Size matters: Grounding quantifiers in spatial perception
Pauw, S.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)
Modifiers such as *many* and *few* are known to be syntactic hybrids, acting alternatively as quantifiers and adjectives. It has been argued that this duality in syntax is the result of grammaticalization, where these modifiers originate as adjectives and later become quantifiers. The present chapter describes an agent-based computational model that combines these linguistic insights with results from cognitive studies. Using this model it is shown that existing quantifiers can create attractor positions for other modifiers to grammaticalize into and that *many* and *few* follow this path in search of reducing cognitive effort. Thus, arguing that the shift from qualifying to quantifying expression has a cognitive motivation. This chapter is a version of a forthcoming publication: Pauw (2013a).

### 8.1 Introduction

Words such as *many, few, much* and *little* (henceforth *gradable quantifiers*) are known to be syntactic hybrids: though traditionally analyzed as quantifiers, they can also act like adjectives (Solt, 2009). One of the main reasons for their classification as quantifiers is their position in a noun phrase: just like more canonical quantifiers, such as *all* or *three*, they cannot follow an adjective (e.g., we can say “many big houses”, but not “big many houses”). However there are many constructions where gradable quantifiers behave like adjectives. Examples are “too many” and “fewer”: we can say “too big” or “smaller”, but not “too three” or “aller”. Why do gradable quantifiers have such a fuzzy syntax? What is the cognitive basis for this syntactic duality?

It has been argued that the dual syntax of gradable quantifiers might be the result of a grammaticalization process where they originate as adjectives and later become quantifiers (Solt, 2009). In Chapter 7, I showed that gradable quantifiers...
are indeed likely to emerge as adjectives due to their cognitive relation with size predicates such as \textit{big} and \textit{small} (perception of quantity relies on the same cognitive mechanisms as perception of size). Gradable quantifiers are not only cognitively related to size predicates but also to quantifiers such as \textit{all} or \textit{three} (they are used to describe sets of objects, not individual objects). In the current chapter I show that this cognitive relation will invite the (initially adjectival) gradable quantifiers to grammaticalize into quantifiers. So together with Pauw (2013b) this chapter shows that the syntactic duality of gradable quantifiers might reflect an underlying cognitive duality.

This chapter describes an \textit{evolutionary linguistic} treatment of this grammaticalization process. In evolutionary linguistics, language is seen as a complex adaptive system that is subject to selectionist principles similar to those of biological evolution. The main agenda of this field is to identify the changes of language that led up to the current state of language and expose the cognitive mechanisms that are responsible for these changes. Lacking empirical data, computational modeling is the most common methodology in this field (See Kirby and Hurford, 2002; Steels, 2005). The objective of such computational experiments is to take changes observed by historical linguists and recreate them using (robotic) multi-agent simulations. This allows us to carefully study the cognitive processes that are required by the agents to produce the simulated language change. Of course it is not feasible to simulate the entire evolution of language from its genesis to the current state of all languages. Instead, it is common practice to focus on one specific aspect of language change and expose the cognitive mechanisms responsible for that particular change (see Bleys et al., 2009; van Trijp, 2012; Spranger, 2012; Pauw and Hilfer, 2012; Beuls and Steels, 2013, for recent examples).

As mentioned above, this chapter focuses on the grammaticalization path of gradable quantifiers from adjectives to quantifiers. It describes a series of experiments, in which a community of robotic agents is provided with a fully operational initial language that allows them to describe objects and sets of objects using nouns and noun phrases such as “two big blocks” or “many boxes”. The gradable quantifiers \textit{many} and \textit{few} in this language are defined as adjectives. The agents are provided with a \textit{reanalysis mechanism} that allows them to change the grammatical categories of lexical items (Bisang, 1998). It could, for example, change \textit{many} into a quantifier.

The proposed reanalysis mechanism is a very general mechanism that could change the grammatical category of any lexical item into any category that has been provided by the grammar. However, in the experiments presented in this chapter, only gradable quantifiers grammaticalize into quantifiers, the grammatical categories of all other lexical items stay the same. The reason that only gradable quantifiers are affected by the reanalysis mechanism is due to the cognitive relation between gradable quantifiers and quantifiers such as \textit{all} or \textit{three}. Unlike adjectives such as \textit{big} and \textit{small}, these gradable quantifiers can only be
used to describe sets of objects. It is shown that this cognitive relation makes
gradable quantifiers more prone to grammaticalization into quantifiers than other
lexical items.

The remainder of the chapter is structured as follows: The next section dis-
cusses the computational framework and the implementation of the semantic and
syntactic model. Sections 8.3 and 8.4 describes a series of experiments that show
how languages in the different stages of the grammaticalization process perform.
In a final experiment, described in Section 8.5, the agents are provided with the
reanalysis mechanism, and it is shown that this indeed gives rise to the hypothe-
sized grammaticalization path.

