Multilevel Alignment Maintains Language Systematicity

Remi van Trijp

1Sony Computer Science Laboratory Paris (France)
remi@csl.sony.fr

Luc Steels1,2

2ICREA-Institut de Biologia Evolutiva (CSIC-UPF), Barcelona (Spain)
steels@upf.edu

May 29, 2012

This paper is the authors’ draft and has now been officially published as:

Abstract

The question how a shared vocabulary can arise in a multi-agent population despite the fact that each agent autonomously invents and acquires words has been solved. The solution is based on alignment: Agents score all associations between words and meanings in their lexicons and update these preference scores based on communicative success. A positive feedback loop between success and use thus arises which causes the spontaneous self-organization of a shared lexicon. The same approach has been proposed for explaining how a population can arrive at a shared grammar, in which we get the same problem of variation because each agent invents and acquires their own grammatical constructions. However, a problem arises if constructions reuse parts that can also exist on their own. This happens particularly when frequent usage patterns, which are based on compositional rules, are stored as such. The problem is how to maintain systematicity. This paper identifies this problem and proposes a solution in the form of multilevel alignment. Multilevel alignment means that the updating of preference scores is not restricted to the constructions that were used in the utterance but also downward and upward in the subsumption hierarchy.
1 Introduction

There is a lot of evidence that language users redundantly store many usage patterns that they frequently need, and out of these usage patterns new grammatical structures may arise. Some evidence comes from corpus studies, which show that language is full of ‘collocations’ (i.e. words that occur together in utterances more often than expected by chance; Sinclair, 1991). Moreover, collocations are more frequent in spoken than in written language, suggesting that speakers turn to ready-made phrases more rapidly when they are faced with the pressures of real-time conversation (Leech, 2000). This effect becomes even more significant in the case of ‘smooth talkers’ such as auctioneers and sportscasters, who have to speak very rapidly in high-demanding communicative situations (Kuiper, 1996).

Other evidence for stored usage patterns comes from the study of dialogs. Speakers and hearers negotiate in the process of conversation ways of expressing meaning and they keep reusing combinations of constructions because they know they have been successful in previous interactions (Garrod and Anderson, 1987; Pickering and Garrod, 2006). Consequently their utterance production and understanding becomes faster and more reliable and it shows less hesitation and communication failure. There is also ample evidence from psycholinguistics, specifically from recall experiments, suggesting that frequent collocations are stored and processed holistically. For example, Tremblay et al. (2011) show that humans process sentences faster and remember them better when they contain collocations (or ‘lexical bundles’) than when they don’t.

Storing frequent usage-patterns helps to cope with the speed of language communication. But it is also helps to achieve robustness. Utterance fragments are often understood, even if some parts are ungrammatical, badly pronounced, or obscured by noise. Compositional parsers tend to break down in these conditions. However, when frequent usage patterns are stored as such, enough information may be present in the input to still retrieve the pattern and then reconstruct in a top-down fashion the missing or ungrammatical parts of the input.

Although redundant storage is beyond doubt, it raises an issue for theories of language evolution, which has not been fully recognized and which is the main topic of the present paper. Language is a complex adaptive system that is socially coordinated through dialog. (Partial) coherence is maintained because dialog participants align themselves at all levels: When an utterance has been successful, all constructions and conceptualizations participating in it are re-enforced and competing constructions are de-enforced. This creates a positive feedback loop between use and success, which gradually leads the whole population towards convergence and hence a more efficient construction inventory (Steels, 1995; Baronchelli et al., 2006; De Vylder and Tuyls, 2006).

However when human language users employ a mixture of stored usage patterns
and grammatical compositional rules, there is a risk that systematicity gets lost. The bi-directional association of meaning and form is called systematic if the same meaning is always expressed using the same form. For example, the plural of a noun is always achieved by adding the morpheme “-s”, rather than sometimes having no explicit marking (as with “deer”) or use “-en” (as with “oxen” or “children”), or change the stem of the vowel (as with “foot/feet” or “mouse/mice”). The latter cases are all remnants of earlier systematic rules that have only survived because they are stored usage patterns. This example shows that if the link between a usage-pattern and a compositional rule is not maintained, the usage-pattern becomes idiosyncratic if a language changes.

Some loss of systematicity is unavoidable and clearly happening in languages. For example, there are many idiomatic phrases in the form of stereotypical usage patterns whose meaning cannot be derived in a compositional fashion, such as kick the bucket (‘to die’). Handling such idiomatic utterances has been a serious headache for strictly compositional approaches to language (Nurnberg et al., 1994; Weinreich, 1969), leading to the rise of usage-based approaches to language (Langacker, 2000), which advocate that large part of a speaker’s grammatical inventory is stored, even if this is in principle redundant.

Another example of systematicity loss is seen when a grammatical system becomes fossilized and is overtaken by another one. The marking of plural is one example. Another example is the past tense in English, currently based on adding the morpheme “-ed” to the stem of the verb, as in “walk/walked”. This system has overtaken an earlier one that was based on changing the main vowel of the verb stem, as in “bid/bade”. Although the “-ed” system is now dominant, remnants of the earlier system still abound in English, particularly for highly frequent verbs, which implies that they are memorized as special cases and then used as such. Another example of fossilization comes from case systems, as found in languages such as German and Latin, which mark the event structure of utterances. There are countless studies of idiosyncratic case marking (often referred to as ‘quirky case’) across the world’s languages. Quirky case ranges from semiregular patterns to verb-specific markings. For instance, the subjects of some experience verbs in Italian are marked for dative instead of nominative case, as in Mi piacciono le ciliegie (‘I like cherries’, lit. ‘cherries are pleasing to me’). An example of a verb-specific case marking pattern is the verb ladieG- (‘listen’) in the Caucasian Ingush language, which takes a different case pattern than all other verbs in the language (Malchukov, 2010). Often the behavior of those quirky cases can be traced back to earlier patterns in the language.

