# Proceedings of the ESSLLI 2007 Workshop on

# Language, Games, and Evolution

August 2007, Dublin, Ireland

Anton Benz, Christian Ebert, Robert van Rooij (eds.)

# EDITOR'S NOTE

This proceedings volume presents the contributions to the *Workshop on Language, Games, and Evolution*, held from August 6th through 10th, 2007 as part of the *19th European Summer School in Logic, Language and Information (ESSLLI-2007)* at Trinity College, Dublin, Ireland.

We received eighteen submissions, of which seven were selected for presentation at the workshop. Those seven abstracts were elaborated into the articles of the present proceedings. In addition, it contains three articles of alternate speakers and an article co-authored by Josef Hofbauer and Simon Hutegger, one of our invited speakers. We would like to thank him and Rohit Parikh, our second invited speaker, for accepting our invitation and their contributions. We would also like to thank them as well as Gerhard Jäger for their thoughtful reviews.

We gratefully acknowledge the support for this workshop by Project P13 on optimality theory at the ZAS Berlin; the SFB 673 "Alignment in Communication" at the University of Bielefeld, funded by the Deutsche Forschungsgemeinschaft DFG; and Robert van Rooij's VIDI project "The Economics of Language".

Berlin, Bielefeld, Amsterdam, June 2007

Anton Benz
Christian Ebert
Robert van Rooij

# CONTENTS

# An agent-based model of linguistic diversity

Pieter de Bie
AI department
Rijksuniversiteit Groningen, the Netherlands
p.de.bie@ai.rug.nl

Bart de Boer
Institute of Phonetic Sciences
Universiteit van Amsterdam, the Netherlands
b.de.boer@ai.rug.nl

### Abstract

Research in the field of language diversity mostly uses models where emerging language patterns visible in the real world can be hard to find. An example of this is the dialect continuum we see between the Netherlands and Germany, or the language border we find on the Dutch-French-line in Belgium.

The agent based model described here introduces language diversity as a consequence of language mutation. Agents adopt these mutations from other agents based on social impact theory, which states that less common varieties have a relatively bigger influence per individual speaking the variety. Using this model, it was possible to show different language patterns existing at the same time.

## 1 Introduction

Linguistic diversity is truly amazing. Not just in the infinite richness with which the approximately 6000 different languages of the world can express meaning, but also in the fact that it exists at all. Although there are indications that dialects and diversity also exist in animal communication systems – chimpanzees and lyrebirds (among other birds) appear to have dialects, and humpback whales appear to change songs regularly – nowhere is the diversity as pronounced as in human language. Although *why* language is so diverse is an interesting question, in the work presented here, we are concerned with the question how language can become diverse.

As with many aspects of language, language diversity can be studied from the perspective of individual behavior and from the perspective of the language as a whole. At the level of the individual, two opposing behaviors can be observed. When infants acquire a language, they learn a language that is very close to the language of their parents and their peers. Much closer, in fact, than is needed for successful communication. Although speakers of different closely related dialects have no trouble communicating with each other, they have equally little trouble in recognizing their interlocutor as the speaker of a different dialect. The opposite behavior to this tendency to conform, is a tendency to innovate. Speakers of language, especially young ones, are constantly renewing their language. This can be a conscious process (when inventing new, cool ways of saying things, for example) or an unconscious process (when adopting subtle variants of pronunciation to differentiate oneself from the old farts of the group, for example).

On the collective level, the level of the language of a population, this leads to slow but sure language change. This language change leads to two rather different kinds of linguistic diversity. One is called a *dialect continuum.* This occurs when languages gradually merge into each other over an extended geographical range. An example is the range of German and Dutch dialects. A very different situation occurs when neighboring languages are very different: this is called a *language border.* The border between French and Dutch in Belgium is a classical example and the more salient because there is no geographical or political barrier to support it, and because it has been stable for hundreds of years .

The existence of language borders without accompanying political or geographical boundaries indicates that such boundaries cannot always be part of the explanation of linguistic diversity. Migration can also not always be relied on as an explanation. The language border between Spanish and English in North America can obviously be explained as the result of the influx of populations that spoke a different language, but such straightforward explanations are not always possible. Speakers on both sides of the French-Dutch language border are closely related genetically, indicating extensive intermixing of populations, but the language border has remained stable for a long time nevertheless.

### 1.1 Previous work

It is clear that linguistic diversity is a phenomenon that needs to be explained as the result of an interaction between behavior on a collective level (the languages and corresponding cultural groups) and the behavior of individuals. The dynamics of such interactions can become extremely complex and this is why agent based computer models are a useful tool for investigating the emergence of complexity.

When one can assume geographical barriers, language borders can be explained straightforwardly, and can even be investigated mathematically (Patriarca and Leppanen, 2004). When such barriers cannot be assumed, the situation becomes more difficult, because contrary to for example biological evolution, small mutations do not tend to get carried on by offspring. (Nettle, 1999a) has identified two problems. One is the *averaging problem*, and the other is the *threshold problem.* The averaging problem occurs when the language learned by an individual is a mixture of the languages spoken around it. Any initial difference between languages (or differences caused by linguistic innovation) then tends to average out. Even when one assumes discretely inherited features of language, language change has difficulty spreading because of the threshold problem. Because of random drift in the population, variants of the language that are infrequent tend to disappear, thus hampering the spread of innovation.

That this is a serious problem is illustrated by early work to model linguistic (or cultural) diversity. Axelrod (1997), for example, has created a spatial model of cultural diversity that starts with maximum diversity. He has shown that agents that tend to interact with and copy from like agents converge to a situation where there are sharp borders between a small number of identical groups. At the same time, diversity tends to decrease. The same is true for a model by Barr (2004) that starts with high diversity and converges to few near-homogenous groups.

In order to overcome these problems, Nettle (1999a,b) has made use of social impact theory.

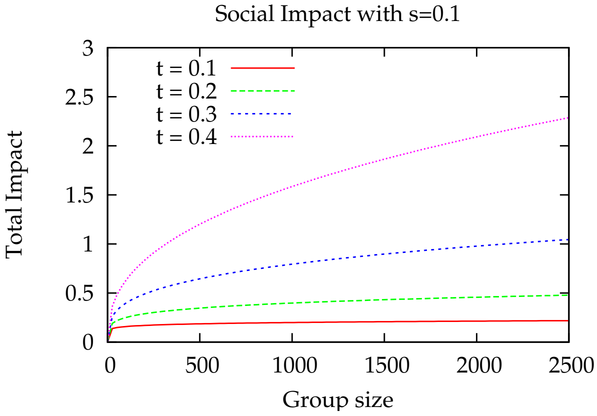


Figure 1: Sociale impact as a function of total group size, with different values of $t$

This social impact theory, originally devised by Latané (1981), models the influence of a social event given the number of people involved in the event. Latané's findings were that the total impact of the group is a power function in the form:

$$I = sN^t \tag{1}$$

where $I$ is the total social impact, $N$ is the group size and $s$ and $t$ are used for fitting the curve. As can be seen in figure 1, the curve tends to flatten out, thus the individual pressure of participants in the group tends to *decrease* with group size increase.

Nettle (1999b) uses this theory to explicitly boost the rare variant in a language modeled as either variant $p$ or $q$. Using this, or the use of super-influential individuals, he shows that language diversity can be obtained. Livingstone (2002), finally, has shown that dialect continua can emerge in groups of agents that are situated along a one-dimensional continuum. The bottom line, however, is that none of these models can generate both dialect continua and language borders.

## 2 Model

We propose a simple spatial model in which both dialect continua and language borders can emerge. This model is based on a population of agents that prefer to interact with agents that speak like them, and sometimes change their language to be more similar to their interlocutors. The agents prefer language variation that are considered a minority, making use of social impact theory.

Languages are represented as strings of binary features (16 in all simulations presented here). The distance between two languages is defined as the number of all features that are different between the two languages, divided by the number of features per language. The meaning of the features is not defined; they may be compared to Nettle (1999a)'s *linguistic items*, and can include words, but also sentence structure or pronounciation hits.

**Algorithm 1** Communication between two agents

```
for $\forall a \in N$ do {for all agents}
  for $\forall b \in Neighbourhood(a)$ do
    if random $\leq p_i(a,b)$ then
      $i \leftarrow$ random vector index
      if $a_i \neq b_i$ then
        if random $\leq p_a(i,b_i)$ then
          $a_i \leftarrow b_i$
        end if
      end if
    end if
  end for
end for
```

### 2.1 Communication and language adoption

In each time step all agents can interact with all their direct neighbors (using 4-connectivity). The whole process of communication can be seen in algorithm 1. The individual steps will be explained further.

The probability of interaction between agents $a$ and $b$ depends on the distance $d(a,b)$ between their languages as follows:

$$p_i = \alpha + (1-\alpha)10^{-\beta \cdot d(a,b)}$$

where $\alpha$ is a minimal interaction probability and $\beta$ determines how fast $p_i$ drops off with increasing distance.

When an agent interacts with an other agent, it selects a random feature from its language, and when it is different, adopts it from the other agent with a probability:

$$p_a = \max\left(1, \gamma\left(\frac{N_{i,v}}{N}\right)^{\kappa-1}\right)$$

where $\gamma$ is a minimal adoption probability, $N_{i,v}$ is the total number of agents that use variant $v$ of feature $i$, and $N$ is the total number of agents. The parameter $0 \leq \kappa \leq 1$ determines how fast $p_a$ drops off with the number of speakers of variant $i$. This probability causes rare variants to be more desirable.

This formula is in fact based on the social impact theory as in equation 1. However, since the communication proceeds on an individual level, the formula calculates individual social pressure and not total group pressure.

### 2.2 Language introduction and stability

Each time step, an agent mutates its language with a probability $p_m$ set to a fixed value in the range $[10^{-5} : 10^{-3}]$. A mutation means the flipping of a single feature in the agents linguistic vector.

Agents can learn for the first ten time steps of their lives. After this period, they can only function as language providers, thus providing some stability in the system. Agents are susceptible to die from their 15th time step, after which they have a probability of dying of 50% each time step.

After agent death, a new agent replaces the old one. This agent will speak the language most common among his direct neighbors. The alternative to this choice is to pick the most common variant of each individual language feature. Since this could result in a mix of languages, where the child is not able to speak with any of its parents, this seems implausible, and thus the idea was discarded.

## 3 Results

The described model was implemented in the Repast framework(Railsback et al., 2006). For the random functions, the Mersenne Twisteer MT19937 (Matsumoto and Nishimura, 1998) was used.

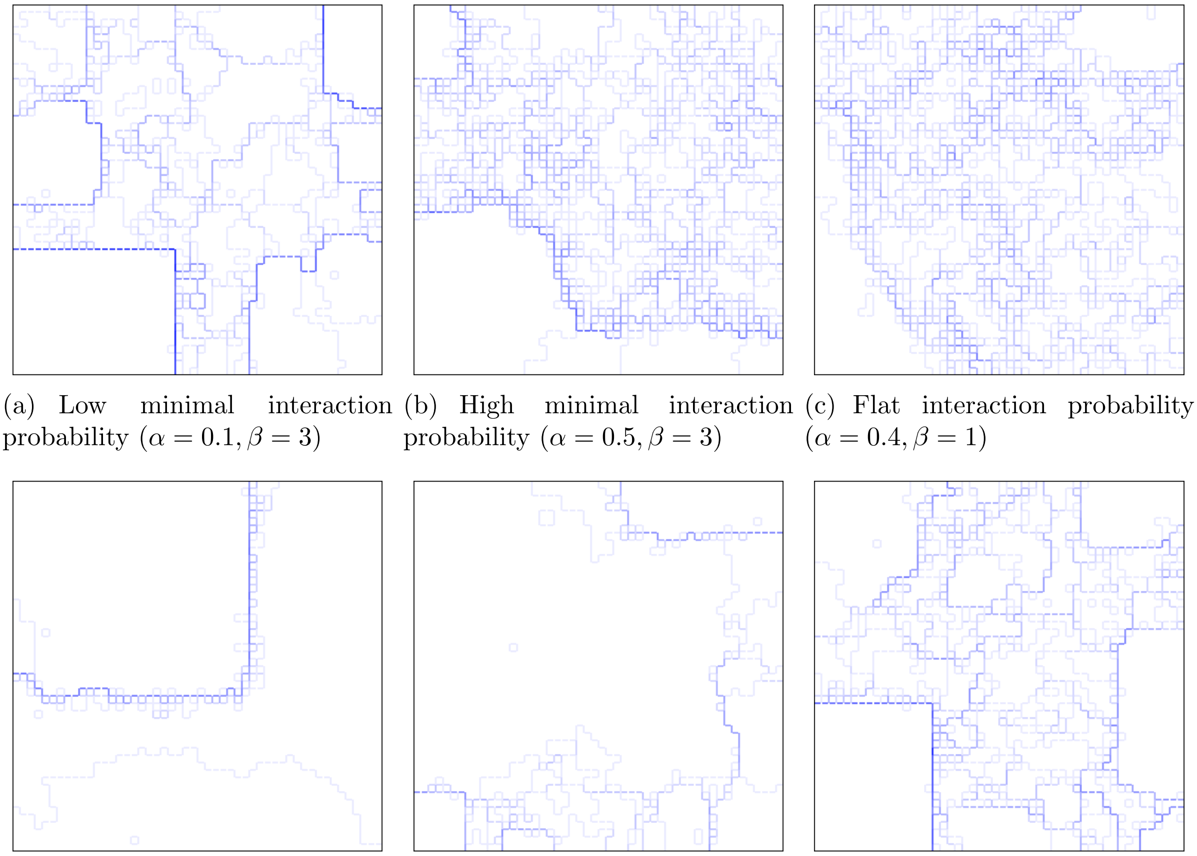
Population size was set to 2500 individuals, implying a space size of $50 \cdot 50$ points. The feature space size was key constant to 16 binary possibilities.

Increasing the minimal interaction probability $\alpha$ resulted in less contained groups and a more freely flowing language continuum, as can be seen in figures 2(a) and 2(b). Increasing the drop-off speed actually decreases the amount of language fading, and increases the amount of language borders, as can be seen in figures 2(c) and 2(d).

A higher minimal adoption probability unsurprisingly increases the amount of variety in the system. The slope of the adoptation probability is reflected in the 'sharpness' of the borders; a steeper slope will result in clearer borders, as can be seen in figures 2(e) and 2(f).

To test the usefulness of the adoption probability (and thus, the social impact theory), a null-model was devised. In this case, the adoption probability function was discarded in favor of a constant value. This model was tested under a wide variety of parameters, but never was the model able to maintain variety. This can be explained by the same experience Axelrod (1997) had. He calls the effect 'random walk with absorbing barriers', which means that in a situation where there is enough room to mutate, the less used mutation will probably be absorbed by the more frequent one.

Surprisingly, the agent death had a big influence on the model. Disabling agent death and birth, and always allowing agents to learn from each other, the results changed dramatically. The model was not able to generate any kind of language border. Neither is there any form of group creation. This can be explained by the fact that agent death has a stabilizing function in the system, by temporarily introducing non-changing agents, as well as allowing the model to converge, by introducing new agents that speak the same language as its immediate neighbors. Figure 3 shows cases of the experiment without agent death, for both a high and a low mutation introduction chance. As can be seen, under both conditions there is no grouping.



(a) Low minimal interaction probability ($\alpha = 0.1, \beta = 3$)

(b) High minimal interaction probability ($\alpha = 0.5, \beta = 3$)

(c) Flat interaction probability ($\alpha = 0.4, \beta = 1$)

(d) Steep interaction probability ($\alpha = 0.4, \beta = 11$)

(e) Flat adoption probability ($\gamma = 0.5, \kappa = 0.3$)

(f) Steep adoption probability ($\gamma = 0.5, \kappa = 0.5$)

Figure 2: Comparison of conditions under which realistic diversity emerges. Shown are the agent spaces (50x50 agents). Dark lines indicate large distances between agents languages, light lines small distances. The middle frame shows a simulation with both preference for rare linguistic features and agent birth and death. Note the emergence of both language borders and dialect continua. The left frame shows a simulation without a preference for rare features, and the right frame a simulation without agent birth and death. Only uninteresting variation emerges.

## 4 Conclusion

Given the right parameter settings, this model results in linguistic (or cultural, depending on how one would interpret the feature strings) diversity that shows both dialect continua and language borders. Dialect continua appear almost for all parameter settings, but language borders only appear when the probability of accepting changes drops sufficiently quickly with the frequency of these changes. In other words agents must have a preference for rare utterances.

Interestingly, it also appears necessary that there is a flux of agents in the population. If agents do not die and are replaced with younger languageless agents, interesting linguistic diversity does not develop. This is an interesting result, which cannot be compared to earlier models: Nettle (1999b,a); Livingstone (2002) all used agent death in their models, but failed to include data on the case without death.

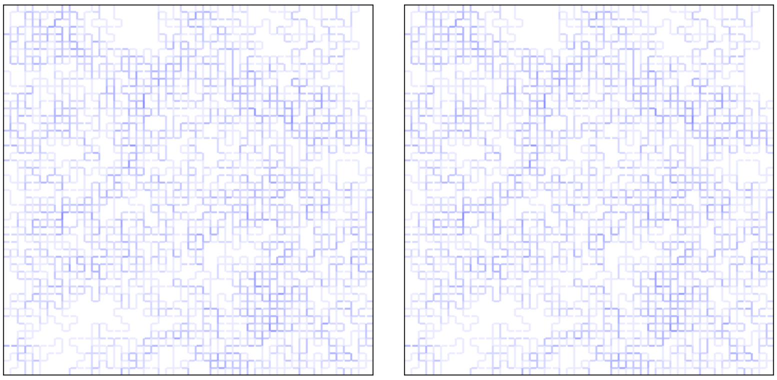


Figure 3: Results of experiment runs without agent death.

Although the analysis of these results is still in progress, and more experiments are necessary to clarify the exact circumstances under which diversity emerges in the model, they nevertheless show that under realistic conditions diversity can emerge without barriers and without unrealistic assumptions about social influence of single individuals.

**References**

Robert Axelrod. The dissemination of culture: A model with local convergence and global polarization. *The Journal of Conflict Resolution*, 41(2):203–226, apr 1997. ISSN 0022-0027.

Dale J. Barr. Establishing conventional communication systems: Is common knowledge necessary? *Cognitive Science*, 28(6):937–962, November-December 2004. doi: 10.1016/j.cogsci.2004.07.002.

B. Latané. The psychology of social impact. *American Psychologist*, 36(4):343–356, 1981.

Daniel Livingstone. The evolution of dialect diversity. In Angelo Cangelosi and Domenico Parisi, editors, *Simulating the Evolution of Language*, chapter 5, pages 99–118. Springer Verlag, London, 2002.

Makoto Matsumoto and Takuji Nishimura. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 8(1):3–30, 1998.

D. Nettle. *Linguistic Diversity*. Oxford University Press, 1999a.

Daniel Nettle. Using social impact theory to simulate language change. *Lingua*, 108(2-3): 95–117, June 1999b. doi: 10.1016/S0024-3841(98)00046-1.

Marco Patriarca and Teemu Leppanen. Modeling language competition. *Physica A: Statistical Mechanics and its Applications*, 338(1–2):296–299, July 2004. doi: 10.1016/j.physa.2004.02.056.

S.F. Railsback, S.L. Lytinen, S.K. Jackson, and J.S. Computing. Agent-based Simulation Platforms: Review and Development Recommendations. *Simulation*, 82(9):609, 2006.

# Independence and Decision-Contexts for Non-Interference Conditionals[*]

Michael Franke
Institute for Logic, Language and Computation
Universiteit van Amsterdam

m.franke@uva.nl

**Abstract**

Non-interference conditionals are conditional sentences whose antecedent and consequent are not conditionally related, e.g.: "If you are hungry, there are biscuits on the shelf." Previous accounts struggled to explain two phenomena in connection with these sentences. Firstly, non-interference conditionals seem to imply their consequents. This can be explained, if we evaluate conditionals with respect to the epistemic state of a speaker and spell out what it means to say that a speaker believes in the conditional independence of propositions. Secondly, non-interference conditionals seem to be conditional assertions. This, too, can be explained even with a standard semantics for conditionals, if we employ a rich context-model, i.e. evaluate propositional information against the background of a decision problem of the hearer.

## 1 Non-Interference Conditionals

During one of his recorded shows the American comedian Demetri Martin told the following joke, much to the amusement of his audience:

> She was amazing. I never met a woman like this before. She showed me to the dressing room. She said: "If you need anything, I'm Jill." I was like: "Oh, my God! I never met a woman before with a *conditional identity*." [Laughter] "What if I don't need anything? Who are you?" — "If you don't need anything, I'm Eugene." [More laughter]
> (Demetri Martin, *These are jokes*)

Martin's joke is possible because of a peculiarity of certain conditional sentences. Some conditional sentences relate propositions that have no conditional relationship. This is by no means contradictory or paradoxical. The sentence

(1) If you need anything, I'm Jill.

links the clauses "you need anything" and "I'm Jill" in a conditional construction, but semantically we may naturally perceive the propositions expressed by these clauses as conditionally unrelated; the name of the woman does not depend on whether the addressee needs anything or not. To humorously misapprehend such conditional sentences, as Martin does, is to pretend to see a conditional relationship where none exists.

[*] Thanks to Robert van Rooij, Sven Lauer, Anton Benz and Tikitu de Jager for help and discussion.

Examples that would lend themselves to similar joking have been discussed as a special case of conditionals from a variety of angles under a variety of names (see Siegel (2006) for a recent treatment and more references). I will speak of *non-interference conditionals* (NICs), because of the intuitive conditional independence of antecedent and consequent in such sentences. However, not all sentences that are NICs in this sense are exactly alike. A reasonable distinction is given by Günthner (1999) based on a corpus study of spoken German: she lists meta-communicative conditionals (2a) and discourse-structuring conditionals (2b), next to relevance conditionals like (1).

(2) a. If I'm honest, I actually like Heidegger.
b. If we now turn to the last point of order, the fund cuts have been tremendous.

Meta-communicative conditionals like (2a) have antecedents which relate in some fashion to communicative rules or the actual linguistic conduct of the speaker. All sorts of politeness hedging would file under this heading. Witness, for instance, phrases like "if I may say so", "if you ask me" or "if I may interrupt". Discourse-structuring conditionals like (2b) seem to introduce or change a topic. These are, perhaps, the least likely variety to be found in written language. If spoken, they show a characteristic pause after the *if*-clause, which nearly exclusively occurs sentence-initially. Relevance conditionals like (1) and (3) are constructions whose antecedent intuitively gives a sufficient — and often also necessary — condition for the *relevance* of the information given in the consequent.[1]

(3) a. If you are are hungry, there are biscuits on the shelf.
b. If you want some, there are biscuits on the shelf.

It is clear that NICs are quite unlike ordinary conditionals such as (4).

(4) a. If it does not rain, we will eat outside.
b. If the butler has not killed the baroness, the gardener has.

Still, NICs are very natural things to say; such sentences are frequent in spontaneous speech and this for good reasons. One consequence of this is that it is easy to imagine a lively context of utterance with rather distinct properties when confronted with naked examples of NICs even in the laboratory context of a linguistic paper such as this. Consequently, the main idea behind this treatment of NICs is to explain some of the characteristic differences between NICs and ordinary conditionals *pragmatically*, and that is to say not in terms of, e.g., an assumed grammatical difference between kinds of conditionals, but rather in terms of (sharpened formalizations of) our rich and solid intuitions about the contexts in which NICs would naturally be used. The thesis to be advanced and defended here is that NICs can be given a standard semantic analysis, yet still their elaborate contextual meanings can

[1] Strictly speaking, two more kinds of sentences are NICs in the given sense: for one, there are so-called monkey's-uncle conditionals like (i) with an obviously false consequent as an emphatic way to deny or reject the antecedent; for another, there are concessive conditionals with an 'even-if' reading like (ii).

(i) If that's true, I'm a monkey's uncle.

(ii) That match is wet. If you strike it, it won't light.

For the present discussion, both of these are of no interest. I exclude them when I speak of NICs.

be explained. We just have to make proper use of our understanding of the contexts in which NICs would be used.

What are then the differences between NICs and ordinary conditionals? One frequently discussed difference is this: NICs seem to imply the truth of their consequent proposition. For instance, any relevance conditional in (3) somehow conveys that there are indeed biscuits on the shelf, no matter whether this information is relevant or not. Of course, this is different for ordinary conditionals like (4) and therefore a theory of NICs should explain how the feeling of entailment of the consequent comes about. The main aim of section 2 is to account for this perceived entailment relation.

Another peculiar meaning intuition to be taken seriously and explained is the following. Somehow NICs appear very much like *conditional assertions*: not the truth of the consequent but the assertion thereof depends on the antecedent. If the antecedent is false, then, intuitively, the speech-act associated with the main clause is somehow not feasible: for meta-communicative conditionals (2a) some politeness rule or felicity condition would be infringed; for discourse-structuring conditionals (2b) the assertion would be off-topic; for relevance conditionals (1) and (3) the assertion would be irrelevant. The raison d'être of NICs is therefore, the intuition continues, to protect speakers from acting inappropriately; speakers perform speech-acts only conditionally, if they are uncertain whether these acts are feasible.

Many theories of NICs have taken this intuition at face value. In order to explain how a conditional sentence like (3a), for instance, can be interpreted as a conditional assertion, linguists and philosophers have variously advanced theories that, in crude outline, either postulate an elliptical performative (5a) or some abstract illocutionary force operator (5b).[2]

(5) a. If you are hungry, (I hereby say to you that) there are biscuits on the shelf.

b. If you are hungry, ASSERT("there are biscuits on the shelf").

Even when we neglect the intricacies of individual proposals, their merits and flaws, it is safe to say what is unappealing about any such account. For one, conditional speech-acts, if taken seriously, are very peculiar entities — where else in life do you perform your actions conditionally? — whose properties can only be assessed via exactly those sentences' meanings whose meaning they are to explain. For another, any account that strives to analyze NICs as conditional assertions along the lines of (5), even if that account is more carefully elaborated as in Siegel (2006), postulates hidden linguistic structure of some kind, be that performatives, speech-act operators or else. Clearly, where possible, a simple standard semantics of conditional sentences would be preferred. I argue in section 3 that we can treat NICs as standard conditionals and still account for the intuition that assertions are somehow conditionalized, as long as we assume a rich enough notion of context.

[2] A particularly outspoken proposal of the performative analysis in (5a) is van der Auwera (1986). The perhaps most extreme version of the conditional speech-act hypothesis is the theory proposed by DeRose and Grandy (1999) who argue that *all* conditionals are such conditional assertions.

## 2 Epistemic Independence

At the outset we characterized NICs as conditional surface structures that relate clauses whose meanings are intuitively not conditionally related. I would like to try to pin down what it means for two propositions not to be conditionally related. Intuitively, some cases are clear: normally we would not expect the truth or falsity of propositions such as

you are hungry ($P$) & there are biscuits on the shelf ($Q$)

to depend on one another. These two propositions should, in some sense, be independently true or false. But where? Not in the actual world where it is fixed whether $P$ and whether $Q$. But perhaps rather in the mind of an agent: it is very plausible to assume that under normal circumstances a rational agent simply does not believe in any conditional connection between these propositions. A belief in a conditional connection between $P$ and $Q$ would be a belief that links the truth or falsity of one proposition to the truth or falsity of the other. Put the other way around, we might then say that $P$ and $Q$ are *epistemically independent* for an agent (in a given epistemic state) if learning one proposition to be true or false (where this was not decided before) is not enough evidence to decide whether the other proposition is true of false (where this was not decided before).

To make this idea more intelligible, gentle formalization helps. Take a set $W$ of possible worlds, propositions $P, Q \subseteq W$ and an agent's epistemic state $\sigma \subseteq W$ of worlds held possible. We write $\overline{P}$ for $W \setminus P$, the negation of proposition $P$. We say that the agent holds $P$ possible and write $\Diamond_\sigma P$ or, dropping the obvious index, $\Diamond P$ iff $\sigma \cap P \neq \emptyset$. Now, suppose we assume a strict conditional analysis and say that a conditional $P \rightarrow Q$ is held true in $\sigma$ iff $\sigma \cap P \subseteq Q$. We could then say that an agent does not believe in any sort of conditional relationship between $P$ and $Q$ if he does not hold true any conditional $X \rightarrow Y$ for $X \in \{P, \overline{P}\}$ and $Y \in \{Q, \overline{Q}\}$. This would mean that an agent does not see a conditional relationship between $P$ and $Q$ iff

$$\forall X \in \{P, \overline{P}\} \; : \; \forall Y \in \{Q, \overline{Q}\} \; : \; \Diamond(X \cap Y). \tag{1}$$

This is indeed the standard notion of *logical independence* between propositions. For our purposes, however, this standard notion is too strong, for it commits the agent to believe in the possibility of $P$, $\overline{P}$, $Q$ and $\overline{Q}$. This does not seem necessary: I may know that there are biscuits on the shelf, but still doubt any conditional relationship between that fact and your appetite. (We will soon see another formal reason why we should relax this criterion.)

So here is another attempt. We said that, for $P$ and $Q$ to be epistemically independent, learning that $P$ is true or false should not decide whether $Q$ is true or false. So, for $X \in \{P, \overline{P}\}$ and $Y \in \{Q, \overline{Q}\}$, if $Y$ is held possible, then it should also be held possible after $X$ is learned. That is to say:

$$\Diamond Y \rightarrow \Diamond(X \cap Y).$$

But again this only holds if $X$ itself is at least possible, for otherwise, we again smuggle in a belief in the possibility of $X$ via the definition of epistemic independence. We also aim for symmetry, so we take

$$(\Diamond X \wedge \Diamond Y) \rightarrow \Diamond(X \cap Y) \tag{2}$$

as our intuitive definition of epistemic independence of propositions $P$ and $Q$: if $P$ was not believed true or false before, then learning $P$ is not enough to establish a belief in $Q$ or $\overline{Q}$ where there was no such belief before.

## 2.1 Epistemic Independence and Conditionals

We noted as an explanandum that NICs seem to entail their consequent proposition. We are now in a position to account for this intuition. Let's boldly assume a material or strict implication analysis of conditionals even for NICs, but let's evaluate the conditionals on the epistemic state $\sigma$ of a speaker. So, if a speaker says 'If $P$, $Q$', we may infer that, if he spoke truthfully, his epistemic state is such that $\sigma \cap P \subseteq Q$. But if we have reason to assume that at the same time the same speaker actually does not believe in a conditional relationship between $P$ and $Q$, we may infer even more, namely that the speaker either believes in the falsity of $P$ or the truth of $Q$. This is so, because if $\Diamond P$ and $\Diamond\overline{Q}$, then by epistemic independence we have $\Diamond(P \cap \overline{Q})$ which contradicts $\sigma \cap P \subseteq Q$. Consequently, if we furthermore have reason to assume that the speaker considers it at least possible that the antecedent proposition is true, as seems uncontroversial for all cases of NICs that I have seen so far, we may conclude that the speaker actually believes $Q$. Whence, I propose, the feeling of entailment: a speaker who (i) speaks truthfully in asserting 'If $P$, $Q$', (ii) considers $P$ and $Q$ epistemically independent and (iii) considers $P$ at least possible *must* believe in $Q$.