8.2 Experimental Framework

The experiments I describe in this chapter are based on communicative interac-
tions between humanoid robots (see (Steels, 2012a) for specific details). Figure 8.1
shows an example scene with two robotic agents interacting in a shared environ-
ment. I use a setup that has been used in many other experiments (Spranger,
2012; Pauw and Hilfery, 2012; Pauw and Spranger, 2010). Each robot perceives
the world through its own on-board sensors, e.g., a camera and proprioceptive
sensors. From this multimodal sensory input, the robots build a world model,
which reflects the robot’s current belief about the state of the environment of
gradable quantifiers from adjectives to quantifiers.

Conducting experiments with actual robots is time consuming. I therefore re-
use recorded data from actual robotic interactions to speed up our experiments
while at the same time preserving the realism of physical robot interactions. Every
scene contains two robots. The recorded data of the scene comprises the world
models that both robots have built from their visual input using standard machine
vision algorithms. In the world model of a scene, every object is represented as a
list of features such as their width, height, color and position.

Using this data, I let a community of agents play language games (Steels,
2012a). The type of language game the agents play depends on the particular
research question. For the purpose of this chapter they play a Multi-Word Guess-
ing Game (Steels, 2012b). In this game there are two agents, one of them takes
the role as a speaker, the other as hearer. The speaker has to pick an object or a
group of objects (the referent) and find an utterance to describe it to the hearer.
The hearer in turn has to interpret this utterance and point at the referent he
thinks the speaker intended. If the hearer points correctly, the game is a success.
It is a failure otherwise. If the game is a failure, the speaker points to the referent
to show the hearer what he meant.

When the speaker cannot find an appropriate utterance to describe the ref-
erent, or the hearer does not point correctly, the agent is provided with a set of
repair strategies to resolve the situation. These repair strategies allow an agent
Figure 8.1: An example of a robotic interaction. Two robots, a speaker and a hearer, are placed in an environment containing different types of objects. This scene contains two yellow blocks, one box, and the two robots themselves. The world models of the robots are shown on the right. The world as perceived by the speaker is shown on top, the world model of the hearer is shown below. The arrows represent the robots. The direction of the arrow marks the orientation of the robot. They place themselves in the origin of their respective world models, looking in the direction of the x-axis. The circles represent the blocks and the blue square represents the box.
to adapt the language they use to the particular communicative needs of the situation. They are essential to this approach because they allow us to investigate language as an ever-adapting fluid system rather than a static given. In the case of the experiments presented in this chapter, the agents are provided with a repair strategy that allows them to reanalyze the grammatical categories of existing lexical items. In the remainder of this chapter I will show how this reanalysis operator gives rise to the grammaticalization of gradable quantifiers.

8.2.1 Model

Before looking at the experimental data, I will briefly discuss the communicative machinery of the agents. The experiments reported in this chapter are certainly not built from scratch. They rely on mechanisms that have been developed and tested in other related experiments (Spranger et al., 2010b; Pauw and Spranger, 2010). The model that is described in this section has previously been used in Pauw (2013b) and is an extension of an existing model that has been used to study spatial language (Spranger, 2012) and quantification (Pauw and Hilfery, 2012).

For conceptualization this model uses a degree-based semantics that is built on top of a prototype system (Rosch et al., 2004; Lakoff, 1987). This combination of prototype theory and degree-based semantics has proven particularly successful in dealing with the problems that are inherent to conceptualizing real-world perception (see Spranger and Pauw, 2012; Pauw and Spranger, 2012, for an in-depth discussion). This semantics is implemented in a formalism called Incremental Recruitment Language (IRL) (Spranger et al., 2012b; Van Den Broeck, 2008; Steels, 2000b), a procedural semantics that is designed for semantic planning in robotic language games. For parsing and production of language the agents rely on Fluid Construction Grammar (FCG) (Steels and De Beule, 2006). IRL and FCG have been co-developed specially for the kind of grounded language games as described in this chapter. The rest of this section discusses the different technical aspects of this model in more detail.

As mentioned in the introduction, the experiments reported do not study the evolution of language from its genesis. The experiments presuppose a language stage in which nouns, adjectives and quantifiers have developed into independent grammatical categories. This language allows the agents to use noun phrases (NP, e.g., “two big blocks”) and common nouns (CN, e.g., “big blocks”). This language is an implementation of a basic context-free grammar in FCG. A CN consists of a zero or more adjectives (ADJ) followed by a noun. A NP consists of a quantifier (QN) followed by a CN.

