = 0.14 , $p = 0.13$).

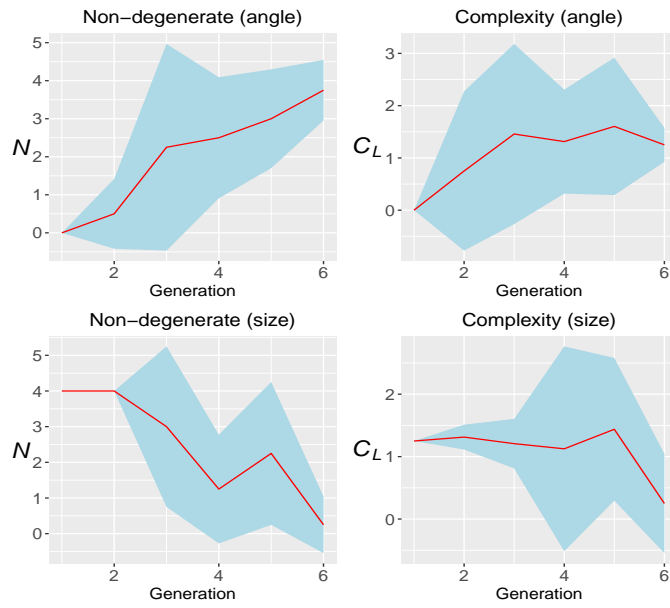


Figure 9: Experimental results for condition 1 for the average complexity of the size and angle categorisations and total number of non-degenerate words in the size and angle categorisations.

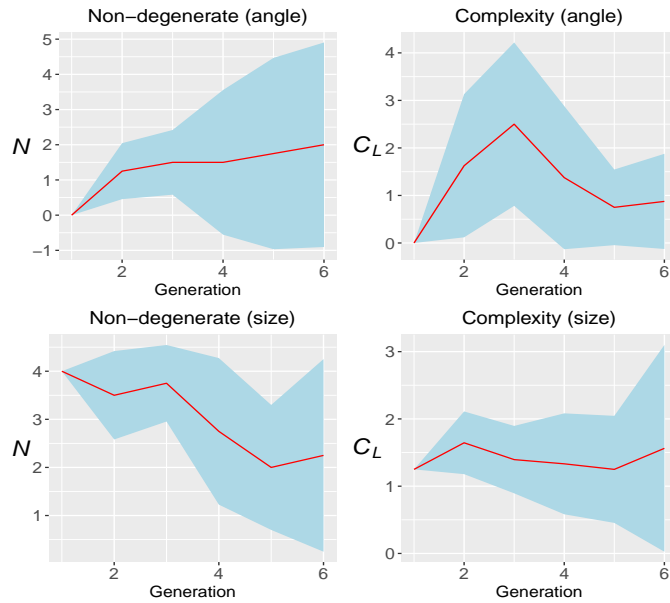


Figure 10: Experimental results for condition 2 for the average complexity of the size and angle categorisations and total number of non-degenerate words in the size and angle categorisations.

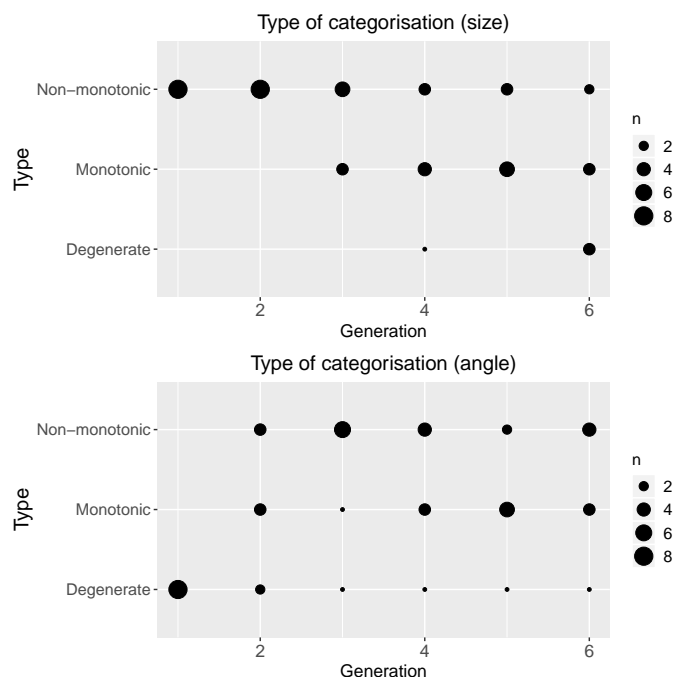


Figure 11: Experimental results for both conditions for the type of categorisation of the size and angle categorisations.

