Simulating the Emergence of Grammatical Agreement in Multi-Agent Language Games

Katrien Beuls
Vrije Universiteit Brussel
Brussels, Belgium
katrien@arti.vub.ac.be

Sebastian Höfer
Vrije Universiteit Brussel
Brussels, Belgium
mail@sebastianhoefer.de

Abstract

Grammatical agreement is present in many of the world's languages today and has become an essential feature that guides linguistic processing. When two words in a sentence are said to "agree", this means that they share certain features such as "gender", "number", "person" or others. The primary hypothesis of this paper is that marking agreement within one linguistic phrase reduces processing effort as phrasal constituents can more easily be recognized. The drive to reduce processing effort introduces the rise of agreement marking in a population of multiple agents by means of an incrementally aligned mapping between the most discriminatory features of a particular linguistic unit and their associative markers. A series of experiments compare feature selection methods for one-to-one agreement mappings, and show how an agreement system can be bootstrapped.

1 Introduction

How can a population of software agents agree on a solution for a problem that is of such a high complexity that even humans tend to struggle with it? Learning a language requires mastering its means to indicate constituent structure, that is how words are linked together in bigger units such as phrases. Many languages do this by relying on so-called *grammatical agreement*. This means that information of a single linguistic unit (*source*) can reappear on another unit (*target*), that is, grammatical information is *percolated* from source to target [Corbett, 2006].

It is exactly this language strategy, that is, the constituent marking strategy, that the agents follow in the experiments that are reported in this paper. The experiments explore the hypothesis that an agreement system might emerge in order to optimize communication by reducing the cognitive load needed for referent identification, increasing expressivity and avoiding search during processing. This paper shows exactly the type of learning operators, invention and alignment dynamics that are needed in order to bootstrap an agreement system in a multi-agent population.

All experiments involve agents playing *description games* about real-world referents [Steels, 2004]. In a description game, two randomly picked agents observe a scene in which two referents are being portrayed. The speaker has to produce a full description of both referents. The game is a success if the hearer agrees with that description or a failure if the hearer disagrees. The goal of the games is to reach optimal communicative success, which will in turn be translated into the establishment of one set of agreement markers that is shared by all agents in the population.

The outline of this paper is as follows: First we introduce the linguistic processing machinery and general setup of the experiment in Section 2. We explain the invention and alignment dynamics that permit the agents to converge on a shared set of markers and illustrate by means of an example game. Section 3 contains the results of a series of experiments in which agents use the constituent marking strategy to reduce processing costs. We report on the different parameters that were used for the invention dynamics and the influence the environment has on the alignment dynamics. A discussion of the results and a first evaluation are included in Section 4.

2 Experimental Set-up

2.1 Linguistic Processing

The information that is percolated is usually packaged in terms of *agreement features*. Agreement features are elements into which linguistic units, such as words, can be broken down. Commonly used features are `number` (e.g., singular, plural, dual), `person` (1st, 2nd, 3rd) and `gender` (e.g., masculine, feminine, neuter). Less clear features include `definiteness` and `case` [Corbett, 2006]. A feature value (or a feature value bundle) is instantiated by means of a morphological affix that marks the linguistic units. It is important to note that we only handle so-called *internal agreement*, that is, agreement within the same noun phrase. The experiments reported in this paper make use of three features, with each two or three values:

- Number: singular, plural
- Gender: masculine, feminine, neuter
- Definiteness: definite, indefinite.

A typical example of internal agreement is found in the Spanish nominal phrase (NP) in (1). The gender and number feature values of the NP's head "chica" (`(gender feminine); (number singular)`) are here repeated on

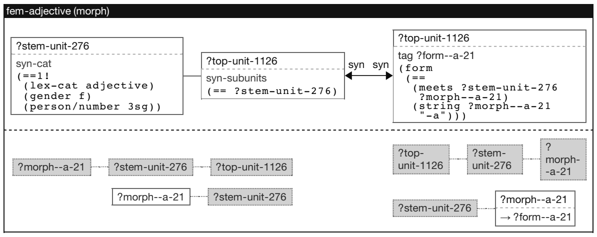


Figure 1: Visualization of the Spanish adjectival "-a" suffix construction. Features are included on the left pole and map into a string on the right pole or reversely.

the determiner "la" and the adjective "estudiosa" so that the complete NP is marked for both features.