Semantically, the function of a NP or CN is to assign a score to every object in the environment that reflects how well the object is described by the utterance. Every noun, adjective, and quantifier is associated with a prototype. IRL compares the prototype of a lexical item to every object in the world model of the agent, which results in a confidence score. When an utterance consists of multi-
ple lexical items these scores are combined by taking the minimum (see Zadeh, 1965; Pauw and Hilfery, 2012, for details). This establishes a confidence score for every object in the environment. During interpretation, IRL looks for the object that has the highest score for the given utterance. During language production (formation), IRL finds the utterance that has the highest score differential for the referent (i.e., the utterance that maximizes the score difference between the referent and all the other objects in the environment). Pauw (2013b) explains the prototype/degree-based approach in more detail and Spranger and Pauw (2012) discusses how this can be implemented in IRL.

It should be mentioned that the model of quantification in the present model is very different from most existing approaches. In most existing approaches quantifiers have a different functional status than adjectives. For example, in Generalized Quantifier Theory (Barwise and Cooper, 1981) adjectives are simple predicates, but quantifiers are relations between predicates. Also earlier similar computational experiments assumed that quantifiers and adjectives are functionally different (Pauw and Hilfery, 2012; Spranger and Pauw, 2012). For the present experiment however, it is essential to let go of this functional difference because this allows gradable quantifiers to emerge as adjectives (as shown in Pauw (2013b)). To be able to model quantifiers in the same way as adjectives, it is assumed that there is a fixed set of object groups in the world model of the agent. The groups are computed directly from the visual data using a standard clustering algorithm called agglomerative clustering (Mitchell, 1997). The algorithm creates a hierarchy of groups based on their spatial arrangement (see Section 5.3 of this thesis). The resulting object groups are represented in the same way as any other object in the environment: as a list of features. Most of the features representing an object group are the same as those for single objects (e.g., location, dimensions, color, ...). This similar representation makes it possible to reuse the concepts that are used to describe single objects (e.g., big, far, etc.) for describing object groups. However, there are some features that are only present in object groups (such as density and number of constituents). This makes it possible to have concepts that can only modify object groups (e.g., two or many).

In sum, the present model permits the agents to use NP’s and CN’s to describe object sets. The NP’s have a determiner position that is always filled by a quantifier. The difference between quantifiers and adjectives has both a syntactic and a semantic side. Semantically, quantifiers are limited to the application to object groups, whereas adjectives can modify both groups and single objects. Syntactically, quantifiers can only occur at the head of the utterance (due to the NP-rule), whereas adjectives can be at any position before the noun (due to the CN-rule).
8.3 Baseline Experiment

As mentioned above, in previous experiments it was shown that agents are likely to derive gradable quantifiers from adjectives (Pauw, 2013b). In these experiments agents did indeed develop gradable quantifiers that are similar to the English \textit{many} and \textit{few}. However they emerge as adjectives, not as quantifiers. In this chapter we look at how they grammaticalize into quantifiers. To this end it is no longer assumed that the grammatical categories are fixed, but that the agents have some mechanism that reanalyzes the grammatical category of lexical entities called a reanalysis operator.

Before describing the precise mechanics of this operator, I'll take a small detour and first speculate on its possible consequences. This section discusses a baseline experiment in which I compare the performance of the initial language (where \textit{many} and \textit{few} are adjectives) to the hypothetical target language (where \textit{many} and \textit{few} are quantifiers).

The initial language is a copy of the language that came out of the experiment described in Pauw (2013b). For this language, the lexicon consists of the adjectives \textit{left}, \textit{right}, \textit{front}, \textit{back}, \textit{far}, \textit{near}, \textit{big}, \textit{small}, \textit{few}, and \textit{many}; the nouns \textit{block}, \textit{box} and \textit{robot}; and the numeral quantifiers \textit{one} to \textit{twelve}. Additionally, the lexicon contains the CN and NP rule as described in the previous section. The target language is identical, but the gradable quantifiers \textit{few} and \textit{many} are defined as quantifiers. As a consequence, in the target language the gradable quantifiers are not parsed using the CN rule, but rather using the NP rule. Syntactically this means that \textit{many} and \textit{few} can only occur at the head of the noun phrase, whereas in the initial language they can occur on any adjective position. Semantically, this means that the application of these gradable quantifiers is restricted to only groups (sets) of objects.

8.3.1 Baseline Communicative Success

We can now compare the performance of these two languages in a language game experiment. This experiment follows the script as described in Section 8.2, but the agents do not do any reanalysis yet. I let the agents play 5000 interactions and monitor the average communicative success. The experiment was repeated for both languages (the initial language and the target language). The graph in Figure 8.2 shows the results. The graph shows three different results: 1) the average communicative success of all the language games where the speaker picked a single object as referent, 2) the average communicative of the language games where speaker picked a group of objects, and 3) the two combined (the average communicative success of all the language games).