Figure 11 shows how the amount of different types of categorisation (non-monotonic, monotonic, degenerate) change over generations. Figures 12 and 13 show the same change for each of the two conditions separately. Here, again, (non)-monotonicity is used to describe the categorisation as a whole. A non-monotonic categorisation could thus in addition to non-monotonic categories also contain monotonic and degenerate categories. A multinomial logistic regression analysis was performed to test for an effect of generation and condition and an interaction effect between generation and condition on the monotonicity of a categorisation of size. Over generations, there is an increase in the amount of monotonic (estimate = 3.91, $p = .94$) and degenerate (estimate = 1.10, $p = .25$) categorisations of size in relation to the amount of non-monotonic categorisations of size. However, these effects are not significant. Furthermore, although these effects are not significant either, this increase is smaller for the second condition than for the first (degenerate: estimate = -0.78 , $p = .97$; monotonic: estimate = -0.14 , $p = .80$)

For the categorisation of angle, there is an increase over generations in the complexity (estimate = .48, $p = .06$) and number of non-degenerate words (estimate = 1.18, $p < .001$). For complexity, however, this effect is not significant. There is a significant interaction effect of generation and condition for the number of non-degenerate words (estimate = -0.43 , $p = .02$). Again, the same effect is present but not significant for complexity (estimate = -0.23 , $p = .15$). Though a categorisation of angle cannot be described in terms of monotonicity, it is insightful to compare the categorisations that emerge here to the categorisations of size. Therefore, the same multinomial logistic

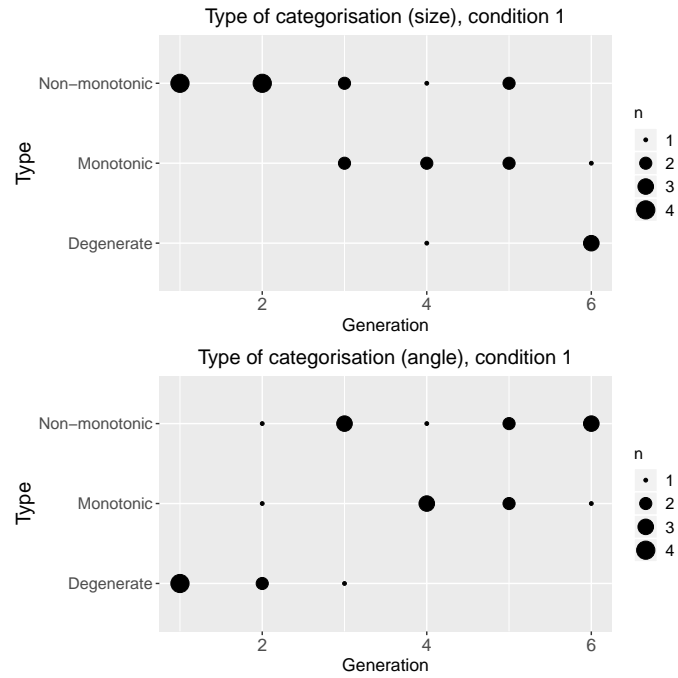


Figure 12: Experimental results for condition 1 for the type of categorisation of the size and angle categorisations.

regression analysis was performed to test for an effect of generation and condition and an interaction effect between generation and condition on the ‘monotonicity’ of a categorisation of angle. Over generations, there is an increase in the amount of monotonic (estimate = 3.96, $p = .06$) and non-monotonic (estimate = 4.28, $p = .04$) categorisations. For the monotonic categorisations, however, this effect is not significant. This increase in monotonic and non-monotonic categories is smaller for the second condition than for the first (non-monotonic: estimate = -2.00 , $p = .06$; monotonic: estimate = -1.69 , $p = .12$), although these effects are not significant.