## 2.2 Epistemic Independence and Probabilistic Independence

It is perhaps necessary to address an obvious worry. It might seem as if the notion of epistemic independence as defined in (2) appears out of thin air; an arbitrary formal tool shaped and designed to do exactly what we want it to. This is not so. In fact, it is the exact non-probabilistic counterpart of the notion of probabilistic independence of events in probability theory. In probability theory two events or propositions $P$ and $Q$ are said to be *probabilistically independent* given a probability distribution $\Pr(\cdot)$ iff

$$\Pr(P \cap Q) = \Pr(P) \times \Pr(Q). \tag{3}$$

If we identify the probability distribution $\Pr(\cdot)$ with the agent's subjective probabilistic beliefs over worlds $W$

$$\Pr : W \to [0;1], \quad \Pr(P) = \sum\nolimits_{w \in P} \Pr(w), \quad \Pr(W) = 1$$

and therefore equate the epistemic state $\sigma$ of the agent with the support of the probability distribution $\Pr(\cdot)$ as usual

$$\sigma = \{w \in W \mid \Pr(w) \neq 0\}$$

we can show that probabilistic independence of propositions $P$ and $Q$ as in (3) entails epistemic independence as in (2).[3]

[3] Proof: First, we establish that if $\Pr(P \cap Q) = \Pr(P) \times \Pr(Q)$, then for arbitrary $X \in \{P, \overline{P}\}$ and $Y \in \{Q, \overline{Q}\}$ it holds that $\Pr(X \cap Y) = \Pr(X) \times \Pr(Y)$. From the three arguments needed, it suffices to give just one, as the others are similar. So assume that $\Pr(P \cap Q) = \Pr(P) \times \Pr(Q)$ and derive

The converse, however, is not the case. Epistemic independence does not entail probabilistic independence. It may be the case that proposition $P$ is not enough (evidence, support, information) to decide whether $Q$ is true or false, but still learning that $P$ is true, for instance, makes $Q$ more or less likely. That is to say that the suggested notion of epistemic independence is indeed the exact non-probabilistic counterpart of probabilistic independence: we only care about believing and holding possible, not likelihoods of propositions.

This makes for a further argument why the stronger notion of logical independence of propositions in (1) may indeed be deemed too strong also on formal grounds. Instead of defining probabilistic independence of $P$ and $Q$ as in (3) we could, perhaps, have required that for $X$ and $Y$ as before

$$\Pr(X|Y) = \Pr(X) \quad \& \quad \Pr(Y|X) = \Pr(Y). \tag{4}$$

This is only defined if all $X$ and $Y$ have non-zero probability. This stronger notion entails, but is not entailed by the standard notion of probabilistic independence given in (1). It is obvious that logical independence as defined in (1) is the non-probabilistic counterpart of this stronger, non-standard definition of probabilistic independence in (4), in the same sense that epistemic independence (2) is the non-probabilistic counterpart of standard probabilistic independence (3).[4]

## 3 Information in Decision Contexts

It remains to be explained why the antecedents in NICs apparently serve to classify or conditionalize a speech-act whose felicity, in particular its relevance, is guaranteed only under the premiss that the antecedent is true. As stated earlier, I would like to defend that we do not have to assume implicit performatives or abstract speech-act operators in the semantics to account for this intuition. I can only sketch my arguments for this thesis in rough outline here. The brief indications I will give apply in an obvious manner to relevance conditionals like (1) and (3), but do not extend directly to meta-communicative (2a) or discourse-structuring conditionals (2b). Still, I believe that a similar story can be told for these too.

On a certain level of abstraction, we may think that most, if not all, reasonable communication takes place against the background of a decision problem of the hearer. A *decision problem* is a quadruple $\delta = \langle W, u, \Pr, Act \rangle$ where $W$ is a set of relevant situations or world states, $u$ is a utility function from $W$ to reals, $\Pr(\cdot)$ is a probability distribution over $W$ and $Act$ is a partition of $W$ (a finite set of propositions) representing the hearer's possible future actions. We can be very liberal interpreting what actions are. Actions may be very abstract *epistemic actions*: interpretations, here conceived as the (deliberate) adoption of

that $\Pr(P \cap \overline{Q}) = \Pr(P) \times \Pr(\overline{Q})$: $\Pr(P \cap \overline{Q}) = \Pr(P) - \Pr(P \cap Q) = \Pr(P) - (\Pr(P) \times \Pr(Q)) = \Pr(P) \times (1 - \Pr(Q)) = \Pr(P) \times \Pr(\overline{Q})$. Next, assume that $\Pr(X \cap Y) = \Pr(X) \times \Pr(Y)$ and that $\Diamond P$ and $\Diamond Q$. That means that $\Pr(X), \Pr(Y) > 0$. Hence, $\Pr(X \cap Y) > 0$, which is just to say that $\Diamond(X \cap Y)$. $\Box$

[4] A final remark in defense of the proposed notion of epistemic independence: van Rooij (2007) noticed that epistemic independence is equivalent to Lewis's (1988) notion of orthogonality of questions whether $P$ and whether $Q$. In the same paper, van Rooij uses epistemic independence/orthogonality to account for the strengthening of conditional presuppositions to unconditional ones. The underlying idea is the same as the one presented here.

beliefs about what was meant with an utterance. Thus conceived we may think of a background decision problem as an abstraction over the notion of a question under discussion: reasonable talk exchange addresses an issue "what shall I do?" of which a special case is the question "what shall I believe?".

In a decision problem $\delta$, a rational agent is predicted to choose an action $A \in \mathit{Act}$ that maximizes expected utility which is defined as

$$\mathrm{EU}(A) = \sum\nolimits_{w \in A} \Pr(w) \times u(w).$$

Information that proposition $P$ is true is processed via Bayesian update to yield an updated decision with $\Pr(\cdot|P)$ instead of $\Pr(\cdot)$. The expected utility of an action $A$ after learning that $P$

$$\mathrm{EU}(A|P) = \sum\nolimits_{w \in A} \Pr(w|P) \times u(w).$$

may differ from $\mathrm{EU}(A)$, of course. Also, the actions with the highest expected utility before learning and after learning need not be identical. This is neither new nor surprising. It simply means that in a concrete decision problem mere information flow may have a particular impact on the decision of the hearer, i.e. relate to the question under discussion.

The interesting question then is, when and how is $P$ *relevant* to the decision problem $\delta$? In other words, when does $P$ count as an answer to the question under discussion? I suggest a weak notion of answerhood: $P$ is relevant to decision problem $\delta$ iff it is a non-trivial argument for some action in $\delta$. For formalization of the notion of an argument, define the *change of expected utility* of action $A$ given $P$ as

$$\mathrm{CEU}(A, P) = \mathrm{EU}(A|P) - \mathrm{EU}(A)$$

and say that $P$ is an argument for $A$ iff

(i) $\forall B \in \mathit{Act}\ \mathrm{CEU}(A, P) \geq \mathrm{CEU}(B, P)$ and

(ii) $\exists B \in \mathit{Act}\ \mathrm{CEU}(A, P) > \mathrm{CEU}(B, P)$.

Equivalently, we could say that $P$ is relevant iff there are actions $A, B$ in $\mathit{Act}$ such that $\mathrm{CEU}(A, P) \neq \mathrm{CEU}(B, P)$. Yet the lengthier formulation makes clear in which sense relevant information is an answer to the question under discussion: if $P$ is relevant it sets apart at least one action.

In real life we would certainly not want to assume that a speaker at all times knows exactly what drives a hearer, what he is concerned with and what he wants. In other words, the background decision problem of the hearer may at times not be known by the speaker, either in its entirety or in some relevant detail. Consequently, it may not always be the case that a piece of information is relevant to all the decision problems that the speaker might consider possible issues for the hearer. Similarly, a hearer may not be sure in the context of which speaker-conceivable decision problem he should interpret statements by the speaker. Enter relevance conditionals like (1) or (3).

If I'm in a shop, the information that the name of the woman helping me find the dressing room is Jill may be relevant in a million ways. (Does she want me to ask for her phone number?) If you pay me a visit and I just say out of the blue that there are biscuits on the shelf, your most natural reaction is perhaps a stunned: "So what?". Yet, with the NICs in

(1) and (3) the case is entirely different. We now know much better in which way we ought to process the information in the consequent, in which sense it is relevant.

How can this intuition be made a little more precise? Suppose that the hearer who interprets the NIC "If $P$, $Q$" adds to his stock of knowledge, Ramsey-style, the proposition $P$. He realizes that the information that $Q$ is true is not restricted to $P$-worlds, because $P$ and $Q$ are epistemically independent etc. But he also realizes that the speaker may believe that all the $P$-worlds are worlds in which the hearer has a particular decision problem. That means that the hearer, after adopting $P$ for the evaluation of $Q$, is in a particular context in which, plausibly, the information $Q$ is indeed relevant: if you are hungry, the information that there are biscuits on the shelf simply increases the expected utility of one action (going over and having some biscuits) more than some other action (ordering a pizza) and therefore is an argument for that action, hence relevant for that decision problem, no matter what exactly the beliefs and utilities of the hearer are.

Though sketchy, this outline should make clear that we can dispense with conditional assertions in the case of relevance conditionals. Instead of saying that, for reasons of unclear relevance, the consequent proposition is asserted if (and, presumably, only if) the antecedent is true, we say that the antecedent (dynamically, if you wish) takes us to a context in which the consequent proposition is relevant. Being relevant then means: relating or arguing to a point. The twist is that relevance does not apply to assertions, but to information.

So what if the antecedent is false? In that case we predict that still a normal conditional has been asserted from which the hearer learns via independence that the consequent proposition is true, but this information may not be relevant, or it is relevant in a different (kind of) decision problem where it relates to and argues for a different point. Hence, by indicating which context the consequent information is to be understood in the speaker expresses that this is the effect (think: point, argument) of the information in the consequent; if the necessary context is not actual, this effect simply doesn't come about. What sounds like a conditional speech-act, is just information processed in the context of a decision problem.

## References

Johan van der Auwera. Conditionals and speech acts. In Elizabeth Closs Traugott, Alice ter Meulen, Judith Schnitzer Reilly, and Charles A. Ferguson, editors, *On Conditionals*, pages 197–214. Cambridge University Press, 1986.

Keith DeRose and Richard E. Grandy. Conditional assertions and 'biscuit' conditionals. *Noûs*, 33(3):405–420, 1999.

Susanne Günthner. Wenn-Sätze im Vor-Vorfeld: Ihre Formen und Funktionen in der gesprochenen Sprache. *Deutsche Sprache*, 3:209–235, 1999.

David Lewis. Relevant implication. *Theoria*, 54:162–174, 1988.

Robert van Rooij. Strengthening conditional presuppositions. To appear in Journal of Semantics, 2007.

Muffy E. A. Siegel. Biscuit conditionals: Quantification over potential literal acts. *Linguistics and Philosophy*, 29:167–203, 2006.

# Horn strategies and optimization in Russian aspect[*]

Atle Grønn
Dept. of European Languages (ILOS)
University of Oslo

atle.gronn@ilos.uio.no

### Abstract

I propose to explain the diachronic development of the aspectual system in Russian in terms of Horn strategies (partial blocking). The analysis is based on Blutner's bidirectional OT, which has a strong diachronic dimension. I also suggest that a context-sensitive version of bidirectional OT can play a role in explaining the synchronic situation for Russian aspect.

## 1 Horn strategies in the lexicon

Standard examples of Horn's iconicity principle – (un)marked forms are mapped with (un)marked meanings – come from lexical pragmatics (Horn, 1984), (Blutner, 1998). In the spirit of Blutner (2000) and his implementation of partial blocking in bidirectional OT, Dekker and van Rooy (2000) came up with graphical representations like the following:

$$\begin{tikzcd} & m_1 & m_2 \\ f_1 & \bullet & \circ \arrow[l] \\ f_2 & \circ \arrow[u] & \bullet \arrow[l] \arrow[u] \end{tikzcd}$$

Table 1: (Weakly) optimal pairs in a 2x2 game

The speaker – i.e. the vertical arrows – has a preference for short, unmarked forms (e.g. 'kill' ($f_1$) > 'cause to die' ($f_2$)) and the hearer – i.e. the horizontal arrows – prefers stereotypical, unmarked meanings (e.g. direct killing ($m_1$) > indirect killing ($m_2$)). Horn's division of pragmatic labor predicts that from two forms with underspecified meanings, the simple form 'kill' is mapped with the meaning of direct/canonical killing, while the more complex form 'cause to die' will be used to denote an indirect killing. The latter is demonstrated in (1), which suggests that the sheriff's death came about after some unusual event, perhaps an accident.

(1) Black Bart caused the sheriff to die.

The algorithm of weakly bidirectional OT (Jäger, 2002) starts with the optimal pair $<f_1,m_1>$ (●). Both pairs which are beaten by this pair in *one* direction (○) are then removed from the tableau, and this leaves us with the *weakly* optimal pair $<f_2,m_2>$ (●). The weakly optimal

[*] I would like to thank Hanne Eckhoff and Dag Haug for their comments on the diachronic data.

pair survives despite the presence of the optimal pair. True, there is a better form ($f_1$), but *not* given meaning $m_2$. Similarly, there is a better meaning ($m_1$), but *not* given form $f_2$.

The Horn strategy (partial blocking) is arguably rational for the speaker and hearer, and represents the optimal way of resolving two conflicting interests in communication: economy (the speaker's I/R principle in neo-Gricean pragmatics) vs. diversification (the hearer's Q-principle). In this paper, I will discuss whether this kind of rationality plays any role in the development of a grammatical category: viewpoint aspect in Russian, that is, the opposition between imperfective aspect (IPF) and perfective aspect (PF). Aspect in Russian is morphologically independent of tense and finiteness and occurs obligatorily in all tense/mood configurations. In this respect, Russian aspect is a lexical phenomenon.

## 2 The meaning of aspect

There is a close relationship between (a)telicity, i.e. lexical aspect, and grammatical aspect in Russian. PF is indeed almost exclusively restricted to telic predicates (Vendlerian accomplishments/achievements). At the same time, the Russian IPF:PF distinction, like the French aspectually loaded past tenses Imparfait:Passé simple, is semantically speaking a viewpoint aspect. For instance, accomplishments can be denoted by imperfective verbs with a progressive interpretation (giving rise to the famous imperfective paradox). For the purposes of the following discussion, I distinguish between three meanings, which form a continuum:

- $m_0$: atelic activities
- $m_1$: (telic accomplishments, where) the speaker's viewpoint is restricted to the preparatory process
- $m_2$: (telic accomplishments, where) the speaker's viewpoint includes the end point of the event

For short, I will refer to $m_0$ and $m_1$ – which both, in a certain sense, satisfy the subinterval property – as *incomplete event interpretations*, while $m_2$ represents a *complete event interpretation*. Thus, so far, we have two forms, IPF and PF, which are to be mapped with three meanings $m_0$, $m_1$ and $m_2$. But let's start with the beginning.

## 3 The emergence of aspect

Bidirectional OT is a powerful explanatory principle in diachronic linguistics, since pragmatic bidirectionality creates special interpretations that can become conventionalized (Blutner and Zeevat, 2004). Many synchronic semantic and syntactic facts can be analyzed from an evolutionary perspective as 'frozen pragmatics' (Blutner, 2006). This strong diachronic dimension of BiOT suggests that we should start with the emergence of the aspectual system itself, i.e. the PF:IPF opposition which came to replace the old Indo-European tenses Aorist:Imperfect.

The earliest Russian texts (Old Church Slavonic, Old Russian) contain morphologically simplex verbs, such as *čitat'* – *to read.* Used intransitively, simplex verbs ($f_0$) denote atelic activities. In its transitive use, with a quantized object, the VP ($f_1$) denotes an accomplishment, just like its Germanic counterparts: *čitat′ pis′mo – to read the letter*. However, accomplishments can also be referred to by a prefixed verb ($f_2$), which in this case is formed by adding the preposition/prefix *pro – through* to the simplex verb: *pročitat′ pis′mo – to read through the letter* (lit. 'through-read'). In hindsight, we know, of course, that the emerging form-meaning pairs are the following:

| | | |
|---|---|---|
| čitat′ ($f_0$) | ⇔ | atelic activities ($m_0$) |
| čitat′ pis′mo ($f_1$) | ⇔ | 'progressive' accomplishments ($m_1$) |
| pročitat′ pis′mo ($f_2$) | ⇔ | 'non-progressive' accomplishments ($m_2$) |

Table 2: The interpretation of the Russian VPs 'read (through) (the letter)'

Is this 1-1 mapping a result of bidirectional optimization?

In terms of complexity of forms, the following ranking suggests itself: $f_0 > f_1 > f_2$. Since $f_1$ and $f_2$ are obviously excluded from being paired with $m_0$, and $m_0$ is presumably the simplest meaning from a conceptual point of view, the pairing $<f_0,m_0>$ is straightforward semantics and requires no further pragmatic/optimality theoretic reasoning. The question is whether we can find a principled explanation for why $f_1$ is mapped to $m_1$, and $f_2$ to $m_2$.

### 3.1 Underspecification

In BiOT, a prerequisite for partial blocking to obtain, is that the candidate forms start out with an underspecified semantics (but see (Benz, 2006) for a critical discussion of this point). I assume here that both $f_1$ and $f_2$ were originally compatible with both $m_1$ and $m_2$. The pair $<f_1,m_2>$ is still viable in contemporary Russian, where it is known as the *factual IPF* (Grønn, 2004):

(2) Ne vyzyvaet somnenij i to, čto Stalin sam **čital$^{IPF}$ pis′mo** Bulgakova. (google)
There is also no doubt that Stalin himself **read the letter** from Bulgakov.

On the other hand, after the process whereby prefixation turned into perfectivization (PF), the pair $<f_2,m_1>$ was no longer grammatical in Russian. However, initially, there is no reason to believe that the Russian VP *pročitat′ pis′mo* behaved differently from its direct counterpart in contemporary German:

(3) Als ich **den Brief durchlas**, den meine Freundin Katja für ihre Tochter Anna schrieb, musste ich weinen. Dieser Brief hat mich so zu Tränen gerührt, ich konnte es kaum aushalten! (google)
As I **was reading the letter** (lit.: 'the letter through-read'), which my friend Katja had written to her daughter Anna, I had to cry. This letter moved me to tears, such that I could hardly bear it!

Although the German VP in (3) typically denotes a complete event, this particular example shows that the prefixed transitive accomplishment predicate is compatible with an incomplete event interpretation. Why would the Russian equivalent *pročital pis′mo* develop into perfective aspect, i.e. loose its compatibility with an incomplete event interpretation? There are two conceivable lines of explanation – the Gricean and the neo-Gricean, which I will briefly present below.

### 3.2 Gricean reasoning

It has been argued that telic predicates by default have a ‘perfective’, complete event interpretation (Bohnemeyer and Swift, 2004). This gives rise to an implicature, which, at **stage 1**, can be exploited by the speaker, who relies on the hearer to infer the complete event interpretation of *pročitat′ pis′mo*. If communication is successful (repeatedly), the implicature can turn into an entailment at **stage 2**.[1]

This is Gricean, but not neo-Gricean reasoning. Nothing here hinges on the presence of $f_1$. The BiOT perspective could still play a role, but only at a later stage: Given the optimal pair $<f_2,m_2>$, the pair $<f_1,m_1>$ becomes weakly optimal at **stage 3**. One possible problem with this kind of analysis is that it blurs the intuitive relationship between $f_0$ and $f_1$. If there is no link between $f_0$ and $f_1$, and no interaction between $f_1$ and $f_2$ at stage 1 and 2, one may ask why the telic VP $f_1$ did not develop into a purely perfective marker.

On the other hand, if we assume that $f_1$ indeed inherits a feature, say [+subinterval property], from its close relative $f_0$, then we can apply the Gricean reasoning once more and abandon the OT perspective: $f_1$ is prototypically used with the implicature of $m_1$, which develops into an entailment. However, this is not satisfactory for the simple reason that IPF is not a grammaticalization of atelicity and/or the subinterval property, and $f_1$ does not entail $m_1$ (cf. example (2) above).

### 3.3 Neo-Gricean reasoning

Let’s now have a look at a neo-Gricean approach, where the diachronic change of *pročitat′ pis′mo* from a complete event implicature to an entailment is explained by looking at the prefixed verb not in isolation but in the competitive environment of forms and meanings.

According to (Benz, 2006), Horn’s derivation scheme proceeds roughly as follows:

1. The speaker used a marked expression f′ containing ‘extra’ material ... when a corresponding unmarked expression f, essentially coextensive with it, was available.
2. The ‘extra’ material must have been necessary, i.e. f could not have been appropriately used.
3. ... Therefore, the unmarked alternative f tends to become associated (by use or – through conventionalization – by meaning) with the unmarked situation m.
4. The marked alternative f′ tends to be associated with the complement of m with respect to the original extension of f/f′ ...

Benz points out some problems with Horn’s use of his own principle. Most importantly, it remains unclear why f′ had to be used in the first place, given that f and f′ were essentially

[1] Thanks to Dag Haug on this point.

coextensive. Benz argues that the procedure should start with the unmarked form being associated with the stereotypical situation m through *learning* – producing the optimal pair <f,m> at stage 1. In order to achieve this, Benz introduces associative learning as a third component (besides Zipf's principles of economy and diversification) in diachronic BiOT. Then, at the next stage, f′ is paired with m′ (weakly optimal pair).

Back to Russian aspect, there is one complication: why should $m_1$ be ranked as more stereotypical/normal than $m_2$? In fact, the above-mentioned generalization in (Bohnemeyer and Swift, 2004) suggests the opposite ranking for accomplishments: $m_2 > m_1$. If we keep the straightforward ranking on forms suggested by morphological complexity, $f_1 > f_2$, this seems to imply that Russian aspect develops from an anti-Horn strategy:

$$\begin{tikzcd}
 & m_2 & m_1 \\
f_1 & \circ & \bullet \arrow[l] \\
f_2 & \bullet \arrow[u] & \circ \arrow[l] \arrow[u]
\end{tikzcd}$$

Table 3: Anti-Horn strategy

The two winning form-meaning combinations in table 3 are marked with •. Although the anti-Horn strategy is evolutionary stable (van Rooy, 2004a), it is hardly attested in natural language. This is not surprising since this equilibrium is clearly more costly than the Horn strategy, where the shortest form is associated with the most frequent meaning. I think the problem with the above line of reasoning – which leads to the anti-Horn strategy – is the tacit assumption that competition at this stage is restricted to telic predicates with their progressive/non-progressive interpretations. A more appropriate ranking on forms and meanings would be the following, where we do not distinguish between $f_0/f_1$ and $m_0/m_1$:

- (in)transitive simplex verbs > prefixed verbs
- incomplete event interpretations > complete event interpretations

The outcome is now the expected Horn strategy in table 4:

$$\begin{tikzcd}
 & m_{0,1} & m_2 \\
f_{0,1} & \bullet & \circ \arrow[l] \\
f_2 & \circ \arrow[u] & \bullet \arrow[l] \arrow[u]
\end{tikzcd}$$

Table 4: Horn strategy for Russian aspect

Also from the point of view of associative learning (Benz, 2006), table 4 makes sense: At **stage 1**, the form $f_1$ is not perceived as aspectually different from $f_0$. The 1-1 mapping between $f_0$ and atelic activities is the external factor which triggers $f_1$ to be associated with incomplete events (progressivity, the subinterval property etc.). Through associative learning, the pair $<f_1,m_1>$ gets strengthened at **stage 2**. If the speaker then, at **stage 3**, wants to emphasize the complete event interpretation, he should choose the marked form $f_2$. At **stage 4**, this invites a strengthening of the pair $<f_2,m_2>$. Finally, at **stage 5**, prefixation develops into *perfectivization*, giving rise to a new aspectual system.

## 4 Markedness and secondary imperfectivization

As a result of the first round optimization described above, PF is grammaticalized. This changes the status of *pročitat′ pis′mo*, which is no longer underspecified, but gets a uniform, invariant semantics: <PF, complete event>. In BiOT, perfectivization leads to a reduction of GEN:

GEN = F × M – {<PF ($f_2$), incomplete event ($m_1$)>}

This creates a new situation. Even if we consider the emergence of the aspectual system to be bidirectionally optimal, it is far from clear that BiOT can play any role *after* grammaticalization.[2] What happens to the status of $f_2$ (now: PF) in terms of ranking on forms? From the application of BiOT in lexical pragmatics, we are used to treating lexicalized items as unmarked forms (e.g. 'kill'). Indeed, markedness theory is tricky (and controversial) in the domain of Russian aspect. I think we have to distinguish between a broad perspective and a narrow perspective: IPF is unmarked in the grammar of Russian (broad perspective), but PF is the default, most salient and frequent choice in the domain of accomplishments/achievements (narrow perspective).

The grammaticalization of PF creates a series of morphological gaps in the verbal paradigms. To take one example: the imperfective simplex verb *kryt′ – to cover* can be prefixed with *ot*, which has the basic meaning of motion/action away from a given point. The result is the perfective accomplishment verb *otkryt′ – to open (uncover)*. Clearly, in this case, the prefixed verb does not form an aspectual correlate to the simplex verb: their lexical semantics are not compatible. This gap is filled at **stage 6** by the productive morphological device of secondary imperfective suffixes, which in this case produces the following aspectual pair: *otkryt′*$^{PF}$ – *otkry**va**t′*$^{IPF}$.

Note that morphological complexity cannot any longer be the crucial factor for ranking of forms at stage 6 since this would not produce a linearly ordered ranking of PF and IPF: čitat$'^{IPF}$ ($f_0$) > pročitat$'^{PF}$ ($f_2$) > pročityvat$'^{IPF}$ ($f_3$).

## 5 Deblocking and context-sensitivity in modern Russian

In the optimality theoretic reasoning in the previous section, the pair <IPF, incomplete event> came out optimal or weakly optimal, depending on the choice of forms/meanings under consideration and the corresponding constraints and rankings. Is this situation confirmed by the synchronic data?

As argued in (Grønn, to appear), a complete event interpretation is never available for IPF in a context which licenses a progressive interpretation. Hence, the possibility of an incomplete event interpretation will effectively block the factual IPF. In examples like (2) above, the complete event interpretation of IPF is *deblocked* since no misunderstanding can

[2] van Rooy (2004b) notes some shortcomings of BiOT which in principle could be problematic for an application to Russian aspect. Consider the following situation: f is a lighter expression than f′, f > f′; and m′ is more stereotypical than m, m′ > m. If the meaning of f is underspecified, while f′ can only mean m′, BiOT predicts that m cannot be expressed. It turns out that the pair <f,m> is *not* weakly optimal. van Rooy argues that certain game-theoretical approaches to communication are better suited to handle such cases.

occur: The sentence and utterance context of (2) do not provide any reference times which could be used to 'zoom in' on the event. The interaction of past tense and aspect produces a temporal configuration, whereby the complete event is located at some proper subinterval of the whole past preceding the utterance time.[3]

Consider also the following example of aspectual competition and deblocking of the complete event interpretation of IPF:

(4) $$\text{Kto} \left\{ \begin{array}{l} \textbf{otkryl}^{PF} \\ \textbf{otkryval}^{IPF} \end{array} \right\} \text{okno?}$$

Who **opened** the window?

PF is the default and expected choice in (4). Importantly, however, the use of IPF in (4) still gives rise to a complete event interpretation since no competing incomplete/progressive interpretation is available for similar reasons as in (2). What we observe in these contexts, is in fact a second round of partial blocking following the deblocking of the factual IPF. If we isolate contexts which only allow for complete event interpretations[4], the partial blocking in table 5 emerges:

| | canonical complete events | | non-canonical complete events |
|---|---|---|---|
| PF | ● | ← | ○ |
| | ↑ | | ↑ |
| IPF | ○ | ← | ● |

Table 5: Partial blocking in contexts of deblocking of the factual IPF

A canonical complete event interpretation is arguably a complete event which produces a relevant *result*. Thus, if the window is open at the utterance time, the speaker should use PF in (4). As a result of pragmatic strengthening and polarization, the factual Ipf can have any complete event interpretation where the result state is not relevant or cancelled. A nice example is the use of IPF in (4), from which the hearer typically infers that the window is closed at the utterance time.

## 6 Conclusion

This paper is a first step towards an analysis of Russian aspect in bidirectional OT. The emphasis here has been on the diachronic process which leads to the grammaticalization of perfective aspect. Many issues are still open, such as the ranking on forms and meanings at various stages, and the question as to which forms/meanings should be considered competitors in each round of optimization. The synchronic situation also requires further research, notably concerning deblocking of the complete event interpretation of imperfective aspect.

The phenomenon of partial blocking raises the issue of diachronic vs. synchronic explanations. The two approaches may coexist in BiOT, as argued recently by Blutner (2006). A

[3] See (Grønn, 2004) for details.

[4] In order to model this properly we would need a context-sensitive version of BiOT, perhaps along the lines of (Benz, 2001). See also (Grønn, to appear) for some suggestions and discussion.

difference is still worth pointing out: When the Horn strategy applies in cases of deblocking, the coordination game of the speaker and hearer does not seem to be fully conventionalized. For instance, in (4), different non-canonical complete event interpretations of IPF may be possible. On the contrary, the division of labor which brought about perfective aspect became completely grammaticalized.

**References**

Anton Benz. Towards a framework for bidirectional optimality theory in dynamic contexts, *ms.*, 2001.

Anton Benz. Partial blocking and associative learning. *Linguistics and Philosophy*, 29: 587–615, 2006.

Reinhard Blutner. Lexical pragmatics. *Journal of Semantics*, 15:115–62, 1998.

Reinhard Blutner. Some aspects of optimality in natural language interpretation. *Journal of Semantics*, 17:189–216, 2000.

Reinhard Blutner. Embedded implicatures and optimality theoretic pragmatics. In T. Solstad, A. Grønn and D. Haug, editors, *A Festschrift for Kjell Johan Sæbø*, pages 11–29. Oslo, 2006.

Reinhard Blutner and Henk Zeevat. Editors' introduction: Pragmatics and optimality theory. In R. Blutner and H. Zeevat, editors, *Optimality Theory and Pragmatics*, New York, 2004. Palgrave Macmillan.

Jürgen Bohnemeyer and Mary Swift. Event realization and default aspect. *Linguistics and Philosophy*, 27(3):263–296, 2004.

Paul Dekker and Robert van Rooy. Bi-directional optimality theory: An application of game theory. *Journal of Semantics*, 17:217–242, 2000.

Atle Grønn. *The Semantics and Pragmatics of the Russian Factual Imperfective*, volume 199 of *Acta Humaniora*. Unipub, dr.art thesis, Oslo, 2004.