The problem of systematicity maintenance is particularly acute when the grammar is in a high state of flux, which is undoubtedly the case during its early stages. Different compositional rules are then competing with each other and if
usage-patterns based on these rules are stored, they may be based on incompatible choices. Consequently any theory of grammar emergence needs a way to tackle how systematicity can arise and be maintained. The present paper first circumscribes the problem more sharply and then proposes a solution in the form of multilevel alignment.

2 The Standard Alignment Strategy

We first recap results in lexicon acquisition using the standard alignment strategy in a language game experiment whereby a population of agents plays Naming Games about a shared perceived situation. The Naming Game assumes that utterances consist of single words and that their meaning can be uniquely identified by the hearer when he does not yet know it.

2.1 Game setup

A standard predicate calculus representation will be used for the situation model and for the meaning of an utterance. A possible situation model $s_k$ is assumed to be grounded in the world through perception using techniques as described in Spranger et al. (2012). It consists of a set of facts expressed as predicates with arguments. A possible situation model $s_0$ could be:

1. see($ev1$, $obj1$, $obj2$) ; there is a seeing event identified as $ev1$ involving $obj1$ as seer and $obj2$ as seen.
2. boy($obj1$) ; $obj1$ is a boy
3. ball($obj2$) ; $obj2$ is a ball

Here we side-step the issue how the situation model is derived or how the speaker chooses which part of the situation model to express. We also assume that the situation model is the same for speaker and hearer.

A possible meaning of an utterance $m_i$ similarly consists of a list of predicates. Variables are denoted by symbols with a question mark in front of them and they are assumed to be governed by universal quantification. For example, a possible meaning $m_0$ in the Naming Game could be:

boy($?obj$) ; the referent is constrained to be a boy

or

see($?event$, $?seer$, $?seen$) ; there is a see event
A possible interpretation of the meaning \( m_i \) of an utterance is a coherent set of bindings for all variables, where a binding is usually denoted as a pair \( (\text{variable}, \text{binding}) \), such that if the variables are substituted for their bindings, \( m_i \subset s_j \). For the example above, the set of bindings \( \{ (\text{?event}, \text{ev1}), (\text{?obj}, \text{obj1}), (\text{?obj2}, \text{obj2}) \} \) is a possible interpretation for the given situation model because if these variables are substituted for their bindings we obtain:

\[
\text{see}(\text{ev1}, \text{obj1}, \text{obj2})
\]

which is a subset of \( s_0 \).

A possible word \( w_m \) is a sequence of characters.

An expression \( e_k = \{ w_1, \ldots, w_n \} \) consists of an unordered set of words. In the Naming Game \( n = 1 \) but later on expressions will be longer. A form \( f_m \) is an (unordered) subset of words. For example, a word might be “see”, an expression “see girl box” and a form might be “girl box”. Word order does not play a role in the present context.

An agent’s language system \( L_a \) consists of a set of associations: \( L = \{ (m_1, f_1, \sigma_1), \ldots \} \) where \( m_i \) is a possible meaning, \( f_i \) a possible form, and \( 0.0 \leq \sigma_i \leq 1.0 \) is the strength of the association between \( m_i \) and \( f_i \), further called the preference score. An association is also called a construction \( \text{cxn}_i \).

To play a Naming Game, agents go through the following steps:

1. The speaker chooses a fact or set of facts, which is a subset of the shared situation model, and which will then act as the meaning of the utterance.

2. The speaker looks up the word associated with the predicates occurring in these facts in his memory and utters the word. If there is no word found, the speaker invents a new word and stores it.

3. The hearer looks up the word in his own memory. This yields a set of predicates with variables.

   (a) If the hearer already knows the word, he matches the meaning against the facts in memory and finds an interpretation in the form of possible bindings for all the variables. If this is the case, the game is a success. Otherwise a failure.

   (b) If the hearer does not know the word or if the game was a failure, the speaker provides non-verbal information or more context so that the meaning of the word can be uniquely determined. The hearer then stores the new association in his memory.

In multi-agent populations, linguistic variation naturally arises as all members of the population may contribute local innovations. Specifically, in the case of
the Naming Game they may invent new words. One of the main challenges in models of cultural language evolution is therefore how such a population can ever reach linguistic consensus and arrive at a shared language. This is why we need alignment. In this paper, we use the classical lateral inhibition alignment (Steels, 1995), which has been used in a large range of experiments. Alternative alignment mechanisms exist but the issues and results discussed in this paper hold for them equally well.

Alignment based on lateral inhibition, as proposed by Steels (1995), works as follows:

1. Each construction (i.e. association of meaning and form) in the linguistic inventory has a preference score $\sigma_i$.

2. After each game, the interacting agents update their linguistic inventories. Besides storing new constructions that were invented or acquired during the game, the agents also update the preference scores of the relevant constructions:
   
   (a) Increase the score $\sigma_i$ of all eligible constructions.
   