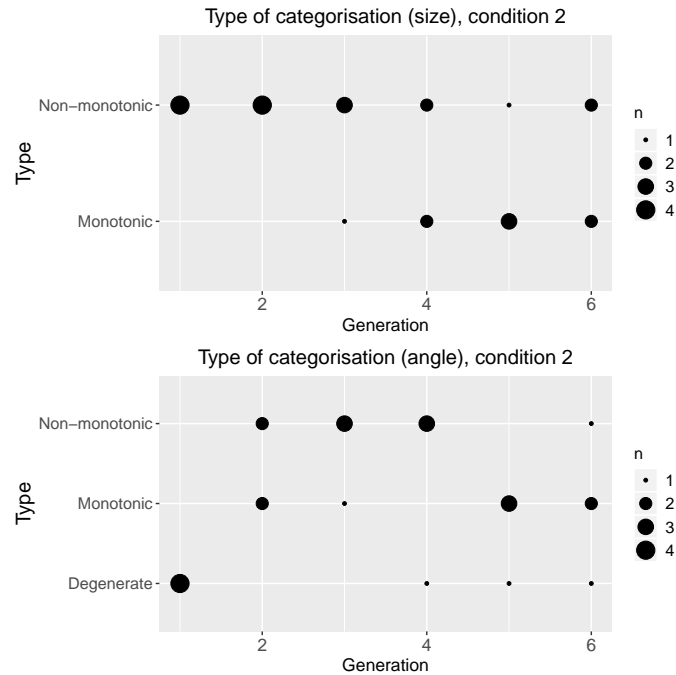


Figure 13: Experimental results for condition 2 for the type of categorisation of the size and angle categorisations.

5 Discussion

The aim of this study was to find out which biases and influences are present in the learning of scalar categorisations, and in what way these biases influence the evolution of scalar categorisations. To explore this, I simulated the evolution of a scalar category system in an Iterated Learning experiment. My hypothesis was that monotonic categories would emerge due to the combined pressures for simplicity and informativeness and pragmatic reasoning.

First I should note that according to my predictions, languages would either become degenerate altogether (i.e. collapse into one single category with one label remaining) or retain some form of categorisation according to size. I did not expect languages to be categorised according to angle instead of size. In fact, however, this is exactly what had happened by the last generation in most cases. For the remainder of this discussion, when I mention ‘degenerate categorisations’, this means that they are degenerate on the property I am discussing at that point. Languages that are categorised according to angle are not entirely degenerate, and all four labels might still be present, but the categorisation according to size, in which I am interested, is degenerate.

5.1 Differences between condition 1 and 2

For the categorisations of size, the linear regression analysis indeed shows an effect of the pressure for simplicity, as languages tend to become less complex over generations. However, this analysis also shows that over generations, the amount of non-degenerate categories decreases significantly. The multinomial logistic regression analysis furthermore shows that the increase in degenerate categorisations over generations is greater than the increase in monotonic categorisations. This means more and more categorisations become degenerate. The linear regression analysis also shows that the decrease in non-degenerate categories is significantly smaller for condition 2 than for condition 1. Although not significant, the analyses also show that the decrease in complexity, and the increase in monotonic and degenerate categorisations are smaller for condition 2 than for condition 1. The difference in increase of these categorisations between condition 1 and 2 is the largest for the degenerate categories. Combined with the smaller decrease in complexity and non-degenerate words for condition 2, these results show that less categorisations of size become degenerate in condition 2. The plots for types of categorisations for condition 1 and 2 in Figures 12 and 13 indeed show that this is the case. In fact, no categorisations of size become degenerate in condition 2, but are rather divided evenly into monotonic and non-monotonic categorisations in the last generation. The plot for condition 1 on the other hand shows a clear dip towards degenerate categorisations in the last generation.

The results for the individual languages shown in Table 1 also show the differences in the amount of degenerate categorisations of size for condition 1 and 2. Whereas in condition 1 all labels can be used to describe each size in almost all languages (i.e. all languages lean towards a categorisation of angle rather than size), in condition 2 there are two languages which lean towards or converged on a categorisation of size. Table 1 furthermore shows that these categorisations are monotonic. One of these languages (language 6 in Table 1) has a two-way categorisation, where the label ‘gex’ is used for the smallest size and the label ‘wud’ is used for all other sizes. This is an extreme case of monotonicity, which one could compare to the adjectives ‘closed’ and ‘open’ respectively. The other language categorised according to size (language 7 in Table 1) showed striking similarities to the hypothesised categorisation consisting of three monotonic categories with one super category. In this language, the label ‘zox’ is used uniquely for the two largest sizes. The labels ‘jef’ and ‘buv’ are used interchangeably for the other sizes, with no apparent hierarchy in size between the two. There is some overlap, with the label ‘zox’ also being used sporadically for the two middle sizes, but overall the picture is quite clear. These categories all have one boundary, and are therefore all monotonic. One non-monotonic category is still present, but it is used for only two meanings and therefore insignificant compared to the other three categories. In this language, the non-monotonic middle category has thus virtually disappeared, and it is likely that it would have disappeared if the experiment was run for one more generation. Instead of the super category covering the largest four categories, it covers the smallest four categories. However, it cannot be called a super category in the way I designed it, as ‘jef’ and ‘buv’ both cover roughly the same part of the meaning space and therefore neither of them

covers the other.