We can see that there is no difference in communicative success between the two languages for any of those conditions. For condition 2 this result is hardly surprising: The only semantic difference between the two languages is that in the
Figure 8.2: Baseline communicative success for the initial language (\textit{many} and \textit{few} as adjectives) and the target language (\textit{many} and \textit{few} as quantifiers). The bars represent the average communicative success taken over 5000 interactions. The results shows three different conditions for the referent type: 1) the success of interactions where the referent that is described by the speaker is a single objects; 2) the success of interactions where the speaker describes only groups of objects; and 3) both combined. There is no difference in communicative success for any of the conditions between the initial and the target language.
target language *many* and *few* are quantifies, which means that they can only apply to groups of objects. For condition 2 the referent is always a group of objects anyway, so for this condition there can not possibly be any difference.

But, it is perhaps less evident that there is no difference for single objects either. This is due to the way these gradable quantifiers are used: Gradable quantifiers can not make meaningful distinctions in single objects. It is, for example, meaningless to say something like “a many block,” even if it would be grammatically correct. In other words, pragmatically, the gradable quantifiers can only be applied to groups of objects not to single objects, even if this is not explicitly marked by the grammar. In the present model this intuition is captured by the fact that gradable quantifiers, when applied to a single object, will always return a score of 0. This means that gradable quantifiers do not create any contrast between single objects and are therefore not helpful in describing them. So, for the communicative success it does not matter if the grammar considers gradable quantifiers to be adjectives or quantifiers.

In other words, in terms of communicative success there is no difference in performance between the two language. This means that communicative success is not likely to be an incentive for the proposed category shift. So why would agents reanalyze the grammatical category? If communicative success is not an incentive, what is?

### 8.3.2 Baseline Cognitive Effort

The main difference between the grammatical category of quantifiers and adjectives in the present grammar (syntactic patterns aside), is the fact that for quantifiers the application is restricted to sets of objects. In the previous section we saw that for the gradable quantifiers *few* and *many* this difference does not affect the communicative success.

It does, however, affect the search space. Before starting interpretation, any entity in the environment (single objects and groups of objects alike) is a candidate referent. The role of the defined language is not only to compute the referent, but also to do so efficiently. As soon as a quantifier is encountered, the semantic system knows it can ignore single objects and reduce the evaluation to groups of objects. Which reduces the search space. The effect is especially strong for the present language because quantifiers are the head of the noun phrase, making it the first item that is evaluated. Thus reducing the search tree right from the start.

So, in order to understand the effects of the quantifier category, we have to consider this processing efficiency. To this end, I adopt a common approach in AI to use the size of the search tree as a measure for *cognitive effort* (Steele and Wellens, 2006) (also sometimes referred to as *processing effort* (van Trijp, 2012)). The search tree size is linearly proportional to the amount of objects that have to be considered for evaluation. The precise values are not relevant. In order to make
Figure 8.3: Baseline cognitive effort for the initial language (many and few as adjectives) and the target language (many and few as quantifiers). The bars represent the average cognitive effort taken over 5000 interactions. The values are normalized between 0 and 1. The results show that for single objects, there is no difference in cognitive effort. However, when the speaker describes a group of object, for the target language (which treats gradable quantifiers as quantifiers), the cognitive effort is much lower.

The result comparable across language games, the cognitive effort is normalized to yield a score between 1 and 0, where 1 is the maximum amount of objects that the agent could consider in a specific scene and 0 the minimum. In order to avoid any confusion that may arise, it should be emphasized that cognitive effort in this context is defined as a function of the semantical search space. The search that is required for parsing the sentence, is not taken into consideration.

Figure 8.3 shows the results of the same experiment as above, but showing the cognitive effort instead of the communicative success. We can see a dramatic difference in cognitive effort for object groups (the second condition). The amount of search that the agents have to do for describing object groups is much lower.
when gradable quantifiers are modeled as quantifiers. And, although there is no difference in communicative success between the initial and target language, we see that, overall, the cognitive effort is much lower for the target language (where gradable quantifiers are quantifiers). So, unlike communicative success, reducing cognitive effort might be a good incentive for the agents to grammaticalize gradable quantifiers into quantifiers.

8.3.3 Baseline Lateral Stability

The previous two sections established the particular communicative pressure that might trigger the reanalysis operator to convert many and few into quantifiers. Yet, before going into the details of a reanalysis operator, there is another question we have to tackle: Why does a mechanism that changes the grammatical categories of many and few leave all other lexical items alone? For example, why aren’t the categories of big and small not altered in the same way as many and few? Or why doesn’t the reanalysis mechanism change left into a noun, or three into an adjective? If we want make any claims of generality at all we can not simply assume that somehow the reanalysis operator is limited to affect only gradable quantifiers.