Atle Grønn. Russian aspect as bidirectional optimization, to appear.

Lawrence Horn. Towards a new taxonomy of pragmatic inference: Q-based and R-based implicature. In D. Shiffrin, editor, *Meaning, Form, and Use in Context: Linguistic Applications*, pages 11–42. Georgetown University Press, Washington, 1984.

Gerhard Jäger. Some notes on the formal properties of bidirectional optimality theory. *Journal of Logic, Language and Information*, 11(4):427–451, 2002.

Robert van Rooy. Evolution of conventional meaning and conversational principles. *Synthese (Knowledge, Rationality and Action)*, 139:331–366, 2004a.

Robert van Rooy. Signalling games select Horn strategies. *Linguistics and Philosophy*, 27: 493–527, 2004b.

# Selection-Mutation Dynamics of Signaling Games With Two Signals

Josef Hofbauer
Department of Mathematics, University of Vienna
Nordbergstraße 15, A-1090 Vienna, Austria

josef.hofbauer@univie.ac.at

Simon M. Huttegger
Konrad Lorenz Institute for Evolution and Cognition Research
Adolf Lorenz Gasse 2, A-3422 Altenberg, Austria

simon.huttegger@kli.ac.at

**Abstract**

In signaling games the replicator dynamics does not almost surely converge to states of perfect communication. A non-negligible portion of state space converges to components of Nash equilibria which characterize states of partial communication. Since these components consist of non-hyperbolic rest points, the significance of this result will depend on the dynamic behavior of specific perturbations of the replicator equations. This paper presents some results on the selection-mutation dynamics of signaling games, which may be considered as one plausible perturbation of the replicator dynamics.

## 1 Introduction

We will focus on signaling games with a very simple structure. The strategies of such a signaling game can be introduced by using the sets $E$ and $S$ where $|E| = |S| = n$. $E$ is the set of events and $S$ is the set of signals. A (pure) strategy of the sender is a function from $E$ into $S$ and can be represented as an $n \times n$ matrix $P$ with each row containing exactly one 1, the other entries being zero. Thus if $p_{ij} = 1$ then the sender sends signal $j$ given that state $i$ has occurred. Similarly, a strategy of the receiver is a function from $S$ to $E$ and will be represented by an $n \times n$ matrix $Q$ with $q_{ij} = 1$ for some $j$ and $q_{ik} = 0$ for all $k \neq j$. If $q_{ij} = 1$ then the receiver associates event $j$ with signal $i$. It is assumed that the sender and the receiver get the same payoff 1 if $p_{ij} = 1 = q_{ji}$ given that state $i$ has occurred. If $(P, Q)$ is a profile of strategies, then the overall payoff for the sender and the receiver is given by

$$\sum_{i,j} p_{ij} q_{ji} = \mathrm{tr}(PQ). \tag{1}$$

Note the presumption that every state is weighed equally for computing the payoffs. This may not hold if, for instance, some states in $S$ occur with higher probability than others. This case may be included into the current framework by introducing the diagonal matrix $V = nW$ where each diagonal entry of $W$, $w_i$, is the probability of state $i$. Thus $w_i \geq 0$ and $\sum_i w_i = 1$. The payoff resulting from strategy profile $(P, Q)$ is then given by

$$n \sum_i w_i \sum_j p_{ij} q_{ji} = \mathrm{tr}(VPQ). \tag{2}$$

Suppose for instance that $n = 2$. Then the pure strategies are given by

$$P_1 = Q_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, P_2 = Q_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, P_3 = Q_3 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, P_4 = Q_4 = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}. \tag{3}$$

Using equation (1) yields all relevant payoffs. For example, $\mathrm{tr}(P_1Q_1) = 2$, but $\mathrm{tr}(P_1Q_2) = 0$ while $\mathrm{tr}(P_3Q_i) = 1$ for all $i$.

The main source of problems for analyzing this class of games is the number of strategies. The number of sender strategies is given by the number of all functions from $E$ into $S$. Hence there are $n^n$ sender strategies when there are $n$ signals. Analogously, there are $n^n$ receiver strategies. The number of strategies is thus already considerably high for a modest number of signals such as $n = 10$.

## 2 Replicator Dynamics

Signaling games have been studied extensively using the replicator dynamics

$$\dot{z}_i = z_i((C\mathbf{z})_i - \mathbf{z} \cdot C\mathbf{z}) \tag{4}$$

(see e.g. Skyrms, 1996; Nowak and Krakauer, 1999; Nowak et al., 1999; Komarova and Niyogi, 2004; Huttegger, 2006; Pawlowitsch, 2006). The dynamics (4) applies to the symmetrized version of the asymmetric signaling game between sender and receiver which was described above. In this symmetrized signaling game each player can be in the role of the sender or in the role of the receiver. A strategy in the symmetrized game is thus a pair of strategies of the asymmetric game where each player has $n^{2n}$ strategies. The roles of the players are chosen randomly. Accordingly, the payoffs are the expected payoffs a player gets in each role (for more on symmetrization of asymmetric games see Cressman, 2003).

We will focus instead on the asymmetric signaling game. The appropriate analogue to (4) is the two-population replicator dynamics

$$\dot{x}_i = x_i((A\mathbf{y})_i - \mathbf{x} \cdot A\mathbf{y}) \tag{5a}$$
$$\dot{y}_j = y_j((B\mathbf{x})_j - \mathbf{y} \cdot B\mathbf{x}) \tag{5b}$$

where $A$ is the $n^n \times n^n$ payoff matrix of the sender and $B = A^T$ is the $n^n \times n^n$ payoff matrix of the receiver. The dynamics (5) lives on $S_m \times S_m$, where $S_m$ is the unit-simplex in $\mathbb{R}^m$ and $m = n^n$.

We study the dynamics of signaling games in terms of (5) rather than in terms of (4) mainly for technical reasons. Since the dimension of the state space of a signaling game will usually be very high, anything that reduces the dimension in special cases will be useful. For $n$ signals the dynamics (5) lives on a $(2n^n - 2)$-dimensional manifold while the dynamics (4) lives on a $(n^{2n} - 1)$-dimensional manifold.

The most relevant results for the asymptotic properties of the dynamics of signaling games have been established by Huttegger (2006) and Pawlowitsch (2006). These results are formulated in terms of (4), but they basically carry over to the dynamics (5).

Local analysis of rest points for (4) and (5) shows that the only asymptotically stable states are given by the strict Nash equilibria of the signaling game. Strict Nash equilibria, in turn, are special pairs of one-to-one functions from $E$ onto $S$ and from $S$ onto $E$, respectively: if $P$ is the one-to-one sender strategy, then the receiver matrix $Q = P^T$ (where $T$ denotes the transpose of a matrix). Since they correspond to states of perfect communication, such strategy profiles will be called *signaling systems* from now on. For $n$ signals there are $n!$ signaling systems. $(P_1, Q_1)$ and $(P_2, Q_2)$ are examples of signaling systems. Another example is given by the following pair of strategies:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Being strict Nash equilibria, signaling systems are local attractors. This does not imply, however, that the set of signaling systems attracts almost all solutions of the dynamics (4) (in the sense that the set of solutions which do not converge to a signaling system have Lebesgue measure zero.) This is in general not the case. Consider the following mixed strategy profiles

$$\bar{P} = \begin{pmatrix} \lambda & 1-\lambda & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \bar{Q} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & \mu & 1-\mu \end{pmatrix} \tag{6}$$

where $0 \leq \lambda, \mu \leq 1$. $(\bar{P}, \bar{Q})$ is clearly not a signaling system, but a Nash equilibrium of the signaling game. These profiles form a two-dimensional linear manifold $M$ on the boundary of $S_{27} \times S_{27}$. The average fitness on $M$ is $\text{tr}(\bar{P}\bar{Q}) = 2$. Since signaling games are partnership games, average fitness is a strict Liapunov function and $M$ consists thus entirely of rest points (see Hofbauer and Sigmund, 1998). The transversal eigenvalues of the Jacobian at a rest point $\mathbf{p}$ in the interior of $M$ are then given by $(C\mathbf{p})_i - 2$, where $i$ refers to a strategy not in the support of $\mathbf{p}$. It is easy to see that $(C\mathbf{p})_i < 2$ for all $i$ if $0 < \lambda, \mu < 1$. Since $M$ consists entirely of rest points, this implies that $\mathbf{p}$ is Liapunov stable. Furthermore, the solutions converging to some point in the interior of $M$ have positive Lebesgue measure (for details see Huttegger, 2006; Pawlowitsch, 2006). A similar argument applies to the dynamics (5).

This result generalizes to signaling games with more than three signals. Moreover, the average fitness on components such as $M$ can get arbitrarily small as the number of signals grows. To see this just consider the following mixed strategy profile:

$$\begin{pmatrix} \lambda_1 & \lambda_2 & \cdots & \lambda_{n-1} & 0 \\ 0 & 0 & \cdots & 0 & 1 \\ \vdots & & & & \vdots \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & \mu_1 & \cdots & \mu_{n-1} \end{pmatrix} \tag{7}$$

The profile (7) is also Liapunov stable according to the characterization of Pawlowitsch (2006). But $\text{tr}(PQ) = \sum_i \lambda_i + \sum_j \mu_j = 2$ irrespective of $n$.

## 3 Perturbed Dynamics

What is the significance of this result? The existence of a linear manifold of rest points like $M$ implies that the dynamics is degenerate and not structurally stable (Guckenheimer and

Holmes, 1983). This means that if the dynamics (4) or (5) are slightly perturbed, then the qualitative behavior of the dynamics near $M$ will change. Some perturbations might lead to a rest point close to $M$ that is asymptotically stable or linearly unstable. Or the rest points of $M$ might disappear altogether. Thus, one way to assess the significance of the convergence of the replicator dynamics to manifolds like $M$ consists in looking at plausible perturbations of those dynamics.

One plausible perturbation are selection-mutation dynamics (Hofbauer, 1985; Hofbauer and Sigmund, 1998). The heuristics behind selection-mutation dynamics is that in each point in time the change in the relative frequency of one type is not only given by selection but also by a term which describes the rate of mutation from one type to another. Let us assume that each type is equally likely to mutate into any other type. Then one can derive the system of differential equations

$$\dot{z}_i = z_i((C\mathbf{z}) - \mathbf{z} \cdot C\mathbf{z}) + \epsilon(1 - lz_i) \tag{8}$$

where $l = n^{2n}$ and $\epsilon$ is a small positive real number.

A similar perturbed dynamics can be introduced in case of the two-population replicator dynamics (5). This perturbed dynamics is given by

$$\dot{x}_i = x_i((A\mathbf{y})_i - \mathbf{x} \cdot A\mathbf{y}) + \epsilon(1 - mx_i) \tag{9a}$$

$$\dot{y}_j = y_j((B\mathbf{x})_j - \mathbf{y} \cdot B\mathbf{x}) + \epsilon(1 - my_j) \tag{9b}$$

where $m = n^n$. For $\epsilon = 0$ the dynamics (8) reduces to (4) and (9) reduces to (5). In a biological context $\epsilon$ can be interpreted as uniform mutation rate. In a non-biological context $\epsilon$ can be understood as the rate of experimentation or as the rate of agents switching strategies by mistake.

The systems (8) and (9) share some properties with (4) and (5), respectively. In particular, they are still gradient systems with respect to the Shashshahani metric. The potential function for (8) has been found by Hofbauer (1985) and is given by

$$\frac{1}{2}\mathbf{x} \cdot C\mathbf{x} + \epsilon \sum_{i=1}^{l} \log x_i.$$

A similar potential function exists for (9):

$$V(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot A\mathbf{y} + \epsilon \sum_{i=1}^{m} (\log x_i + \log y_i) \tag{10}$$

The first term of $V$ is just the mean fitness in the sender population and the receiver population, which must be the same since $\mathbf{x} \cdot A\mathbf{y} = \mathbf{y} \cdot A^T\mathbf{x} = \mathbf{y} \cdot B\mathbf{x}$. From the second term of $V$ it is clear that if $x_i \to 0$ or $y_j \to 0$ then $V \to -\infty$ . This implies that there can be no rest points on the boundary of $S_m \times S_m$.

The existence of a potential function allows us to exclude cyclic behavior of the dynamics (9). Moreover, we may conclude that all orbits converge to the set of rest points.

Before we turn to the special case of two signals, let us state two simple but important facts about the location of rest points of (9) for small $\epsilon$ (compare the proof of Theorem 13.4.1 in

Hofbauer and Sigmund, 1998). First, if $(\mathbf{p}(\epsilon), \mathbf{q}(\epsilon))$ is a rest point of (9) then

$$\begin{aligned}(A\mathbf{q}(\epsilon))_i - \mathbf{p}(\epsilon) \cdot A\mathbf{q}(\epsilon) - m\epsilon &= -\frac{\epsilon}{\mathbf{p}(\epsilon)} < 0 \\ (B\mathbf{p}(\epsilon))_j - \mathbf{q}(\epsilon) \cdot B\mathbf{p}(\epsilon) - m\epsilon &= -\frac{\epsilon}{\mathbf{q}(\epsilon)} < 0.\end{aligned}$$

This implies that as $\epsilon \to 0$ every accumulation point $(\mathbf{p}, \mathbf{q})$ of $(\mathbf{p}(\epsilon), \mathbf{q}(\epsilon))$ must satisfy

$$(A\mathbf{q})_i - \mathbf{p} \cdot A\mathbf{q} \leq 0 \quad \text{and} \quad (B\mathbf{p})_j - \mathbf{q} \cdot B\mathbf{p} \leq 0, \tag{11}$$

i.e. $(\mathbf{p}, \mathbf{q})$ must be a Nash equilibrium of the signaling game. This allows us to conclude that *there are no perturbed rest points of (9) close to rest points of (5) which are not Nash equilibria of the underlying game.*

The second general fact concerns the *existence and asymptotic stability of perturbed signaling systems.* To see this set

$$\begin{aligned}F_i(\mathbf{x}, \mathbf{y}, \epsilon) &= x_i((A\mathbf{y})_i - \mathbf{x} \cdot A\mathbf{y}) + \epsilon(1 - m x_i), \\ G_j(\mathbf{x}, \mathbf{y}, \epsilon) &= y_j((B\mathbf{x})_j - \mathbf{y} \cdot B\mathbf{x}) + \epsilon(1 - m y_j).\end{aligned}$$

Let $(\mathbf{p}, \mathbf{q})$ be a signaling system. Let $H = (F_1, \ldots, F_m, G_1, \ldots, G_m)$. $H$ is defined in an open set containing $(\mathbf{p}, \mathbf{q}, 0)$ and $H(\mathbf{p}, \mathbf{q}, 0) = 0$. The Jacobian matrix of $H$ with respect to $\mathbf{x}$ and $\mathbf{y}$ evaluated at $(\mathbf{p}, \mathbf{q}, 0)$ is invertible. Hence, by the implicit function theorem, there exists a unique smooth function $\phi : (-\epsilon_0, \epsilon_0) \to \mathbb{R}^m \times \mathbb{R}^m, \phi(\epsilon) = (\mathbf{p}(\epsilon), \mathbf{q}(\epsilon))$ such that

$$\phi(0) = (\mathbf{p}, \mathbf{q}) \quad \text{and} \quad H(\phi(\epsilon), \epsilon) = 0.$$

This establishes the existence of perturbed signaling systems for small $\epsilon > 0$. The asymptotic stability of perturbed signaling systems follows from the fact that the entries of the Jacobian matrix for (9) are continuous in $\epsilon$. Therefore, each of the $n!$ perturbed signaling systems is a local maximizer of the potential function (10).

There are no similar general statements about perturbations of equilibria other than signaling systems. But we can use degree theory and Morse index arguments to infer the existence of further rest points of (9). We will illustrate this in the simple case of two signals. The case of three or more signals needs more space and will be treated elsewhere.

## 4 Two Signals

### 4.1 Uniform Weights

A simple case has been partially analyzed in Huttegger et al. (2007). Consider the signaling game with strategies as given by (3). Let $x_i$ be the relative frequency of $P_i$ and $y_j$ be the relative frequency of $Q_j$. The Nash equilibria are given by the two signaling systems $(P_i, Q_i)$, $i = 1, 2$ and a component of Nash equilibria given by

$$\begin{pmatrix} \lambda & 1-\lambda \\ \lambda & 1-\lambda \end{pmatrix}, \begin{pmatrix} \mu & 1-\mu \\ \mu & 1-\mu \end{pmatrix}. \tag{12}$$

That these matrices describe Nash equilibria follows from Theorem 5.1 of Trapa and Nowak (2000). The matrices in (12) are generated by many points in $S_4 \times S_4$ since the sender and receiver matrices look like this:

$$\begin{pmatrix} x_1 + x_3 & x_2 + x_4 \\ x_2 + x_3 & x_1 + x_4 \end{pmatrix}, \begin{pmatrix} y_1 + y_3 & y_2 + y_4 \\ y_2 + y_3 & y_1 + y_4 \end{pmatrix} \tag{13}$$

Thus, the matrix pairs (12) constitute a four dimensional component of Nash equilibria given by the condition $x_1 = x_2$ and $y_1 = y_2$. Let us denote this component by $N$.

To see that there are no other Nash equilibria, suppose first that $P$ (or $Q$) has a zero column. Then $(P, Q)$ is a Nash equilibrium only if $Q$ (or $P$) is of the form (12) for some $0 \leq \mu \leq 1$ (or some $0 \leq \lambda \leq 1$). If $P$ (and $Q$) has no zero column and if it is not one-to-one, then by Theorem 5.1 of Trapa and Nowak (2000) $(P, Q)$ must be of the form (12) in order to be a Nash equilibrium.

From the first general fact stated at the end of Section 3 we know that, for small $\epsilon$, rest points of (9) must lie near a signaling system or near the component $N$. The second general fact implies that, for small $\epsilon$, there are two perturbed and asymptotically stable signaling systems for (9).

The average payoff on the component $N$ is 1. Hence the first term of the potential function (10) will remain constant. The second term of $V$ attains its unique maximum at the barycenter $\mathbf{b}$. Hence if $\epsilon > 0$ the component $N$ collapses into $\mathbf{b}$. Since there cannot be any other rest points of the perturbed dynamics, $\mathbf{b}$ must be linearly unstable having one positive eigenvalue. This follows from the Morse inequalities (Milnor, 1963): a gradient system with two asymptotically stable rest points must have at least one rest point having one positive eigenvalue.

Hence we see that from almost all initial conditions solutions of (9) converge to one of the perturbed signaling systems. For uniform weights this result is already true for the unperturbed dynamics (5) (see Huttegger, 2006, Theorem 9). However, as we shall see in the next section, in the case of non-uniform weights the dynamics (5) and (9) are indeed qualitatively different.

### 4.2 Non-Uniform Weights

We now consider signaling games with two signals and a non-uniform weight matrix $W$. Suppose without loss of generality that

$$W = \begin{pmatrix} p & 0 \\ 0 & q \end{pmatrix}$$

where $q = 1 - p$ and $p > q$. The structure of the equilibrium components is different in this case. To find the relevant rest points let us analyze the pure Nash equilibria first (see Figure 1). To simplify matters we will look at a truncated version of the game shown in Figure 1. $Q_4$ is strictly dominated by $Q_3$ and can be ignored in the further analysis. The pure strategies $P_3$ and $P_4$ correspond to mixtures of $P_1$ and $P_2$. Thus we will study the reduced game given by the $2 \times 3$-matrix with $P_1$ and $P_2$ as sender strategies and $Q_1$, $Q_2$ and $Q_3$ as receiver strategies.

| | $Q_1$ | $Q_2$ | $Q_3$ | $Q_4$ |
|---|---|---|---|---|
| $P_1$ | 2 | 0 | $2p$ | $2q$ |
| $P_2$ | 0 | 2 | $2p$ | $2q$ |
| $P_3$ | $2p$ | $2q$ | $2p$ | $2q$ |
| $P_4$ | $2q$ | $2p$ | $2p$ | $2q$ |

Figure 1: Signaling game. If $p = q = 1/2$ then the signaling game has uniform weights. If $p \neq q$ then the signaling game has non-uniform weights.

A picture of the state space is shown in Figure 2. $(P_1, Q_1)$ and $(P_2, Q_2)$ are still strict Nash equilibria. There is another component of rest points given by

$$\begin{pmatrix} \alpha & 1-\alpha \\ 1-\alpha & \alpha \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \quad 0 \leq \alpha \leq 1. \tag{14}$$

Not all points in this line are Nash equilibria, however. If $\alpha = 1/2$ (i.e. $x_1 = x_2$) then the payoffs for $Q_1$ and $Q_2$ are 1 while the payoff to $Q_3$ is $2p$. On the other hand, the payoff to either $P_1$ and $P_2$ at $Q_3$ is $2p$. Hence the mixed strategy of (14) with $\alpha = 1/2$ is a Nash equilibrium. The same holds for strategies with $\alpha$ close to 1/2 (see the bold line in Figure 2). Hence the component of rest points (14) contains a component of Nash equilibria $K$, given by $q \leq \alpha \leq p$. The size of $K$ relative to (14) depends on $p$. The average payoff along $K$ is $2p$.

There can be no rest point if $Q_1$, $Q_2$ and $Q_3$ are present. To see this one just has to note that the payoff to $Q_3$ to any mixture of $P_1$ and $P_2$ is always $2p$ while the payoffs to $Q_1$ and $Q_2$ cannot be $2p$ at the same time. Hence, the only remaining rest points are given by the corners of state space and by the point $x_1 = x_2 = y_1 = y_2 = 1/2$. Among these only the two signaling systems are Nash equilibria.

For the unperturbed replicator dynamics (5) the two signaling systems are asymptotically stable. The component $K$ is not asymptotically stable, but the points in the relative interior of $K$ are Liapunov stable (like the points in the relative interior of the component $M$ of Section 2). Points in $K$ are (non-strict) local maximizers of the potential function $\mathbf{x} \cdot A\mathbf{y}$. Therefore $K$ attracts an open set of initial conditions under (5). (Note that a component like $K$ does not exist for the uniform case by the remarks at the end of Section 4.1.)

What happens under the perturbed dynamics (9)? We know that, like in the uniform case, there exist two perturbed rest points of (9) close to the signaling systems for sufficiently small $\epsilon$. Both rest points are asymptotically stable. Index or degree theory (see e.g. Section 13.4 in Hofbauer and Sigmund, 1998) tells us that there is at least one further rest point. From the analysis above it is clear that this rest point can only lie near $K$. Using a perturbation expansion one can show its uniqueness. By the Morse inequalities, this must be a saddle point with one positive eigenvalue. We have also used Newton's method to numerically estimate the location of perturbed rest points close to $K$ by using Mathematica. Doing so always leads to one rest point close to $K$ which is hyperbolic and has one positive eigenvalue. This situation is shown in Figure 2.

A very similar analysis can be given for the whole game of Figure 1 since nothing essential concerning the Nash equilibria changes (except that the component $K$ gets bigger). Hence we may conclude that *the dynamics (9) converges to a signaling system from almost all initial conditions.*

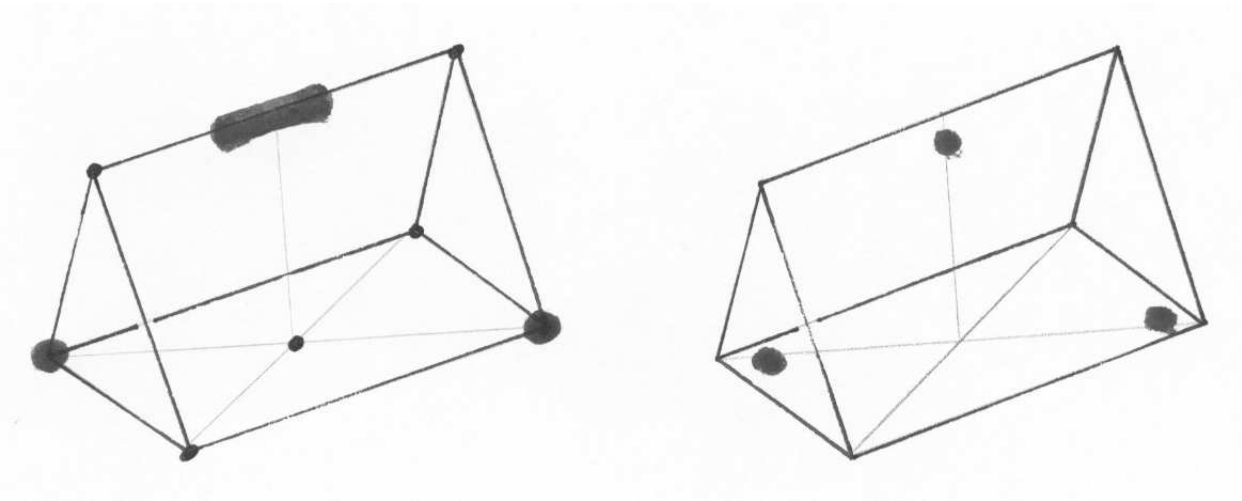


Figure 2: The left-hand side shows the stable rest points for the unperturbed dynamics in bold lines and dots, respectively. The right-hand side shows the rest points of the perturbed dynamics. The perturbed equilibrium close to the component is a saddle point.

## References

R. Cressman. *Evolutionary Dynamics and Extensive Form Games.* MIT Press, Cambridge, MA, 2003.

J. Guckenheimer and P. Holmes. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields.* Springer, New York, 1983.

J. Hofbauer. The Selection Mutation Equation. *Journal of Mathematical Biology*, 23:41–53, 1985.

J. Hofbauer and K. Sigmund. *Evolutionary Games and Population Dynamics.* Cambridge University Press, Cambridge, 1998.

S. M. Huttegger. Evolution and the Explanation of Meaning. Forthcoming in *Philosophy of Science*, 2006.

S. M. Huttegger, B. Skyrms, R. Smead, and K. Zollman. Evolutionary Dynamics of Lewis Signaling Games: Signaling Systems vs. Partial Pooling. Technical Report, Institute for Mathematical Behavioral Sciences. Paper 56. http://repositories.cdlib.org/imbs/56, 2007.

N. L. Komarova and P. Niyogi. Optimizing the Mutual Intelligibility of Linguistic Agents in a Shared World. *Artificial Intelligence*, 154:1–42, 2004.

J. Milnor. *Morse Theory.* Princeton University Press, Princeton, 1963.

M. A. Nowak and D. C. Krakauer. The Evolution of Language. *Proceedings of the National Academy of Sciences*, 96:8028–8033, 1999.

M. A. Nowak, J. B. Plotkin, and D. C. Krakauer. The Evolutionary Language Game. *Journal of Theoretical Biology*, 200:147–162, 1999.

C. Pawlowitsch. Why Evolution Does not Always Lead to an Optimal Signaling System. Working Paper, University of Vienna, 2006.

B. Skyrms. *Evolution of the Social Contract.* Cambridge University Press, Cambridge, 1996.

P. E. Trapa and M. A. Nowak. Nash Equilibria for an Evolutionary Language Game. *Mathematical Biology*, 41:172–188, 2000.

# Quantity implicature and speaker expertise in signalling games[*]

Tikitu de Jager
Institute for Logic, Language and Computation
Universiteit van Amsterdam

S.T.deJager@uva.nl

### Abstract

This paper treats implicatures due to Grice's first submaxim of Quantity; a typical case is the interpretation of "John and Mary" (in answer to "Who went skiing last weekend?") as meaning "John and Mary and nobody else". Two variations are standard in the literature: the weak epistemic reading "...and *as far as I know*, nobody else" and the exhaustive reading "...and *I know that* nobody else". These two interpretations have been related to speaker expertise, in the sense that the exhaustive reading can be derived from the weak epistemic reading by adding an assumption of speaker expertise. We show that the two interpretative strategies are rational —in the standard game-theoretic sense of Nash equilibria— respectively when the speaker is known to be inexpert (weak epistemic reading) and known to be expert (exhaustification). Furthermore we characterise the weak epistemic reading as the only interpretative strategy appearing in a Nash equilibrium profile under a particular structural constraint on strategies.[1]

## 1 Introduction

Grice's well-known maxims of cooperative conversation (Grice, 1967) are used to derive pragmatic IMPLICATURES from semantic meaning: under the assumption that a speaker is behaving cooperatively, the hearer can conclude that the speaker intended to communicate more than simply the semantic meaning of the expressions she used. In this paper we will make use of a simplified version of the maxim of Quality, and we will be mainly concerned with the first submaxim of Quantity:

**Definition 1** (Quality)**.** Say only what you know to be true.

**Definition 2** ($\mathsf{Quantity}_1$)**.** Make your contribution as informative as is required (for the current purposes of the exchange).

The first observation based on $\mathsf{Quantity}_1$ is that if a speaker utters $\varphi$ and not some stronger $\psi$ (such that $\psi \models \varphi$ and $\varphi \not\models \psi$), then she does not intend to communicate $\psi$.[2] Strictly

[*] I would like to thank Robert van Rooij for his substantial contribution to this work, Michael Franke and Remko Scha for extremely helpful discussions, and two anonymous reviewers of the manuscript for their comments.

[1] The results presented here are given a more extensive formal treatment by De Jager and Van Rooij (2007), which is however rather technical. We focus here on giving the motivation behind the models, and refer the reader to the other paper for proofs and further formal exposition.

[2] A precise formulation requires that we consider only a restricted set of alternative expressions as candidates for $\psi$; see the following section for details.

speaking, the strongest conclusion this licenses (in combination with Quality) is that she *does not know* that $\psi$: the WEAK EPISTEMIC reading. However in practice we are likely to conclude something stronger: the EXHAUSTIVE INTERPRETATION, that she *knows that not* $\psi$. Various authors (see for instance Spector, 2003; Van Rooij and Schulz, 2004) have recently proposed a two-stage approach: first the hearer draws the weak epistemic conclusion, and then this is strengthened by the assumption that the speaker is EXPERT in the matter at hand.

In this paper we implement both interpretative strategies in a signalling games setting. Lewis (1969) introduced signalling games to provide a non-circular account for the developement of linguistic conventions. A Lewisian signalling game typically allows for many equilibria, any of which is potentially a convention; Lewis described various non-linguistic means by which one of these might be selected. Here however we use signalling games to model the pragmatic refinement of existing semantic conventions. In this case we would prefer that the game-theoretic analysis picked out a *single* strategy, since $\mathsf{Quantity}_1$ implicature is not usually considered a matter of convention. We will see that this is only achieved with an extra constraint in addition to the standard signalling game model.

Section 2 gives formal definitions of weak epistemic $\mathsf{Quantity}_1$ implicature (Definition 5) and exhaustification (Definition 6) modelled on those of Van Rooij and Schulz (2004) and Schulz and Van Rooij (2006). Section 3 then introduces the framework of interpretation signalling games which we will use to analyse these implicatures. Section 4 gives a characterisation within this framework of expert and inexpert speakers, and shows that of the two strategies, exactly one is selected for each type of speaker. However this does not achieve our aim of selecting a single pragmatic interpretation, since many pathological strategies are not ruled out by the signalling games framework. Section 5 describes and motivates a structural constraint on strategies which does, indeed, characterise $\mathsf{Quantity}_1$ implicature as the only interpretative strategy giving rise to a Nash equilibrium in the game with inexpert speaker.