The results are remarkable, given my hypothesis. I expected that the non-monotonic category would take longer to disappear in condition 1, where acceptability judgments are transferred. However, it disappeared more quickly in this condition, as did the three monotonic categories. Furthermore, although I expected no degenerate categorisations in this condition, all languages became degenerate in their categorisation of size. Conversely, in the condition where acceptability judgments were not transferred, languages did *not* become degenerate in their categorisation of size. I predicted that transferring the results from the acceptability judgment task would induce a pressure for informativeness which would compete with the pressure for simplicity, making the pressure for simplicity less strong. However, based on the higher amount of degenerate languages and loss of information in condition 1, it seems the pressure for simplicity was stronger instead of weaker in this condition. Why the categorisations of size lose their communicative function quicker in the condition where I least expected it, is an interesting question.

5.2 Categorisations along size and angle

In condition 1, in all four chains the language evolved into a categorisation of angle instead of a categorisation of size. What could have caused this switch to a categorisation of angle? In Carr *et al.*'s (2018) Iterated Learning experiment, most chains converged on a categorisation of angle as well. According to Carr *et al.* (2018), both the categorisation of angle and the categorisation of size would emerge due to a pressure for simplicity, and one was therefore not more likely than the other, given this bias. They did not further address the question why more languages converged on a categorisation of angle. To counter the effect that participants tended to converge on a categorisation of angle, I started from languages that were already categorised along size. Even so, over the course of five generations, this often turned into a categorisation along angle. It seems that participants prefer to categorise the meaning space that I used along angle. This might be because in the Shepard circles, the angle is a more salient property than the size. One participant who performed a test round of the experiment reported that he had not noticed there even was a difference in size between the pictures, although he also admitted to have been slightly distracted. I used the Shepard circles because they had been used by Carr *et al.* (2018) and therefore served as the right stimuli for comparison. Furthermore, these Shepard circles had been found to be useful as stimuli for Iterated Learning experiments (Canini *et al.*, 2014).

However, even if angle is a more salient property in the Shepard circles, this does not explain why the languages in condition 1 all changed into a categorisation of angle, losing their communicative function for the categorisation of size quicker than in condition 2, where some languages retained the categorisation of size. Perhaps this effect is caused by the pressure for informativeness that the acceptability judgments introduce in the experiment. It could be that due to the more salient nature of the angle property, categorisations of angle are much more informative than categorisations of size in this experiment. This would mean that in condition 1, languages do in fact evolve to be both

more simple and more informative, and the best way to be informative is to categorise along angle instead of along size. However, this does not immediately become clear from the results of the Iterated Learning experiment by Carr *et al.* (2018). They found that the bias for simplicity was stronger than the bias for informativeness in semantic category learning. Even with this ‘weak’ bias for informativeness, categorisations of angle were more frequent than categorisations of size in Carr *et al.*’s (2018) experiment. Perhaps the bias for informativeness was strong enough for the participants to favour angle over size in their experiment, but the argument is not convincing enough to explain why the same preference for angle was present in my experiment.

Another possibility is that transferring responses from the acceptability judgment task introduced more noise into the overall interpretation of the language than it increased informative value. For instance, participants sometimes accepted three or even all four labels for a picture. If this picture was then selected for training, the labels would quickly be scattered across the meaning space. A label that was first used to describe only the smallest sizes could then additionally be used to describe the largest size, creating a non-monotonic category which contains a ‘gap’ in the meaning. Such categories do not exist in natural language (i.e. there does not exist an adjective meaning ‘smaller than 3cm and bigger than 5 cm’). Therefore, to create order, it may be that participants joined the two detached parts of the meaning of such a category together by including the sizes in between as well. In doing so, the category becomes degenerate in terms of their size meaning. If this happened often enough, the meanings could have been dispersed so much that the categorisation of size disappeared altogether, and it would make sense to switch to a categorisation of angle instead.

In the condition where acceptability judgments were not transferred, I still gained insight into the participants’ interpretations of the language. However, these judgments did not influence the language in the next generation, and therefore did not create as much noise as occurred in the other condition. The categorisation of size that I designed therefore remained more stable. However, in this condition, one language also became completely degenerate, with no apparent categorisation of size nor of angle, and one language was categorised according to angle. The results are thus not so straightforward for condition 2 either.