Consider, for example, the adjectives big and small. What would happen if they are defined as quantifiers? Remember that any concept that can be used to describe single objects can also be used to describe object groups. So it is perfectly consistent to define big and small as quantifiers, they are already used by the agents to describe object groups. The difference resulting from defining them as adjectives is that their application is restricted to only object groups. To see the consequences, I have implemented a hypothetical (and, in this case, undesirable) language in which big and small are defined as quantifiers, and compared it to the initial language. Figure 8.4 shows the result for cognitive effort, and Figure 8.5 shows the results for communicative success.

We can see just as for the gradable quantifiers, that the cognitive effort is lower when big and small are defined as quantifiers. The effect is much smaller than for the gradable quantifiers, but it’s there. However, if we now look at Figure 8.5, we see that this minor decrease in cognitive effort comes at a very high cost: a very strong decrease in communicative success. The reason for this decrease is that if big and small are defined as quantifiers, their application is restricted to sets of objects. For gradable quantifiers this was not a problem, since they do not make meaningful distinctions between single objects anyway. But, for big and small this does not hold; they do make meaningful distinctions for single objects, and therefore restricting them to sets seriously hampers the language. So, if the reanalysis mechanism is to take both cognitive effort and communicative success into account, it is not likely to affect big and small.
Figure 8.4: Baseline cognitive effort for the undesirable language (where big and small are treated as quantifiers), the initial language (many and few as adjectives) and the target language (many and few as quantifiers). The bars represent the average cognitive effort taken over 5000 interactions. The results show that there is a slightly lower cognitive effort for the undesirable language. Although the effect is not nearly as dramatic as with the target language.

Figure 8.5: Baseline communicative success for the undesirable language (where big and small are treated as quantifiers), the initial language (many and few as adjectives) and the target language (many and few as quantifiers). The bars represent the average communicative success taken over 5000 interactions. The results show a much lower communicative success for the undesirable language than for the initial and target language.
8.4 Selection Experiment

The previous section showed that a combination of communicative success and cognitive effort might provide the right selective pressures for the agents to reanalyze *many* and *few* as quantifiers, but leave *big* and *small* unbothered as adjectives. To this end I manually created two languages to compare with the initial language. There are, however, many more linguistic variations imaginable that could be tested in the same way. In fact, there are way too many possible languages to test all of them individually (with the 10 modifiers and 2 categories of the example language we can already define $2^{10} = 1024$ different languages). So instead, I let the agents select the optimal languages themselves in a selection experiment. In the selection experiment the agents use a very general selectionist mechanism to find the optimal language given a set of constraints. In this case, linguistic variation is created by providing the agents with both an adjective and a quantifier versions of every modifier (any lexical item that is not a noun). So, every modifier has two competing versions. The task for the agents is for every modifier to select the fitter version of the two.

This is done by assigning a confidence score (or construction score) (between 0 and 1) to every lexical item, that reflects its performance. After every successful interaction the hearer will increase the confidence score of every lexical item that was involved in that particular interaction. If the interaction was not successful the confidence score of all the lexical items involved is decreased. The scores are updated using the following update function:

$$S_{n+1}(l) = S_n(l) + \lambda S_n(l)(1 - S_n(l))$$

Where $S_n(l)$ is the original score, $S_{n+1}(l)$ is the updated score of lexical item $l$, and $\lambda$ is a rate at which a single update influences the score. For successful interactions $\lambda = 0.1$ for unsuccessful interactions $\lambda = -0.1$. The function is a numerical approximation of an S-curve. This means that when the score gets closer to its extreme values (0 or 1), the rate at which the score changes gets lower. Update functions with this property have proven to be more stable than linearly updating scores for similar experiments; i.e., a false positive or negative has less influence on the confidence score.

Since there are multiple versions of every lexical item in the lexicon of the agent, they need some way of deciding which one they use. This is done by picking one randomly, using the scores of the lexical items as a distribution. So, lexical items with a high score have a bigger chance of being used than the ones with a low score.

Using the given initial variation and the update function, I let the agents play a series of language games. Figure 8.6 shows the results. It shows the average scores of different groups of lexical items. (Lexical items are grouped together for the sake of readability.) The graph clearly shows that for the lexical items, *big*, *small*, *left* and *right*, the agents converge on their adjectival variant. This
is not surprising: The quantificational variants are much less successful. They are limited to describe object groups whereas the adjectival variants can be used both to describe single objects and groups of objects, so the adjectival variants are being used successfully much more often.

For the words many and few, on the other hand, Figure 8.6 does not show any convergence on either their adjectival or quantificational variant. This is due to the problem discussed in the previous section: The gradable quantifiers many and few do not make meaningful distinctions for single objects. Therefore, the grammatical category they belong to does not influence the communicative success. And, since the update function is only affected by communicative success, neither variant will prevail over the other.

