Modeling Language Change

Quentin Feltgen, Benjamin Fagard, Jean-Pierre Nadal

▶ **To cite this version:**

Quentin Feltgen, Benjamin Fagard, Jean-Pierre Nadal. Modeling Language Change: The Pitfall of Grammaticalization. Language in Complexity – The Emerging Meaning, Springer, pp.49-72, 2017, 978-3-319-29481-0. 10.1007/978-3-319-29483-4_3 . hal-01481919

HAL Id: hal-01481919
https://hal.science/hal-01481919

Submitted on 3 Mar 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

Feltgen, Q., B. Fagard & J.-P. Nadal. 2017. Modeling Language Change: The Pitfall of Grammaticalization. In F. La Mantia, I. Licata & P. Perconti (eds), Language in Complexity – The Emerging Meaning. New York : Springer, 49-72.

Abstract: Language evolution is the subject of various theoretical studies, following two main paths: one, where language is viewed as a code between meanings and forms to express them, with a focus on language as a social convention; and the other defining language as a set of grammatical rules governing the production of utterances, evolution being the outcome of mistakes in the acquisition process. We claim that none of the current models provides a satisfactory account of the grammaticalization phenomenon, a linguistic process by which words acquire a grammatical status. We argue that this limitation is mainly due to the way these models represent language and communication mechanisms. We therefore introduce a new framework, the "grammatheme," as a tool which formalizes in an unambiguous way different concepts and mechanisms involved in grammaticalization. The model especially includes an inference mechanism triggering new grammaticalization processes. We present promising preliminary results of a numerical implementation and discuss a possible research program based on this framework.

3.1. Introduction

Since the 1990s, understanding the mechanisms underlying language emergence and evolution is the subject of many studies (Christiansen and Kirby 2003; Fitch 2005; Steels 2011). Yet, it is not an easy task to account for these changes: timescales of interest are overlapping (Croft 2013), data are often sparse and/or uneasily processed (how to meaningfully present data remains an open question), and above all, every insight one may have about these changes drastically depends on the ideas one has about language: these ideas constitute an unavoidable prism whose specific biases are imparted to the interpretation and the understanding of linguistic phenomena.

Several attempts to model these changes have been proposed; they are generally due to the merging of efforts from linguists and other scientists coming from fields where modeling has already become a habit: biologists, physicists, population geneticists, economists, mathematicians. Such models combine linguistic hypotheses on language and language change with a quantitative formulation, which may lead to numerical simulations or mathematical considerations (Vogt and De Boer 2010). They may even lead to implement algorithms of language learning and communication in artificial intelligence, as has been done for the Sony robots (Steels 1999; Steels and Belpaeme 2005).

Two main directions have been thoroughly explored so far: one focusing on the social, conventional nature of language, mainly addressed with agent-based models, and one relying on transformational grammar, where language is seen as a set of rules learned by children, whose mistakes result in language change. However, neither manages to successfully model grammaticalization, although this phenomenon may be regarded as one of the major mechanisms of language change and language evolution (Meillet 1912; Peyraube 2002).

Grammaticalization (Meillet 1912; Hopper and Traugott 2003) is the process whereby a word (or any other linguistic element) acquires new grammatical functions and grammatical characteristics. It has been thoroughly explored and detailed in numerous works (cf. Prévost 2003 for a discussion, and Heine 2003 for an overview).

Nonetheless, both modeling approaches have proven rich and enlightening: our goal is merely to stress their limitations, and ideally to shed some light on the dynamics of grammaticalization phenomena with a more adequate model. We will first briefly review some of the works which have been proposed to model language change; these approaches will be then discussed and we will try to understand why they fail to satisfactorily account for grammaticalization. From what we will have learned, we will derive a few principles which may serve as a guideline in the modeling task. At last, we will lay the basis for a proposal of a model which, we hope, will account for some essential features of grammaticalization.

3.2. Two Roads to Language Change

Despite the fact that the field is very young, the literature concerning language change models is already impressive, both by its size and its diversity. It is not our purpose here to review in detail this polymorphic body of literature. In this section, we only aim to briefly recall the main paths which have been explored in these attempts to model language change. More extensive reviews can be found elsewhere (Castellano et al. 2009, Sect. V; Loreto et al. 2011).