Spanish (own example)

(1) la chica estudios-a
DEF.F.SG girl(F).SG diligent-F.SG.
'The diligent girl.'

Features translate into agreement markers, such as the "-a" ending in example (1). The operationalization of these markers (as morpho-syntactic constructions) and the basic phrasal architecture supporting the experiments is implemented in Fluid Construction Grammar (FCG) [Steels, 2011a]. More information on the use of morpho-syntactic constructions and grammatical meaning in FCG is available in [Gerasymova, 2011]. An example FCG marker for the Spanish adjectival "-a" ending has been included in Figure 1. Further references to markers in this paper always refer to constructions that map a feature to a string and reversely.

FCG is a bi-directional unification-based grammar formalism that has been especially designed with the purpose of flexible language processing. This implies that processing problems are diagnosed and repaired in a meta-layer of the regular processing pipeline. Agents learn from problems they encounter and adapt their construction inventories according to their experiences. A series of learning operators support this process. Learning operators consist of diagnostics and repairs.

2.2 Example Game

The interaction script of the a single game proceeds as follows: Assume a *population P* of agents, and a *world W* consisting of a set of individual objects. Two members are randomly selected from the population and take on the roles of speaker and hearer respectively. The *context C* contains a subset of the world W, more precisely, it contains two individual objects each attributed by a determiner ("the", "a") and an optional adjective (e.g. "tall", "red", "old", . . .). The contexts are automatically generated by a script in such a way that each feature is equally probable to be discriminatory. The interaction then proceeds as follows:

1. The speaker and the hearer jointly perceive C, which always consists of two objects, e.g.
`(tall id-1) (unique id-1) (men id-1)`
`(old id-2) (unique id-2) (cheese id-2).`
2. The speaker conceptualizes C and produces an utterance that looks like an English pidgin-like language: e.g. "tall the men cheese the old".
3. The speaker then parses his own utterance to verify the degree of cognitive (processing) effort [Steels, 2003]. Cognitive effort (CE) is measured based on this result and averaged over an interval of $n - m$ games.

$$\text{Parsing result}_i = 1 - \frac{\#\text{ possible hypotheses}}{\#\text{ found hypotheses}} \quad (1)$$

$$\text{CE} \begin{array}{c} n \\ m \end{array} = \frac{1}{(n-m)} \sum_{i=m}^{n} \text{Parsing result}_i \quad (2)$$

Found hypotheses are final nodes in the search tree that pass the predefined goal tests. Possible solutions are all combinations of the linguistic strings that match the context predicates: "the smart the man boy" → 2 combinations: (i) the man/ the smart boy, (ii) the boy/ the smart man. The sum of the parsing results (i stands for the interaction number) is divided by the size the interval into a single number between 0 and 1. The interval in all the reported simulations is set to 100.

4. When cognitive effort is *diagnosed* by the speaker, he tries to solve it by following the constituent marking strategy. If there are discriminatory features that distinguish the two objects, a *repair strategy* introduces corresponding agreement markers. Markers are always invented pairwise and function to minimize cognitive effort: e.g. "tall-bu the-bu men-bu cheese-ba the-ba old-ba", where "-bu" can either stand for `(gender m)` or `(number pl)` and "-ba" for `(gender n)` or `(number sg)`. Please note that in the experiments reported in this paper we only consider a one-to-one mapping of markers to features, i.e. each marker is associated with exactly one feature value bundle.
5. The hearer parses the utterance.
6. When the hearer *diagnoses* marker strings that are either new to him or incompatible with the earlier features he assigned to them, a *repair strategy* reconceptualizes the context and discriminatory features are calculated, which are then associated with the unknown/incompatible markers.
7. Both agents update their marker scores depending on the outcome of the game (success/failure). The hearer signals a failure when he cannot parse the speaker's description. Communicative success (CS) is measured based on this result and averaged over an interval of $n - m$ games.