As mentioned in Section 2.1, angle is not a scalar property. Therefore, although the language has been categorised a certain way, it is not the categorisation that I am interested in. In Section 3.5, I briefly touched upon the possibility that participants view the angles in a scalar fashion, with ‘1 o’clock’ being higher than ‘11 o’clock’. If this is the case, then monotonic categories and categorisations could in fact be possible for the categorisation of angle. The same pressures for simplicity and informativeness and pragmatic reasoning could then lead to monotonic categorisations of angle instead of size. However, as can be seen in Table 1, the categorisations of angle often have three or four categories, some of which cover the middle part of the meaning space, and thus have two boundaries (e.g. languages 1, 3 and 4). If angle was indeed viewed as a scalar property, such a categorisation would not be monotonic. There are some categorisations of angle which are monotonic in this sense (e.g. languages 2 and 8), but the results do

not show a preference for such categorisations.

This is in clear contrast with the only two languages which were categorised along size (Languages 6 and 7 in Table 1). These two are in fact both made up of monotonic categorisations. This suggests that when a true scalar property such as size is involved, participants do indeed favour monotonic categories. If angle was also viewed as a scalar property, I would then also expect to find more monotonic categorisations of angle in the results. Although monotonic categorisations of size are present in only two of the eight languages, it is an interesting finding. Perhaps in future research, I could find out whether there is indeed a preference for monotonic categorisations of size or whether this is simply a coincidence in my experiment. For this, I would need to focus solely on the setup of condition 2, and test on more chains to have enough data points to produce meaningful, significant results.

5.3 Pressures for implicit and informativeness biases and pragmatic reasoning

In the models by Carcassi *et al.* (2019), monotonicity arose due to the combined pressures for simplicity and communicativity and pragmatic reasoning. How can we be certain that the same pressures apply in this experiment? Many Iterated Learning experiments have shown that languages in these experiments become simpler due to a pressure for simplicity (Kirby *et al.*, 2014; Little & Shiffrin, 2009). Indeed, in this experiment, the complexity of the categorisations of size became simpler over generations. In many languages the categorisations of size gave way to categorisations of angle, which first increased in complexity. This is likely due to the scattering of the categories involved in the switch from size to angle. However, as can be seen from Fig. 8, the complexity of the categorisations of angle decreased as soon as this categorisation had settled in. It is therefore certain that the pressure for simplicity was present in this experiment, and led to less complex categorisations.

I cannot be as certain a pressure for informativeness was present in this experiment. As mentioned before, I introduced the acceptability judgment task as a novel way to institute a pressure for informativeness in this experiment. The idea was that the acceptability judgment task would give more insight into participants' intuitions about the language and therefore increase informativity as well as create a drive for participants to be more communicative by supplying more information about the language. However, it is most likely that transferring the responses for the acceptability judgment task led to too many perturbations in the overall categorisations, rather than increasing informativeness.

In an Iterated Learning experiment by Kirby *et al.* (2015), a pressure for informativeness was introduced by having two participants communicate about the language in a type of game. This indeed led to a pressure for informativeness, as languages which first evolved towards degeneracy now retained separate words for separate objects and were thus communicatively useful. It would therefore be interesting to see how the starting categorisation from this experiment would evolve in an Iterated Learning experiment if a pressure for informativeness was added through a communication game rather than through the acceptability judgment task used in this experiment.

Could there be a pressure for informativeness in my experiment despite the acceptability task’s lack of success? After all, according to Iterated Learning theories, a lack of pressure for informativeness leads to *completely* degenerate languages (Kirby *et al.*, 2015; Silvey, Kirby & Smith, 2015; Perfors & Navarro, 2014; Carcassi *et al.*, 2019). The languages in my experiment did not become completely degenerate. On average, three categories would still remain, describing the angle of the Shepard circles (Fig. 8). The languages are therefore still communicatively useful. Perhaps, if run for more generations, the categories would disappear due to the pressure for simplicity until only one remains. Alternatively, it could be that humans learning categories are influenced by a pressure for informativeness by default, as with the pressure for simplicity. Carstensen, Xu, Smith & Regier (2015) performed an Iterated Learning experiment in which categories evolved towards increased informativeness, rather than simplicity, in contrast to Carr *et al.*’s (2018) experiment. Their experiment did not contain an element which induced a pressure for informativeness (such as a communication game). Thus, perhaps it is possible that a pressure for informativeness is indeed present by default and this stopped languages from evolving towards complete degeneracy. However, Carr *et al.* (2018) suggested that the categories in Carstensen *et al.* (2015) *appeared* to be informative, but this was in fact a by-product of a pressure for simplicity. The most simple categorisation just happened to be an informative one as well.