## 2 $\mathsf{Quantity}_1$ implicature formalised

Formalising $\mathsf{Quantity}_1$ requires that we choose the alternative expressions carefully: if "John and Mary *and no-one else*", is an alternative to "John and Mary", then the former cannot be an implicature of the latter (for if the speaker intended the stronger statement she should have said so). One standard solution (taken by Horn (1972); Gazdar (1979); Levinson (2000), among others) is to take only expressions arranged along a linear scale by entailment as alternatives. This SCALAR IMPLICATURE approach works for numerical expressions such as "John has $n$ children" (in the "at least" reading), but is insufficient for answers to wh-questions, since these cannot be ordered linearly by entailment. The appropriate alternative expressions in this case are the POSITIVE SENTENCES:

**Definition 3** (Positive sentences)**.** A positive sentence contains only positive atoms, conjunction and disjunction. Given a finite domain of objects and a predicate $Q$, we define the finite set of POSITIVE $Q$-EXPRESSIONS by choosing a shortest exemplar from each class of logically equivalent positive sentences using only the predicate $Q$.[3]

[3] Taking shortest exemplars is motivated by technical considerations, but it also means we are not required to make predictions about the interpretation of utterances such as "John or John and John", which are aberrant according to the Gricean maxims we have not included in the model.

For convenience we always take the speaker to be answering a wh-question given by the predicate $Q$ (for instance "Who went skiing in the weekend?"), over a finite domain; $W$ is the full set of worlds differing in the extension of $Q$, and if $w$ is a world then $V_w(Q)$ represents the extension of $Q$ in $w$. In the examples we take a domain containing individuals John, Sue, and Mary; in the world $\mathsf{JmS}$ John and Sue went but Mary did not; in $\mathsf{jmS}$ only Sue went, and so on.

For modelling the $\mathsf{Quality}$ maxim we lift the semantic denotation function $[\![\cdot]\!]$ to information states, given the symbol $(\!|\cdot|\!)$.[4] Where the standard denotation of $\varphi$ is the set of worlds in which $\varphi$ is true, the lifted notion is the set of *information states* in which $\varphi$ is *licensed* (that is, in which the speaker knows that $\varphi$ is true). Formally:

**Definition 4 (**Lifted semantic denotation**).**

$$(\!|\varphi|\!) \overset{\text{def}}{=} \{i \subseteq [\![\varphi]\!]\ ;\ i \neq \varnothing\}.$$

The following definitions are adapted from Van Rooij and Schulz (2004), where they were given in a modal logic setting.

The first definition is for the weak epistemic interpretation (first introduced under the name "Grice" because it incorporates both the Gricean maxims we are concerned with): an utterance $\varphi$ is interpreted as conveying the set of minimal models with respect to speaker knowledge of the positive extension of $Q$ ($\mathsf{Quantity}_1$) where $\varphi$ is known to hold ($\mathsf{Quality}$):

**Definition 5 (**Minimising unstated positive knowledge**).** Let $i, i' \subseteq W$ be information states for the speaker. We say she has no more positive knowledge of $Q$ in $i$ than $i'$, $i \leq^{\mathrm{K}}_{Q} i'$, when $\forall w' \in i' : \exists w \in i : V_w(Q) \subseteq V_{w'}(Q)$.

$$\begin{aligned}\textsc{Grice}(\varphi) \overset{\text{def}}{=} &\{i \in (\!|\varphi|\!)\ ;\ \forall i' \in (\!|\varphi|\!) : i \leq^{\mathrm{K}}_{Q} i'\}\\ = &\{i \in (\!|\varphi|\!)\ ;\ \forall w' \in [\![\varphi]\!] : \exists w \in i : V_w(Q) \subseteq V_{w'}(Q)\}.\end{aligned}$$

Van Rooij and Schulz (2004) show that the 'strong epistemic' implicature of exhaustification can be correctly derived by maximising speaker expertise[5] *after* applying Grice:

**Definition 6 (**Maximising expertise**).** Let $i, i' \subseteq W$ be information states for the speaker. We say that she is no more expert about $Q$ in $i$ than in $i'$, $i \leq^{\mathrm{E}}_{Q} i'$, when $\forall w' \in i' : \exists w \in i : V_{w'}(Q) \subseteq V_w(Q)$.

Applying this principle on top of Grice gives us exhaustification:

$$\textsc{Expert}(\varphi) \overset{\text{def}}{=} \{i \in \textsc{Grice}(\varphi)\ ;\ \neg\exists i' \in \textsc{Grice}(\varphi) : i <^{\mathrm{E}}_{Q} i'\}$$

(where "$<^{\mathrm{E}}_{Q}$" is derived in the obvious manner from "$\leq^{\mathrm{E}}_{Q}$" given above).

Now we will see how these definitions relate to rational behaviour in signalling games.

[4] The notation is due to Michael Franke, p.c.

[5] The term used by Van Rooij and Schulz (2004) is "competence", however to avoid confusion with the standard notion of linguistic competence (as opposed to performance) we refer instead to "expertise".

## 3 Interpretation signalling games

A SIGNALLING GAME is a game of partial information between a Sender and a Receiver. The sender observes the state of the world and sends a message to the receiver, who in turn must choose an action purely on the basis of the message, but various actions might be more or less appropriate depending on the state of the world.

The game is usually given as a tuple $\langle S, M, A, u\rangle$ where $S$ is a set of *states* (sometimes referred to as *types* of the sender), $M$ is a set of *messages* with no pre-arranged meaning, $A$ is a set of *actions* the receiver of the message might take, and $u\colon S \times A \to \mathbb{R}$ gives the *utility* of each action in each state. Typically a different action is optimal in each state, and the messages acquire meaning by being systematically associated with a particular state and hence inducing a particular action.

Here the model is rather more complex, since we wish the structure of the game to already fix the conventional semantic meaning of the messages being exchanged. The game is between Nature and the Sender and Receiver of the standard system, with Nature providing a random distribution over the OBSERVATIONS (sets of worlds, or INFORMATION STATES) given to the sender. Rather than actions we take INTERPRETATIONS (sets of information states) as the moves of the receiver, and most importantly we give a relation (semantic meaning) between the messages and the sender observations.

**Definition 7** (Interpretation signalling game)**.** An INTERPRETATION SIGNALLING GAME is a tuple $\langle Q, W, M, [\![\cdot]\!], \text{utility}_n, P\rangle$ with the following properties:

- $Q$ is a one-place predicate;
- $W$ is the full set of possibilities ("worlds") differing in the extension of $Q$;
- $M$, the MESSAGES available to the sender, is the set of positive $Q$-expressions;
- $[\![\cdot]\!]\colon M \to \wp(W)$ is the usual semantic denotation function: $[\![\varphi]\!]$ gives the set of worlds in which $\varphi$ holds (from which we derive the lifted notion $(\!|\cdot|\!)$ as usual);
- $\text{utility}_n\colon \wp(W) \times \wp(\wp(W)) \to \mathbb{R}$ is a function parameterised by $n$ (see below) giving a numerical payoff dependent on the *observation* and the *interpretation*: what the sender actually meant to convey, and how the receiver interpreted the message; and
- $P$ is a probability distribution on information states, such that all states in $\wp(W)$ occur with positive probability.

A PLAY of the game is a triple $\langle i, \varphi, \mathcal{Y}\rangle$ where $i \subseteq W$ is an information state (the OBSERVATION) given according to $P$ to the sender, $\varphi \in M$ is a MESSAGE (sent by the sender to the receiver), and $\mathcal{Y} \subseteq \wp(W)$ is a set of information states (the INTERPRETATION of the receiver).

The PAYOFF (or UTILITY) of a play $\pi = \langle i, \varphi, \mathcal{Y}\rangle$ is given by $\mathrm{U}(\pi) = \text{utility}_n(i, \mathcal{Y})$, defined by

$$\text{utility}_n(i, \mathcal{Y}) \overset{\text{def}}{=} \begin{cases} P(\{i\} \mid \mathcal{Y}) & \text{if } i \in \mathcal{Y}, \\ -n & \text{otherwise,} \end{cases}$$

where $n$ is a natural number, the PENALTY for misinterpretation; $P(\{i\} \mid \mathcal{Y})$ is the standard notion of conditional probability (recall that $\mathcal{Y}$ is a *set* of information states).

Given a game $G = \langle Q, W, M, \text{utility}_n, P\rangle$, we write $G_m$, $G^{P'}$ and $G_m^{P'}$ for, respectively, $\langle Q, W, M, \text{utility}_m, P\rangle$, $\langle Q, W, M, \text{utility}_n, P'\rangle$ and $\langle Q, W, M, \text{utility}_m, P'\rangle$.

The information asymmetry of the game is encoded in the notion of a STRATEGY: strategies $\sigma$ for the sender are functions from observations to messages, while receiver strategies $\rho$ are functions from messages to interpretations (sets of information states). In particular, we define two receiver strategies corresponding to the two interpretative principles:

$$\rho_{\mathrm{G}}(\varphi) \overset{\text{def}}{=} \textsc{Grice}(\varphi); \qquad \rho_{\mathrm{E}}(\varphi) \overset{\text{def}}{=} \textsc{Expert}(\varphi).$$

A play $\langle i, \varphi, \mathcal{Y}\rangle$ is ACCORDING TO THE STRATEGIES $\sigma$ and $\rho$ if $\sigma(i) = \varphi$ and $\rho(\varphi) = \mathcal{Y}$.

The penalty value in the payoff function models the intuition that communicative failure is always the worst outcome, no matter how much effort is saved in arriving *efficiently* at a wrong interpretation. Without such a penalty, if the same message $\varphi$ is sent in two different information states the optimal response by the receiver is to interpret it as the singleton set containing only the highest-probability observation giving rise to $\varphi$. This in turn gives rise to UNINTERPRETABLE observations, which do not occur in the interpretation of any message, guaranteeing communicative failure whenever the speaker in fact makes such an observation. Setting a numerical value on the penalty also turns out to be crucial for modelling speaker expertise, as we will see in the next section.

First, however, we need to clarify what "optimal response" means in the previous paragraph. We call a pair of strategies $\langle\sigma, \rho\rangle$ a LANGUAGE, and define the expected utility of a language in a particular game as usual: each observation induces a unique play according to $\sigma$ and $\rho$, so we take the expected payoff of the induced plays over the observation probabilities given by $P$. A sender strategy $\sigma$ is a BEST SENDER RESPONSE to a receiver strategy $\rho$ if $\sigma$ maximises expected utility given $\rho$, and similarly for best receiver responses. A language is a NASH EQUILIBRIUM if its strategies are mutual best responses, and a STRICT NASH EQUILIBRIUM if they are mutual *strict* (that is, unique) best replies.

Now we can use a standard game-theoretic solution concept: a pair of rational agents will play a Nash equilibrium strategy, because if the language is *not* an equilibrium then at least one player has a payoff incentive to change strategies. Next we will see how speaker expertise influences which strategies take part in Nash equilibria.

## 4 Quantity implicature depends on speaker expertise

First we must say how we will represent speaker expertise in the interpretation signalling game. The intuition is that an expert speaker makes precise observations, while an inexpert speaker is more likely to be in a state of uncertainty. The distribution $P$ over speaker observations gives us the representation we want, but a question remains: *how unlikely* is an uncertain observation by an expert speaker? Any quantitative answer would be arbitrary, but it seems too strong to say that such a speaker *always* makes precise observations.[6] Here we make use of the penalty value in the payoff function: for any arbitrary 'expertise threshold' on the probability distribution we can find a corresponding penalty value giving rise to the interpretative behaviour we expect. The bundle of expertise threshold and penalty

[6] Indeed, if this is the case then "*John or Mary* went skiing..." is predicted to be uninterpretable; the disjunction represents exactly the speaker's uncertainty about the precise state of affairs.

value[7] is taken to represent a single 'cultural parameter' of language use, corresponding to the notion of 'sufficient certainty' for stating something as fact.

**Definition 8** (Inexpert speaker)**.** Take $\delta \in [0,1]$ a non-negligible value. The distribution $P$ represents an INEXPERT SPEAKER (with respect to $\delta$) if $\forall i \subseteq W : P(i) \geq \delta$.

**Proposition 9.** *Let $P_\delta$ represent an inexpert speaker for some given value of $\delta$. Then by choice of $n$ we can always construct a game $G_n^{P_\delta}$ with the following property:*

*For any sender strategy $\sigma$ utilising all messages, the* unique *best receiver reply $\rho_{\mathrm{BR}(\sigma)}$ is given by*

$$\rho_{\mathrm{BR}(\sigma)}(\varphi) = \sigma^{-1}(\varphi) \overset{def}{=} \{i \subseteq W \; ; \sigma(i) = \varphi\}.$$

We call such a game a GAME WITH INEXPERT SPEAKER.

**Theorem 10.** *Let $G$ be a game with inexpert speaker. Then in $G$,*

1. *$\rho_{\mathrm{G}}$ has a unique best sender reply $\sigma_{\mathrm{G}}$, and $\langle \sigma_{\mathrm{G}}, \rho_{\mathrm{G}} \rangle$ is a strict Nash equilibrium; and*
2. *for no sender strategy $\sigma$ is $\langle \sigma, \rho_{\mathrm{E}} \rangle$ a Nash equilibrium.*

That is, when the speaker is *in*expert (frequently makes imprecise observations), GRICE is a rational interpretative strategy while EXPERT is not.

**Definition 11** (Expert speaker)**.** Take $\epsilon \in [0,1]$ a value 'reasonably close' to zero. The distribution $P$ represents an EXPERT SPEAKER (with respect to $\epsilon$) if for all $i \subseteq W$, $P(i) < \epsilon$ just in case $\exists w, w' \in i : V_w(Q) \subset V_{w'}(Q)$.

(This definition is a coarse-grained version of the expertise order given in Definition 6; each low-probability information state is lower in the ordering than some high-probability information state. Note however that this notion of precision is compatible with the sort of uncertainty expressed by "John or Mary": ignoring Sue for the moment, the information state $\{\mathsf{Jm}, \mathsf{jM}\}$ will be given high probability; on the other hand the *denotation* of "John or Mary", $\{\mathsf{Jm}, \mathsf{jM}, \mathsf{JM}\}$, is given low probability.)

**Proposition 12.** *Let $G_n$ be a game with penalty $n$. We can find a positive probability $\epsilon$ with the following property:*

*For* all *everywhere-nonzero probability distributions $P$, in $G_n^P$ the following holds for each receiver strategy $\rho$: If for any $i, i' \subseteq W$ and for some message $\varphi$ we have $\{i, i'\} \subseteq \rho(\varphi)$ and $P(i) < \epsilon \leq P(i')$, then $\rho$ is not the best response to any sender strategy in $G_n^P$.*

A game $G_n^P$ is known as a GAME WITH EXPERT SPEAKER if there is such an $\epsilon$ for $G_n$ such that $P$ represents an expert speaker with respect to $\epsilon$.

**Theorem 13.** *Let $G$ be a game with expert speaker. Then in $G$,*

1. *for no sender strategy $\sigma$ is $\rho_{\mathrm{G}}$ a best response; and*
2. *there is at least one $\sigma$ which is a best sender response to $\rho_{\mathrm{E}}$, and for* every *such $\sigma$, $\rho_{\mathrm{E}}$ is in turn the unique best receiver response.*

Unlike the case for an inexpert speaker, we cannot speak simply of a strict Nash equilibrium: since some of the information states are uninducable, there will be a *set* of sender strategies

[7] In fact for technical reasons we need two threshold values: an upper and a lower bound.

that differ only on how they behave on these information states, all equally optimal. The formulation is a special case of the notion (standard in evolutionary game theory) of an evolutionarily stable *set* of strategies; roughly speaking this means that payoff-neutral drift may occur within the set, but movement outside the set will result in strictly lesser payoff.

What these two theorems mean is that in each of the cases of an inexpert and an expert speaker, only one of the two interpretative strategies GRICE and EXPERT is rational. However there are a number of alternative strategies that are not ruled out; in particular, any interpretative strategy that assigns each information state to only one message will give rise to a strict Nash equilibrium strategy in just the same way that GRICE does, if the speaker is inexpert. In the following section we show how this proliferation of strategies (some extremely odd-looking) can be reduced by a structural restriction, leaving only GRICE as a rational choice.

## 5 Characterising $\mathsf{Quantity}_1$

**Theorem 14.** *Let $G$ be a game with inexpert speaker. Let $\rho$ be a receiver strategy with the following properties:*

1. *$\forall\varphi \in M : [\![\varphi]\!] \in \rho(\varphi)$ ("Faithfulness"), and*
2. *$\forall i, i', i'' \subseteq W : \forall\varphi \in M : i \subseteq i' \subseteq i''$ & $i, i'' \in \rho(\varphi) \Rightarrow i' \in \rho(\varphi)$ ("Convexity").*

*Then there exists a sender strategy $\sigma$ (obeying* Quality*) such that $\langle\sigma, \rho\rangle$ is a strict Nash equilibrium in $G$ if and only if $\rho(\varphi) = \text{GRICE}(\varphi)$.*

These conditions alone are obviously not sufficient to characterise GRICE. The game-theoretic setting, however, ensures that in a Nash equilibrium language obeying Quality the interpretation of each message $\varphi$ includes a particular information state, the set of $Q$-MINIMAL WORLDS from $[\![\varphi]\!]$, written $\min_Q([\![\varphi]\!])$ (these are the worlds in $[\![\varphi]\!]$ where the denotation of $Q$ is minimal; for a formal definition see De Jager and Van Rooij (2007)). Between this minimal element (provided by the signalling game) and the maximal element (provided by faithfulness), convexity fills out the same interpretation given by GRICE.

The two conditions are structural restrictions on the form of acceptable strategies. The first, faithfulness, can be thought of as a basic requirement before we can justifiably call $[\![\cdot]\!]$ a 'semantic denotation' function and $\rho$ a 'pragmatic interpretation' rule. If the 'semantic meaning' of some message is *never* in the interpretation by a receiver, it could be described as somewhat disingenuous to continue to call it 'semantic meaning'; the faithfulness requirement then ensures that our terminology remains honest.

The convexity condition is more speculative, but by no means unfamiliar in natural language semantics. Closure properties in general are motivated by simplicity considerations; they make rules (in this case rules of interpretation) easier to describe and learn, since apparently complex sets can be given a compact representation. Here the pragmatic interpretation of a message $\varphi$ can be represented compactly by just giving $\min_Q(\varphi)$, $[\![\varphi]\!]$, and the closure condition $i \in \text{GRICE}(\varphi) \Leftrightarrow \min_Q(\varphi) \subseteq i \subseteq [\![\varphi]\!]$. Convexity constraints of similar form have been used for describing linguistic universals in generalised quantifier theory (Thijsse, 1983; Van Benthem, 1986) and cognitive semantics (Gärdenfors, 2000), and are given an independent game-theoretic motivation in a forthcoming paper by Jäger and Van Rooij.

## 6 Conclusions

We have shown that Gricean $\mathsf{Quantity}_1$ implicature is rational (in a signalling games setting) when the speaker is known to be inexpert in the matter at hand; when the speaker is known to be expert, the standard exhaustive interpretation instead becomes rational. Furthermore, under a natural convexity constraint on interpretative strategies $\mathsf{Quantity}_1$ implicature is *uniquely* rational given an inexpert speaker.

## References

Johan F. A. K. van Benthem. *Essays in Logical Semantics.* Reidel, Dordrecht, 1986.

Peter Gärdenfors. *Conceptual Spaces: The Geometry of Thought.* MIT Press, Cambridge, Massachusetts, 2000.

Gerald Gazdar. *Pragmatics.* Academic Press, London, 1979.

H. P. Grice. Logic and conversation. The William James Lectures, delivered at Harvard University. Republished with revisions in Grice (1989), 1967.

H. P. Grice. *Studies in the Way of Words.* Harvard University Press, Cambridge, Massachusetts, 1989.

Laurence R. Horn. *The semantics of logical operators in English.* PhD thesis, Yale University, 1972.

Gerhard Jäger and Robert van Rooij. Language structure: psychological and structural constraints. *Synthese*, to appear.

Tikitu de Jager and Robert van Rooij. Deriving quantity implicatures. In *Proceedings of Theoretical Aspects of Rationality and Knowledge*, Brussels, June 2007.

Stephen C. Levinson. *Presumptive Meanings. The Theory of Generalized Conversational Implicatures.* MIT Press, Cambridge, Massachusetts, 2000.

David K. Lewis. *Convention.* Harvard University Press, Cambridge, 1969.

Robert van Rooij and Katrin Schulz. Exhaustive interpretation of complex sentences. *Journal of Logic, Language and Information*, 13:491–519, 2004.

Katrin Schulz and Robert van Rooij. Pragmatic meaning and non-monotonic reasoning: The case of exhaustive interpretation. *Linguistics and Philosophy*, 29:205–250, 2006.

Benjamin Spector. Scalar implicatures: exhaustivity and Gricean reasoning? In Balder ten Cate, editor, *Proceedings of the Eighth ESSLLI Student Session*, Vienna, Austria, August 2003.

Elias Thijsse. On some proposed universals of natural language. In Alice G. B. ter Meulen, editor, *Studies in Modeltheoretic Semantics*, pages 19–36. Foris Publications, Dordrecht, 1983.

# Sometimes a signal is just a cigar: utterance and context in coordination experiments[*]

Mark Jeffreys
Dept. of Behavioral Science & Integrated Studies
Utah Valley University

jeffrema@uvsc.edu

**Abstract**

This presentation looks to contribute to an interdisciplinary exchange between pragmatics, game theory, behavioral economics, and evolutionary anthropology, one that focuses not on game-theoretic predictions but the challenges of designing signalling games that can isolate the relative contributions made by different types of signals, whether embedded in utterances or contexts, to the degree of coordination among participants. At Utah Valley we have begun employing signalling-game experiments to try to tease apart the precise roles of linguistic, paralinguistic, and nonlinguistic signals when humans coordinate with nonkin partners despite material conflicts of interest. Here we review the reasons and premises behind this nascent research program and then give a brief introduction to three lines of experiments in varying stages of development.

## 1 Introduction

Language, as practiced by humans, operates conjointly with systems of communication primitive and derived. Those systems knit together individual and social adaptations, ancient and recent, ranging from involuntary vocalizations and affective facial expressions to inferential reasoning and theories of mind, from chemical and visual signals that alter our behaviors unawares to synthetic perfumes and dyes refined over centuries to elicit behaviors we desire, and from universally recognizable postures of dominance or submission to highly localized symbolic complexes indicating subtleties of membership, status, skill, kinship, ideological orientation, and so forth, all of which can take years or decades of cultural initiation properly to interpret. All these items together only begin to indicate the variety of systems by which human beings communicate.

Many of these systems of communication we share by homology or analogy with other speciesmore of them, in fact, than we once believed and perhaps still many more that we have yet to recognize. Nonetheless, a few of our means of communicating with each other do appear quite recently derived and unusual, and may yet prove unique to our species, even when compared to extinct humans. Because one defining trait of modern humans is our keen interest in highlighting and celebrating the uniqueness of our respective in-groups, we tend to assign higher values to these few communicative capacities that would seem to define our uniqueness as a species, and thus particularly to language, the long-favored

[*] I would like to thank my students Jeff Barth, Carmell Hoopes-Clark, Rob Jorgenson, and Scott Martin

candidate for the quintessentially human system of communication (Cheney and Seyfarth 2007, p. 24).

But what does language do for us, exactly, as an adaptive trait, and how does it fit in to all the other signal traffic going on between humans in every social context, all the time? At least one serious entanglement continues to bedevil the most sophisticated accounts of the evolutionary origins of language: the interdependence of our capacity for language with our psychological heuristics for cooperation in nearly all our social interactions. It is this precise combination that, for example, enables humans routinely to initiate non-kin cooperation with conspecifics who were total strangers to them only moments earlier (Christiansen and Kirby 2003). In the terminology of pragmatics, this underscores the fundamental Gricean prinicple that language is a cooperative activity (Huang 2007). Unfortunately from a Darwinian point of view, the principle presents a classic chicken-and-egg type of dilemma for any evolutionary account, one made all the more problematic by the fact that we are far from the only cooperative, smart and social species. Why and how, then, did the need for cooperation also entail selection for fully syntactic languages, alongside of and addition to our many other, functional systems of communication, and that apparently only in our lineage?

## 1.1 Reasoning and Premises of the Utah Valley Experiments

Whenever we frame questions about the origins of behaviorally modern humans in terms of what makes us essentially unique, compared to any other species, or indeed of what makes human language essentially unique, compared to any other communication system, we have already biased our research. This is not all bad, of course. Bias in favor of searching for what makes humans and human languages unique has proved abundantly productive at discovering unexpected ways in which neither we nor our communication systems are unique after all (Seyfarth and Cheney 2007), just as the early game-theoretic assumptions of rational agency actually, if unintentionally, helped to inspire the global, laureate-accumulating industry of behavioral economic theory as we know it today (Camerer 2003).

However, the biological uniqueness of humans simply is not all that consequential. Or, rather, our consequences themselves are our uniqueness. We have been reshaping ecosystems, literally from the ground up,on every major landmass and most archipelagoes, eventually extending our eco-engineering or niche construction (Odling-Smee et al. 2003) down through most if not all of the worlds major bodies of water as well, and we began doing so tens of millenia ago (Anderson 2005).

The premises of the fledgling research program we have begun in Utah Valley, therefore, are as follows.

First, that human ecological impact can be considered a coordination problem in which the success rate can be quantified as the fractional, collective efficiency of resource extraction of any group, dyadic or larger, given a quantified pool of resources and some form of a conflict of interest between individual members of the groups or between subgroups when it comes to extracting those resources. The fraction of the total pool depleted at the end of the experiment is an operational proxy for the many ways in which humans coordinate to remake ecosystems, from setting wildfires to domestication to genetic engineering and beyond.

Second, that the established methodologies of experimental, behavioral economics, including many well-studied, off-the-shelf economic coordination and cooperation gamescan be adapted to create experiments measuring collective efficiencies of resource extraction within the given experiments.

Third, that while we agree that it is a matter of great scientific interest and importance to establish the sequence of selection events, including perhaps cultural as well as biological pressures, leading up to the emergence of the modern human languages, we feel that the full spectrum by which humans communicate, included derived and widely shared communication systems, need to be considered alongside of language in experiments testing how modern humans manage to coordinate these collective efficience of resource extraction. That is, rather than assume that, given capacities for cooperation and language, human impact automatically emerges, or that utterance has a known boundary with context when designing these signalling-game coordination experiments, we need to make an effort to identify the exact roles played by various elements of human communication in the relative success or failure of solving coordination problems as measured by fractional, collective efficiencies of resource extraction. Or, in other words, what impact does controlling for a given signalling variable have on a particular type of coordination problem?

## 1.2 The Experiments So Far

The remainder of this presentation describes three ongoing sets of experiments either under way or under design in Utah Valley, each at a different stage of development, and all of which attempt to utilize the just-stated premises as guiding principles. We realize that it may seem odd to some to tackle the evolution of language through the proximate problem of how the various forms of signalling that contemporary humans deploy contribute to effective or failed coordination, but the idea here is that a better dissection of the traits of modern human coordination could usefully constrain the possible narratives for the origins of those traits. Game theory is not primarily a predictive, modeling tool in these experiments, but games are a laboratory tool of sorts, and we are most concerned with the methodological challenge of using games in such a way as to be able to distinguish and quantify, say, the amount of coordination due to a nonverbal costly signal from that of a paralinguistic cue or a syntactic utterance.

This research is in its infancy and the other purpose in bringing this presentation to this workshop is the hope of speeding its development and of gaining advice that will steer it away from some of the many possible dead ends toward which infant research programs tend to wander.

The furthest developed of the three sets of experiments along these lines involves a simple, pairwise social-dilemma game nicknamed ChickenHawk, in which the usual scenario of defect or cooperate game of this nature has been altered, somewhat strangely, so that neither member of the participant pair benefits from mutual cooperation. It is also true that neither gains from mutual refusal to cooperate. If and only if one member of the pair cooperates by choosing to sacrifice the opportunity for material gain while the other participant seizes the opportunity, then the so-called defector will indeed gain. It may be compared to a simultaneous toss of a pair of fair coins, in which Heads wins, but only if the other persons coin comes up Tails. If both are Heads or both are Tails nobody wins anything. And one never wins with Tails.

The object of the game design was to remove any individual material incentive to cooperate. Success was measured, in each treatment, by the percentage of pairs in that treatment who managed to match one defector against one cooperator, so that at least one individual did win a material reward. Assuming that a prosocial pair of participants wished to succeed in having at least one of them get a reward, then if they were lacking any ability to signal their intentions to each other in any way, their best strategy would be to each mentally flip a coin. Totally random choice of cooperate or defect would average the highest fraction, that is, half, of pairs matching one cooperator to one defector.

To clarify, maximum possible collective efficiency would mean that, in every pair, one participant gained something even though the other participant gained nothing. Randomized collective efficiency would result in half of all pairs matching one defector to one cooperator, while a quarter of pairs would either end up as both cooperators who got nothing or both defectors who also got nothing.

The usual technique to control for social signalling confounds in laboratory economic experiments is to isolate the participants from each other completely and go to some length to reassure them of their anonymity. If cheap talk promises before the game are permitted, they may take the form of a click on an icon on a computer screen. The general idea is to subtract everything but language. Unfortunately, natural language is a good deal richer than the click of a statement on a computer screen.

In the ChickenHawk experiments, participants have been recruited and test in familiar social situations in which they are allowed to interact face-to-face. Thus, there is no attempt to remove social context. Rather, it is the opportunity for thirty seconds of natural language which is permitted or removed during the critical period of decision-making.

So far, these experiments have been conducted with both a highly homogeneous college cohort in Utah Valley and a considerably more diverse population of shoppers at a downtown outdoor mall in nearby Salt Lake City. In both sets of experiments, there was one set of treatments involving a condition allowing the introduction of thirty seconds of natural-language cheap talk before anonymous choices to cooperate or defect were made. Defectors who were matched to cooperators only ever knew if they had succeeded by a check in the mail. For both populations, this introduction of a short opportunity for natural-language cheap talk among nonkin resulted in higher matching rates of defectors to cooperators than did any of the conditions in which no pre-decision chatter occurred. In fact, in cases in which participants could signal their choices semaphorically with red (defect) or blue (cooperate) cards, the results were an excess of mutually cooperative pairs who therefore knew neither would receive anything.