   (b) Decrease the score $\sigma_{j\neq i}$ of all competitors through lateral inhibition.

Constructions are eligible if they have been applied in the game. For example, the construction for the word “boy” is eligible if the utterance contained the word “boy”. A construction is considered to be a competitor of an eligible construction if it either has the same meaning but a different form, or if it has the same form but a different meaning. The competitor was not used to derive the meaning (in parsing) or the form (in production) because it had a lower score.

We use the following update learning rule, with $\gamma$ as the learning rate (set to 0.1):

$$\sigma'_i = \sigma_i(1-\gamma) + \gamma$$

$$\sigma'_{j\neq i} = \sigma_j(1-\gamma)$$

(1)

This learning rule causes agents to interpolate between the previous preference score of any eligible construction and 1, and between the previous score of any competitor and 0. Previous research has demonstrated that this approach has the desired properties for the population to reach convergence (De Beule et al., 2006). In the experiments reported here only the listener updates the preference score (following De Vylder, 2007).
2.2 Experimental Results

The effect of lateral inhibition is illustrated in Figures 1a and 1b, which compare the communicative success and inventory sizes for two populations of five agents that play Naming Games about 15 predicates. The X-axis represents the number of language games on a sequential time scale, which means that only one pair of agents is interacting at each time point \( t_s \). The Y1-axis (left) represents the running average of communicative success (\( rcs \)), which is calculated using a Simple Moving Average that computes the mean of the \( n \) latest data points for communicative success (with \( n = 10 \) in all experiments reported in this paper). This yields a number between the interval \([0, 1]\). Each data point (i.e. \( cs(t_s) \); the communicative success of a language game at a particular time step \( t_s \)) has the value 1 if the game at time step \( t_s \) was a success, and 0 if it was a failure:

\[
cs(t_s) = \begin{cases} 
1 & \text{game}(t_s) \text{ successful} \\
0 & \text{game}(t_s) \text{ failure}
\end{cases}
\]  \hspace{1cm} (2)

\[
rcs(t_s) = \frac{1}{n} \sum_{i=0}^{n-1} cs(t_s - i)
\]  \hspace{1cm} (3)

The average inventory size measure (shown on the Y2-axis on the right of each chart) plots the average inventory size \( I \) at each time step \( t_s \). \( I(t_s) \) is calculated by dividing the sum of the length of the language system of each agent at that time step (\( L_{a(t_s)} \)) by the population size \( N \):

\[
I(t_s) = \frac{1}{N} \sum_{a=1}^{N} |L_{a(t_s)}| 
\]  \hspace{1cm} (4)

Using these measures, 10 independent experimental runs (henceforth called ‘series’) were conducted. The charts in Figure 1 show the average of the recorded values. The error bars are calculated over the 10 series and are two standard deviations of the mean (one above and one below the graphs).

The agents of the first population (left chart) do not have an alignment strategy, whereas the agents of the second population (right chart) perform alignment based on lateral inhibition. Results have been obtained through 10 series of 3000 language games for each population. As can be seen in Figure 1a, the population that does not perform alignment reaches 100% communicative success, but agents need to memorize every linguistic variant for doing so and consequently end up with an inventory size of about 30 words for the 15 predicates. Figure 1b shows that, with alignment, there is an initial ‘overshoot’ in which competing words for
Figure 1: These two charts compare communicative success and inventory size on a sequential time scale between two populations of size \( N = 5 \), which develop names for 15 predicates by playing Naming Games. The left graph shows a population without lateral inhibition dynamics. As can be seen, the agents reach communicative success but they develop an inventory size of more than 30 words for 15 predicates. The right chart shows a population equipped with the Standard Alignment Strategy. Here, the agents succeed in aligning their inventories and converge on an optimal lexicon size of 15 words.

The 15 predicates propagate in the population, after which the agents succeed in aligning their inventories to settle on an optimal inventory size. Besides achieving a coherent and shared language, the alignment strategy makes the vocabulary easier to learn, requires less memory and less lookup time in processing.

## 3 Breakdown in Systematicity

The previous section demonstrated that a lateral inhibition strategy leads to a convergent lexicon. Now we turn to the question what happens when agents start building composite constructions. A composite construction is built from elements that already exist (in the present case known words) and the construction stores those elements as part of the composite. We also say that the composite subsumes the constructions which it has absorbed. Thus a composite construction captures a usage pattern and makes it available for direct matching in the linguistic inventory of speakers and hearers.
3.1 The game setup

To experiment with this, we need first of all an extended language game where agents use multi-word utterances, but we want this to be as simple as possible, and therefore use a Description Game. A description game requires that the speaker describes a situation to the hearer and it succeeds if the hearer agrees that the description is valid for the present situation. A description typically involves an event and various participants, as in “the girl sees the ball”, which involves a seeing event and two participants: a girl and a ball. The speaker is free to decide how many participants he expresses. For example, he could say “the girl sees” or “the ball was seen” or simply “there was a seeing event”. We want to minimize the use of syntax, stripping away determiners, tense markers, and so on. So possible descriptions could be “girl see ball” (meaning the girl sees the ball) or “girl see” (meaning the girl sees), “ball see” (meaning the ball was seen) or “see” (meaning there was a seeing event). Word order does not play any role in the experiment: “girl see ball” is equivalent to “ball see girl” or “girl ball see”.