$$\text{Game result}_i = \begin{cases} 1 & \text{if game}_i\text{is successful} \\ 0 & \text{if game}_i\text{is successful} \end{cases} \quad (3)$$

$$\text{CS} \begin{array}{c} n \\ m \end{array} = \frac{1}{(n-m)} \sum_{i=m}^{n} \text{Game result}_i \quad (4)$$

Score	Setting
δ_{init}	0.5
$\delta_{success}$	$+0.1$
δ_{fail}	-0.1
$\delta_{inhibit}$	-0.2

Table 1: Lateral Inhibition Settings

2.3 Alignment Dynamics

Since the invented markers should not only spread through the population but the agents should at the same time also align their marker inventories, an alignment strategy is added to the game dynamics. The alignment dynamics specify operations (e.g. in terms of scoring) that allow the agents to reach convergence on one final marker set after multiple interactions. In this paper, alignment has been implemented by means of lateral inhibition, a machine learning technique that rewards successfully used markers and inhibits their competitors. This happens only for the hearer at the end of each successful game. After a failed game, the markers used by the speaker are also punished.

Each marker has a score $s \in [0.0; 1.0]$ and enters the lexicon with the initial score δ_{init}. The lateral inhibition scores used in the experiment are summarized in Table 1.

3 Results

3.1 Baseline

The baseline run shows the performance of the agents in a game set-up that lacks the constituent marking strategy. This means that the repair strategy that catches a high degree of cognitive effort in the parsing of an utterance (see Step 4 of the interaction script above), is not available to the agents and no agreement markers will be invented. Also the hearer's learning operators (Step 6) are not accessible. Figure 2 shows the result of such a baseline run after on average 500 games have been played per agent (on a population of ten agents). Two measures are plotted here in function of the number of games: cognitive effort (see equation (2)) and communicative success (see equation (4)).

Cognitive effort is situated around 60% over 500 games. Depending on the contexts, the number of possible predicate combinations varies, which explains the fluctuations in the cognitive effort curve. By having only a lexicon and phrasal constructions that lack agreement at their disposal, the agents are constantly experiencing a considerable degree of cognitive effort in parsing the utterances they encounter. Communicative success, on the contrary, stays up at 100% since the games are after all successful. Since the context is available for the hearer, he can always map the heard utterance to the context and try out all possible solutions.

3.2 The Constituent Marking Strategy

The results of the baseline run have shown the need that arises when talking about multiple objects to mark their attributes consequently so that the hearer is guided in his interpretation process. A language that is more expressive is mostly easier to decode. Figure 3 shows the same number of games (and the same number of agents) this time ran with the constituent marking strategy and its learning operators activated. This means that the full interaction script is available in this run.