Figure 8.7 shows what happens if we also take cognitive effort into account. The λ in the update function is not statically defined as above, but takes the cognitive effort (C) into account by letting it influence the scoring rate (λ = 0.2 · (1 − C) for success and λ = −0.2 · C for failure\(^1\)). Now we can see that for the words many and few, the agents converge on their quantificational variant in about 80% of the time.

In sum, the grammatical categories of the lexical items were not fixed in the initial language, but the agents were provided with all options. Using a basic selectionist mechanism, the community automatically converged on a fixed categorization of all the lexical items. The modifiers in the lexicon that predicate both single objects and object groups (e.g., big, small, left and right) are categorized as adjectives, and the modifiers that only predicate object groups (e.g., many and few) are more likely to be categorized as quantifiers. Where the latter only happens if the cognitive effort is taken into account.

### 8.5 Grammaticalization Experiment

The previous section illustrated the the working of the selectionist mechanism that allows the agents to converge on an optimal categorization provided a given variation. This section describes an experiment where agents have to create this variation themselves.

The agents will start out with the same initial language as described in Section 8.3 (i.e., few, and many are defined as adjectives). But, now we provide the agents with a reanalysis operator: a mechanism that allows them to reanalyze the grammatical categories of the lexical items in their lexicon. The reanalysis operator is used when an interaction is not successful or the cognitive effort is too high. It introduces a copy of the item under analysis but with a different grammatical category. So after applying the reanalysis operator, there are two competing versions of the same lexical item.

\(^1\)Note that the cognitive effort is normalized to average out on \(C = 0.5\), so on average \(\lambda = 0.1\) — the same as for the previous graph.
Figure 8.6: The selection of lexical items. The graphs follow the scores of specific lexical items over 1000 interactions. Also, the communicative success and the cognitive effort are shown. The experiment is repeated 20 times, the results are averaged. Every modifier has two variants: one adjectival (ADJ), and one quantificational (QN). They all start with a score of 0.5. The selection mechanism only regards the communicative success. After few hundred interactions it becomes clear that for the modifiers left, right, big, and small, the adjectival variants are selected. However, the community of agents does not converge on either the QN or ADJ variant of many and few. The cognitive effort and communicative success are stable over the course of the experiment.
Figure 8.7: The selection of lexical items while minimizing cognitive effort. These graphs follow the scores of specific lexical items over 1000 interactions. Also, the communicative success and the cognitive effort are shown. The experiment is repeated 20 times, the results are averaged. Every modifier has two variants, one adjectival (ADJ), and one quantificational (QN). They all start with a score of 0.5. The selection mechanism regards both the communicative success and the cognitive effort. After a few hundred interactions it becomes clear that for the modifiers left, right, big, and small, the ADJ variants are selected. For the modifiers many and few the agents converge on the QN variant. As a result, the cognitive effort decreases over the course of the experiment. The communicative success remains the same.
8.5. Grammaticalization Experiment

8.5.1 Reanalysis Operator

The reanalysis operator is based on the notion of an attractor pole as presented in Bisang (1998) and Sommerer (2012). The general idea is that if a slot in an utterance correlates with a specific grammatical category, it encourages lexical items that occur at that slot to grammaticalize into it. For example, in our grammar, quantifiers occur at the head of an utterance. If an adjective like many occurs in the same position, it might be invited to grammaticalize into a quantifier as well.

A lexical item is reanalyzed with respect to a problematic utterance (i.e., an utterance that was used in an unsuccessful interaction or that required a high cognitive effort in parsing). Figure 8.8 provides an outline of the working of the reanalysis operator for the lexical item many in the utterance “many blocks”. The operator requires, first of all, a memory of earlier conversation. To keep things simple, the agents keep track of every utterance they ever used. Now, to reanalyze many in “many blocks”, first the pattern “X blocks” (where X can be any single lexical item) is looked up in the memory of the agent. This way, the agents establish a list of co-occurring lexical items. In the present example the agents heard the utterances “big blocks”, “three blocks” and “five blocks” before. So, the lexical items big, three, and five co-occur with many for the slot X in “X blocks”.

Second, a distribution over grammatical categories for that slot is established:
All these lexical items have a specific grammatical category. By simply counting these grammatical categories the agent can establish a distribution over grammatical categories for the slot. In this case, slot $X$ has a $\left[\frac{1}{3}, \frac{2}{3}\right]$-distribution over adjectives and quantifiers respectively (since big is an adjective, and three and five are quantifiers).