This suggests that languages that evolve due to a pressure for simplicity alone need not necessarily evolve towards complete degeneracy. In Carr *et al.*’s (2018) experiment, languages did not become completely degenerate either, even though here too there was no element which induced a pressure for informativeness in the experimental design. Perhaps the reason categories did not disappear completely in Carr *et al.*’s (2018) and my own experiment is simply that all possible category labels were presented as options. When multiple category labels are presented, participants might be less likely to choose the same label every time. Categories therefore take much longer to disappear from a language, and the language might reach convergence before all labels have disappeared. It would be interesting to see how a categorisation would evolve if labels are reproduced from memory by participants, rather than having the participants reproduce them through selecting the correct one from a multiple choice list. The words used for the category labels might change slightly, which could first lead to more categories, but I expect that category labels will quickly become more similar and languages will evolve towards complete degeneracy quicker, unless a communicative element is added to the experiment.

Carr *et al.* (2018) do not touch upon this last point, and rather conclude that the types of categorisations found in their experiment (and which can also be found in my experiment) are an effect of the pressure for simplicity, rather than of the experimental design. They suggest that a pressure for informativeness would have led to categorisations which are compact (Fig. 3). Such categorisations were not found in Carr *et al.*’s (2018) Iterated Learning experiment, nor in my own. Following this line of reasoning, there is no concrete proof of presence of the pressure for informativeness in my experiment.

The lack of informativeness in my experiment could explain why languages did not evolve towards monotonic categorisations, but it also makes it difficult to compare my results to the model results from Carcassi *et al.* (2018). It is therefore not possible to back up their findings. However, it is still valuable to consider whether pragmatic reasoning, the final necessary ingredient for the evolution of monotonic categorisations, is in place in my experiment. In section 3.1, I argued that participants reason pragmatically during the production trials, as they do in natural conversations. Since the production trials are the main type of trial, this makes it more difficult to test the effect of adding pragmatic reasoning, as we cannot compare scenarios in which pragmatic reasoning is present with scenarios where it is absent. Again, the acceptability judgment task was introduced to aid in this matter, shedding more light on how pragmatic reasoning influences the evolution of these categorisations. This was based on the idea that humans behave more literally in an acceptability judgment task than when producing utterances (Ellis, 2004). In condition 1, transferring the responses for the acceptability judgment task would mean the literal, non-pragmatic meaning would stay visible to the next participant, whereas in condition 2, this would be lost, causing the categorisation to lose its monotonic nature in the process. However, this intended effect of the acceptability judgment task does not become apparent from the results. The most interesting results can be found in condition 2, which means it is probably best to ignore the acceptability judgment task for the remainder of this discussion as it is a factor of disturbance.

Did participants then in fact reason pragmatically? The lack of success of the acceptability judgment task and the switch to categorisations of angle make it difficult to provide a conclusive answer. However, given there was no element in the main trial (i.e. the production trial) which decreased or ‘switched off’ pragmatic reasoning in participants, I argue participants reasoned pragmatically, as they would normally do. Furthermore, pragmatic reasoning was the essential ‘ingredient’ which led to monotonic categories in scalar adjectives in the models by Carcassi *et al.* (2019). In my experiment, the two languages categorised along a scalar property are both monotonic. If this is more than a simple coincidence, this could show participants indeed reasoned pragmatically. According to Carcassi *et al.* (2019), a pressure for simplicity alone is not enough for monotonic categorisations to evolve, nor is a combined pressure for simplicity and informativeness: pragmatic reasoning is essential, as are the two pressures. However, the models did not test how a language would evolve under the influence of only a pressure for simplicity and pragmatic reasoning, because pragmatic reasoning was added to the model after a pressure for informativeness had already been added. Therefore, it could be possible that a pressure for simplicity and pragmatic reasoning are sufficient for monotonic categorisations to evolve, and a pressure for informativeness is not needed.