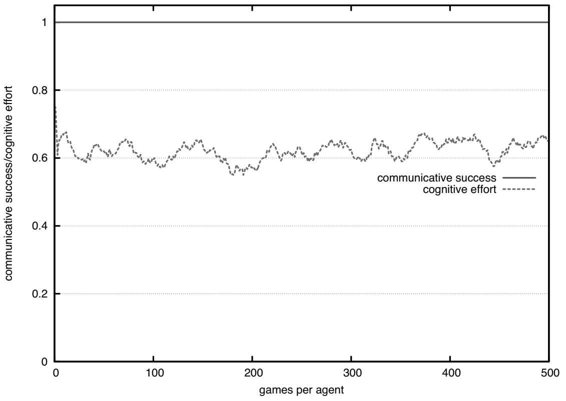


Figure 2: Without agreement marking

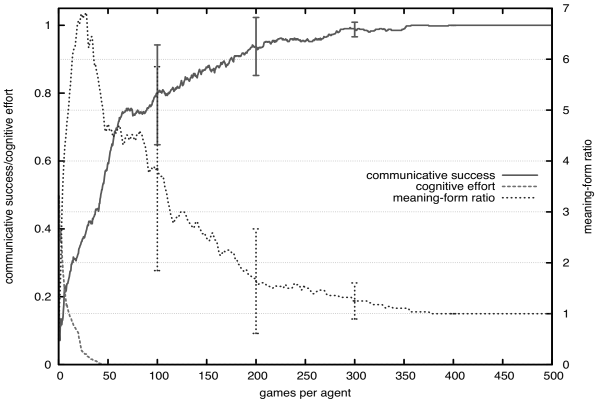


Figure 3: With agreement marking

An additional measure has been plotted since we are for the first time dealing with the presence of agreement markers: the meaning to form ratio. This ratio expresses the number of forms, that is the marker strings (similar to right left pole in Figure 1)), for every single marker meaning (or feature specification), e.g. `(gender f)`. When the agents have aligned their marker inventories, this ratio should be equal to 1: one form for one meaning.

Unlike in the baseline run, communicative success starts off very low due to the high number of failed games when the hearer could not parse unknown agreement markers. The communicative success rises slowly until it reaches 100%. This moment coincides with the alignment of the marker inventories of the agents (convergence), which happens when the meaning to form ratio curve reaches 1 (on the y2 axis). The large error bars on the communicative success and the meaning-form ratio curves signal the occurrence of statistical outliers.

The degree of cognitive effort, around 60% in the baseline run, drops sharply within the first series of games. This is due to the inherent semantics of the measure: the number of solutions found in parsing (before interpretation[1]) starts to decrease as soon as words carry endings that mark their relations. Even if the internal meaning of these endings is still left open, cognitive effort as defined in (2) has disappeared. Surely, another type of cognitive effort has replaced the first measure. This new layer of ambiguity is reflected by the communicative success curve. By the time cognitive effort has already dropped down to zero, communicative success has only reached 30%. With the invention of more and more markers and the agents' gradual alignment communicative success rises up to 100%.

3.3 Feature Selection Methods

Until now, we have largely ignored one crucial aspect of the interaction script: the procedure to select discriminatory features. The decision made in Steps 4 and 6 as to which discriminatory feature an agent selects when he adds a new marker to his inventory becomes problematic as soon as there is more than one discriminatory feature. Since we follow a one-to-one feature-form mapping, every suffix string can only map to one single feature that is extracted from the objects in the context. Therefore, a choice needs to be made. We have implemented three possible feature selection methods:

1. Single feature selection: The speaker and the hearer automatically select the same feature. There is thus some kind of 'feature transfer' implemented that prevents meaning ambiguity.
2. Random feature selection: The agents each select a discriminatory feature at random.
3. Emergent feature selection: The agents have a preference towards features they have selected in the past. All agents keep an internal (individual) scoring of all features which is independent of the marker-meaning pair scoring.

The latter method has been used in the runs shown in Figure 3. The dynamics of the single feature selection method are very similar to those of the traditional Naming Game [Steels, 1995; 2011b]. This means that there is no ambiguity in identifying the meaning of a particular marker string since the feature choice of the speaker is automatically transferred to the hearer. On the other hand, form ambiguity has not disappeared from the games. Since interactions are distributed across all agents in the population, a local decision of one pair of agents is not transferred to another pair of agents talking. Pair A might therefore invent "-a" for `(gender f)`, while pair B invents "-e" for the same feature. This is reflected in the meaning to form ratio in Figure 4b.

The differences between the random and the emergent (default) feature selection method also become visible in the same Figure. The maximum number of forms for one meaning is about 25% higher in the random method. In the emergent method the agents evolve a tendency towards selecting the same feature over time. Therefore, the number of forms for a particular feature will not rise that high since alignment can start earlier.

3.4 Robustness

The performance and the outcome of the strategies are heavily influenced by the statistical properties of the perceived contexts. In fact, the agents should only converge on a small subset of the available features that are discriminatory in more contexts than the others. This would give an explanation for the fact that natural languages only choose a small subset of features for agreement.