To simplify the experiment further, particularly the problem of guessing the meaning of relational predicates such as ‘see’, situation models are designed such that there is never any uncertainty about the role of objects in event. For example, the event can only involve an animate object as the Actor and an inanimate object as the Undergoer (as in “John threw the ball”). It is of course well known that in most events there is uncertainty about the possible roles and this is the main reason why almost all human languages have evolved some form of argument-structure constructions (van Trijp, 2008). Here we want to avoid this issue to focus on the question of systematicity.

To play a Description Game, agents go through the following steps:

1. The speaker chooses a fact or set of facts, which is a subset of the shared situation model. This will then act as the meaning of the utterance.

2. The speaker looks up the minimal set of constructions that covers the complete meaning. If there are uncovered meanings, the speaker falls back on a Naming Game in order to establish words for those meanings.

3. The hearer uses his own inventory to find the minimal set of constructions that covers all the words in the utterance. This yields a set of predicates with variables.

   (a) If the hearer has successfully found constructions covering the complete utterance, he matches the meaning against the facts in memory and finds an interpretation in the form of possible bindings for all the variables. If this is the case, the game is a success. Otherwise a failure.
(b) If the hearer encounters unknown words or if the game was a failure, speaker and hearer fall back on a Naming Game to acquire the meaning of these unknown words.

Step [2] and [3] involve a search process and the present experiments use a classical depth-first heuristic search. The search space is formed by the application of constructions. The constructions are ordered from larger composite constructions (i.e. constructions that handle more meanings and forms in one step) to smaller constructions (for instance single words) in order to consider first constructions that provide ready-made solutions. Constructions of the same length are themselves ordered from the highest preference scores to the lowest preference scores. This ordering of constructions is updated after each language game. Each node in the search space has a confidence score (reflecting how confident the agent is with the coverage so far), which is calculated by averaging the scores of all linguistic rules that were applied until then. The branch with the highest confidence score is explored further. The agent picks the first found solution, unless that solution has a confidence score below 0.5. Only in that case does the agent check whether a better solution can be found by exploring another branch in the search tree.

There are two further actions that are taken by speakers and hearers as part of a game. The first one concerns the updating of the preference scores of the constructions. The same lateral inhibition dynamics will be used as before:

1. Increase the score $\sigma_i$ of all eligible constructions.
2. Decrease the score $\sigma_{j \neq i}$ of all competitors through lateral inhibition.

based on the update rule defined in (2). Eligible constructions are again those that have effectively been used in formulating or interpreting the utterance. Competitors are constructions covering the same meaning with an alternative form (for the speaker) or covering the same form with an alternative meaning (for the hearer). The same type of self-organization now occurs at the level of constructions as seen earlier at the lexical level.

The second action concerns the building of composite constructions. There is a single learning rule for making new constructions which is based on an analysis of the search path:

If more than one step in the search space is required, then a new construction is built which combines all the meanings and all the forms used in the utterance in a new construction.

For example, if the agent has a word for “girl” and one for “see”, and now needs to produce “girl see” then a new construction is built. It contains the two words on the syntactic pole and the following meaning on the semantic pole:
1. see(?event, ?seer, ?seen)
2. girl(?seer)

If later on the agent needs to produce "girl see boy", then he will use this construction as well as the construction for "boy" and then make a new composite construction that will now have three words covering the following meaning:

1. see(?event, ?seer, ?seen)
2. girl(?seer)
3. boy (?seen)

The composite construction subsumes the compositional constructions used to build it.

### 3.2 The Problem of Systematicity

The Standard alignment strategy has proven to be successful for lexical languages in which there is a one-to-one competition between linguistic variants. However, as soon as the linguistic inventory exhibits subsumption relations, a loss of systematicity may occur. Figure 2 is helpful for understanding the problem. This Figure shows an example of a linguistic inventory in which an agent knows four words, of which two words (girl and ingenue) are competitors because they cover the same meaning. Besides the single words, the agent has also created and memorized two competing two-word patterns, in which girl and ingenue are combined with see. Finally, the agent has also memorized two competing three-word patterns girl see box and ingenue see box. The lines between the boxes in the Figure represent the subsumption relations.

The Standard alignment strategy is problematic in the following way. Since ‘competitors’ are defined as constructions that have either the same meaning or the same form, competition always occurs locally. This means that it is perfectly possible that the single word girl may win the competition from ingenue, but that ingenue is retained in a pattern if it occurs with see, and yet another competitor might win in a different chunk. In other words, there is a danger of loss of systematicity. We will show soon results from simulations showing this effect.

One might object that natural languages are not fully systematic either, as they contain pockets of exceptions and subregularities. However, the experiments described in section 4 demonstrate that the problem of systematicity is so strong that the Standard Alignment Strategy can never scale up to achieve the level of systematicity observed in natural languages.
Figure 2: An example linguistic inventory of an agent. The agent knows four single words, two of which are competing for covering the meaning of \([\text{GIRL ?x}]\). Both competitors are connected to a two-word pattern, in which they are combined with “see” and to a three-word pattern. The two two-word patterns and two three-word patterns are competitors of each other. The lines indicate the subsumption relations between constructions.

3.2.1 Four Alignment Strategies

The challenge addressed in this paper is now clear: how to establish and maintain systematicity while still storing usage-patterns? Clearly what should happen is that the subsumption links between constructions should be exploited to reward successful usage. And this can happen either in a top-down (from composite constructions to those they subsume), bottom-up (from subsumed constructions to their composites), or multi-directional manner (bottom-up and top-down), so that we end up with four possible strategies:

1. **Standard alignment**: This strategy rewards only the applied constructions and punishes its direct competitors (i.e. constructions with either the same
meaning or the same form). For example, if pattern-1 of Figure 2 is applied, its score is increased and the score of its direct competitor pattern-2 is decreased. This is the strategy studied in the previous section.