Using this distribution the agents pick a new category for the lexical item. If this is a new category for that lexical item, it is added to the lexicon of the agent. For the present example, many has a two out of three chance of being reanalyzed as a quantifier. If this happens, the agent adds the new version of the lexical item to the lexicon with an initial confidence score of 0.5. So, after application of the reanalysis operator, the lexicon contains two versions of the lexical item many, one as an adjective, and one as a quantifier. At this stage, both versions of the lexical item are available for parsing and production.

It should be noted about this mechanism that it is extremely general. Nothing is put into it to make it focus specifically on turning adjectives into quantifiers. It simply reanalyzes based on emerging surface patterns. It can assign any grammatical category to any lexical item if the produced utterances would give rise to it. In fact, the reanalysis operator is so general that it is completely agnostic to the grammar that is being used. If we would define a different grammar, with, for example, more grammatical categories (e.g., verbs or prepositions), these categories would be automatically become available to the reanalysis operator. Deciding on which reanalyses are useful and which are not is not up to the present operator, but to the selectionist mechanism as described in the previous two sections.

### 8.5.2 Experiment

Now we can put everything together. In Section 8.4 it was shown how a community of agents can use a selectionist framework to find an optimal language given a specific set of linguistic variations. This framework encourages variants of lexical items that yield a high communicative success and require a low cognitive effort. In the previous section I illustrated the workings of the reanalysis operator that creates this variation. What happens when we put the selection mechanism and the reanalysis operator together?

In this last experiment I provide a community of agents with the initial language as described in Section 8.3 (i.e., many and few are defined as adjectives). Using this language I let the agents play a series of language games. The language games are the same as described in the previous sections, but now we provide the agents with a reanalysis operator. Every time, a language game is not successful or the parsing of the involved utterance requires a high cognitive effort, the reanalysis operator is called.

The learning parameters are identical to the ones described in Section 8.4: After every successful interaction, the hearer increases the scores of the involved lexical items. If the interaction was not successful the scores are decreased. The
rate at which the scores increase or decrease is determined by the cognitive effort. This experiment is repeated 10 times for 4000 interactions.

Figure 8.9 and 8.10 show the results. They show the cognitive effort, the communicative success, and the scores of the lexical items over time. The figures show the same results, but Figure 8.9 shows individual lexical items and Figure 8.9 shows the lexical items grouped together in semantic groups (many and few—gradable quantifiers; big and small—adjectives; front, back, left and right—adjectives).

The scores of the lexical item at the first interaction correspond to the initial language: many and few start out as adjectives. The first thing to note about the graph in Figure 8.9 is that the scores of many and few for their adjectival variants systematically decrease over time, while the scores of their newly created quantificational counterparts increase at the same time. The figure shows that the quantificational variant of few overtakes the adjectival one at around interaction 450, and for many the same happens between interaction 500 and 900. After 3000 interactions we see the scores stabilize. The lexical items many and few are now quantifiers.

Furthermore, the graph shows a lot of variation being created. Not only many and few are being reanalyzed, but all lexical items are at some point in the experiment. These other variations, however, do not persist. The graph in Figure 8.10 shows this more clearly. The quantificational variants of other adjectives are being tentatively created by the agents, but their scores stay low. Thus, adjectives other than many and few stay adjectives.

The graphs further show, that the grammaticalization of the gradable quantifiers into quantifiers has no effect on the communicative success. The communicative success stays stable around 80% over the 4000 interactions. But, it does have a very strong effect on the cognitive effort. Reanalyzing many and few reduces the search space dramatically.

These results are not surprising; they essentially present a corollary of the results presented in Section 8.3 and 8.4. Since gradable quantifiers do not make meaningful distinctions for single objects, they can be grammatically confined to sets of objects. Doing so reduces the cognitive effort. So the quantifier versions of these gradable quantifiers are preferred over their adjectival ones. The lateral stability (the fact that other adjectives do not grammaticalize into quantifiers) is due to the fact that these other adjectives do make meaningful distinctions for single objects, and therefore, confining them to sets by turning them into quantifiers would make the language less successful.
Figure 8.9: The grammaticalization of quantifiers. This graph shows the progression of scores of the lexical items in the lexicon of the one agent over 4000 interactions. The experiment is repeated 10 times. The results are averaged. Figure 8.10 shows a more comprehensible versions of this graph. The agents starts out with the initial language described in 8.3 (where the modifiers left, right, big, small, many, and few are all adjectives). The graph shows that the reanalysis mechanism introduces quantificational variants of all these modifiers at some point. However, most of them are not successful. Only the quantificational versions of many and few increase in score and overtake their quantificational variants between interaction 400 and 1000. Their adjectival variants gradually disappear from the lexicon of the agent. As a result, the cognitive effort shows a steady decrease, while the communicative success stays the same.
Figure 8.10: The grammaticalization of quantifiers. This graph shows the progression of the average score of the gradable quantifiers and the average score of other modifiers over 4000 interactions. The experiment is repeated 10 times. The results are averaged. This is graph is a more comprehensive version of the graph in Figure 8.9. The reanalysis mechanism introduces quantificational variants of all modifiers. But, only the quantificational variants of gradable quantifiers become successful. As a result, the cognitive effort shows a steady decrease, while the communicative success stays the same.
8.6 Discussion