In order to investigate how non-balanced contexts influence the dynamics of the system with respect to the different strategies, we introduce an artificial feature bias by making the `determination` feature more likely to be discriminatory. More precisely, the probability of experiencing a context where one referent can be distinguished by `(determination definite)` and the other by `(determination indefinite)` is set to $p = 0.7$. It is important to note that this does not mean that determination is the *only* discriminatory feature in the context.

The results for the random and the emergent feature selection method in this setting are shown in Figure 5. The average marker score for the three features (determination, number and gender) is plotted in function of the number of games that are played. In both settings, it is the `determination` feature that first reaches a stable maximum score. There are two important results we should mention here. First, the speed at which the maximum score is reached is double as fast in the random method. Second, the remaining features also reach this score much faster when features are selected at random.

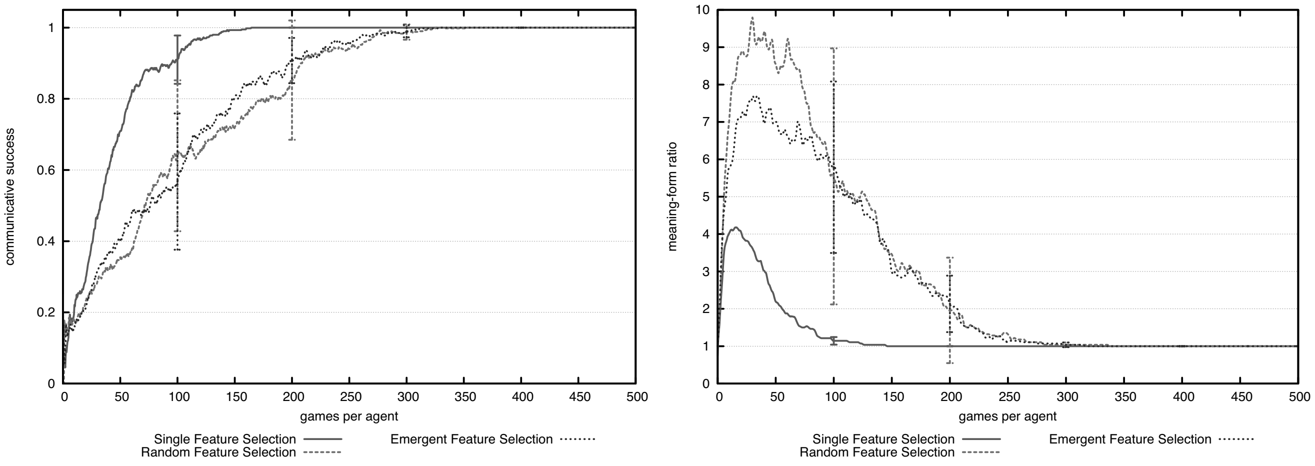
These two results are interrelated. It seems that even though convergence is reached about the same time in both methods (see Figure 4b), it takes the markers in the emergent selection twice as long as the randomly selected markers. This is due to the implications of the evolving feature preferences. Once the majority of the agents starts to prefer `determination` as their favorite feature, they will use markers expressing this feature more than other markers. Therefore, the remaining markers are only used when determination is *not* a discriminatory feature, which happens only in 30% of all contexts. The reason why the `gender` feature is in both runs the last feature to reach 1, needs to be searched in the fact that it is the only non-binary feature, consisting of three possible values: masculine, feminine and neuter.

4 Discussion and Conclusion

We have argued that agreement may have arisen in order to reduce cognitive effort in communication and help conversation partners to identify objects in a jointly perceived context more easily. Experiments were presented where agents were able to recruit features in order to create markers for establishing agreement. Following from the implemented innovation and alignment dynamics, markers were invented according to a feature selection method and spread efficiently across the population.

All the experiments included in this paper have focused on *semantic* agreement features, rather than more syntactic ones.

[1] Interpretation is here defined as the process of mapping back the parsed meaning to the context.



(a) Communicative success

(b) Meaning to form ratio

Figure 4: Comparing three feature selection methods: single feature selection, random feature selection and emergent feature selection (default). The plots show the results of four runs of on average 500 games per agent (2500 runs with 10 agents).