5.4 Similarities with Brochhagen *et al.* (2016)

If this last point from section 5.3 is true, then the effects which led to monotonic categorisations are similar to the effects which were present in the models by Brochhagen *et al.* (2016). In the models by Brochhagen *et al.* (2016), discussed in section 2.1, a pressure for simplicity and pragmatic reasoning were the only factors which influenced

the evolution of categories. Their models showed a preference for monotonic categories over non-monotonic categories. In a certain sense, the setup of my experiment is similar to their models. In Brochhagen *et al.*'s (2016) models, agents could choose between languages constituting categorisations consisting of two meanings, some of which had upper bounds and some of which did not. In my experiment, participants could also choose between competing meanings: they could either choose the lower-level category or the super category for meanings at the top part of the scale. The lower level category has an upper bound, while the super category does not. Choosing the lower-level category means choosing the non-monotonic categorisation, whereas choosing the super category means choosing the monotonic categorisation. In one of the languages which converged on a monotonic categorisation of size (language 6 in Table 1), the categorisation is very similar to the original categorisation consisting of just the bottom category and the super category (i.e. a two-category system). It did not evolve directly from the starting categorisation, since the labels do not correspond directly to the labels used in the starting categorisation. However, in essence, it is the same kind of categorisation as the bottom category together with the super category in the starting categorisation. The evolution of this language shows similarities with the evolution in the models by Brochhagen *et al.* (2016), in the sense that the monotonic super category was preferred over the non-monotonic middle category, and the super category took the place of this middle category and the lower-level top category. It is worth exploring how the models by Brochhagen *et al.* (2016) relate to behavioural experiments in future research. Carcassi *et al.* have pointed out that the languages in the models by Brochhagen *et al.* (2016) cannot distinguish between degenerate and monotonic meanings. This presents a problem for studying the biases that influence language evolution, because it cannot be determined whether a pressure for simplicity alone will lead to monotonic or degenerate languages. Carcassi *et al.* (2019) do suggest that the simpler models by Brochhagen *et al.* (2016) are suitable for looking at various cases of scalarity. In this sense, their models are relevant, and it would be interesting to further explore the relation between my results and the results by Brochhagen *et al.* (2016), and to adapt my experimental design to better reflect the model choices by Brochhagen *et al.* (2016).

5.5 Other factors influencing the results

There are some other unwanted factors which could have influenced the results in my experiment. As mentioned before, a participant who performed a test round of the experiment admitted to have been slightly distracted. To avoid cases of distraction during the real experiments, I measured performance accuracy on the training items. If a participant scored too low on the training items, they would be excluded from the experiment and the experiment would be repeated for that generation and chain. Performance accuracy was measured using the checks after each training item, where participants were required to repeat the label that they just learned. However, in practice, participants could obtain a very high performance accuracy score by only looking at the label and repeating that for the next question, ignoring the picture entirely. This would mean they had not actually learned the language going into the test phase. Participants who

obtained high performance accuracy scores would sometimes go on producing very different labels for a categorisation that was already fairly stable. This also explains the sudden jumps in complexity in Figure 8. Carr *et al.*'s (2018) Iterated Learning experiment avoided this situation by introducing a 'mini-test' after every fourth training trial, and by implementing a reward system for correct responses. In my experiment, however, there were insufficient resources to implement these additions, but the fact that participants could score very high on the training items and still show low proficiency in the test phase shows that these additions are necessary.

Finally, most languages had not converged yet after 5 generations. Only one language had converged after 3 generations. One language remained identical for one generation, but changed drastically in the next generation. Since the experiment started out with a language that already showed structure instead of starting out with a completely random language, I expected that less generations would be needed in order to reach convergence. Therefore, I opted to let the chains run for 5 generations instead of the usual 10. However, this expectation was based on the idea that the languages would build upon the structure that was already present. Because many participants instead abandoned the categorisation along size and formed a new structure, the languages underwent a phase in which they were less structured than they were at the start. Because of this, the amount of generations was insufficient to let the new structure settle in. This means that the final languages after 5 generations often still had some labels in places where they clearly did not belong according to the bigger categorisation that was present in that language. Had the languages been let run for more generations, the languages might have been more structured and the overall results might have been much clearer. For future experiments, the chains should therefore be let run for at least 10 generations, and it should be ensured that the languages converge.

6 Conclusion

The aim of this study was to explore what biases are present in learning and producing scalar adjectives and scalar adjective categorisations, and how these biases could explain the evolution towards monotonic scalar adjectives. In particular, I was interested in the role that pragmatic reasoning plays in this evolution. I proposed that pragmatic reasoning would be essential in the emergence of monotonic categories, and that this shows the importance of implementing pragmatic reasoning in AI models to improve human-AI interaction and to avoid false beliefs and confusion in AI agents.