2. Top-Down alignment: In this strategy, constructions are not only eligible for reward if they are applied, but also if they are subsumed by the applied constructions. For example, if pattern-1 is applied and rewarded, lex-1 and lex-3 get rewarded as well. The direct competitors of these three constructions (pattern-2 and lex-2) are punished.

3. Bottom-Up alignment: This strategy is similar to the previous one, but works the other way round. Constructions are not only eligible when they are applied, but when they subsume the applied constructions as well. For example, if pattern-1 is applied and rewarded, the score of pattern-3 is increased as well. The scores of the competitors of both constructions (pattern-2 and pattern-4) are decreased through lateral inhibition.

4. Multilevel alignment: The fourth strategy rewards all constructions that are either applied in successful interactions, or that have any subsumption relationships with those applied constructions. For instance, if pattern-1 is applied and rewarded, lex-1, lex-3 and pattern-3 are rewarded as well. All competitors of the eligible constructions are punished (lex-2, pattern-2 and pattern-4).

In the case of Bottom-Up and Multilevel alignment, it may happen that some constructions that meet these eligibility criteria are in fact competitors of each other. For example, suppose that lex-3 is successfully applied, then all four patterns should be rewarded as well even though they are in conflict over which form can cover the meaning for girl. Whenever there are such conflicts, only the eligible constructions with the highest score gets rewarded. In our example, this would be pattern-2 (because its score is higher than the one of pattern-1 and both three-word patterns (because they have the same score). Pattern-1 is not punished, but cannot profit from the application of lex-3 either.

4 Experimental Results

All experimental results described in this section have been obtained by conducting ten series of 10000 language games in populations of size \( N = 5 \) for each alignment strategy. The size of the meaning space in all experiments is 215 meanings, consisting of 15 individual meanings (5 actors, 5 actions and 5 objects), 75 combinations of two meanings and 125 combinations of three meanings. Each language game, a
meaning is selected randomly from the meaning space with equal probability for each meaning.

4.1 Communicative Success

The first important measure for all experiments on language evolution is communicative success, which measures whether the agents attain their communicative goals through language. The measure for communicative success is the same one as given in (3) in section 2. Figure 3 compares Running Communicative Success for each of the four alignment strategies on a sequential time scale. The chart only shows the first 1200 games, and the insert chart zooms in to the first 250 games. The results show that all strategies allow the agents to achieve 100% communicative success, but that they differ in the amount of language games it takes the agents to attain the maximum value.

Figure 3: This chart shows running communicative success for each alignment strategy in a population of 5 agents. The insert zooms in on the first 250 games. All strategies reach 100% communicative success, but the Multilevel alignment strategy is faster than the other strategies in reaching a high degree of success.

Standard and Top-Down alignment show the same evolution over time, and require about 1200 games for achieving maximum success (which amounts to an...
average of 480 games per agent). Bottom-Up alignment rises slightly slower in the first 250 language games, then catches up with Standard and Top-Down alignment and then demarcates itself after 420 games. This strategy then steeply rises to 100% success in 600-620 language games, which is about 240-250 language games on average per agent. The Multilevel Alignment strategy is very fast in reaching more than 90% success before the increase slows down because of the ceiling effect when success reaches its maximum value of 100%. The ceiling effect also allows the Bottom-Up strategy to catch up with Multilevel alignment, making both strategies twice as fast as the other two in the current experimental set-up.

Given a meaning space of 215 meanings, the rise to 100% communicative success seems very fast in all four strategies. This speed is explained by the fact that there are only 15 individual meanings. Once the agents have observed all variants in the population for those 15 meanings, they can learn and create new patterns while maintaining their communicative success. Moreover, in the case of multi-word utterances, new meaning-form pairs can often be inferred from the context without failure.

### 4.2 Meaning-to-Form Coherence

Communicative success measures whether the agents achieve their communicative goals, but mutual understanding does not automatically imply that they also converge on the same linguistic behavior. The *Meaning-to-Form Coherence* measure therefore tracks how coherent the linguistic behavior is within a population. A language is considered completely incoherent if all agents have their individual way of expressing a particular meaning. In such a language, success can still be achieved if the agents learn to understand the individual languages of the other agents (as was demonstrated in Figure 1a). In a fully coherent language, all agents have converged on the same meaning-to-form mappings.

Meaning-to-form coherence is measured as follows. For each meaning \( m_i \), an expression list \( e_{l_j} = \{e_1, ..., e_k\} \) is kept that consists of the \( k \) latest expressions observed in the population for uttering \( m_i \) through words. In all experiments reported in this paper, \( k \) was set to 10. The Meaning-to-Form coherence for each meaning \( m_i \) is calculated by dividing the number of occurrences of the most frequent expression \( e_f \in e_{l_j} \) by \( k \). Population-wide Meaning-to-Form Coherence averages over all individual coherences, which yields a number in the interval \([0, 1]\).