The assignment of a particular feature value to a given object was always guided by an ontology that was given to the agents at the start of the games. Since all non animate objects in the ontology were assigned a neuter gender, noun phrases that expressed such an object were automatically marked for neuter gender. Many languages today assign grammatical genders to non animate objects: a table is feminine in French, masculine in German. Experiments that investigate the rise of grammatical gender in nouns parallel to the rise of agreement markers are published in [Beuls and Höfer, 2011].

There were some methodological assumptions that had to be made in the implementation of these language games. First, it is implicit that the hearer knows the unknown string is a marker that expresses one semantic feature that characterizes the noun phrase as a whole. This choice implies on the one hand the use of a one-to-one feature mapping, which does not allow more that one feature to be selected. On the other hand, the possible feature range is predefined, since feature values need to be extracted from the ontology.

Second, it should be noted that the agents always invent immediately as soon as they experience cognitive effort in terms of indistinguishable referents. Uncertainty is quite common in natural language, but it does not immediately lead to invention though. This can be accounted for by introducing a threshold for invention, i.e., agents only invent a marker for a certain feature if they face cognitive effort very often without disposing of a marker for that feature. For the sake of simplicity, the results of such an experiment have not been considered here.

The resulting set of markers can be analyzed and evaluated at the end of a series of games: seven constructions each characterized by a certain string and an agreement feature, e.g. "-a" = `(gender f)`. In this sense, this set is very similar to a lexicon where every word is a mapping between a string and a meaning. It is therefore important to revise the results obtained in the experiments in the light of experiments that typically deal with lexicon formation by means of language games (see [Steels, 1995; Van Looveren, 2001; Wellens *et al.*, 2008] among others). The results of the current experiments are hereby not invalidated but should be interpreted as being a first step into the direction of grammatical language games that build a robust agreement system.

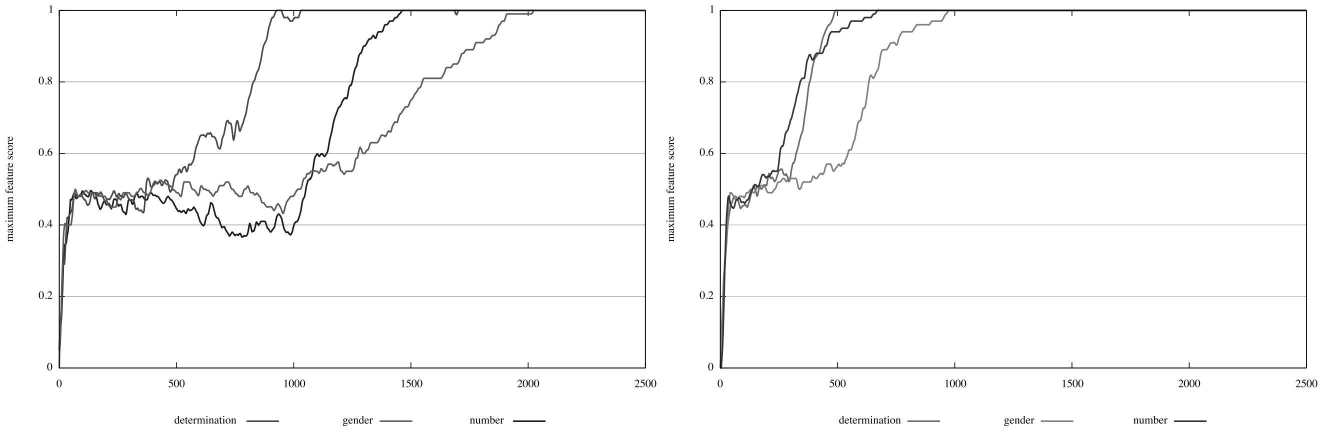
The use of distinctive feature matrices [van Trijp, 2011] could fill in this gap. Instead of having a single feature `(gender m)`, a marker string could be mapped onto a *feature matrix* that situates the `(gender m)` into the agreement system as follows:

```
(agreement
 ((number - -)
  (gender + - -)
  (definiteness - -)))
```

Every row in this matrix represents a feature while every column contains the value of a feature value. The number values are `singular` and `plural`; the gender values `masculine`, `feminine` and `neuter` and the definiteness values `definite` and `indefinite`. Such a matrix can more easily link a single marker to other markers that express the same or a different feature.