These results hint at a role for language in allowing contractual negotiation that subverts or allows social partners to work around the problems of appearing cooperative or preserving a reputation for submission or generosity, and so forth. Another possibility, is that language also establishes a finely graded, costly signal indicating degree of in-group membership-in-good standing, which then allows a space for cheap signals such as words or semantic gestures to carry on the business of negotiating a conflict of interest in good faith. In the ChickenHawk experiments these natural-language conditions allowed spoken utterances, which meant in turn that elements of slang and accent, which are hard to fake, could have played a role in successful pairwise negotiations. In fact, in the more homogeneous, campus population, these conditions were significantly more effective than in the treatments with the more linguistically and socially heterogeneous shopping mall participants. Yet in both populations, the difference in performance did not carryover into conditions where people could see each other or know what decision each made, but not speak, and in both populations the availability of a brief conversation still led to the highest percentages of collective efficiency of any treatment in that population.

However, the number of participants, replications, and culturally distinct populations involved in these ChickenHawk experiments are still quite small to date. The usual caveat about the necessity for further replications with different population applies.

The second set of experiments, conducted by my student Scott Martin, have just begun. In these experiments, we chose to used recorded materials to test peoples ability to predict other peoples economic decisions based on those recordings. To wit, one set of participants plays an All-or-Nothing Dictator Game, or ANDG, involving the simple decision to keep 10 dollars or give it away to an anonymous stranger whom they will never meet. Either immediately before or immediately after making this decision, half of these ANDG participants are digitally recorded against a neutral background, reading the text of Mary Had a Little Lamb. The other half of the ANDG participants pose against the same background for a still digital photo, again randomized before or after they make their ANDG decision.

Scott is now showing additional sets of participants, at locations around the state of Utah, randomized clips of the ANDG participants on a laptop in a portable cubicle. Each of these viewing participants sees a randomized sequence of five video clips and five stills, with the stills held on screen for 30 seconds to match the time of the video clips and a 5-second blackscreen between each clip or still. For each person they view, the participants check whether they believe that the person they just viewed, either reading Mary Had a Little Lamb or in a still photo, Kept the 10 dollars or Gave away the 10 dollars.

Early results suggest an emerging pattern in which viewing participants do significantly better than chance at predicting whether the ANDG participants kept or gave away the money if they get to see the video clip, but not if they only see a still. If this pattern holds with a large enough sample, our next step will be to randomly alternate video clips of the readers with audio clips only, again asking for the same decisions. Depending on the results, we may move to removing color from the video clips, analyzing responses by gender, and testing across visually and phonetically distinct cultural groups, and so forth. The latter test may help to confirm or disconfirm the hypothesis suggested by the ChickenHawk experiments that hard-to-fake phonetic differences among subgroups speaking the same native language may serve as cues for socially coordinating behavior under conditions of material, individual conflicts of interest.

At some point, we would also hope to replicate whatever results we get from the above experiments with tests involving different texts. In particular, we want to test how peoples predictive powers react to framing effects of texts which are semantically at odds with the actual decision. That is, if we continue to find that people do better than chance at predicting givers or keepers when all read from the same, familiar text, will we find that the capacity vanishes if people view givers reading texts about the value of selfishness and keepers reading texts extolling altruism?

Although quite different from the directly interactive ChickenHawk experiments, the aim with these experiments involving ANDG recordings and predictions by viewers of the ANDG participant decisions is to try to isolate, channel by channel, the contributory channels to the ability to recognize a prosocial partner in a situation in which coordination to extract more resources collectively would require some sacrifice of immediate individual interest. Ideally, we would like to also weight the relative contributions of various signalling channels to the collective efficiencies of resource extraction.

The third set of experiments is still in its design phase. It takes its inspiration from the familiar problem of traffic merges. On crowded roads in which individuals have severely curtailed means of making normal human social contact, and in which individual physical differences are often rendered irrelevant or even reversed by differences in the size or status signified by their motor vehicles, breakdowns of ordinary civility and incidents of road rage have become accepted as a normal price of contemporary society in many cities and suburbs around the world. The situation, however, provides an interesting model for testing the pragmatics of coordinative signalling under conditions of individual conflicts of interest and with a quantifiable common resource at stake. What is best for the smooth flow of traffic may not be best for the individual rushing to a hospital with news of an injured family member, but the ability of that individual to negotiate terms for special dispensation from the ordinary rules will be severely limited in a motor vehicle as compared to making excuses for, say, missing work due to that same hospitalized family member.

These experiments will be more conventional, in terms of experimental economics, than the first two. For one thing, the initial game will be set up within a stationary, on-campus computer lab with cubicles in which participants play at computer screens. The aim is to adapt the platform of a Public Goods Game in such a way as to mimic the varying degrees to which individual drivers of automobiles are able to communicate with each other. Icons of vehicles will represent costly, real investments made by participants from their initial funding. Merge rates will vary. Small individual costs will be incurred for such actions as allowing a vehicle to merge in front of the driver, but those costs will spread in a wave passed on to other drivers. Larger costs, including elimination from the game will be paid for collisions caused by multi-round mutual standoffs. Undergraduate IT majors at Utah Valley with sufficient technical competency are still being recruited, so the degree of simulation of the trial runs may span the gamut from highly stylized, stem-and-leaf type software choices making no references involving vehicles or traffic at all, up to possible iconography of vehicles, lanes, and even gestures.

The more important aspect of these planned merging traffic experiments is that, however abstracted the interface ends up being, participants in different treatments will have access to varying types of signalling capacities. The simple goal again is to identify the signalling capacities that matter most to managing the environment in question so as to maximize the collective efficiency of resource extraction from it, and then ideally to weight the relative positive and negative contributions of those capacities.

## 2 Conclusion

Public goods and common property resources are conventionally seen as a policy problem in which the individual interest is pitted against the interests of the whole society. However, one of the reasons humans face such problems in the first place is that we are extraordinarily capable, as societies, at stripping ecosystems down to their foundations by coordinating our conflicting individual interests to maximize our collective efficiencies. Indeed, one groups prosocial, cooperative, self-sacrificing behavior is often another groups internal or external antisocial threat. One needs think only of the etymologies of words like collusion and conspiracy, or of the traditions of honor and silence in organized crime families around the world. The very terms used for these sorts of coordinated social-group activities, whether we view them from the outside as threats or from the inside as our own public good, are terms suggestive of spoken languages peculiar powers of negotiationputting heads together and whispering, then keeping the communication private and secret from all outsiders. This economy of conspiracy may turn out to be the fabled Rubicon of human evolution into modern human cultural capacities. But first we need to better isolate and measure the relative contributions to our capacities for collective efficiencies of resource extraction of as many human signalling systems as possible, ancient or recent, before we can know what language really brought us, much less why and how.

## References

M.K. Anderson. *Tending the Wild.* University of California Press, Berkeley, 2005.

Colin Camerer. *Behavioral Game Theory: Experiments in Strategic Interaction.* Princeton University Press, Princeton, 2003.

Dorothy L. Cheney and Robert M. Seyfarth. *Baboon Metaphysics: The Evolution of a Social Mind.* Chicago University Press, Chicago, 2007.

Morton H. Christiansen and Simon Kirby, eds. *Language Evolution.* Oxford University Press, Oxford, 2003.

K. Laland F. Odling-Smee and W. Feldman. *Niche Construction.* Princeton University Press, Princeton, 2003.

Yan Huang. *Pragmatics.* Oxford University Press, Oxford, 2007.

# Production Pressures and Syntactic Change: Towards a New Perspective on Syntactic Evolution?[*]

Ruth Kempson, Miriam Bouzouita
Philosophy Dept.
King's College London

{ruth.kempson/miriam.bouzouita}@kcl.ac.uk

Ronnie Cann
Linguistics and English Dept.
University of Edinburgh

ronnie@ling.ed.ac.uk

### Abstract

With syntax seen as the progressive construction of semantic representations (Kempson et al., 2001; Cann et al., 2005), some syntactic changes can be seen as the result of production pressures on dialogue, with transitions due to routinizations, some of which consolidate minimal effort considerations on production. The case study is the diachronic development of clitic pronouns in the modern Romance languages from Latin. Two consequences emerge in this account: a shortening in the evolution of natural-language systems relative to more conventional grammar formalisms, and the tight coordination of parsing and production redolent of other displays of coordination in cognitive tasks.

## 1 Introduction

This paper argues that persistent pressure to minimize production costs in all conversational dialogue plays an important role in the relative early placement of pronouns in Latin, and then through sequences of routinization, in the early ordering of clitics within a clausal sequence in Medieval Spanish (MedSp) and the subsequent shift in clitic positions via Renaissance Spanish (RenSp) to the Modern Spanish (ModSp) system. In other words, complex phenomena at the morpho-syntactic interface will be shown to emerge from production pressures in dialogue. The formal framework within which this account is set out is Dynamic Syntax (DS) (Kempson et al., 2001; Cann et al., 2005), in which parsing and production are by definition tightly coordinated. The paper closes with reflections on the consequences of this stance for language evolution.

## 2 Towards a Dynamic Syntax of Latin

DS is a parsing-directed grammar formalism, in which a decorated tree structure representing a semantic interpretation for a string is incrementally projected following the left-right sequence of the words, from a starting point with just a rootnode and a requirement for

[*] This paper stems from work on dialogue, for which we are grateful to Matt Purver for substantial input in formulation. We would also like to thank Stelios Kyriakidis, Eleni Gregoromichelaki, Jieun Kiaer, Aiko Yamanaka, Andrew Gargett for helpful input as these ideas have emerged. Normal disclaimers apply.

some propositional value, to an endpoint which is a fully decorated binary branching tree structure encoding functor-argument structure of a familiar sort:[1]

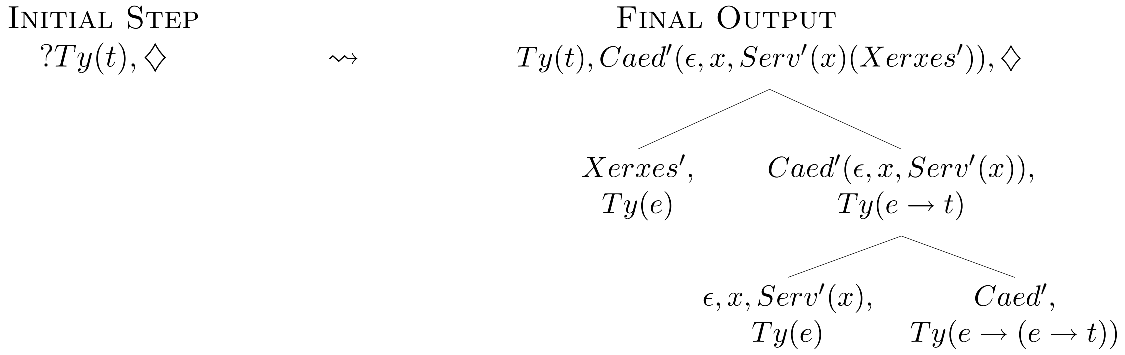

Initial Step
$?Ty(t), \Diamond$
$\rightsquigarrow$
Final Output
$Ty(t), Caed'(\epsilon, x, Serv'(x)(Xerxes')), \Diamond$
$Xerxes'$, $Ty(e)$
$Caed'(\epsilon, x, Serv'(x))$, $Ty(e \to t)$
$\epsilon, x, Serv'(x)$, $Ty(e)$
$Caed'$, $Ty(e \to (e \to t))$

Figure 1: Parsing *Xerxes servum cecidit* 'Xerxes killed a slave'

The process of tree-growth is the sole basis of syntactic explanation: a sentence is defined to be well-formed just in case there is at least one possible route through that process (with parse states defined as triples of $\langle \phi, T_\phi, A_\phi \rangle$, $\phi$ a word-sequence, $T_\phi$ a partial tree under construction from $\phi$, and $A_\phi$ actions used to induce $T_\phi$). Central to this process is the concept of requirement $?X$ for any decoration X representing a type, formula or treenode address. For example, decorations on nodes such as $?Ty(t)$, $?Ty(e)$, $?Ty(e \to t)$ etc. express requirements to construct formulae of the appropriate type on the nodes so decorated, $?\exists \mathbf{x}.Fo(\mathbf{x})$ a requirement to provide a fixed formula specification. The underpinning formal system is a logic of finite trees (LOFT: Blackburn and Meyer-Viol (1994)). There are two basic modalities, $\langle \downarrow \rangle$ and $\langle \uparrow \rangle$, such that $\langle \downarrow \rangle \alpha$ holds at a node if $\alpha$ holds at its daughter, and its inverse, $\langle \uparrow \rangle \alpha$, holds at a node if $\alpha$ holds at its mother. Function and argument relations are distinguished by defining two types of daughter relation, $\langle \downarrow_0 \rangle$ for argument daughters, $\langle \downarrow_1 \rangle$ for functor daughters (with their inverses $\langle \uparrow_0 \rangle, \langle \uparrow_1 \rangle$). Domination relations are then definable through Kleene star operators, e.g. $\langle \uparrow_* \rangle Tn(a)$ for some node identified as dominated by treenode $Tn(a)$.

Requirements drive the subsequent tree-construction process. One such driving property is the specification of case, which can be defined as imposing filters on update: e.g. for accusative case $?\langle \uparrow_0 \rangle Ty(e \to t)$, or for nominative $?\langle \uparrow_0 \rangle Ty(t)$. Another is the license to construct relatively weak tree-relations with a node introduced as "unfixed", described simply as $\langle \uparrow_* \rangle Tn(a)$, with an associated requirement for update, (with $\langle \uparrow_0 \rangle \langle \uparrow_1^* \rangle Tn(a)$ as a locally restricted variant). Transition steps between partial trees along a monotonic growth process are determined by general computational actions and lexical actions triggered by parsing words in the order in which they are presented in some string.[2]

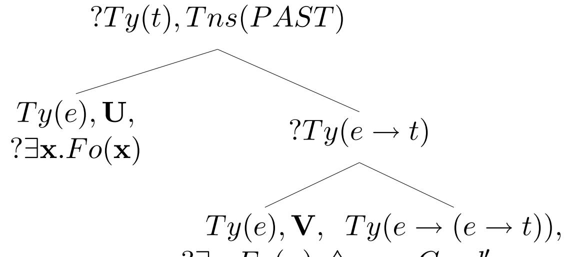
Crosslinguistic variation is expressed in terms of such lexical actions. For example, with its relatively free word order and possibility of pro-drop, the parsing of a Latin verb induces

[1] $Fo$ is a predicate that takes a logical formula as value, $Ty$ a predicate that takes logical types as values, $Tn$ a predicate that takes tree-node addresses as values, e.g. $Tn(0)$ being the rootnode.

[2] Quantification is expressed in terms of variable-binding term operators, so that quantifying NPs like all other NPs are of type $e$. The underlying logic is the epsilon calculus, the formal study of arbitrary names, with term-expressions whose internal structure is made up of an epsilon binder, $\epsilon$, a variable, and a restrictor: e.g. $\epsilon, x, Man'(x)$. Since in Latin, nouns project full specification of terms, the structure defined to be projected by *servum* would be a subtree of which the quantifying term is the topnode, dominating a subtree decorated with binder, variable, and restrictor specification. We leave all details on one side.

a propositional structure whose argument nodes are decorated with metavariables: place-holders that stand for some real value to be assigned from the context, capturing the effect of null pronouns without the assumption that these exist as parts of a *linguistic* string:[3]

```
        CECIDIT
IF      $?Ty(t)$
THEN    put($Tns(PAST)$);
        make($\langle\downarrow_0\rangle$); go($\langle\downarrow_0\rangle$);
        put($Ty(e), Fo(\mathbf{U}), ?\exists\mathbf{x}.Fo(\mathbf{x})$); go($\langle\uparrow_0\rangle$);
        make($\langle\downarrow_1\rangle$); go($\langle\downarrow_1\rangle$); put($?Ty(e \rightarrow t)$);
        make($\langle\downarrow_1\rangle$); go($\langle\downarrow_1\rangle$);
        put($Fo(Caed'), Ty(e \rightarrow (e \rightarrow t)), [\downarrow]\bot$);
        go($\langle\uparrow_1\rangle$); make($\langle\downarrow_0\rangle$); go($\langle\downarrow_0\rangle$);
        put($Fo(\mathbf{V}), Ty(e), ?\exists\mathbf{x}.Fo(\mathbf{x})$)
ELSE    Abort
```


$?Ty(t), Tns(PAST)$
$Ty(e), \mathbf{U},$
$?\exists\mathbf{x}.Fo(\mathbf{x})$
$?Ty(e \rightarrow t)$
$Ty(e), \mathbf{V},$
$?\exists\mathbf{x}.Fo(\mathbf{x}), \diamondsuit$
$Ty(e \rightarrow (e \rightarrow t)),$
$Caed'$

Lexical projection of propositional structures then interacts with the early construction of nodes with only a weak structural relation to the dominating node, as driven by constructive use of case: in particular the output-filter restriction of case-specifications can be implemented at any time to update an unfixed node to a fixed relation as each such unfixed node is introduced, thus licensing the parse of NP sequences before the verb.[4] The actions of the verb then fill out the remainder of the propositional structure to yield the appropriate output tree:[5]

| (1) | *Serv-um* | *Xerxes* | *cecidit* |
|---|---|---|---|
| | slave-ACC | Xerxes-NOM | killed.3SG |

‘Xerxes killed a slave.’

This specification of verbs as inducing full propositional structure equally applies in cases where its associated metavariable argument annotations are provided from context. Such a case occurs in the building of paired ‘linked’ trees, a process defined in DS for construal of relative clauses, clausal adverbials, and also external topic constructions. Such secondary structures have an attendant requirement that the newly introduced proposition-requiring tree have somewhere within it a copy of that term (specified as $?\langle\downarrow_*\rangle Fo(\alpha)$: see Cann et al. (2005) for details).[6] With options to build an unfixed node within an individual tree or transitions from one tree to another to yield pairs of ‘linked’ trees, there are several strategies at the outset of building structure for any single string-interpretation pair such as (1), an intrinsic property of a parsing-directed grammar formalism.

[3] With context defined like parse states as a sequence of triples $\langle W, T, A\rangle$, anaphora construal is identifiable from formula or action values.

[4] Formally, nodes in a tree are uniquely defined by their relation to other such nodes, with consequent restriction to only one unfixed node of a type at a time in any partial tree: duplication induces immediate collapse of any such pair of nodes.

[5] Unlike two case-distinguished unfixed nodes, either subject or object nodes induced by actions of the verb harmlessly collapse with those introduced as unfixed and updated through constructive use of case Nordlinger (1998), as annotations provided by the verb are compatible with those provided by computational actions used in parsing the NPs.

[6] The process of inducing such pairs of semantic trees is permitted by defining an additional modal operator in the tree logic $\langle L\rangle$, and its inverse $\langle L^{-1}\rangle$; and a rule is defined to yield a transition from an arbitrary node in one tree across a LINK relation to the topnode of a new propositional tree.

### 2.1 Production

In production, the same rules apply (Purver et al., 2006): the only difference is that while the parser may not know in advance the interpretation to be constructed, the producer in contrast must do so, the formal reflex being the requirement in generation, that each update step licensed by the parsing mechanism has to constitute an enrichment towards completing a 'goal tree' representing the interpretation to be conveyed. Formally a subsumption relation is required to hold between the parse tree and the goal tree (Purver and Otsuka, 2002). Given the incrementality of parsing, carried over to production, this task is computationally expensive: incremental search in the lexicon word by word threatens to make production cognitively non-viable. However, production is just as context-dependent as parsing, and use of items in context allow massive reduction in full-lexicon search, hence in cognitive cost: any element in context that can be identified as adding appropriately to the tree may not require words to be uttered, as long as the effect of adding it as a tree update matches the subsumption condition. The context-dependency of anaphora is relatively well understood, hence also for argument nodes projected by the verb. However, this minimization of cognitive costs extends beyond using elements in context to identify construal of anaphoric devices, to choice of words, structure, and actions. Once a word or sequence of actions has been used in parsing/producing a string, these actions can be re-used with very considerable cost benefits, and this constitutes the basis for lexical and structural alignment effects (Pickering and Garrod, 2004):

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| (2) | *Te,* | *dea.* | *Te* | *fugiunt* | *vent-i.* | *Te* | *nubil-a* | *coel-i* |
| | you | goddess | you | flee.3PL | winds-NOM | you | clouds-NOM | heaven-GEN |

'You goddess, the winds flee from you, the clouds of heaven (flee from you).' (Lucretius, De Rerum Natura 1.6)

Minimizing production costs affects word order even without alignment. Though in Latin, there may be no need of a pronoun, anaphoric expressions serve a purpose in the linearization task as they enable argument terms to be identified independently of processing the verb. This consideration, in conjunction with the parallelism of parsing and production and general cognitive constraints such as relevance, helps to explain their preferred early positioning. In relying on context, speakers/hearers need the search for a substituend to be as small as possible by general relevance considerations. Accordingly, unless there is reason to the contrary, the position of an anaphoric expression requiring context-identification will be as early as possible in the setting out of propositional structure. This is of course no more than a relevance-based explanation of the wellknown given-before-new ordering. However, pronouns in Latin may be used to provide some initial term which constitutes a point of departure for what follows, or to provide a contrast, an update to what follows. In both such cases, the expression is set out initially, in order to be identifiably separate from the structure to be constructed from what follows:[7]

| | | | | | | |
|---|---|---|---|---|---|---|
| (3) | A: *Tibi* | *ego* | *dem?* | B: *Mihi* | *hercle* | *uero.* |
| | you.DAT | I.NOM | give.1SG | me.DAT | by Hercules | in truth |

'A: Am I to give it to YOU?' 'B: Yes, by god, to ME.' (Plautus, Pseudolus 626 apud Adams (1994))

[7] The pronouns noted in (3) are taken by Adams (1994) to be illustrative of an emphatic use "often marked by placement of the pronoun at the head of its clause"(pg. 104).

Such uses of so-called strong pronouns as in (3) are analysed as involving the projection by the pronoun of a term decorating a node at the left-edge of a propositional boundary, i.e. as a separate linked structure, or an unfixed node. There are in addition so-called weak uses of pronouns, which serve solely for anaphoric purpose. Being by definition complementary to strong use of pronouns, this remainder of the set of pronouns will not be associated with those very structural devices which serve to identify some initiation of an emergent propositional structure. Nevertheless, like their "strong" counterparts, positioning of pronouns under this use will be driven by relevance considerations. That is, once an emergent propositional structure is identified by some *other* expression, we can expect weak pronouns to occur as closely following as possible, decorating some locally unfixed node duly updated through its case specification.[8]

## 3 Towards Diachrony

We now have everything in place to explain why clitic pronouns cluster at some early position in a string. The weak pronouns of Latin occur as close to the left-edge of a clause as possible, but not quite at the edge. Rather, they follow those devices which define an emergent propositional boundary, i.e. following focussed elements, expressions containing a negative element, complementizers, relative pronouns, subordinate temporal adverbials, and verbs: indeed this is the only property common to this structurally heterogeneous set. We illustrate with relatives, complementizers, and initial verbs:[9]

| | | | | | |
|---|---|---|---|---|---|
| (4) | *quae* | ***tibi*** | *nulla* | *debetur* | [relative-pron. + pronoun] |
| | which.NEUT-PL | you.DAT | no.NEUT-PL | is-owed | |

'nothing of which is owed to you.' (Cicero, In Actilinum 1.16 apud Adams (1994))

| | | | | | | |
|---|---|---|---|---|---|---|
| (5) | *rogo* | *ut* | ***mi*** | *mittas* | *dalabram* | [complementizer + pronoun] |
| | ask.1SG | that | me.DAT | send.2SG | mattock | |

'I ask you to send to me a mattock.' (Terentianus 251.27 apud Adams (1994))

| | | | | | |
|---|---|---|---|---|---|
| (6) | *delectarunt* | ***me*** | *tuae* | *litterae* | [verb + pronoun] |
| | delighted.3PL | me | your | letter.NOM | |

'I was delighted with your letter.' (Cicero, Ad Familiares IX 16.1 apud Adams (1994))

In the subsequent Medieval Spanish system the clitic pronouns share this distribution (for a detailed account see Bouzouita, 2002, in preparation; Bouzouita and Kempson, 2006):

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| (7) | *Esto* | *es* | *el* | *pan* | *de* | *Dios* | *que* | ***vos*** |
| | this | be.3SG | the | bread | of | God | that | CL |
| | *da* | *a* | *comer* | | | | [rel-pron. + pronoun] | |
| | give.3SG | to | eat.INF | | | | | |

'This is the bread of God that he gives you to eat.' (Granberg, 1988, pg. 35)

[8] Following Sperber and Wilson (1995), if there are specific inferential effects to justify commensurate enlargement of the context to be searched, this would explain the lack of tightness of fit that Adams (1994) notes of weak pronoun positioning in Latin, even assuming that the effects are clause by clause (or "colon" by "colon"), to use his terminology.

[9] For visibility reasons, we have highlighted in bold the weak pronouns/clitics under consideration.

(8)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *e* | *dizie* | *que* | ***lo*** | *tenie* | *del* | *prior* | *de* | |
| and | said.3SG | that | CL | had.3SG | of-the | prior | of | |
| *Sancti* | *Johannis* | | | | | | | [compl. + pronoun] |
| of-Saint | Johan | | | | | | | |

'And he said that he got it from the prior of Saint John.' (Granberg, 1988, pg. 46)

(9)

| | | |
|---|---|---|
| *Connocio-**la*** | *Jacob* | [verb + pronoun] |
| recognised.3sg-CL | Jacob | |

'Jacob recognised her.' (Fazienda de Ultramar: 51)

Such left-peripheral items may however be a sequence of NPs (Devine and Stephens, 2006), a pattern which recurs in Medieval Spanish associated specifically with clitic pronouns:

(10)

| | | | | | | |
|---|---|---|---|---|---|---|
| *caseum* | *per* | *cribrum* | *facito* | *transeat* | *in* | |
| cheese | through | sieve | make.2SG | go-through.3SG | in | |
| *mortarium* | | | | | | [scrambled NP pair] |
| bowl | | | | | | |

'Make the cheese go through the sieve into the bowl.' (Cato 76.3 apud Devine and Stephens (2006))

(11)

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Et* | *los* | *dioses* | ***me*** | *quisieron* | *mal* | *e* | ***me*** | ***lo*** | *quieren* |
| And | the | gods | CL | wanted.3PL | harm | and | CL | CL | want.3PL |

'And the gods wanted to harm me and they still want to.' (Granberg, 1988, pg. 235)

Thus proclisis and enclisis effects in finite clauses for Latin weak pronouns and the clitic pronouns of Medieval Spanish can be described by a single generalization as a minimizing of context search, given the signal of the need to construct a distinct propositional domain.

## 4 Alignment, Routinization and Change

Without an explanation of the change, this is not yet a diachronic account; but dialogue effects go further than mere use of anaphoric devices and alignment. Dialogue participants, having set up a parse sequence of actions, quickly set up routines for retrieving a stored sequence of actions encompassing more than one word (Garrod and Doherty, 1994), yet another saving on cognitive costs since it involves retrieval from the lexicon of only one sequence of actions for a multiple string. Given the availability of more than one strategy to yield the same output tree for any string, there is more than one alternative choice of actions for individual participants in a dialogue, allowing flexibility, hence harmless mismatch between speaker/hearer choice of actions in certain syntactic environments. For either party though, choices will be constrained by processing-cost minimization.

Production, storage, and language change can now be seen as going hand in hand in the shift from Latin in the development of Spanish clitic pronouns. One form of pronoun gets progressively phonologically reduced in virtue of predictability of its construal from context. Given increasing phonological dissimilarity, separate clitic forms get encoded, being defined to follow any one of the heterogeneous set of triggers previously established through pragmatically induced production constraints.[10] The first observable step of encoding this set

[10] The strong pronouns subsequently come in Modern Spanish to be restricted to decorating linked structures, necessitating clitic doubling (Cann et al., 2005):

of triggers is a step of economy in combining computational/lexical actions as one lexical macro of actions. But, in all such cases, the commonest pattern is for the verb immediately follow the clitic(s) (Adams, 1994). A natural subsequent step of routinization is to store the actions of the verb along with those of the clitic, a re-routinization which again reduces processing effort on a two-for-the-price-of-one basis. We achieve the effect of re-bracketing, often observed. With such routinization, restrictions on proclisis collapse, the heterogeneous set of triggers defining the environment licensing construal of a clitic not being an appropriate basis for subclassifying verbs; and we get the intermediate stage of Renaissance Spanish, when all constraints on pre-verbal positioning of the clitics drop (Bouzouita and Kempson, 2006; Bouzouita, 2002, in preparation).

From this point on, the Romance languages, with the disappearing free constructive use of case, face the problem of confronting a ban on more than one unfixed node at a time on NP construal. Preserving case distinctions, divergent clitic routinizations emerge to side-step the problem. Some clitics directly induce the construction of the requisite fixed structural relation (e.g. French *le*). Others induce the building of a locally underspecified tree relation, hence underspecified with respect to the two discrete object construals (e.g. French *me, te*, Castilian Spanish *le*). And in some cases a phonologically distinct composite clitic form is introduced that induces a single unfixed relation from which are constructed two argument nodes (e.g. Italian *glielo*, Spanish *se lo*). It is notable that each of these possibilities corresponds to actions put to use as strategies for projecting argument nodes prior to the verb (scrambling), albeit at this point in time stored as lexical sequences of actions. Even the last of these alternatives is the local analogue of actions underpinning multiple long-distance scrambling as in (10) (Kiaer, 2007). Seen in processing terms, the clitic-template phenomenon is thus a freezing of scrambling strategies, hence explicable as a progressive shift, each change in lexical specification reflecting only one of a set of strategies for early NP placement. Furthermore, because each such transition is motivated by the ever-present pressure on production imposed by relevance-driven constraints, the process is one that take place gradually across a time-span for all speakers within a language community.

With the grammar now seen as a system which induces progressive growth of semantic representation, this opens the way for a novel perspective on language evolution. First, the length of the evolutionary chain from iconic use of signals (with one-one correspondence of input-interpretation) to a natural language system (with one-many correspondence of input-interpretation) is much shorter than formalisms with distinct components of morphology, syntax and semantics. On this perspective, the emergence of language systems lies largely in recognition that a signal or sequence of signals underspecifies its construal on any occasion of use, requiring enrichment from context in interaction with co-occurring signals. Secondly, in constituting one model of such systematization, the DS account sustains coordination at two levels: (i) at the individual level as between the activities of speech and understanding, (ii) at the level of the groups which the language or individual traits within that language define - ensured by the persistent pressure to minimize costs for all users of the language. This approach thus opens the way for developing accounts of language

| | | | | |
|---|---|---|---|---|
| (1) | Le | hablaron | a | ella |
| | CL | spoke.3PL | to | her |

‘They spoke to her.’

evolution that are simultaneously cognitively grounded (Arbib, 2005), yet also a reflection of multi-level selection pressures (Sober and Wilson, 2002).

## References

J. Adams. Wackernagel's law and the position of unstressed personal pronouns in classical latin. *Transactions of the Philological Society*, 92:103–178, 1994.