8.6.1 Conclusion

The experimental results show how the development of the gradable quantifiers *many* and *few* in a community of robotic agents follow the speculated grammaticalization path: the grammatical category of gradable quantifiers moves from adjective to quantifier due to the cognitive overlap between quantifiers and gradable quantifiers. The resulting language has shown to be more efficient in processing than the initial language while preserving the communicative success. It is this increased efficiency (or lower cognitive effort) that provides the incentive for the gradable quantifiers to become quantifiers.

Other adjectives (such as *big* and *small*) on the other hand, do not grammaticalize into quantifiers. In principle the reanalysis mechanism could also turn these adjectives into quantifiers, and this would also reduce the cognitive effort. But, this would be at the cost of a much lower communicative success, which remains the most important selection criterion. For the same reason other category shifts (e.g., common noun to adjective or quantifier to common noun) are not observed in this experiment.

These results do not stand on their own. In the previous chapter it was shown how gradable quantifiers are likely to emerge as adjectives in a community of agents due to the cognitive relation between number and size. Overall, together with the present chapter, these results show how gradable quantifiers can evolve in a language. They also highlight how cognitive overlap between number motivates their dual nature; i.e., the fact that *many* and *few* can act like both quantifiers and adjectives.

On a more general note, this research is part of a recent effort that studies known grammaticalization patterns, and uses situated interaction games to find a cognitive motivation for those patterns (van Trijp, 2012; Beuls and Steels, 2013). These studies, including the present experiment, suggest how the inclusion of cognitive constraints may illuminate some of the grammaticalization patterns found in human languages.

One can wonder how general these results are. Don’t they depend too much on model-specific assumptions? Of course, there are a number implicit and explicit assumptions in the model. For example, the use of a prototype theory to model the semantics, the pre-established joint attention, the fact that the goal of the game is shared knowledge, and the assumption that categories are not fixed but that they can change based on the surface patterns they occur in. Some of these assumptions play an important role in the outcome of this experiment (e.g., the working of the reanalysis mechanism). And it is of course impossible to defend that the model mimics human cognition perfectly.

However, the model is inspired by cognitive studies. The representation of number approximation is at least reminiscent of the human representation in
that it relies on features of density and size. Secondly, by keeping the model as simple and transparent as possible and breaking down the experiment into the steps that are described in this chapter, I have shown how the individual assumptions influence the outcome of the experiments. Furthermore, it should be emphasized that the crucial parts of the experiment are very general, non-domain-specific, mechanisms. For example, the selection mechanism is a general mechanism that has been applied to many evolutionary experiments in different linguistic domains. Also, the reanalysis mechanism was not biased to target gradable quantifiers. It can change the grammatical category of any lexical item into any other category. So, nothing in the reanalysis mechanism or the selectionist model is tailored towards the specific question of gradable quantifiers. It is important to realize that therefore, the fact that only gradable quantifiers are affected and no other lexical items, is not a particularity of the present model. It can only be attributed to the linguistic maxims; i.e., a priori any lexical item could get any grammatical category, but only the grammaticalization of gradable quantifiers provides a communicative advantage.

8.6.2 Future research

One essential assumption was not addressed in this article: the existence of a quantifier category. The reanalysis mechanism relies on existing grammatical categories. If the agents did not have any other quantifier in their lexicon, they would never grammaticalize gradable quantifiers into it. To my opinion, the most interesting topic for further research is the emergence of the grammatical category of quantifiers itself. So, how does a language go from having no quantifiers at all, to at least one quantifier? Which hopefully would shed some light on the more general question: Where do grammatical categories come from?

The reanalysis mechanism relies on another assumption that is open to future research. The reanalysis is, as mentioned before, agnostic to the particular grammar used. But, for the present experiment it is assumed that looking at surface patterns is enough. For the relatively simple grammar as presented in Section 8.2 this works fine. But, for more elaborate grammars, a deeper structural analysis might be required. It would be interesting to see if existing formalisms (such as DOP (Bod et al., 2003)) could be used to create a more general reanalysis mechanism.

Studying more complex grammars would also open the door to other interesting aspects of quantification. This research focuses on noun phrases that describe either objects or sets of objects. But, in more complex sentences, the role of quantifiers becomes more than just these descriptions. For example, quantifiers play an important role in inference. These uses of quantifiers have been extensively studied in the field of formal semantics. Another interesting topic of further investigation would be to see how this well documented inferential use of quantifiers emerges in a language.