Although I still believe this to be true, the results from my experiment do not shed more light on the importance of pragmatic reasoning. The results are inconclusive. Many languages evolved in such a way that I could not properly assess their monotonicity, becoming categorisations of angle rather than size. Despite my efforts to counter this effect, the saliency of the angle property of the Shepard circles appeared much stronger than the size property. Furthermore, the acceptability task proved ineffective as a way to induce a pressure for informativeness, which means not all necessary biases were present in my experiment. However, there were some interesting findings in the results. The

only two categorisations that remained categorised along the scalar property of size, are indeed both monotonic, suggesting that for a scalar property such as size monotonic categorisations are indeed the most simple and informative, given the ability to reason pragmatically. This is a result which is worth noting, but which in itself is not enough to conclude that pragmatic reasoning indeed plays a role in the evolution of categorisations towards monotonicity. As a result, I cannot conclude from this experiment that pragmatic reasoning is essential in AI to improve AI-human communication and make language in AI more human-like. Expanding and improving my experiment would be a good first step in exploring the importance of pragmatic reasoning for AI. Apart from this, it would also be interesting to see if and how monotonic categories could evolve on their own in computational models. In the models by Carcassi *et al.* (2019) and Brochhagen *et al.* (2018), agents could choose from a set of pre-designed categorisations. Perhaps using logics such as pragmatic logics or default logics, an AI agent could create a monotonic categorisation on its own, by learning from data and improving the categorisation to be more learnable and informative. If all circumstances in learning and producing categorisations are equal to those in human learning, monotonic categories should evolve on their own. This would show pragmatic reasoning indeed makes AI language more human-like, thus aiding communication between AI agents and humans.

References

- Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*, 149-178.
- Bell, J. (1991). Pragmatic Logics. In *KR '91*, pp. 50-60.
- Bell, J. (1999). Pragmatic reasoning: Inferring contexts. In *International and Interdisciplinary Conference on Modeling and Using Context* (pp. 42-53). Springer, Berlin, Heidelberg.
- Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of memory and language*, *51*(3), 437-457.
- Brochhagen, T., Franke, M., & van Rooij, R. (2016). Learning biases may prevent lexicalization of pragmatic inferences: a case study combining iterated (Bayesian) learning and functional selection. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, pp.2081-2086.
- Canini, K. R., Griffiths, T. L., Vanpaemel, W., & Kalish, M. L. (2014). Revealing human inductive biases for category learning by simulating cultural transmission. *Psychonomic Bulletin & Review*, *21*(3), 785-793.
- Carcassi, F., Schouwstra, M., & Kirby, S. (2019). The evolution of adjectival monotonicity. In *Proceedings of Sinn und Bedeutung*, Vol. 23, No. 1, pp. 219-23.

- Carr, J. W., Smith, K., Culbertson, J., & Kirby, S. (2018, July 1). Simplicity and informativeness in semantic category systems. <https://doi.org/10.31234/osf.io/jkfyx>
- Carstensen, A., Xu, J., Smith, C., & Regier, T. (2015). Language evolution in the lab tends toward informative communication. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 303–308). Austin, TX: Cognitive Science Society.
- Chemla, E., Buccola, B., & Dautriche, I. (2019). Connecting content and logical words. *Journal of Semantics*, *36*(3), 531-547.
- Davis, E., & Marcus, G. (2015). Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, *58*(9), 92-103.
- De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental psychology*, *54*(2), 128-133.
- Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, *171*(3968), 303-306.
- Ellis, R. (2004). The definition and measurement of L2 explicit knowledge. *Language learning*, *54*(2), 227-275.
- Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998-998.
- Grice, P. (1975). Logic and conversation. In *Studies in the ways of words* (pp. 22–40). Cambridge, MA: Harvard University Press.
- Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681-10686.
- Kirby, S., & Hurford, J. R. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In *Simulating the evolution of language* (pp. 121-147). Springer, London.
- Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87-102.
- Little, D. R., & Shiffrin, R. (2009). Simplicity bias in the estimation of causal functions. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 31, No. 31).
- Perfors, A., & Navarro, D. J. (2014). Language evolution can be shaped by the structure of the world. *Cognitive science*, *38*(4), 775-793.
- Regier, T., Kemp, C., & Kay, P. (2015). Word Meanings across Languages Support Efficient Communication. *The handbook of language emergence*, *87*, 237.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of mathematical psychology*, 1(1), 54-87.

Silvey, C., Kirby, S., & Smith, K. (2015). Word meanings evolve to selectively preserve distinctions on salient dimensions. *Cognitive science*, 39(1), 212-226.

Steinert-Threlkeld, S., & Szymanik, J. (2019). Learnability and semantic universals. *Semantics and Pragmatics*, 12, 4.

Von Fintel, K., & Matthewson, L. (2008). Universals in semantics. *The linguistic review*, 25(1-2), 139-201.