M. Arbib. From monkey-like action recognition to human language: an evolutionary framework for neurolinguistics. *Behavioral and Branic Sciences*, 28:105–167, 2005.

P. Blackburn and W. Meyer-Viol. Linguistics, logic and finite trees. *Bulletin of Interest Group of Pure and Applied Logics*, 2:2–39, 1994.

M. Bouzouita. *Clitic Placement in Old and Modern Spanish.* Msc, Kings College London, 2002.

M. Bouzouita. *Clitic Placement in the History of Spanish.* Phd, King's College London, in preparation.

M. Bouzouita and R. Kempson. Clitic placement in old and modern spanish: a dynamic account. In O. Nedergaard Thomsen, editor, *Competing Models of Linguistic Change.* John Benjamin, 2006.

R. Cann, R. Kempson, and L. Marten. *The Dynamics of Language.* Elsevier, Oxford, 2005.

J. Devine and L. Stephens. *Latin Word Order: Structured Meaning and Information.* Oxford University Press, Oxford, 2006.

S. Garrod and G. Doherty. *Cognition*, 53:181–215, 1994.

R. Granberg. *Object Pronoun Position in Medieval and Early Modern Spanish.* Phd, University of California Los Angeles, 1988.

R. Kempson, W. Meyer-Viol, and D. Gabbay. *Dynamic Syntax.* Blackwell, Oxford, 2001.

J. Kiaer. *Processing and Interfaces in Syntactic Theory: The case of Korean.* Phd, KCL London, 2007.

R. Nordlinger. *Constructive Case.* CSLI, Stanford, 1998.

M. Pickering and S. Garrod. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169–226, 2004.

M. Purver, R. Cann, and R. Kempson. Grammars as parsers: meeting the dialogue challenge. *Research on Language and Computation*, 4:289–326, 2006.

M. Purver and M. Otsuka. Incremental generation for dialogue. *ACL workshop proceedings*, 2002.

E. Sober and D. Sloan Wilson. *Unto Others: A Study of Altrusim.* Harvard University Press, Harvard, 2002.

D. Sperber and D. Wilson. *Relevance: Communication and Cognition.* Blackwell, Oxford, 1995.

# Evolving Lexical Networks. A Simulation Model of Terminological Alignment

Alexander Mehler
Bielefeld University
Alexander.Mehler@uni-bielefeld.de

**Abstract**

In this paper we describe a simulation model of terminological alignment in a multiagent community. It is based on the notion of an *association* game which is used instead of the classical notion of a *naming* game (Steels, 1996). The simulation model integrates a small world-like agent community which restricts agent communication. We hypothesize that this restriction is decisive when it comes to simulate terminological alignment based on lexical priming. The paper presents preliminary experimental results in support of this hypothesis.

## 1 Introduction

The approach to interactive alignment in communication (Pickering and Garrod, 2004) postulates two mechanisms of alignment: (i) *priming* as a short-term mechanism of information percolation within the same or between different levels of representation and (ii) *routinization* as a long-term mechanism of expectation driven control of dialogue unfolding. So far, this mechanistic approach has not been tested by a simulation model of the build-up of routines and their underlying priming processes. This paper describes a multiagent model of the emergence of *intralevel* priming relations and, thus, serves as a preparatory study to such a simulation model. It focuses on semantic priming based on lexical associations and postulates that alignment is manifested by the harmonization of contiguity and similarity associations among interacting agents. Further, it hypothesizes that this alignment converges into a lexical network which constrains the lexical choices of interlocutors in order to maintain the success of their communication. As far as language communities are considered (beyond small groups of a couple of interlocutors), this leads to a dynamic understanding of lexical networks which are seen to converge into fluent equilibria in order to serve the communication needs of speakers and hearers.

The present research is in the line of approaches to circumvent models of linguistic conventions (Lewis, 1969) which rely on common ground in terms of knowledge that is shared and known to be shared (Clark, 2000). Recently, Barr (2004) has re-conceptualized simulation models of intra-generational learning in terms of such a circumvention. He argues against simulation models in which frequencies of successful communications are implicitly utilized as shared representations or in which all agents interact with each other. Such an unstructured community model is characteristic of many implementations of the naming game (Steels, 1996; Hutchins and Hazlehurst, 1995; Baronchelli et al., 2006a). Thus, in order to prevent that common ground simply results from an unstructured community model, it is required to be structured. Although we follow this conception, we depart from Barrs simulation model in two respects: Firstly, we consider sense relations of lexical units and, thus, do not implement a classical naming game. Secondly, we consider community models which are more realistic in terms of what is known about social communities. More specifically, we aim at a simulation model which obeys the principles of complex networks (Newman, 2003) and integrates a learning model of lexical associations subject to discourse processing in a multiagent setting. The simulation model integrates three areas: the theory of language evolution (Niyogi, 2006; Jäger, 2006), *Latent Semantic Analysis* (LSA) (Landauer and Dumais, 1997) as a learning theory of lexical associations and *Complex Network Theory* (CNT) which studies constraints of networking in agent *and* in linguistic networks.

The basic idea is to utilize the community model as the *independent* variable and the topology of the evolving linguistic network as the *dependent* one. We predict that if the community model has the *Small World* (SW) property (Newman, 2003) then the linguistic network — which is evolving subject to information flow within this community structure — is similar to *real* linguistic networks (Mehler, 2007). It turns out that this similarity is indicated by the SW-property. Thus, we predict that the SW-property of the community network and the simulated linguistic network correlate positively. The paper presents preliminary results in support of this hypothesis. It is organized as follows: Section 2 outlines a vector space model whose computation can be iterated in a multiagent setting and, thus, serves as a simplified reconstruction of LSA. Section 3 presents an algorithm for the generation of SW-like communities in accordance with well-known results of CNT. Section 4 describes an association game which is designed to simulate terminological alignment in agent communities. Section 5 presents results of a preliminary simulation experiment. Finally, Section 6 gives a conclusion and prospects future work.

## 2 Toward Distributed Semantic Spaces

Approaches to multiagent simulations of language learning and evolution include *inter*-generational simulations in which languages evolve subject to the bottleneck of generation change. The *iterated learning model* is a prime example of this approach (Kirby and Hurford, 2002). On the other hand, *intra*-generational models of language learning and change concentrate on the spread of linguistic forms, meanings, lexicons or even grammars in the same generation of learners (Niyogi, 2006). This paper belongs to the latter framework. However, we will uniformly speak of *multiagent simulation models of language learning* without differentiating inter- and intra-generational approaches. This is done in order to give a more general account of the type of language game we will introduce.

In the area of semantics, simulation models mostly implement a variant of the *signalling* or, more specifically, *naming game* (Steels, 1996). This game starts from a set $E$ of expression units and a set $M$ of meaning units where each agent $a \in P$ of a given population $P$ learns a meaning relation $\|\|_a \subseteq E \times M$. Under this regime, a simulation is said to be successful after a number of iterations $i$ if a “language” converges in $P$, that is, if each agent $a \in P$ has the same meaning relation and retains it after any further iteration. Generally speaking, the naming game relies on a *bipartite* graph model whose instantiations range from a state of complete semantic diversification (a single expression carries several meanings) to complete formal diversification (a single meaning is expressed by several expressions) where the “ideal” of a simultaneous unification (i.e., a one-to-one mapping) lies in-between. Note that under this perspective, $E$ and $M$ are seen to be unstructured (cf., e.g., Baronchelli et al. (2006a) who treat mappings of different objects as independent).

There are many reference points under which implementations of the naming game differ: Firstly, simulation models may (and most of them do) require that $\|\|_a$ is a bijective function so that neither synonymy nor homonymy can emerge in $a$'s memory (Barr, 2004; Baronchelli et al., 2006a). Secondly, simulations may allow for growing sets $E$ and $M$ during iterations (Steels, 1996). Thirdly, the meaning relation $\|\|_a$ may be weighted by means of frequencies so that mappings between meanings and expressions can be conceived as probability functions — this approach leads to a Zipfian notion of signalling and its study in diversification analysis (Altmann, 1985). Fourthly, the topology of the community may vary so that agents are additionally endowed with a neighbor selection function (Barr, 2004).

Although some of these variants are considered in this paper, we depart from the classical setting of the naming game as follows: instead of a bipartite graph model, we use a unipartite one in which vertices denote signs whose association links are object of learning. Hashimoto (1997) proposes a pioneering approach towards multiagent learning of such associations. In such a model, the meaning relation $\|\|_a$ is defined as a homogeneous relation $\|\|_a \subseteq V \times V$ over a set $V$ of signs where, initially, an agent $a$ does not associate any signs, i.e. $\|\|_a^{t=0} = \emptyset$.[1] As $G(a,t) = (V, \|\|_a^t)$ defines a directed graph with vertex set $V$ and edge set $\|\|_a^t$, the memory graph $G(a,t)$ of an agent $a$ at time $t$ can be conceived as a lexical network if $V$ is restricted to lexical units. Now, convergence after a number of iterations means that all agents have learnt the same lexical associations and, thus, the same lexical network. As the number of lexical items and, correspondingly, their candidate associations can get very large, a threshold of community-wide conformity is needed instead. In the naming game of Barr (2004), e.g., four expression and meaning units are considered so that learning has to choose among 24 candidate mappings. If in contrast to this, a set of $|V| = n$ lexical items is considered, there are $2^{n(n-1)}$ directed candidate graphs which can be built out of these $n$ items (for the sake of simplicity, the identity of link weights is disregarded). In order to master this complexity, we redefine the convergence criterium as follows: Let $0 \ll \sigma < 1$ be a threshold. A simulation of an association game (see Section 4) is said to converge after a number of $t \in \mathbb{N}$ iterations if the graph $G(P,t) = (V, E_t, \omega_t)$ with vertex set $V$ and edge set

$$E_t = \left\{ (v,w) \;\middle|\; \frac{|\{G(a,t) = (V, \|\|_a^t, \omega_a^t) \,|\, a \in P \wedge (v,w) \in \|\|_a^t\}|}{|P|} \geq \sigma \right\} \tag{1}$$

has the SW-property according to the Watts-Strogatz-model and the preferential attachement model and does retain it for any further iteration — cf. Bollobás and Riordan (2003) for a formal account of the latter two models. This redefinition of convergence is in the line of approaches which refer to quantitative characteristics (e.g., indices as the cluster coefficient or distributions as some power law) of linguistic systems as reference points of judging simulation quality: as far as the simulating system approximates the corresponding characteristics of the simulated one it is said to be successful (Mehler, 2006). $\omega_t \colon E_t \to \mathbb{R}$ is an association measure (Bock, 1974) appropriately derived from the set of functions $\{\omega_a^t \colon \|\|_a^t \to \mathbb{R} \,|\, a \in P\}$. We call

$$\{(V, \|\|_a^t, \omega_a^t) \,|\, a \in P\} \tag{2}$$

[1] Note that we do not define $\|\|_a$ as a relation over $E$ or $M$, but over a set of de Saussurean articulations of both sets. This opens the door to combine the naming game with the association game developed here.

a *distributed semantic space* which after $t$ iterations of the association game is distributed over the population $P$.

The question arises how to model the association functions $\omega_a^t$. This can be done according to the *Weak Contextual Hypothesis* (WCH) (Miller and Charles, 1991) which says that the similarity of the contextual representations of words contributes to their semantic similarity. The WCH is most prominently implemented by LSA as a model of contiguity and similarity associations (Landauer and Dumais, 1997). It implements semantic spaces (Rieger, 1978) as a geometric model of meaning in which signs are interrelated even if they do not co-occur, but are similar according to the WCH. That is, signs are mapped onto meaning points (e.g. feature vectors) whose geometric distance models their semantic similarity. Although LSA focuses on language learning, it does not deal with the evolution of linguistic systems, nor does it explore constraints which separate "natural" semantic spaces from implausible ones. Rather, LSA is a *single*-agent model leaving out the dynamics of *multiagent* communication.

Empirical evidence which allows deciding the cognitive plausibility of competing models of semantic spaces comes from *Complex Network Theory* (CNT) (Newman, 2003). Recently, Steyvers and Tenenbaum (2005) interpreted the SW-property of association networks as an indicator of efficient information storage and retrieval. They apply CNT which investigates network topologies in terms of their small world characteristics (Watts and Strogatz, 1998): Firstly, compared to random graphs, SW-graphs have a considerably higher amount of cluster formation. Secondly, compared to regular graphs, any randomly chosen pair of nodes of a SW-graph has, on average, a considerably shorter geodesic distance (the geodesic distance of two vertices in a graph is the length of the shortest path in-between). Steyvers and Tenenbaum (2005) show that lexical association networks as well as reference systems like WordNet share these properties. A central implication of these findings is that they question the cognitive plausibility and adequacy of LSA and related models which rely on semantic spaces in terms of *completely connected weighted graphs* as the underlying memory representation format — by refusing this memory model we view the signalling system to be learnt no longer to be unstructured.

Although Steyvers and Tenenbaum (2005) propose a growth model of semantic networks in accordance with CNT, they do not develop a multiagent simulation out of which such networks emerge. One might think that a candidate solution comes from CNT itself where lexical association networks are a prime object of study. But, actually, these co-occurrence networks are far to simple to grasp the kind of lexical associations underlying priming. The reason is that they consider any two items to be linked if they co-occur at least once (Ferrer i Cancho et al., 2007). In contrast to this, an approach is needed which is sensitive to the frequencies of co-occurrences. Although LSA clearly meets this requirement, it is nevertheless much to complex to serve as a learning model in a multiagent setting. This insufficiency is clarified by the following definition:

**Definition 1** *A lexical association measure $\alpha$ is said to be* iteratively computable *if for any sequence $\langle x_1, \ldots, x_n, y\rangle$ of texts the following statement is true for any pair of lexical items $v, w \in V$ and some function $\beta$ which is irrespective of co-occurrences in $x_1, \ldots, x_n$:*

$$\alpha(v, w, \langle x_1, \ldots, x_n, y\rangle) = \alpha(v, w, \langle x_1, \ldots, x_n\rangle) + \beta(v, w, y)$$

Iteratively computable association measures which are sensitive to certain sequences of texts to be processed are indispensable for modeling distributed semantic spaces. The reason is that in such simulation models single agents process different sequences of texts which unfold with the ongoing simulation experiment. Further, agents may process different texts simultaneously so that the set of texts being produced and processed in such experiments is partially ordered — other than predicted by the set-theoretical corpus model of LSA. This also holds for the Vector Space Model (VSM) whose iterative computation already fails because of its weighting scheme based on logarithmic damping.

In order to arrive at iteratively computable association measures we reconstruct the VSM in a way which prevents the usage of logarithms and standardization. Three functions are needed: (i) a function for the iterative memorization of lexical associations, (ii) a function for calculating priming relations of lexical items and (iii) a function for mapping priming relations of textual units:

- For a sequence $S = \langle x_1, \ldots, x_n \rangle$ of $n$ texts, the association of two lexical items $v_i, v_j \in V$ is computed as

  $$\alpha(v_i, v_j, S) = \sum_{k=1}^{n} \left( f_{ik} \frac{k}{F_{ik}} \right) \left( f_{jk} \frac{k}{F_{jk}} \right) = \sum_{k} k^2 \frac{f_{ik} f_{jk}}{F_{ik} F_{jk}} \tag{3}$$

  where $F_{ik}$ is the number of texts out of $\langle x_1, \ldots, x_k \rangle$ in which $v_i$ occurs; $f_{ik}$ is the frequency of $v_i$ in $x_k$. For a given $k$, $\frac{k}{F_{ik}}$ is the larger, the rarer $v_i$'s text frequency $F_{ik}$; this effect is reinforced by $f_{ik}$. $f_{ik}\frac{k}{F_{ik}}$ resembles the TFIDF-weighting scheme of the VSM, but disregards any kind of, e.g., logarithmic damping as well as standardization. This is done in order to meet Definition 1.

- As we assume that agents do *not* memorize any single text, but only text frequencies, the cosine approach of the VSM in measuring lexical similarities is obsolete. Instead of this, a lexical item $v_j$ is said to be primed by an item $v_i$ for an agent $a$ at time $t$ to the degree of $\alpha(v_i, v_j, S_a^t)$ where $S_a^t$ is the sequence of texts $a$ has processed till iteration $t$. Accordingly, we utilize $\alpha(v_i, v_j, S_a^t)$ to build the lexical network $(V, \| \|_a^t, \omega_a^t)$ memorized by agent $a$ at time $t$, that is, $\omega_a^t(v_i, v_j) = \alpha(v_i, v_j, S_a^t)$.

- Finally, a text $x$ is represented as a fuzzy set of lexical items whose membership value is computed by a function of the degrees to which they are primed by the lexical tokens of $x$. These membership values model the degree by which items are primed by *$x$ as a whole*, that is, in terms of *text priming* (Sharkey and Sharkey, 1992).

## 3 Artificial Small World Networks

The majority of models of language evolution or change starts from an unrealistic community model in which for an increasing number of iterations all agents tend to communicate to every other agent to the same degree. This model corresponds to a *Completely Connected Graph* (CCG) where the smaller the agent community, the less iterations are needed to get a CCG. Earlier simulation models of language evolution exemplify this situation as they are based on very small populations (Hutchins and Hazlehurst, 1995; Steels, 1996; Hashimoto, 1997). Conversely, if the number of iterations is small but the population large, agents have the chance to communicate only with a small number of other agents so that a random graph emerges according to the probability function used to choose speakers and hearers. Generally speaking, while in a CCG clustering is maximal and geodesic distances

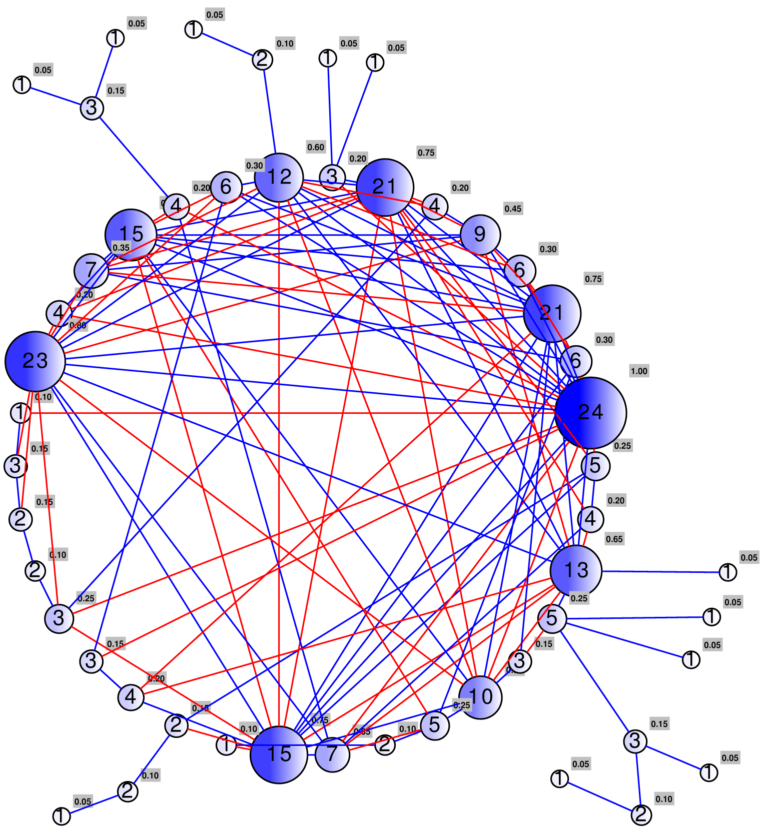


Figure 1: A sample Aswan $G = (V, E)$, $|V| = 50$, in which node centrality is signalled by vertex size. The SW-profile of $G$ is as follows: cluster coefficient $C(G) = 0.25$, average geodesic distance $L(G) = 3.09$, assortativity $A(G) = 0.11$, power law-exponent $\gamma = 1.3$.

are minimal, random graphs assume short, but not minimal average geodesic distances in conjunction with small clustering values (Watts and Strogatz, 1998). Thus, whereas CCGs are unrealistic in terms of the topology of social networks, random graphs of the latter kind disregard, at least, the clustering of social groups.

Barr (2004) compares CCGs with a structured community model. Starting from randomly distributing agents over a plane $\Pi$ each agent $a$ is assigned a Gaussian neighbor selection function $N_a$ which measures the probability by which another agent $b$ is a neighbor of $a$ as a function of their distance in $\Pi$. Under this regime, the more peaked $N_a$, the smaller the agents' neighborhoods, the longer the average geodesic distance of agents mapped by consecutive neighborhoods. As this model does not know short-cuts for linking "remote" agents, geodesic distances tend to be longer than in random graphs. At the same time, the less peaked $N_a$, the more likely two neighbors of $a$ are neighbors on their own. Thus, Barrs model seems to approximate though not being equal to some regular graph of the same size. Regular graphs are known for high cluster values and long geodesic distances (Watts and Strogatz, 1998) — *obviously a likewise unrealistic model of social communities.*

What is missed so far is a community model which integrates efficient information flow by means of short-cuts with clustering as known from social groups together with highly skewed communication links and heterogeneous agent mixing (apart from homogeneously mixing CCGs). Although there are approaches to language evolution which consider related graph models, they either handle them as the *dependent* variable (as done by Gong and

**Input:** number $N$ of vertices, exponent $\gamma$ of power law, fraction p of local links. **Steps:**

1. **Preferential attachment:** Generate a power law $\mathsf{P}[\mathsf{N}, \gamma] = \mathsf{a}\mathsf{k}^{-\gamma}$ and assign each vertex $\mathsf{v_i} \in \{\mathsf{v_1}, \ldots, \mathsf{v_N}\} \in \mathsf{V}$ a degree $\mathsf{d_{v_i}}$.
2. **Random plane:** Randomly place N vertices $\mathsf{w_1}, \ldots, \mathsf{w_N} \in \mathsf{W}$ on a plane and map each of the vertices $\mathsf{v_i} \in \mathsf{V}$ onto W such that the more central $\mathsf{w_j}$ (in terms of the sum of its Euclidean distances to every other vertex in W), the higher the degree of $\mathsf{v_i}$ mapped to $\mathsf{w_j}$. **Variant:** randomly place the N vertices $\mathsf{v_1}, \ldots, \mathsf{v_N}$ on a plane.
3. **Minimal spanning tree:** Construct a *power law-conformant minimal spanning tree* (PLC-MST) which spans all vertices in V and is conformant to $\mathsf{P}[\mathsf{N}, \gamma]$.
4. **Local transitivity-providing links:** link each vertex $\mathsf{v} \in \mathsf{V}$ to its nearest $\mathsf{pd_v} - \mathsf{d'_v}$ neighbors where $\mathsf{d'_v}$ is the number of edges by which v has already been linked in the PLC-MST of Step 3.
5. **Remote short-cuts providing links:** randomly link each vertex to $\mathsf{d_v} - \mathsf{d''_v}$ vertices where $\mathsf{d''_v}$ is the number of edges adjacent to v as generated in Step 3 and 4.

Table 1: An algorithm for the generation of small-world networks.

Wang (2005) who consider the emergence of agent communities as a by-product of language evolution, but disregard the SW-property of the linguistic network), or consider only a selected number of characteristics of social networks (as done, e.g., by Baronchelli et al. (2006b) who concentrate on the preferential attachement model of Barabási and Albert (1999)). What we need instead is a SW-like community model which combines several of these features and can be used as the *independent* variable in simulation experiments.

Table 1 presents an algorithm for the generation of such *Artificial Small-World Agent Networks* (Aswan). It is based on Jin and Bestavros (2006) who generate SW-graphs according to the model of Watts and Strogatz (1998) *and* of Barabási and Albert (1999). Thus, their graphs combine high cluster values with short geodesic distances and skewed degree distributions. However, this model may generate disconnected graphs. In order to prevent this we generate SW-graphs according to Algorithm 1. This algorithm differs from Jin and Bestavros (2006) in that it constructs a minimal spanning tree which guarantees connectedness of the vertices randomly mapped onto a plane. Further, we consider the variant that the degrees $d_v$ mapped to vertices $v$ reflect their position in the plane so that the more central a vertex, the higher its degree. Figure 1 exemplifies a small-world graph generated by Algorithm 1. Figure 2 outlines the behavior of the algorithm. It shows that for increasing values of $\gamma$ the cluster value of the network decreases, while the average geodesic distance increases. The next section utilizes Algorithm 1 as a generative model of small world agent networks.

## 4 Association Games

In this section we describe an association game as part of an intragenerational simulation model of terminological alignment in a multiagent community which has the small world-property. The basic notion is that of an *association game* which replaces the classical notion of a signalling game (cf. Section 2). The simulation model takes as input a small world agent community. It restricts which agents can communicate with each other. Further, the model iterates the association game defined as follows:

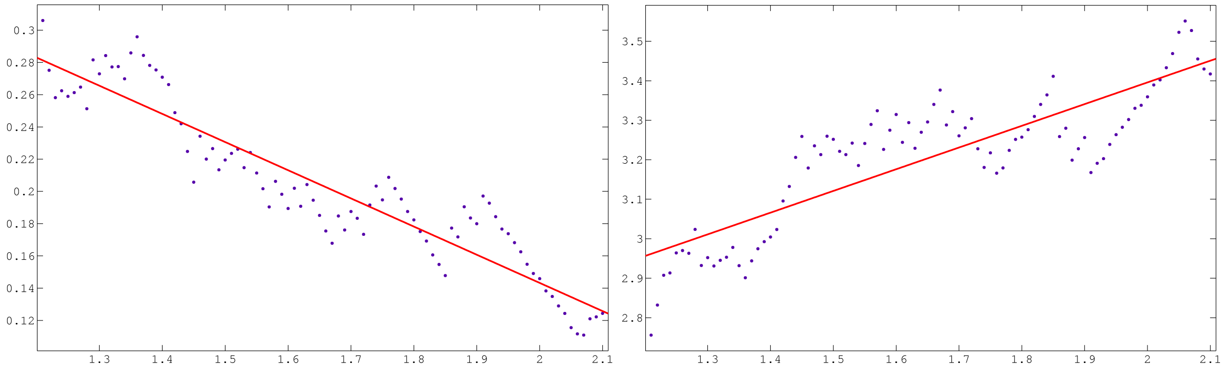


Figure 2: *Left:* $C$ as a function of $\gamma$ (determination coefficient of fitting $\approx .87$), $p = .6$. *Right:* $L$ as a function of $\gamma$ (determination coefficient of fitting $\approx .71$), $p = .6$. Numbers are averaged over 500 runs using 100 agents in each run.

**Community Model** Let $C(P) = (P, E)$ be a SW-community model of population $P$ generated by Algorithm 1. Let further round $t \in \mathbb{N}$ be given. The sender $a_S$ is picked at random who will then communicate to all his neighbors (in the sense of a "classroom situation"). That is, we do not pick the listener $a_L$ at random among the neighbors of $a_S$. The reason is that in a SW-graph with node connectivity according to a power law the chance is high that a purely connected agent is selected as the sender while highly connected agents have a higher chance to be selected subsequently as the listener (cf. Baronchelli et al. 2006b). Under this regime, highly connected agents would be "instructed" more often by lowly connected ones than vice versa — in contrast to what is expected from their topological significance as a sort of hubs.

**Memory Model** Each agent $a \in P$ disposes of a semantic space $G(a, t) = (V, \|\|_a^t, \omega_a^t)$ where the lexicon $V$ is shared by all agents $a \in P$. $\{(V, \|\|_a^t, \omega_a^t) \,|\, a \in P\}$ is a semantic space distributed over $P$ at time $t$ according to Section 2. Note that $G(a, t)$ and $G(b, t)$ will differ for two agents $a \neq b$ subject to the varying communication situations to which they participate.

**Text Generation Model** A lexical prime $v_+ \in V$ is randomly picked and used by the sender $a_S$ to activate a subgraph in his semantic space $G(a_S, t)$. This subspace consists of a subset of the nearest neighbors of $v_+$ in $G(a_S, t)$. As an example, suppose that $v_+$ equals *color* so that its neighborhood consists of words like *red*, *yellow*, *sun* etc. Initially, primed lexical neighbors are picked at random. The resulting neighborhood is used to let $a_S$ produce an output text $x_t$ which consists of $m$ tokens of the lexical items primed by $v$ in $G(a_S, t)$. This procedure generates a multiset in which the same item may recur where tokens are selected randomly. Obviously, this is the entry point for a more elaborate text generation model (cf. Biemann 2007) which generates texts of varying length that obey some well-known quantitative text characteristics.

**Update Model** The output text $x_t$ of the sender is then processed by the listener $a_L$ who utilizes the lexical tokens of $x_t$ to prime an item $v_-$ in $G(a_L, t)$. Next, this item is compared with $v_+$ where the longer the geodesic distance $L(v_-, v_+)$ in $G(a_S, t)$, the less the communication success. That is, sender and listener play an association game in which the sender is masking which item he used to prime the tokens of his output text and where the listener has the task to find out which word the sender had initially in mind when producing

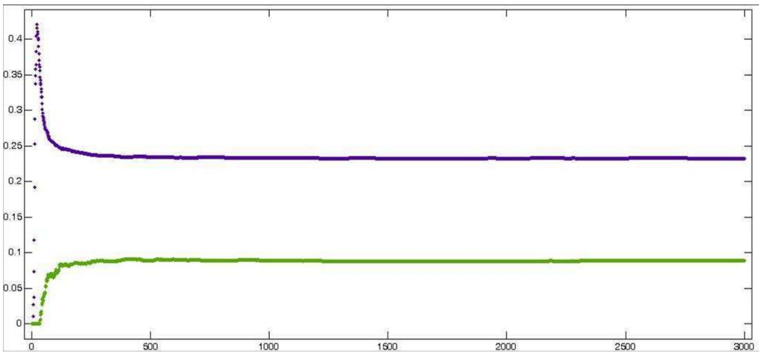


Figure 3: The cluster value of the summary lexical network as a function of the number of iterations. Upper curve: SW-community based run; lower curve: random graph-community based run.

$x_t$. We assume that the listener tells the sender the word, i.e. $v_-$, he has interpreted so that the sender can decide whether he was understood or not. The association game is successful if both sender and listener associate the same or related words with the same input text. In order to implement an update model which goes beyond decision theory, we call the communication of $a_S$ and $a_L$ *successful* if $L(v_-, v_+) \leq r$ for some $r \in \mathbb{N}$ as a further parameter of the model. Otherwise the association game is *unsuccessful*. In successful communication both sender and listener update their memories according to Equation 3, otherwise not. In a more realistic setting it is expectable that the sender is his "first" recipient, whereas the listener will process $x_t$ even if he does not unmask the prime $v_+$ or one of its nearest neighbors correctly. Under this regime, a more appropriate update model is to use $x_t$ to update the sender and listener memory *twice* if communication is successful (reinforcement learning), but only *once* if it is not.