Figure 4 shows the evolution of Meaning-to-Form Coherence for each of the four strategies on a sequential time scale. The insert chart provides a closeup of the first 1000 language games. A similar picture emerges as for communicative success: all four of the strategies lead to coherent languages but differ in the amount of time they need for arriving at a coherent state. After 10000 games (4000 on average per agent), the Standard and Top-Down alignment strategies have not reached 100%
yet, which means that they would require additional time for reaching a maximum value. The Bottom-Up and Multilevel alignment strategies, on the other hand, have attained maximum coherence in most of the simulations.

Multilevel alignment distinguishes itself early on in the experiments and stays ahead until the ceiling effect of reaching the maximum score allows the other strategies to catch up. Multilevel alignment leads to 90% coherence after less than 2500 language games (1000 on average per agent), a degree of coherence that is only reached after more than 4000 games by the Bottom-Up strategy. The latter strategy also only attains the same level of coherence as Multilevel alignment after 8500 language games (3400 games on average per agent).

### 4.3 Meaning-to-Form Systematicity

This paper is mostly concerned with which strategy allows the agents to establish and maintain language systematicity, which means that the same meanings are sys-
tematically expressed using the same forms across linguistic rules. Systematicity will be measured as follows:

1. For each meaning \( m_i \), calculate the individual systematicity of the most frequently used expression \( e_{f1} \) from the expression list \( e_{j1} = \{e_1, ..., e_k\} \), which consists of the \( k \) latest expressions that were used for covering \( m_i \):

   (a) Take all related meanings \( rm_i = \{m_1, ..., m_l\} \) consisting of every meaning \( m_{x \neq i} \) for which holds that \( m_i \cap m_{x \neq i} \neq \emptyset \).

   (b) Calculate raw systematicity by counting 1 for each \( m_{x \neq i} \) that uses the same words for expressing \( m_i \cap m_{x \neq i} \) in its most frequently used expression \( e_{f2} \).

   (c) Divide the raw systematicity by the length of \( rm_i \).

2. Calculate population-wide systematicity by averaging over all individual systematicity scores, which yields a number in the interval \([0, 1]\).

Figure 5: This chart compares population-wide systematicity on a sequential time scale with population size \( N = 5 \). The chart indicates that Standard lateral inhibition dynamics does not guarantee language systematicity. The best strategy is Multilevel alignment.
Figure 5 plots the population-wide systematicity for each of the four alignment strategies in a population of five agents on a sequential time scale of 10000 language games (X-axis), which amounts to an average of 4000 games per agent. As can be seen, the Standard alignment strategy does not achieve systematicity and reaches a maximum of about 60%. Systematicity is still higher than chance because new patterns are always formed by combining the most successful constructions hence there is a small head-start for patterns that are introduced early on.

Top-down alignment improves systematicity but only to 70%. Bottom-Up alignment, on the other hand, performs much better and reaches full systematicity in some of the experimental rounds and averaging to 99% over all ten series. This striking difference begs the question why Bottom-Up performs so much better than Top-Down. The answer lies in the impact that the alignment strategy has on the inventory: a single word may occur in at least 15 patterns, whereas a three-word pattern is connected to at most 6 constructions (three two-word combinations and three single words), hence the top-down impact is smaller than the bottom-up one. The result is that the Top-Down alignment becomes a disruptive force in the alignment of words and two-word patterns, whereas the Bottom-Up alignment is strong enough to push alignment on the pattern-level in sync with the dynamics on the world level for most cases. The best strategy, however, is the Multilevel strategy, which uses both bottom-up and top-down alignment. The Multilevel alignment strategy allows the agents to reach 95% systematicity in less than 2500 language games (1000 on average per agent), after which systematicity slowly goes up to 100%. Since the results of Multilevel and Bottom-Up alignment lie so close together, it is necessary to inspect in more detail whether they show qualitative differences in how systematicity evolves over time in each strategy.

**Systematicity in the early phase.** The chart in the bottom of Figure 6 allows a better view of the early evolution of systematicity by enlarging language games 500 to 750 of Figure 5. Even though there is very little systematicity in any of the four strategies at this point, the chart clearly shows that the Multilevel alignment strategy takes off faster than the other ones. The difference between Multilevel alignment and the second-best (Bottom-Up) strategy becomes significant after 715 language games (286 games on average per agent – measured using the Welch two sample t-test; t = 2.1723, p=0.04951 < .05).

The difference between the two best strategies stays significant until language game 2084. At this point, the increase in systematicity using Multilevel alignment slows down, which allows the Bottom-Up strategy to catch up. The slow-down is caused by the fact that once a certain meaning-form mapping is dominant in the population, it is hard to kick it out of its place even if it is unsystematic with respect to the rest of the language. The resistance of such mappings is due to the
Figure 6: Bottom: Detailed view of Figure 5 for games 500–750. The difference between Multilevel and Bottom-Up alignment becomes significant after 715 language games (286 games per agent). Top: Boxplot of systematicity at game 715.
fact that speakers prefer to use patterns because they require less processing, so they only consider alternatives once a pattern has reached a low confidence score. However, as can be observed in Figure 5, the Bottom-Up strategy shows a loss in systematicity between games 3500 and 5000. As a result, Multilevel alignment becomes significantly better again at game 3807 (1522 games on average per agent; \( t = 2.1659, p=0.04968 < .05 \)).