Many extensions of the presented experiments are conceivable and should be carried out in order to investigate the complexity of agreement systems further. On the one hand scaling is an important issue, that is, the results should be verified with larger populations, lexicons and a bigger amount of available features.

On the other hand, further conceptual alternations of the experiment should be carried out. Most importantly, the one-



(a) Emergent feature selection

(b) Random feature selection

Figure 5: The rate of feature development: the average score of all markers for a given feature plotted in function of the number of played games for two feature selection methods. Due to the biased environment, the determination feature is always the first to converge.

to-one marker-meaning mapping constraint should be weakened by allowing association of one marker with a conjunction of features. As mentioned before, evidence for this phenomenon is largely available in natural language. Also the reuse of a marker for a different feature remains to be investigated in future work. Such research would lead to the emergence of feature matrices [van Trijp, 2011] and open the discussion on how design islands such as declension or conjugation paradigms are built.

Although the presented experiments only dealt with the case that two agents perceive exactly the same scene, internal agreement becomes indeed indispensable in more complex environments: if the agents have a different perception of the scene, or in case of the total absence of a joint perception, communicative success is not guaranteed without means to express internal agreement.

Acknowledgments

This research was conducted at the Vrije Universiteit Brussel, financed by a strategic basic research grant (IWT-489) from the agency for Innovation by Science and Technology (IWT). Additional funding came from the European research project ALEAR (FP7, ICT-214856). We especially want to thank Pieter Wellens and other members of the team for the many discussions on the actual implementation of the experiments reported in this paper.

References

[Beuls and Höfer, 2011] Katrien Beuls and Sebastian Höfer. The Emergence of Internal Agreement Systems. In Luc Steels, editor, *Experiments in Language Evolution*. John Benjamins, Amsterdam, 2011.

[Corbett, 2006] Greville G. Corbett. *Agreement*. Cambridge, Cambridge Textbooks in Linguistics, 2006.

[Gerasymova, 2011] Kateryna Gerasymova. Expressing Grammatical Meaning - A Case Study for Russian Aspect. In Luc Steels, editor, *Design Patterns in Fluid Construction Grammar*. John Benjamins, Amsterdam, 2011.

[Steels, 1995] L. Steels. A Self-Organizing Spatial Vocabulary. *Artificial Life Journal*, 2(3):319–332, 1995.

[Steels, 2003] Luc Steels. Language Re-Entrance and the Inner Voice. *Journal of Consciousness Studies*, 10:173–185(13), 2003.

[Steels, 2004] L. Steels. Constructivist Development of Grounded Construction Grammars. In W. Daelemans and M. Walker, editors, *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 9–19. Association for Computational Linguistic Conference, 2004.

[Steels, 2011a] Luc Steels, editor. *Design Patterns in Fluid Construction Grammar*. John Benjamins, 2011.

[Steels, 2011b] Luc Steels. The Grounded Naming Game. In Luc Steels, editor, *Experiments in Language Evolution*. John Benjamins, 2011.

[Van Looveren, 2001] Joris Van Looveren. Robotic Experiments on the Emergence of a Lexicon. In *Proceedings of the 13th Belgium-Netherlands Conference on Artificial Intelligence (BNAIC '01)*, Amsterdam, the Netherlands, 2001.

[van Trijp, 2011] Remi van Trijp. Feature Matrices and Agreement. In Luc Steels, editor, *Design Patterns in Fluid Construction Grammar*. John Benjamins, Amsterdam, 2011.

[Wellens *et al.*, 2008] Pieter Wellens, Martin Loetzsch, and Luc Steels. Flexible Word Meaning in Embodied Agents. *Connection Science*, 20(2):173–191, 2008.