The association game implements, so to speak, a variant of "*I spy with my little eye, something beginning with ...*".[2] Here, *terminological alignment* means that sender and listener align their lexical associations as they continually communicate so that they finally play the game more and more often successful. This is the local perspective. On the global level we expect that this game leads to a lexical network which has the SW-property subject to the SW-property of the community network $C(P)$ — without any reference to common ground shared by the whole community.

## 5 A Preliminary Experiment

This section exemplifies a simplified version of the association game: for a population of 100 agents we compare two competing community models $C_1(P)$, which is generated according to Algorithm 1 as an Aswan, and $C_2(P)$, which is a random graph. We consider a lexicon $V$ of 500 items and set the threshold $\sigma$ of the summary language network (see Equation 1) to 37.5%. Further, we compute $t = 3,000,000$ iterations of the association game and fall back to a decision-theoretical update model in which sender and receiver memory is updated after each iteration. Figure 3 compares both community models by example of the cluster coefficient $C$. It shows that $C_1(P)$, but not $C_2(P)$, converges into a lexical

[2] In German: "*Ich sehe was, was Du nicht siehst ...*".

network which shows clustering as known from SW-like lexical networks (Mehler, 2007). This picture is confirmed by the average geodesic distance and the higher speed by which a single connected component emerges in the lexical network of the SW-community $C_1(P)$. Evidently, these results have to be substantiated by a thorough parameter study.

## 6 Conclusion

In this paper we have introduced a novel type of language game. Based on the notion of alignment in communication, we defined an association game by which interacting agents may align their lexical associations in sequenced communication situations. Further, we argued for a more realistic and, thus, structured community model as input to simulation experiments which reflects the insights of computational sociology. In order to meet this requirement we proposed an algorithm for the automatic generation of such artificial communities. A first test of the model indicates promising results. Future work will focus on systematically testing the model and elaborating it in terms of a turn-taking model. This will be a further step towards a simulation model of dialogical alignment.

## References

Gabriel Altmann. Semantische Diversifikation. *Folia Linguistica*, 19:177–200, 1985.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286: 509–512, 1999.

Andrea Baronchelli, Maddalena Felici, Vittorio Loreto, Emanuele Caglioti, and Luc Steels. Sharp transition towards shared vocabularies in multi-agent systems. *Journal of Statistical Mechanics: Theory and Experiment*, P06014, 2006a.

Andrea Baronchelli, Vittorio Loreto, Luca Dall'Asta, and Alain Barrat. Bootstrapping communication in language games: strategy, topology and all that. In *Proceedings of the 6th International Conference on the Evolution of Language*, pages 11–18, 2006b.

Dale J. Barr. Establishing conventional communication systems: Is common knowledge necessary? *Cognitive Science*, 28(6):937–962, 2004.

Chris Biemann. A random text model for the generation of statistical language invariants. In *Proceedings of HLT-NAACL-07, Rochester, NY, USA*, 2007.

Hans H. Bock. *Automatische Klassifikation. Theoretische und praktische Methoden zur Gruppierung und Strukturierung von Daten.* Vandenhoeck & Ruprecht, Göttingen, 1974.

Béla Bollobás and Oliver M. Riordan. Mathematical results on scale-free random graphs. In Stefan Bornholdt and Heinz G. Schuster, editors, *Handbook of Graphs and Networks. From the Genome to the Internet*, pages 1–34. Wiley-VCH, Weinheim, 2003.

Herbert H. Clark. *Using Language.* Cambridge University Press, Cambridge, 2000.

Ramon Ferrer i Cancho, Alexander Mehler, Olga Pustylnikov, and Albert Díaz-Guilera. Correlations in the organization of large-scale syntactic dependency networks. In *Proc. of TextGraphs-2 at NAACL-HLT'07, Rochester, New York*, 2007.

Tao Gong and William S-Y. Wang. Computational modeling on language emergence: A coevolution model of lexicon, syntax and social structure. *Language and Linguistics*, 6(1):1–41, 2005.

Takeshi Hashimoto. Usage-based structuralization of relationships between words. In P. Husbands and I. Harvey, editors, *ECAL97*, pages 483–492. MIT Press, 1997.

E. Hutchins and B. Hazlehurst. How to invent a lexicon: the development of shared symbols in interaction. In G. N. Gilbert and R. Conte, editors, *Artificial Societies: The computer simulation of social life*. UCL Press, London, 1995.

Gerhard Jäger. Convex meanings and evolutionary stability. In Angelo Cangelosi, Andrew D. M. Smith, and Kenny Smith, editors, *The Evolution of Language. Proceedings of the 6th International Conference (EVOLANG6), Rome*, pages 139–144, 2006.

Shudong Jin and Azer Bestavros. Small-world characteristics of internet topologies and implications on multicast scaling. *Computer Networks*, 50(5):648–666, 2006.

Simon Kirby and James R. Hurford. The emergence of linguistic structure: An overview of the iterated learning model. In Angelo Cangelosi and Domenico Parisi, editors, *Simulating the Evolution of Language*, chapter 6, pages 121–148. Springer, London, 2002.

Thomas K. Landauer and Susan T. Dumais. A solution to plato's problem. *Psychological Review*, 104(2):211–240, 1997.

David Lewis. *Conventions. A Philosophical Study.* Harvard U.P., Cambridge, Massachusetts, 1969.

Alexander Mehler. Stratified constraint satisfaction networks in synergetic multi-agent simulations of language evolution. In Angelo Loula, Ricardo Gudwin, and João Queiroz, editors, *Artificial Cognition Systems*, pages 140–174. Idea Group Inc., Hershey, 2006.

Alexander Mehler. Large text networks as an object of corpus linguistic studies. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook of the Science of Language and Society.* De Gruyter, Berlin/New York, 2007.

George A. Miller and Walter G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.

Mark E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

Partha Niyogi. *The Computational Nature of Language Learning and Evolution.* MIT Press, Cambridge, 2006.

Martin J. Pickering and Simon Garrod. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169–226, 2004.

Burghard B. Rieger. Feasible fuzzy semantics. In K. Heggstad, editor, *COLING-78*, pages 41–43. ICCL, Bergen, 1978.

A. J. C. Sharkey and N. E. Sharkey. Weak contextual constraints in text and word priming. *Journal of Memory and Language*, 31(4):543–572, 1992.

Luc Steels. Self-organising vocabularies. In C. Langton and T. Shimohara, editors, *Proceedings of Artificial Life V, Nara, Japan*, 1996.

Mark Steyvers and Josh Tenenbaum. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29(1):41–78, 2005.

Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

# Modelling Semantic Change as Equilibrium Shift in a Signalling Game*

Bernhard Schröder
Institute for Communication Sciences
Dept. of Language and Communication
University of Bonn

B.Schroeder@uni-bonn.de

Philipp C. Wichardt
Institute of Economic Theory 3
Dept. of Economics
University of Bonn

philipp.wichardt@uni-bonn.de

**Abstract**

Semantic change is modelled in a game theoretic setting as a transition between Nash equilibria in a signalling game. It is shown how, in this setting, the three major classes of semantic change – generalisation, specialisation, and semantic shift – can be characterised by the types of initial and resulting equilibria. Moreover, the importance of occasional mistakes in strategy choices for the cases of generalisation and semantic shift is emphasised.

## 1 Introduction

This paper demonstrates how a change of meaning for verbal expressions can be illustrated in a simple game theoretic setting. In particular, it is shown how the three main classes of semantic change – generalisation, specialisation, and semantic shift – can be understood as transitions between Nash equilibria in a signalling game (cf. Spence, 1973) triggered by environmental changes.[1] In doing so, the present approach builds on models of communication as developed by Parikh (2001), but emphasises the importance of occasional mistakes or trembles in strategy choices for some of the diachronic changes to come about.

Regarding the classification of semantic changes, we adhere to the following convention: As semantic generalisation we consider the process which leads to a broader extension of a given word, e.g. German *Sache* which orginally meant an issue of dispute and nowadays has a meaning very similar to *thing*. We find a specialisation in the case of German *Mut*, which formerly had a meaning comprising all sorts of frames of mind, cf. English *mood*, and nowadays has the special meaning of audacity. As semantic shift we understand all other processes of semantic change. The extensions of the old and new meaning may overlap or not. A paradigmatic case of non-overlapping semantic change is German *Hammer*. We presuppose that the story of this change can be told as follows. The word meant in Proto-Germanic stone or rock and was transferred to certain newly invented tools consisting of stones (among other parts). Later on there appeared a need to differentiate between the tool and its material. *Hammer* was preserved for the tool in German (and English). The material was designated by alternative words like *Stein (stone)* and *Felsen (rock)*.

*Financial support of the German Research Foundation (DFG) through SFB/TR 15 at the University of Bonn is gratefully acknowledged.

[1] Special cases of semantic change like metaphorisation or pejoration can be treated as subcases of the three major categories with special conditions on the games.

In the sequel, we first present the abstract signalling game used in our later analysis and specify the relevant equilibria (Section 2). We then show how the three forms of semantic change mentioned above can be understood as equilibrium changes triggered through changes in the interaction environment if players make errors (Section 3). The paper concludes with a brief summary and a discussion of the results (Section 4).

## 2 General Set-Up

The basic type of games we consider is a local game of partial information as proposed for the pragmatics of NL communication by Parikh (2001). There are two players, a sender and a receiver. The sender wants to convey a certain information – here a concept expressed by a lexeme – to the receiver. For the sake of simplicity, we adhere to a purely extensional view on concepts, identifying them with sets of individuals. But the approach is compatible with richer concepts of concepts.

The exact content of the information depends on the type $\Theta$, $\Theta \in \{\Theta_i \mid i = 1, \ldots, n\}$, of the sender. This type is determined by a random draw of Nature prior to the interaction and is revealed only to the sender. The receiver knows nothing but prior probabilities, $p_i$ for $\Theta_i$, $p_i \neq 0$, $\sum_i p_i = 1$. Each sender type $\Theta_i$ has two expressions available, $e$ and $a_i$. Expression $a_i$ guarantees the interpretation $\iota$ desired by type $\Theta_i$, denoted by $\iota_i$. We also write $[[a_i]] = \iota_i$. Yet, for some reason $a_i$ is a bit more costly or risky to use (e.g. due to the complexity of $a_i$ or due to a lack of familiarity of $a_i$). Expression $e$ in turn is ambiguous as it can be interpreted differently, i.e. $\iota(e) \in \{\iota_i \mid i = 1, \ldots, n\}$. Yet, if interpreted in the desired way, each $\Theta_i$ prefers using $e$ to using $a_i$. The preferences of the receiver are assumed to be aligned with those of the sender. Payoffs are denoted by $\Pi_i(\iota)$ in case expression $e$ is used and by $\Pi_i(a_i)$ otherwise.

For the subsequent discussion we restrict attention to the case $n = 2$; see Figure 1 for illustration. Moreover, we assume that the payoff difference between $a_i$ and a correctly interpreted $e$ is the same for all types $\Theta_i$, i.e. for all $i$: $\Pi_i(a_i) = \Pi_i(\iota_i) + c$, $c > 0$.[2]

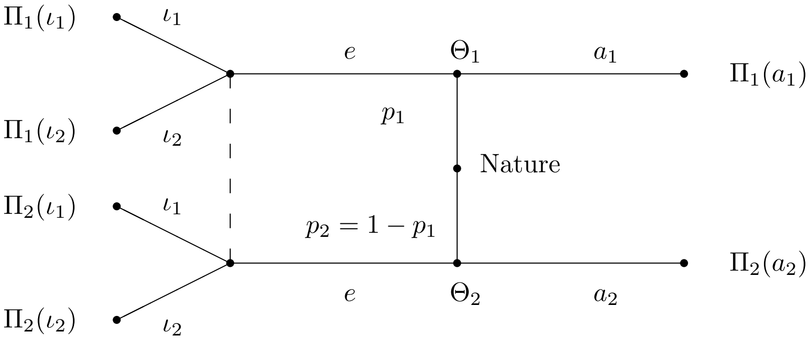


Figure 1: Graphic representation of the strategic situation. The dashed line indicates that the receiver cannot distinguish between the two decision nodes a priori.

In general, there are three pure strategy Nash-equilibria for the above described game which

[2] This assumption is not restrictive and is only made to facilitate later calculations.

are relevant for the ensuing discussion:[3]

**N1** *Separating e1*: type $\Theta_1$ plays $e$, $e$ is correctly interpreted as $\iota_1$, and type $\Theta_2$ plays $a_2$ because $\Pi_2(a_2) > \Pi_2(\iota_1)$.

**N2** *Separating e2*: type $\Theta_2$ plays $e$, $e$ is correctly interpreted as $\iota_2$, and type $\Theta_1$ plays $a_1$ because $\Pi_1(a_1) > \Pi_1(\iota_2)$.

**N3** *Pooling on e*: both sender types play $e$, and $e$ is interpreted as $\iota_1$ if

$$p_1\Pi_1(\iota_1) + (1 - p_1)\Pi_2(\iota_1) > p_1\Pi_1(\iota_2) + (1 - p_1)\Pi_2(\iota_2),$$

and as $\iota_2$ if the reverse inequality holds. Furthermore, for the imperfectly understood sender $i$, being imperfectly understood has to be less costly than sending the more cumbersome signal $a_i$, i.e. $\Pi_i(\iota_{-i}) > \Pi_i(a_i)$.

## 3 Three Cases of Semantic Change

In the following, we show how the three main cases of semantic change can be understood as equilibrium changes in the above described signalling game due to changes in the interaction environment. For the sake of argument, environmental changes will be modelled as changes in type frequencies $p$.[4] The restriction to frequencies, however, is made for reasons of space only. It should be clear from the discussion that similar results can be obtained if players' payoffs change, e.g. because certain incorrect interpretations become more costly.

We begin our discourse with the slightly easier case of a specialisation (3.1). Generalisation and semantic shift then are discussed under one heading (3.2) as they essentially are the consequences of the same process but for different starting conditions.

### 3.1 Specialisation

Specialisations are cases where an initially broader interpretation of an expression $e$ becomes narrower thereby restricting the use of $e$ to fewer individuals. As triggering contexts for the specialisation process we consider upward monotonic contexts $C$, i.e. contexts with

$$[[a]] \subseteq [[b]] \;\Rightarrow\; C(a) \models C(b). \tag{1}$$

The context *John saw a X* is an example of an upward monotonic context because

(2) John saw a dog.

follows from

(3) John saw a yorkie.

*yorkie* being a subconcept of *dog*.

If $b$ is a well-established lexeme for a broader concept and $a$ a more complex and therefore costly expression for a narrower concept, $b$ will be chosen in stead of $a$ in a context where

[3] Knife-edge cases, i.e. cases where one or more of the following inequalities are equalities, are neglected as they are of minor importance for our argument.

[4] The rise of type frequencies may e.g. reflect the grown importance of a certain concept.

the greater specificity of the information conveyed by $a$ does not outweigh its greater costs.[5] This type of context is modelled in the sequel.

Cast in terms of the above formal model, a specialisation is a transition from a pooling equilibrium ($e$ is used by both types) to a separating equilibrium ($e$ is used only by one type). In the sequel, we shall argue how such a transition from one type of equilibrium to the other can result from a change in type frequencies.

Consider an N3-type pooling equilibrium as a starting point. Without loss of generality, assume that in this equilibrium $e$ is interpreted as $\iota_1$ (i.e. $[[e]] = \iota_1$). Thus, we have

$$p_1\ \Pi_1(\iota_1) + (1 - p_1)\ \Pi_2(\iota_1) > p_1\ \Pi_1(\iota_2) + (1 - p_1)\ \Pi_2(\iota_2), \tag{4}$$

i.e. given uncertainty about sender types expected payoffs are such that it is optimal for the receiver to interpret $e$ as $\iota_1$. Moreover,

$$\Pi_2(\iota_1) > \Pi_2(a_2), \tag{5}$$

i.e. given that $e$ is interpreted as $\iota_1$ it is still better for $\Theta_2$ senders to use $e$ than $a_2$.

Varying only type frequencies and assuming payoffs to be fixed, it is an immediate consequence of inequality 5 that only transitions to N2-type separating equilibria (with standard interpretation $\iota_2$) are feasible.[6] For such a transition to occur, two conditions have to be satisfied. On the one hand, the standard interpretation of $e$ has to change to $\iota_2$ (i.e. $[[e]] = \iota_2$). Thus, frequencies have to change in such a way that inequality 4 is reversed. Rearranging inequality 4, this is equivalent to requiring

$$p_1\ \underbrace{(\Pi_1(\iota_1) - \Pi_1(\iota_2))}_{>0\ by\ ass.} < (1 - p_1)\,\underbrace{(\Pi_2(\iota_2) - \Pi_2(\iota_1))}_{>0\ by\ ass.}, \tag{6}$$

which is obviously satisfied if we assume $\Theta_2$ senders to become sufficiently more frequent relative to $\Theta_1$ senders, i.e. if we assume $p_1$ to be sufficiently small. On the other hand, $\Theta_1$ senders have to prefer $a_1$ to an imperfectly interpreted $e$, i.e. payoffs have to be such that $\Pi_1(\iota_2) < \Pi_1(a_1)$. Otherwise, senders always pool on $e$ (irrespective of the actual interpretation of $e$). If both above conditions are satisfied, then a specialisation can occur as a result of changes in type frequencies.

Summing up, starting from a pooling equilibrium where $e$ is interpreted as $\iota_1$, a specialisation can arise from a change in the interpretation of $e$, triggered through a change in type frequencies, if for $\Theta_1$ senders the new interpretation $\iota_2$ of $e$ is worse than using $a_1$.

### 3.2 Generalisation and Semantic Shift

The term generalisation refers to a case where an initial interpretation of an expression $e$ is broadened so as to allow the use of $e$ in more larger variety of contexts; cf. the introductory example of *Sache*. Accordingly, within our model, a generalisation can be viewed as a

[5] This covers especially those cases where the sender relies on the restricting force of the situation. The psycholinguistic preference for basic level concepts (cf. Rosch (1977)) would be modelled here by especially low costs of choosing them.

[6] $\Theta_2$ senders will always prefer to accommodate with an imperfectly interpreted $e$.

change from a separating equilibrium, say N1 (with standard interpretation $[[e]] = \iota_1$), to a N3-type pooling equilibrium (with broader interpretation $[[e]] = \iota_2$).[7] Similar to the specialisation case, we confine our considerations to certain change triggering contexts. These are upward monotonic, and for a $\Theta_1$ sender the choice of $a_1$ is less attractive than the choice of $e$, despite being interpreted in the less specific way of $\iota_2$ after the semantic change.

A semantic shift, in turn, is a situation in which an a priori standard interpretation of an expression $e$ vanishes and a new interpretation emerges; cf. our example of *Hammer* above. Thus, in terms of our model, a semantic shift is a change from one separating equilibrium, say N1 (with standard interpretation $[[e]] = \iota_1$), to the other, i.e. to N2 (with standard interpretation $[[e]] = \iota_2$).[8] In the following, we show how both transitions can be described as equilibrium changes triggered through changes in type frequencies.

To begin with, consider an N1-type equilibrium. Thus, by construction it holds that:

$$\Pi_2(\iota_2) > \Pi_2(a_2) > \Pi_2(\iota_1);$$

i.e. for $\Theta_2$-sender types, sending the costly alternative $a_2$ is preferred to being imperfectly interpreted when sending $e$.

A necessary condition for an equilibrium change from N1 to N2/N3 in this case is for senders and/or receivers to make errors in sending and interpreting signals. This is simply due to the fact that without errors, even a dramatic change in frequencies, resulting in $p_2 >> p_1$, does not set up the N1 equilibrium. If no sender of type $\Theta_2$ ever tries message $e$, the best receivers can do is to interpret $e$ as $\iota_1$, which in turn renders it optimal for $\Theta_1$ ($\Theta_2$) senders to play $e$ ($a_2$). However, if with a small probability $\varepsilon$, $0 < \varepsilon << 1$, people make mistakes - or simply try something different irrespective of the rationally expected consequences - things change.[9, 10]

Consider, for example, the case where, with probability $\varepsilon$, senders do not send their equilibrium message (receivers still make no mistakes). In that case, for N1 to be an equilibrium at all, it has to hold that:

$$p_1(1-\varepsilon)\Pi_1(\iota_1) + (1-p_1)\varepsilon\Pi_2(\iota_1) > p_1(1-\varepsilon)\Pi_1(\iota_2) + (1-p_1)\varepsilon\Pi_2(\iota_2). \tag{7}$$

Otherwise, i.e. if the reverse inequality was satisfied, receivers would do better by interpreting $e$ as $\iota_2$. Yet, more can be said. In particular, rearranging equation 7 gives:

$$\frac{p_1}{1-p_1}\cdot[\Pi_1(\iota_1) - \Pi_1(\iota_2)] > \frac{\varepsilon}{1-\varepsilon}\cdot[\Pi_2(\iota_2) - \Pi_2(\iota_1)]. \tag{8}$$

Now, if relative frequencies change, i.e. if $p_1$ decreases, the sign of the inequality eventually will change. Once that happens, it will be better for receivers to interpret $e$ as $\iota_2$ even if $e$

[7] It will be apparent from the subsequent argument, why the new interpretation of $e$ will emerge in case only type frequency changes are considered.

[8] Unaffected interpretations could be added in terms of further types for whom the strategic situation does not change through the shift.

[9] The idea to account for the possibility of trembles in the equilibrium concept goes back to Selten (1975).

[10] An alternative interpretation of mistakes, which we do not pursue here, would be in terms of variations arising in the process of language acquisition. This would rather correspond to a construal of mistakes as mutations in an evolutionary process.

is the the standard message only for type $\Theta_1$ senders. Forgone payoffs in case of mistakenly sent messages $e$ by $\Theta_2$ senders (which now a far more frequent) outweigh the benefits from correct interpretations of $\Theta_1$ senders. Thus, receivers will change their standard interpretation from $\iota_1$ to $\iota_2$. And, in reaction to that, $\Theta_2$ senders will change from $a_2$ to $e$.

Whether this change in the standard interpretation and $\Theta_2$'s behaviour eventually results in a generalisation, i.e. an N3-type pooling equilibrium, or a semantic shift, i.e. an N2-type equilibrium, depends on the payoffs of the $\Theta_1$ sender.

*Generalisation:* If the payoffs of $\Theta_1$ are such that

$$\Pi_1(\iota_1) > \Pi_1(\iota_2) > \Pi_1(a_1), \tag{9}$$

i.e. imperfect interpretations of $e$ are better than sending $a_1$, then $\Theta_1$'s optimal choice still is to play $e$ and an N3-type pooling equilibrium will obtain. Such a constellation is, for example, possible if $\iota_2$ is a broader interpretation of $e$ which is not exactly the same as $\iota_1$ but which easily accommodates $\iota_1$.

*Semantic Shift:* If the payoffs of $\Theta_1$ are such that

$$\Pi_1(\iota_1) > \Pi_1(a_1) > \Pi_1(\iota_2), \tag{10}$$

i.e. imperfect interpretations of $e$ are worse than sending $a_1$, then $\Theta_1$ senders will change from $e$ to $a_2$ thereby completing the change from the N1 to the N2 equilibrium, i.e. the semantic shift. These conditions are likely to be satisfied, for example, if $\iota_1$ and $\iota_2$ have some kind of conflicting meaning components, so that being misunderstood is comparably costly.

As regards semantic shifts, it is interesting to note from inequality 8 that these are more likely to occur in the direction of an interpretation of $e$ which would cause a higher cost in case of misinterpretation. In particular, if $\Pi_1(\iota_1) - \Pi_1(\iota_2) < \Pi_2(\iota_2) - \Pi_2(\iota_1)$, the above described shift will occur already for some $p_1 > \varepsilon$. If however $\Pi_1(\iota_1) - \Pi_1(\iota_2) > \Pi_2(\iota_2) - \Pi_2(\iota_1)$, then $p_1 < \varepsilon$ is necessary. Accordingly, assuming exogenous conditions to vary over time, we would expect receivers to lock in with interpretations that are most costly if misinterpreted.

## 4 Conclusion

Modelling communication as a simple incomplete information game, we have shown how the three major classes of semantic change can be characterised by different initial and resulting Nash equilibria. For the sake of expositional clarity, we have confined our analysis to the triggering contexts of semantic change. In this stylised case, specialisation starts with a pooling equilibrium and results in a separating equilibrium, generalisation goes the other way round, and semantic shift is a transition between two separating equilibria.[11] While specialisation can be explained as a mere effect of changes in the probabilities of sender types or payoffs, generalisation and semantic shift need the assumption of sender (or more generally communicator) mistakes in addition.

[11] The upward monotonic contexts are considered here as typical contexts in which the semantic change occurs. We do not want to exclude cases where semantic change is triggered by other contexts. Especially negative polarity items can be seen as cases where the typical context is downward monotonic. In these contexts the broader and the narrower concept change their roles.

## References

A. Benz, G. Jäger, and R. van Rooij, editors. *Game Theory and Pragmatics*. Palgrave MacMillan, 2005.

P. Parikh. *The Use of Language*. CSLI, Stanford, 2001.

E. Rosch. Human categorisation. In N. Warren, editor, *Advances in cross-cultural psychology (Vol. 1)*. Academic Press, London, 1977.

R. Selten. Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory*, 4:25–55, 1975.

M. Spence. Job market signaling. *Quarterly Journal of Economics*, 87:355–74, 1973.

# Signalling signalhood: An exploratory study into the emergence of communicative intentions[*]

Thomas C. Scott-Phillips, Simon Kirby, Graham Ritchie
Language Evolution and Computation Research Unit
University of Edinburgh

thom@ling.ed.ac.uk

**Abstract**

The emergence of communication systems in a population of interacting individuals is a phenomenon of interest to researchers in a wide range of disciplines, not least language evolution, for which the advent of a shared symbolic communication system is a crucial explanandum. Most previous work has focused on how signallers and their audiences agree on the forms of communicative protocols, but the associated and deeper problem of how audiences even know that signals are signals in the first place has been neglected. We report a recent experiment with human subjects that sheds light on this question. Our results suggest two things: first, that the question of how audiences even know that signals are signals is not trivial; and second, that in order to signal signalhood the signaller's behaviour must deviate in some way from the intended audiences expectations of the signaller's behaviour. Importantly, these expectations change as communicative protocols develop, with the result that what appears to be a signal in one context will not appear to be a signal in another.

## 1 Introduction

Language is plainly a major component of human social behaviour. One defining feature of language is that it is predicated on the use of *symbols*: arbitrary meaning-form mappings that are shared across a population. As such, an account of how symbolic communication can emerge is crucial to an understanding of how language evolved and indeed to the emergence of human social behaviour more generally (Deacon, 1997). In recent years this question has been studied in a number of different ways. The most common approaches have been computational modelling and evolutionary robotics, but they are by no means the only way to investigate the matter. In particular game theory and experimental psychology have also contributed valuable insights.

However almost all of this work has focused on how audiences infer, or how groups of individuals negotiate upon, the meaning of a given signal. But there is a deeper problem that needs to be overcome beforehand: how do audiences know that a signal *is* a signal? Put another way, how does a signal signal its own signalhood? The difference between this problem and the more commonly studied problem of signal meaning can also be seen in the difference informative and the communicative intentions. The former is the signaller's

[*] TSP and GR are funded by grants from the Arts and Humanities Research Council and the Engineering and Physcial Sciences Research Council respectively. We also wish to acknowledge financial support from the Arts and Humanities Research Council grant number 112105.

intention that the audience understand their signal, while the latter is the signaller's intention that the audience realise that they have an informative intention i.e. the intention that the audience recognise that the signaller is trying to communicate with them.

Hence whereas most previous work has focused on the detection and use of informative intentions, the present work is focused on communicative intentions. The question, then, is how audiences detect communicative intentions in a novel medium, a challenge that humans must have overcome at some point in their evolutionary history. In other words, how do audiences even knows that a signal is a signal? This is not, as our results show, a trivial task. Indeed, some form of inferential reasoning must be required to overcome the problem. We have therefore designed our investigation in such a way that this problem is not circumnavigated in any way, and it is this feature distinguishes our study from almost all other previous work. The one exception to this claim is Quinn's work with situated robots (2001), which we discuss below. The present study is, then, the first to address the problem of how humans signal signalhood in a novel medium.

The next section offers a brief overview of how previous studies have circumnavigated the problem of signalling signalhood. We then describe the game that we have devised to investigate the problem. Results from our exploratory experiment are reported alongside some detailed analysis. The implications of this work and possible directions for future research are considered in the conclusion.

## 2 Background

Previous work has avoided the problem of signalling signalhood in, broadly speaking, (at least) one of three ways. These are, in short, that we may pre-define the meaning and form spaces; we may pre-define the roles of signaller and receiver; or we may pre-define the communication channel. In the first case, if we pre-define the meaning and form space then audiences are immediately able to differentiate between signals and non-signals: anything that resembles items in the form space is likely to be communicative. There is of course the interesting question of how a population agrees on how the individual elements of each space will map to each other, but the deeper problem of how receivers distinguish communicative from non-communicative behaviour cannot be addressed if we pre-define the meaning and form space. Game theoretic models, in which individuals are offered a choice of possible communicative strategies, are a good example of this approach.

In fact game-theoretic approaches typically also pre-define the roles of signaller and receiver. This places a second constraint on any attempt to understand how receivers-to-be even recognise that signallers are signalling: as soon as these roles are defined then the receiver then is primed to interpret the signaller's behaviour in communicative terms. This compromise can also be seen in some experiments with human subjects and in some computational studies.

The final and perhaps most obvious way in which we may circumnavigate the question of how signals are even recognised as such is to pre-define the communication channel. Under this circumstance audiences know in advance that inputs they receive through that channel are highly likely to be communicative. A clear example of this is Galantucci's recent pioneering work on the emergence of communication systems with human participants

(2005). Here, pairs of subjects have to coordinate their behaviour in a simple environment, but that coordination is impossible without some form of communication. To overcome this a convinient communication channel is also supplied. As a result this work sheds light on how a dyad may agree upon a set of arbitrary conventions, but it does not speak to the question of how subjects recognise signalhood. The experiments presented in this paper share some important similarities with Galantucci's work but are differentiated most saliently and most crucially by the fact that we provide no such pre-defined communication channel.