**Systematicity in the middle phase.** The chart in the bottom of Figure 7 zooms in to language games 3500–6500, which includes the dip in systematicity of the Bottom-Up strategy. As a consequence, the difference between Multilevel and Bottom-Up alignment becomes highly significant at language game 4151 (1660 on average per agent – \( t = 3.0237, p=0.009997 < .01 \)). A five-number snapshot of this time-point is offered in the boxplot in the top of Figure 7. The difference between the two strategies remains significant until it evens out due to the ceiling effect when systematicity reaches its maximum value. The results indicate that Multilevel alignment is faster than Bottom-Up alignment for achieving and maintaining language systematicity.

**More details on systematicity.** In order to understand why the Standard and Top-Down alignment strategies do not achieve systematicity, or why the Bottom-Up strategy shows a dip in the middle, a closer inspection of the results is needed. Figure 8 provides a histogram that compares detailed systematicity numbers for each strategy by dividing meaning-form pairs into different bins. All bins have a width of 10% except the last two bins, which have a width of 5%. The histogram shows systematicity after 5000 sequential language games (2000 on average per agent), which is a time-point with roughly the biggest differences between the four strategies. The X-axis shows the number of constructions that are dominant for their respective meaning-to-form mapping at \( t_{s,5000} \).

The Figure shows that the Standard and Top-Down strategies are normally distributed. Most of the dominant meaning-form pairs using the Standard strategy fall in the bin of 50-<60% systematicity, with a few number of constructions that are either unsystematic or highly systematic with respect to other items in the linguistic inventory. For the Top-Down strategy, most items fall in the 70-<80% bin. The Bottom-Up and Multilevel strategies do not show normal distribution because of the ceiling effect when systematicity reaches 100%. The vast majority of constructions using the Multilevel alignment strategy fall into the top bin, indicating that they are fully systematic. The Bottom-Up strategy also performs well with most constructions falling in the top two bins, but the strategy also has more constructions that are distributed over lower bins than is the case for the Multilevel strategy. A breakdown of the bins per construction type allows to inspect
Figure 7: Bottom: View on games 3500–6500 of Figure 5. The Bottom-Up strategy shows a dip in systematicity, which causes a highly significant difference between Multilevel and Bottom-Up alignment at game 4151. Top: Boxplot of the systematicity at game 4151.
what kind of constructions are problematic for which alignment strategy.

**Systematicity per construction type.** Figures 9 and 10 show a detailed histogram per strategy for each construction type after 5000 sequential language games. The X-axis is the same as in Figure 8, while the Y-axis represents the number of dominant meaning-to-form mappings at $t_{s,5000}$ converted to a percentage scale, which allows easier comparison between construction types. The top chart of Figure 9 shows the results for the Standard alignment strategy. As can be seen, the systematicity of two- and three-word patterns shows a normal distribution, with $50\%-60\%$ as the largest bin for three-word patterns and $60\%-70\%$ as the largest bin for two-word patterns. The individual words show a different picture, where the two top bins (which have the same bin width as the other bins
Figure 9: These two histogram charts show the distribution of systematicity per construction type for the Standard strategy (top) and Top-Down strategy (bottom) after 5000 language games.
Figure 10: These two histogram charts show the distribution of systematicity per construction type for the Bottom-Up strategy (top) and Multilevel strategy (bottom) after 5000 language games.
when they are combined) collect the largest group of constructions, which means that about one fifth of the dominant words have become dominant on the level of patterns as well.

The Top-Down strategy is slightly more successful than the Standard strategy. Here, about 35% of the single words have also survived as the dominant forms in two- and three-word patterns. The effect of Top-Down alignment can also be observed in both the two- and three-word patterns, which tend to be in higher bins than was the case for the Standard strategy. However, the large amount of constructions that still fall below 50% systematicity indicates that the competition among constructions on lower levels is in sync with the competition on higher levels, and that the Top-Down forces are even disruptive for achieving population-wide systematicity.

The picture that emerges for the remaining two strategies is completely different. As can be seen in Figure 10 for Bottom-Up (top chart) and Multilevel (bottom chart) alignment, most constructions are situated in the top bin of 95-100% systematicity. The difference between both strategies is that Bottom-Up alignment has a longer slope (or tail) of constructions that fall in lower bins than is the case for Multilevel alignment. This observation can be explained in the same way as the lack of systematicity is explained for the Top-Down strategy: the local competition of higher levels (i.e. between patterns) is not always in sync with the competition on lower levels (i.e. on the level of single words). As stated before, however, the bottom-up force has a strong enough impact on the dynamics to direct the local competition on higher levels towards stronger systematicity. The lack of top-down alignment, however, causes some resistance on the higher level to follow the evolution on lower levels. The success of Multilevel alignment can thus be seen as a self-enforcing loop whereby lower levels influence higher levels, which then in turn influence lower levels so we get a structural coupling between different levels of the language inventory.

### 4.4 Inventory Size and Propagation Phase

The experimental results indicate that the choice of the alignment strategy has a significant impact on the inventory size of each agent, both in terms of the amount of memory that is required as well as in the amount of time an agent requires for acquiring all variants of a meaning-form mapping in the population.

**Maximum Average Inventory Size.** Figure 11 compares the maximum average inventory size of an agent for each of the four alignment strategies, broken down per type of construction (ranging from single words to three-word patterns). The average inventory size is calculated using the same formula as in (4) in section
Figure 11: This stacked bar chart shows the maximum inventory size of an agent for each alignment strategy. As can be seen, agents that use the standard alignment strategy need to memorize more than 600 meaning-form pairs in order to cope with the variation in their population, compared to +/– 550 for the top-down strategy and +/– 410 for the bottom-up strategy. The multilevel alignment strategy again comes out as the best one with a maximum inventory size of +/– 350 meaning-form pairs.