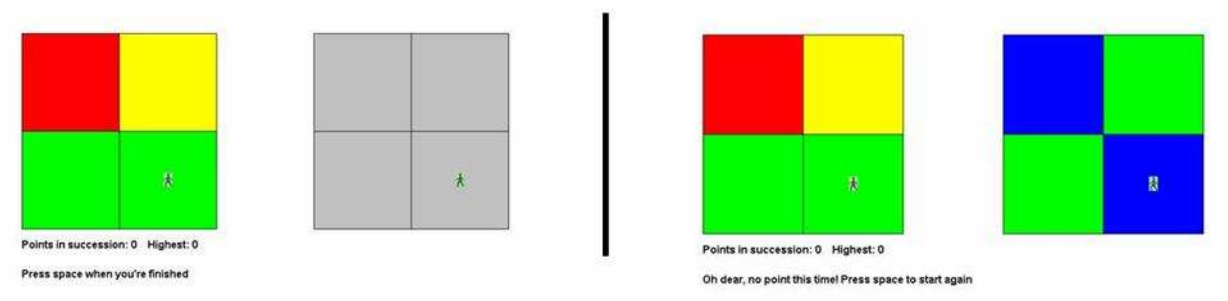
Many other approaches to the study of the emergence of communication systems have done similarly. In particular, nearly all of the work done on the emergence of communication in embodied and situated agents builds in communicative layers to the cognitive apparatus of the agents. That is, some particular aspect of the agent architecture is pre-defined to act as the communication channel. There is, however, one exception to this rule. Quinn's work, *Evolving communication without dedicated communication channels* (2001), can be seen, along with Galantucci's experiments, as the main inspiration behind the present study. Here pairs of situated robots are offered no dedicated communication channel. Neither are the signal and meaning space pre-defined, nor the roles of signaller or receiver. Natural selection, in the form of a genetic algorithm, is then asked to find a solution to a movement task that requires the robots to coordinate their behaviour with each other. This is acheived through the use of the infrared sensors that the robots were equipped with to prevent collision with each other. It should be noted, however, that the solution found by Quinn's robots was iconic, and moreover was one found by natural selection, rather than by specific agents. Yet linguistic signals are arbitrary and are created or learnt individually. How, then, do humans (and, indeed, how did pre-linguistic humans) detect signalhood in real time?

## 3 Methods and design

We have designed an interactive two-player computer game that can be used to investigate this and associated questions. Unlike the work discusses above, it does not pre-define the meaning and form spaces, the roles of signaller and receiver, or the communication channel, and as a result it allows us to study our focus of interest: the signalling of signalhood.

In our game each player is represented by a stick man who may move around a box that contains four quadrants, and each quadrant is assigned a random colour: either red, blue, green or yellow. Both players have their own box with independently assigned colours, with the exception that at least one of the four colours will appear in both players' boxes. Both boxes are visible to both players, but only each player's own box is shown in colour; the other player's box will appear as grey. Hence, each player knows the other's current location, but not their colours.

Each player may move around their box in their own time using the arrow keys. The players do not move freely around the box but rather from the middle of one quadrant to the middle of another, so as to prevent conventional symbols being outlined by the players' movements. At any point the players may press space to end the current turn. Once both players have done so then all the colours are revealed. A point is scored if the two players have managed to finish in same-coloured quadrants. (Recall that at least one colour will



Points in succession: 0 Highest: 0

Press space when you're finished

Points in succession: 0 Highest: 0

Oh dear, no point this time! Press space to start again

Figure 1: The game. On the left is the player's view whilst they are still playing, when they can see their own colours but not the other players'. On the right is the view once both players' have pressed space and the colours are revealed. In this example the players have failed to land on the same coloured square, and so do not score a point.

appear in both boxes, ensuring that success is possible in every round.) The players then play again, with new boxes, and their performance over time is measured by the highest number of points they are able to score in succession. Figure 1 shows the players' view during and after each round.

As should be clear, it is impossible to do well at this game without some way of coordinating behaviour between the two players. However, the players have no way of doing this except through their behaviour within the game. Hence they must create new symbols in the form of particular movements around the box and negotiate on their meaning. In order to do this, they must also signal that that these movements are indeed signals. Furthermore, because iconic solutions are ruled out by the games design, communicative intentions must be signalled in some way that is independent of meaning.

What sort of solutions are possible? One possibility (there are several others, and not just simple variants of this solution) is that we could oscillate along a particular side of the box to indicate our intended destination. For example, left-right-left-right-left along the top of the box could indicate an intention to finish on red, left-right-left-right-left along the bottom could indicate an intention to finish on blue, up-down-up-down-up along the left-hand side could be green and up-down-up-down-up along the right-hand side could be yellow (note that the actual location of the intended colour does not change the signal). We are interested in the question of whether and if so how subjects are able to converge on such a system.

We ran this game with 12 pairs of normally-developed adult subjects, none of which were exposed to the game beforehand. The subjects did not meet their playing partner at any point during the experiment. Having received an initial set of instructions, each pair were given 3 minutes with which to familiarise themselves with the game. After that time they were given more detailed instructions and given the opportunity to ask the experimenter about anything that was not clear. They then had 40 minutes with which to achieve as high a score as possible. At the end the subjects were asked about the communication systems they achieved and/or attempted to get started, and where necessary these self-reports were compared against the game logs. Subjects were paid for their time and in order to ensure motivation a significant additional payment was offered as a prize to each member of the highest scoring pair.

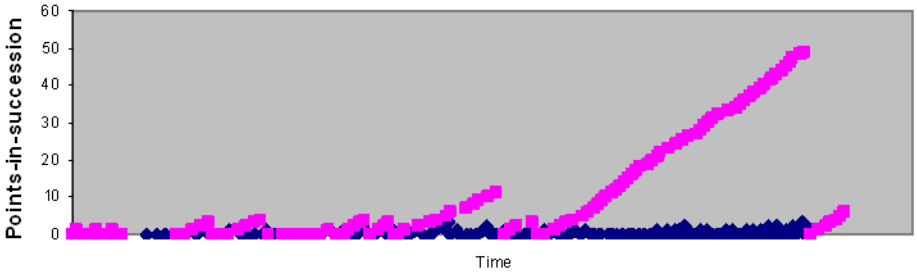


Figure 2: Two pairs' performance. The x-axis is real-time and the y-axis the pairs' points-in-succession. One pair (blue squares) failed to achieve any communication at all (final score of 3), while the other pair (purple squares) achieved a very good score (49). Note that successful run, which was built on a foolproof strategy, came to an end near the end of the experiment due to a lapse in concentration.

## 4 Results and discussion

The most salient result is that *the task is not trivial*. The final scores for the 12 pairs were, in ascending order, 83, 66, 54, 49, 39, 17, 14, *7, 5, 4, 3, 3* (see figure 2 for a detailed view of two of these runs). Italicised scores represent runs in which no communication was achieved and hence where the final score was the result of chance alone. That is, in nearly half (5 of 12) of cases pairs of subjects failed to develop any relevant communication system at all, despite a small number of relevant meanings and a clear understanding of the task at hand (all subjects reported understanding that they needed to find some way of communicating with each other). *Relevant* is used here to refer exclusively to information about the signaller's intended destination colour, for that information is not just necessary and sufficient, it is the only information that is so. Nevertheless, some subjects attempted to and managed to communicate irrelevant information to one another i.e. information that was not useful for the task. Specifically, oscillating movements were often used to mean and/or inferred to mean that the player had the same colour on the two squares between which they were oscillating, and several variants on this theme were observed. Yet this information is irrelevant in our technical sense, for it does not help the players complete the task.

In some cases pairs' failure was because neither subject could work out any way of signalling the relevant information to the other; that is, they reported not knowing how they would be able to communicate such information. In other cases players conceived of communication systems that would work if known by both players but could find no way of sharing the system with the other player. Coupled with the fact that some subjects did manage to conceive of systems that will achieve success and were also detectable by the other player, this suggests that the twin abilities to have and to signal communicative intentions in a previously undefined medium is a cognitively sophisticated task with a wide range of individual variation.

Of further interest is *how* the successful pairs actually developed their communication systems. In five of the seven cases, the pair utilised what we term the *default colour strategy* in order to bootstrap the communication system. This works as follows. In order to achieve any success above chance levels the players simply choose the same colour on every round, assuming that this default colour appears somewhere in their box. Although they have no way of independently agreeing on what that colour should be, it does not take long for the players to note each others behaviour in this regard and to converge on the same default colour. Even though this strategy is not communicative[1] it nevertheless enables them to score at slightly above chance levels. However sooner or later one player or the other will not have the default colour in their box, and the pair will most likely fail to score in that round and hence their points-in-succession score will return to zero. This situation is then resolved when one of the players, faced with a box without the default colour, behaves in an idiosyncratic way, typically through the use of the sort of oscillating movements described earlier, or by or travelling full-circle around the box. This behaviour may be glossed as "No red!", "Not plan A!", "Something's wrong!" or anything similar. The two players then make a random choice of the available colours.

Although their choices in this scenario are unlikely to be the same, the idiosyncratic behaviour that marked the situation as different to the default scenario often comes to take on the meaning of one or the other of the colours chosen by the two players. For example, if one player had circled around the box and then chosen blue, that circular movement, which originally simply meant "Not red!", often comes to mean, after a number of repeated uses, "Blue". Once these two behaviours (the direct movement to a default colour and the idiosyncratic movement that acquires meaning through entrenchment) are established then the players typically find it relatively easy to negotiate on what movements should be used to indicate the final two colours. A typical end-result would be something like the following: for red, take an efficient route to the destination quadrant (this would be the default colour); for blue, take an inefficient route to the destination quadrant; for green, oscillate left and right; and for yellow, oscillate up and down. With a system like this with four unambiguous signals, one for each of the possible destination colours, players are able to negotiate on which colour they will both finish on through simple dialogue and hence achieve success in every round. As a result the points in succession score of such a pair will grow to be as large as the number of rounds they can complete in the allotted time (although lapses in concentration do occasionally occur).

This use of the default colour strategy to bootstrap the communication system offers insights into what qualities are required to signal signalhood. Specifically, we should examine the transition from the pre-communicative state where only the default colour achieves success to the communicative state where one colour (blue in the example above) has an associated arbitrary symbol. Before the players had established the default colour strategy a non-efficient route to a quadrant would have appeared indistinguishable from other, essentially random behaviour. However, once the default colour strategy is in place then any behaviour that deviates from it carries the suggestion that there is some other purpose behind the behaviour. It would appear then that deviation from the expected is necessary for a signal to announce its own signalhood.

As further support for this conclusion, recall that 7 of 12 pairs achieved communication of some sort, but that 5 of these 7 utilised some version of the default colour strategy to

[1] It might be suggested that this behaviour *is* communicative. A full defense of why we disagree deserves a paper of its own. For now, we simply observe that with this strategy the players need not actually pay attention to each other, and this seems to be a crucial feature of communication.

bootstrap the final system. The two pairs that achieved success without this assistance did so through the use of the sort of idiosyncratic movement that we see in the system above for the colours green and yellow; that is, their signals possessed the quality of unexpectedness. The difference between this case and those that used the default colour strategy is that in the absence of the default colour strategy this unexpectedness is required from the very beginning, before any signals or other strategy is in place. Only through this unexpectedness are players able to signal their intended destination to the other without pre-existing protocols about either what a signal may look like or how it is to be differentiated from other behaviour in the world.

In fact this quality of unexpectedness can, on further analysis, be specified even more precisely. One pair (in fact, the highest scoring pair of all) converged on a system that tied the number of movements they made from their starting position to their intended destination: one movement for blue, two for yellow, three for green and five for red. They arrived at this system via the default colour strategy as described above, with the minor difference that part-way through they negotiated a specific symbol - one movement from the starting position - for the previously default strategy. The result is almost identical in terms of signal form to the following system which one player in one of the *failed* pairings composed themselves but which went undetected by their partner: two movements for red, three for blue, four for yellow and five for green. Hence we see that despite near identity in the final system there was a wide divergence in effectiveness of these two systems, and this difference would appear to be due to the manner in which the systems were created: the first system, which was grown organically via the default colour strategy, achieved a very high degree of success, while the latter, which was created by one individual who then attempted to impose it upon their partner, achieved no success at all.

Why did this difference in creation result in such a difference in result? Recall that the signals for green and yellow in the successful strategy (2 and 3 movements from starting position respectively) were detectable only because symbols for the other two colours were already in place and hence the two players had certain expectations of the each other's behaviour. They were thus able to infer the purpose of behaviour that deviated from those expectations. Similarly there are expectations of behaviour in the failed case but here those expectations are different: without any existing protocols the expected behaviour is, by and large, randomness. (As we saw with the two pairs who succeeded without recourse to the default colour strategy, non-random behaviour will be detected as purposeful.) Here, then, we arrive at the crucial distinction: in the failed case, a small number of movements in no particular direction from the starting point appears random, and hence is *not* differentiated with respect to the intended audience's expectations. In the successful case, however, a small number of movements in no particular direction *is* differentiated with respect to the intended audience's expectations. Hence we may define the quality that signals signalhood as *deviation from the intended audience's expectations.* Importantly, those expectations will change as communication systems develop and hence the space of possible new signals will change with them.

## 5 Conclusions

Whereas most previous work into the emergence of communication systems has focused on how audiences comprehend the informational content of signals, we have focused on the deeper problem of how audiences even know that signals are signals, particularly when the signaller makes use of a novel medium. Our results suggest that this is indeed a real problem with no trivial solution. We found that under such circumstances almost half of all dyads failed to achieve any communication at all, and subjects exhibited a large degree of individual variation in performance. Moreover, detailed analysis of how subjects approached the task and how they did or did not achieve success suggests the that a behaviour must deviate from the intended audience's expectations in some way if it is to successfully signal its own signalhood. The immediate implication of this, which we have observed within the game, is that some behaviours will appear communicative in some contexts but not in others. This is because the communicative protocols that develop morph the space of expected behaviours.

These results open up the question of what cognitive abilities are required to communicate one's communicative intention in a novel medium. It seems reasonable to suppose that such cognitive abilities are a pre-requisite for the creation of symbols and hence for the advent of language, a communication system predicated on symbolism. Indeed, in no other species do individuals create and learn their own symbols. In most cases signals are either iconic or indexical, and most other natural communication systems are, with a few exceptions, innate rather than learned.

There are several ways in which this work may be extended. First, the game could be played with a variety of different subjects to establish exactly what cognitive abilities are required to signal one's communicative intention. Second, we have made a specific suggestion about what quality is required for a signal to signal its own signalhood. This could be formally tested using the same game but with a confederate playing one of the roles: in one condition they would use signals have possess this quality; in another signals that do not possess this quality. It would also be of interest to remove the possibility that subjects use the default colour strategy to bootstrap their communication system: do even fewer pairs achieve communication under this restriction? This could be done, for example, by changing the game so that whenever a point is scored on a given colour that colour will not appear in both boxes in the following round (though it may appear in one of the boxes).

Finally, we should observe that the game we have presented here is a truly original tool. Its unique property that there is no *a priori* way for players to distinguish between behaviour that is communicative and that which is not enables the study of one aspect of communication systems that has been hitherto ignored: the emergence of communicative intentions.

## References

Terrence W. Deacon. *The Symbolic Species: The Co-evolution of Language and the Brain.* W.W. Norton, 1997.

Bruno Galantucci. An experimental study of the emergence of human communication systems. *Cognitive Science*, 29(5):737–767, 2005. doi: 10.1207/s15516709cog0000_34.

Matt Quinn. Evolving communication without dedicated communication channels. In J. Kelemen and P. Sosk, editors, *ECAL01*, Lectures Notes in Computer Science, pages 357–366, Prague, September 10-14 2001. Springer.

# Language regulation and dissipation in meaning[*]

Andrew Stivers
Dept. of Economics
Oregon State University

andrew.stivers@oregonstate.edu

**Abstract**

This paper studies the micro-dynamics of language evolution in the presence of speaker and hearer preferences for language use that are simultaneously: fully aligned (as a stylized social context) and fully opposed (as a stylized commercial context). We take a simple world with two, quality-differentiated goods and three possible utterances: a simplex (categorical) word and two morphological derivations, one specifying high quality attributes the other specifying low quality attributes. Taken individually, the social game has two evolutionarily stable strategies for assigning content to the simplex word, and the commercial game has a single evolutionarily stable strategy of assigning low quality content to the simplex word.

Taken together, with players making mistakes in context (whether the speaker is commercial or social), the use of the simplex word to denote low quality in both contexts always survives. The use of the simplex word to indicate high quality in the social context only survives if the expected cost in the commercial setting (of mistaking low for high quality good) is low relative to the value of successful coordination in the social setting. In this case, the content in the commercial setting tends toward one good or the other depending on the difference in value between the goods.

This paper studies the micro-dynamics of language evolution in the presence of speaker and hearer preferences for language use that are simultaneously: fully aligned (as a stylized social context) and fully opposed (as a stylized commercial context). We take a simple world with two, quality-differentiated goods and three possible utterances: a simplex (categorical) word and two morphological derivations, one specifying high quality attributes the other specifying low quality attributes. Taken individually, the social game has two evolutionarily stable, pragmatic strategies for assigning content to the simplex word, and the commercial game has a single evolutionarily stable, pragmatic strategy of assigning low quality content to the simplex word.

Taken together, with players making mistakes in context (whether the speaker is commercial or social), we have a more comprehensive model of language use. The specifics of this model are motivated by the problem of regulating natural, unowned language in the market. The primary and most direct regulations take either the form of "truthtelling" statutes, that attempt to align commercial use with social practice, or of statutory definitions, that provide a very specific mapping between word and object.

The need for commercial language regulation seems to arise from the tension between cooperative social practice and the potentially antagonistic commercial practice. The op-

[*] I would like to thank Max Stinchcombe, Ayca Tekin-Koru for useful comments and discussion. All errors are my own.

positional nature of commercial communication arises from the opposing preferences of the speaker (the seller) and the hearer (the buyer) in the market. Sellers in general will want to claim the highest possible quality for their products while minimizing their costs of production, i.e., claiming to sell a high quality good while actually producing the low. Buyers want to know the true characteristics of the products.

Truthtelling regulations try to solve the informational problem that arises when sellers are free to use any utterance at all in conjunction with their goods. By creating appropriately enforced penalties, that is, by making the costs of language vary monotonically across sellers by the type of good that they sell, they turn the lexicon into a set of signals. This kind of regulation was first formally studied in the economics literature in the early eighties by Milgrom (1981) and Jovanovic (1982), among others. Truthtelling regulations generally define a transgression as one that is deceptive for an ordinary consumer, as does, for example, the US Federal Trade Commission's regulations. These kinds of regulations were the downfall of the term “Chablis” in the US, as the ordinary American consumer believed Chablis to be simply white wine.

There are at least two potential problems with these regulations. First, Stivers (2004) shows that while truthtelling regulations solve the information problem, they do not solve the efficient use problem. That is, the costs or benefits of particular pragmatic strategies arising from framing effects, relevance or availability are ignored. Solving this allocation problem, by defining the categorical terms “chocolate,” “hotdog,” or “diamond” in a particular way, for example, is the ostensive goal of the definitional statutes.

The second problem is that using the beliefs about content and meaning of an ordinary consumer ignores the endogeneity of those beliefs to commercial use. Although the practitioners of language regulation are fully aware that the salience of a regulation depends on the implementation (or lack of implementation) of that regulation, none of the economics literature deals with the question of how social and commercial uses of language interact. Neither does this literature attempt to deal with the regulation of language in an evolutionary context. Intuitively, the ability of a speaker to gain from using natural language in a commercial setting must arise from some combination of commercial and social association. This model attempts a framework to address this interaction and thus provide a framework for considering when and how to implement language regulation.

This work is motivated on a more theoretical level as an attempt to integrate cooperative models of language, as exemplified by Parikh (2001), and the more competitive models of communication that are usual in the literature on market information transmission from economics and game theory. As many authors have noted — e.g. Sally (2005) in a critique of the analogy between “cheap talk” in games and language and Allott (2006) in a critique of game theory and language modeling in general — the usual game theory framework may leave out important aspects of the practice of language. This paper begins to explore more fully how conflicting uses and contexts of language might interact.

## 1 Model

Language users simultaneously play two evolutionary games where beliefs about the content of utterances are replicated according to their relative utility. The primary difference

| | $c \to L$ | $c \to H$ |
|---|---|---|
| $c \to L$ | $v_S - \rho k, v_S$ | $(v_S - k)\rho, \rho v_S$ |
| $c \to H$ | $(v_S - k)(1-\rho), (1-\rho)v_S$ | $v_S - (1-\rho)k, v_S$ |

Table 1: Social game payoffs

between the games, for our purposes, is that one has complementary preferences between speaker and hearer — a stylized social context — and the other has opposing preferences — a stylized commercial context.

We limit the scope of communication to discussion of one of two quality differentiated objects, denoted $H$ and $L$. We assume that the quality difference between the goods in not self-evident, but any consumer that can differentiate will prefer $H$ if both goods are offered at a price close to their cost.

The lexicon is limited to a simplex (categorical) word $c$ and two morphological derivations, one specifying high quality attributes $h$, the other specifying low quality attributes, $l$. Thus, the literal meaning of the simplex word is the set $\{H, L\}$ and the literal meanings of $h$ and $l$ are $H$ and $L$ respectively.

### 1.1 Social channel

Players in the social context meet randomly and are randomly assigned roles as speaker and hearer. The speaker attempts to communicate to the hearer the quality of a good — either $H$ or $L$.

Each player benefits from, and is assumed to use, a pragmatic strategy where the simplex word refers to a specific good, either $H$ or $L$. In addition, each benefits from mutual communication. The relevant strategies in the game for both speakers and hearers are then the choices of how to assign the content of the simplex word: $c \to L$ or $c \to H$.

The payoffs in this game are constructed as follows. The value of successful communication is $v_S$ for both parties, where the subscript $S$ refers to the Social context. The cost to the speaker of using the categorical word is normalized to zero, and the cost of either specific word is some constant $k > 0$. The players' payoffs are given in the normal form game in Table 1 with the speaker as the row player and the hearer as the column player.

If both players assign $c$ to $L$ (or $H$ with costs reversed), then communication is always successful. However, the payoffs to the speaker depend on whether the speaker is referring to good $H$, for a payoff of $v_S - k$, or $L$, with a payoff of $v_S$. The probability that any given communication will refer to $H$ is $\rho$, so that the payoff to both the speaker and the hearer is an expectation over $\rho$. If players assign different contents to $c$, then communication is only successful when $h$ or $l$ is spoken, and thus the higher cost is incurred. From the point of view of the hearer, this event occurs with probability $\rho$ when the speaker assigns $c \to L$ and the hearer assigns $c \to H$, and with probability $(1 - \rho)$ when the players have the opposite beliefs.

For $\rho \in [0, 1]$ these payoffs form a coordination game with two Nash equilibria where either $c \to L$ or $c \to H$ for all players. Shifting $\rho$ up will shrink the basin of attraction for the

$c \rightarrow L$ equilibrium, but will not eliminate it even at $\rho = 1$ because the certainty of what is being communicated makes the hearer indifferent between the two strategies.

Given this coordination game, we construct an evolutionary game where $x_t$ is the proportion of players with strategy $c \rightarrow L$. The expected payoffs for speakers playing $c \rightarrow L$ and $c \rightarrow H$ are then (with superscript $s$ indicating the speaker):

$$\begin{aligned} EU_S^s(c \rightarrow L) &= \rho(v-k) + v(1-\rho)x_t \\ EU_S^s(c \rightarrow H) &= (1-\rho)(v-k) + k\rho(1-x_t). \end{aligned}$$

The expected payoffs to the hearers are:

$$\begin{aligned} EU_S^h(c \rightarrow L) &= v - v\rho(1-x_t) \\ EU_S^h(c \rightarrow H) &= v - v(1-\rho)x_t. \end{aligned}$$

We define a replicator dynamic based on the relative fitness of the population where we assume that the probability of being a hearer to speaker is equal. When the two channels are always distinct, this dynamic is driven by the difference between the expected payoff for believing $c \rightarrow L$ and the average payoff across both types of beliefs:

$$x_{t+1} = \Big(1 + \frac{1}{2}(1-x_t)(k - 3k\rho - (1-\rho)v + (v(2-\rho) + k\rho)x_t)\Big)x_t.$$

We let $R_S(x_t)$ denote the social dynamic (inside the big parentheses), so that we can write $x_{t+1} = R_S(x_t)x_t$.

Taken by itself, there are two evolutionarily stable strategies in the social channel, corresponding to the two pure strategy Nash Equilibria at $x = 1$, where all players believe $c \rightarrow L$ and at $x = 0$ where all believe $c \rightarrow H$.

### 1.2 Commercial channel

Concurrent with the social language game, players participate in a market game with sellers. The sellers are the speakers in this context. Each chooses a word from the lexicon in reference to the good that they have for sale, either $L$ or $H$.

Sellers can freely enter the market and choose to sell either $L$ or $H$, with costs of production $c_H > c_L$. We assume a truth-telling constraint — i.e., a binding regulation on fraud in the market — so that misapplying utterance $h$ to $L$, for example, does not happen. We do not formally elaborate on the market setting in this paper, but in brief, all consumers prefer $H$. Sellers of $L$ will use the word "$c$" and sell at price $c_H$ in order to fool those consumers who believe $c \rightarrow H$. Sellers of $H$ will use the more expensive utterance $h$, and sell at price $c_H + k$. The slightly higher cost of the "true" $H$ pushes the credulous consumers (those believing $c \rightarrow H$) toward the sellers of $L$. Note that speaking "$l$" is dominated since no consumer will buy $L$ when $H$ is thought to be available at cost.

In the commercial channel the proportion of those that believe $c \rightarrow L$ is $y_t$. Payoffs to hearers are given by the difference between the value of the good purchased and the price of that good. The price to those who believe $c \to L$ is $c_H + k$ and the price to those who believe $c \to H$ is $c_H$. The commercial dynamic is then:

$$y_{t+1} = \Big(v_L - c_H + (v_H - v_L - k)y_t\Big)y_t = R_C(y_t)y_t.$$

There is a single equilibrium in evolutionarily stable strategies in the commercial channel at $y = 1$ where all consumers believe $c \to L$. In the case where no confusion is possible between the two channels, skeptical consumers win out over credulous consumers in the commercial channel.

The market dynamic here is such that the seller of a high quality good is always forced to differentiate if they want to successfully compete.

### 1.3 Confusion

Because consumers are required to carry beliefs for two different channels simultaneously, it seems reasonable that they occasionally make mistakes in assigning beliefs. This may be especially true in the case of the commercial channel, where firms are aware that fooling consumers about the context of the communication could result in a more credulous consumer. Because of this, we allow mistakes by agents. An agent always has two, potentially different beliefs about the utterance $c$. The fraction of consumers making a mistake in the commercial channel (assigning their social belief to the commercial context) is $\gamma$ and the fraction of consumers mistakenly assigning their commercial beliefs to the social context is $\lambda$.

Given these mistakes, the growth rate of each population remains the same, but the fraction of offspring that correctly receive the parent's beliefs is changed. The fraction of offspring of $c \to L$ believers that correctly receive the parent's belief is $1 - \lambda$ (the fraction of parents not making a mistake) minus the offspring that are mistakenly given the (opposite) belief from the commercial channel, $\lambda(1 - y_t)$, plus the offspring that are mistakenly given the (correct) belief from the commercial channel. Added to this amount is the offspring of those believing $c \to H$ in the social channel and $c \to L$ in the commercial channel whom are mistakenly given the $c \to L$ in the social. This fraction is given by $(1 - x_t)y_t\lambda$. Taken together, this yields a social channel replicator dynamic of:

$$x_{t+1} = R_S(x_t)\Big(1 - \lambda(2 - y_t)\Big)x_t + \lambda y_t.$$

The dynamic in the commercial channel is similarly affected. The proportion of agents $y_{t+1}$ is reduced by the amount of consumers that would have been part of $y_{t+1}$ but were mistakenly assigned a belief of $c \to H$. The proportion of agents $y_{t+1}$ is then increased by the amount of offspring that would not have been part of $y_{t+1}$ but were mistakenly assigned $c \to L$. The fraction of consumers with $C \to L$ in the commercial channel and $c \to H$ in the social channel is $y_t(1 - x_t)$. This dynamic is then given by:

$$y_{t+1} = R_C(y_t)\Big(1 - \gamma(2 - x_t)\Big)y_t + \gamma x_t.$$

With this system of difference equations, we see two general classes of evolutionarily stable steady states. First, we always see a stable steady state where $x = y = 1$. That is, in both the social and commercial channels the simplex term $c$ is taken to mean the low quality good. This arises because the relatively high benefits to skepticism about the content of $c$ in the commercial channel will completely reverse even unanimous use of $c \to H$ in the social channel.

Second, when the potential loss to a hearer believing $c \to H$ in the commercial channel is low enough (relative to the social value of successful communication), a stable steady state emerges with a high proportion of agents playing $c \to H$ in the social realm. The content assigned to $c$ in the commercial realm can tend to either $H$ or $L$ depending both on the relative value of commercial and social communications, and the probability of a mistake about content falls.

## 2 Implications and further work

While the goal of this paper is limited to setting out a framework for examining the interaction between social and commercial use of language, we point out two directions for further work in the spirit of application to efficient regulation.

One area of particular interest is in thinking about incentives to manipulate the language environment. Sellers might gain, for example, by manipulating the probability of mistaken context. To see this, suppose that there is no interaction between the two channels, and the stable strategies are $c \to H$ in the social channel and $c \to L$ in the commercial. If a speaker can successfully induce mistakes in assigning context, this could lead to either of the two kinds of equilibria detailed above. Because seller profits for fraud increase proportionally with buyer losses, the most likely scenario would lead to a temporary increase in $c \to H$ strategies in the commercial channel. Eventually, however, the higher cost of $c \to H$ in the social channel with mistakes will drive an increase in $c \to L$ in the social channel, which then feeds back into an increase in $c \to L$ in the commercial channel. In this case, the end result is that $c \to L$ in both channels, but sellers get a temporary boost in profits due to the confusion. The effect is to extract value from the high use of $c \to H$ in the social channel.

Another area of interest is in exploring more fully the implications of a “truthtelling” regulation. In this paper we have imposed a regulation that relies only on literal meanings of the available utterances, so that $c$ is left to mean either $H$ or $L$ or both. In practice, simplex words like $c$ are implicitly defined, as in the case of “food,” “genetically modified food,” and “GM-free food” where rulings about the use of “food” has been based on consumer beliefs about what food is. In this case, the introduction of mistakes and the interplay between beliefs could mean a change in consumer beliefs about even a protected commercial meaning through a public relations campaign directed at manipulation in the unprotected social channel.

In its present form, this model merely suggests some of these possibilities, but it does underscore and operationalize the interaction between cooperative and competitive uses of language.

## References

Nicholas Allott. Game theory and communication. In Anton Benz, Gerhard Jäger, and Robert van Rooij, editors, *Game Theory and Pragmatics*, pages 123–151. Palgrave Macmillan, 2006.

Boyan Jovanovic. Truthful disclosure of information. *Bell Journal of Economics*, 13(1): 36–44, Spring 1982.

Paul R. Milgrom. Good new and bad news: Representation theorems and applications. 12 (2):380–391, Autumn 1981.

Prashant Parikh. *The Use of Language.* CSLI Publications, 2001.

David Sally. Can I say "bobobo" and mean "there's no such thing as cheap talk"? *Journal of Economic Behavior & Organization*, 57(3):245–379, July 2005.

Andrew Stivers. Regulating language: Market failure and competition in descriptive signals. *Oregon State University Economics Working Paper*, 2004.