When using the standard alignment strategy, an agent’s linguistic inventory grows to the maximum average size of 607.42 constructions, more than half of which are two-word patterns. As the optimal inventory size is 215 constructions (15 single words, 75 two-word patterns and 125 three-word patterns), the agents need to cope with a maximum variation of about 2.83 competing form variants for each meaning at any given time point. The Top-Down strategy yields a maximum of 549.68 constructions (about 2.56 form variants per meaning). The Bottom-Up strategy performs better with a maximum average inventory size of 404.86 con-
Figure 12: This chart compares the average amount of language games on a parallel time scale necessary before reaching the maximum average inventory size for each alignment strategy, broken down per type of construction and summarized for the entire inventory size. As can be seen, agents with the Standard alignment strategy keep on learning new constructions until almost 3600 games. The other strategies all reduce the time needed for acquiring all variants, with again the Multilevel alignment strategy coming out as the best one.

Contructions (about 1.88 form variants per meaning), but again Multilevel alignment is the best strategy with maximally 332,14 constructions (about 1.54 competing form variants per meaning).

Using the Welch two-sample t-test, the difference between the best and second-best strategies (Multilevel and Bottom-Up) was found to be highly significant, with $p = 0.0005506 < .01$ and $t = 4.4304$. The observed values for the multilevel alignment strategy ranged from 275.8 to 355.2 constructions (mean = 309.74), whereas the values for the Bottom-Up strategy varied between 322 and 473.6 constructions (mean = 394.26). Looking at the types of constructions, the difference is highly significant for two-word patterns ($p = 0.000523 < 0.01$, $t = 4.4862$) and three-word patterns ($p = 0.005722 < 0.01$, $t = 3.284$), and significant for single words ($p = 0.02735 < 0.05$, $t = 2.4014$). The results therefore clearly show that Multilevel
alignment requires the least memory resources of all four strategies.

**Propagation Phase.** The semiotic dynamics of the experiments can be roughly divided into two phases: (a) the *propagation phase*, in which competing variants spread in the population, and (b) the *alignment phase* in which all variants are known and the agents try to converge on the same preferred meaning-form mappings. The time-point (i.e., language game) at which the average maximum inventory size is reached is taken as the cut-off between both phases. Figure 12 shows the length of the propagation phase for each alignment strategy per agent, that is, how many language games each agent needs to play on average before entering the alignment phase. The X-axis shows the number of language games on a parallel time scale. A parallel time scale means that at each point in time, every agent in the population has had the same number of interactions on average, hence $t_p = \frac{2 \times t_s}{N}$. For each strategy, the propagation phase is shown per type of construction and for the total inventory size.

As can be seen, the Standard alignment strategy requires the agents to play almost 2800 language games on average before reaching their maximum inventories. The propagation phase of three-word patterns even takes 3564 games on average per agent, which is twice as long as the amount of time needed by agents that use the Multilevel alignment strategy. Generally speaking, the same trend as in all previous measures is confirmed where the Standard alignment strategy performs worst and increasingly better results are obtained using Top-Down, Bottom-Up and Multilevel alignment.

Surprisingly, however, the overall second-best strategy (Bottom-Up) performs worst when looking only at the propagation phase of single words. This result requires some deeper examination. The observed values for the Bottom-Up strategy in the ten series of experiments show a much larger range than the other strategies, going from as little as 164 language games on average per agent to 884 before reaching the maximum number of words in memory. Closer inspection reveals that the Bottom-Up strategy may lead to a longer propagation phase for individual words because Bottom-Up alignment may cause some unsystematic patterns to reach a high confidence score in some agents early in the simulation. So even if agents have ‘forgotten’ a single word through lateral inhibition, they keep on ‘re-learning’ that word as long as it is observed in patterns.

The comparison between the two best strategies confirms earlier findings. The Welch two-sample t-test shows that the difference in propagation phase length between Multilevel and Bottom-Up alignment is highly significant ($p = 0.001688 < .01, t = 4.0678$). Per type of construction, the same statistical test shows highly significant differences for single words ($p = 0.003615 < .01, t = 3.8476$) and a significant difference for two-word patterns ($p = 0.03425 < .05, t = 2.352$). The
difference between both strategies for three-word patterns is not significant (p = 0.4736, t = 0.7378) because in both strategies three-word patterns are only affected by direct and bottom-up competition as there are no larger patterns. However, there is a trend that Multilevel alignment is nevertheless faster. This trend can be explained by the fact that this strategy is better at reducing the number of variants that need to be propagated in the population, as was shown in Figure 11.

5 Discussion and Conclusion

This paper showed that subsumption relations between constructions lead to a loss of systematicity. A subsumption relation exists when one construction copies part of another construction inside it, for example a frequent usage pattern is stored for fast access or a combination of constructions is chunked into a more encompassing construction. Loss of systematicity means that the same form is no longer expressed by the same meaning and the same meaning by the same form. We also showed that systematicity can be maintained if subsumptions relations are used to implement multilevel alignment, which means that the preference score is increased not only for the construction used to build or interpret the utterance directly but also all the ones related to it in a bottom-up or top-down manner through the subsumption relation. We have shown in computer simulations that multilevel alignment leads to faster convergence and smaller inventories, and most importantly to higher systematicity.

References