We choose to present these different works as a diptych, where each plate is characterized by the representation of language on which these models rest: in the first series of approaches, language is seen as a set of sign-meaning pairs, i.e., a code (henceforth ‘code’ models); in the second series, language is seen as a set of rules (henceforth ‘rule-based’ models). Each of these two representations may vary a lot from one paper to the other; yet, we think that they both unify a sufficient number of works, and that they also differ enough from one to another to provide an efficient discrimination between the two, thus leading to a clear, unambiguous setting of the hinge which separates them.[1]

3.2.1. Code Models

In this first line of models we will review, language is seen as a code, that is, an association between forms (or signs, words, signals, sounds, etc.) and meanings (or linguemes, objects, events, etc.). The underlying assumption of this view is that meanings are clearly identified: they correspond to obvious, prominent features of the environment of the speakers; they are also shared by all speakers.[2]

3.2.1.1. Population-Based Code Models

In these models, the language of agents is characterized, for each user, by two matrices, often labeled P and Q: one is the coding matrix; the other is the decoding one. The production (or active) matrix P relates meanings to forms; the comprehension (or passive) matrix Q relates forms to meanings. Thus, when a speaker wants to communicate a given meaning to a hearer, he utters a form according to the specifications of the production matrix; the hearer, receiving this uttered form, uses his comprehension matrix to decode the form and recover the conveyed meaning. The size of these matrices (i.e., the number of meanings and forms) is the same for all agents.

This coding/decoding process can be either probabilistic or determinist. If it is determinist (Pawlowitsch 2007; Pawlowitsch et al. 2011), then only one form is associated to each meaning in the production matrix (excluding thus synonymy); simultaneously, only one meaning is associated to each form in the reception matrix (excluding polysemy and homonymy).

These models may also account for a protogrammar (Nowak and Krakauer 1999), where meanings are separated into distinct classes, objects, and actions: utterances then amount to the production of an object-action pair.

This framework often leads to the use of population dynamics. A population is described by the number of its individuals speaking each language (for instance, if there are three languages or linguistic variants L_1, L_2, L_3 in the population, this population will be described as the vector (x_1, x_2, x_3) of the proportions of individuals speaking each of these three languages). If an agent uses a language which allows him to communicate successfully with his fellow agents, then he will be able to reproduce more effectively, leading to the production of an offspring speaking the same language. Language thus plays a role similar to that of genome in population genetics. It should be stressed, however, that population dynamics do not describe each individual communicative event between every pair of agents; rather, they consider the averaged communicative potential of different languages.

3.2.1.2. Agent-Based Models

Similar dynamics may also be interpreted on the timescale of a human being (Cucker et al. 2004): agents change their language in order to be able to communicate more successfully with the rest of the population. Models adopting this interpretation may not strongly differ in their mathematical formulation from the previous models; yet, they lead to a very different vision of language evolution, since language changes now within individuals, rather than changing as individuals are replaced.

Most models based on the agent approach are utterance models. Through linguistic interactions, agents record all occurrences of forms expressing a given lingueme in the interactions in which they took a part. They, thus keep in memory in which proportion each different form has been uttered to express the lingueme, rather than the simple fact that such forms have been used in the past to express it. Such an approach is prototypical of the "exemplar" models introduced in phonology (Pierrehumbert 2000). Worth mentioning is the utterance selection model (Baxter et al. 2006) for which a detailed mathematical analysis has been done.

One of the most thoroughly studied models in language change is the naming game. It consists of an agent-based model where all agents cooperate to find an overall convention about meaning-forms association. In this type of model, the inventory of each agent is specified, and each of their interactions is accounted for. If the mechanisms governing the interaction between agents are simple enough, these agent-based models may be turned into population dynamics model (Castelló et al. 2009).

The naming game was first developed by Luc Steels to develop an efficient way of implementing a communicative behavior in robots (Steels 1995). Since then, it has been more

extensively studied by the team of Vittorio Loreto (Baronchelli et al. 2006, 2008). Each agent is characterized by the set of inventories (i.e., list of possible names) it associates with each object. These inventories are modified as agents interact with each other, until all agents agree on the name to be given to each object. An interesting point is that there is no real difference between the naming of a single object and the naming of a plurality of objects, the latter option only leading to a slowing down of the conventionalization process.

Several extensions have been made around this naming game. It is possible to create new specific forms to make the mutual understanding easier (Omodei and Fagard 2013). The category game (Puglisi et al. 2008) considers a continuum of objects (the different light frequencies of visual stimuli) which is both perceptually and linguistically categorized.

To conclude this first subsection, it should be stressed that a majority of these models are devised to study how a linguistically heterogeneous population may reach a consensus, i.e., how all individuals may happen to speak the same language, through their successive interactions for agent-based models, or because one of the linguistic variants yields a higher fitness in population-based models. They generally address the question of the emergence of language, rather than the question of language change.

3.2.2. Rule-Based Models

While, the previous models focus on the lexical component of language, rule-based models approaches insist on the grammatical and structural rules. Language is described as a set of parameters (Chomsky 1981) specifying these rules (I-language), which in turn govern the production of utterances (E-language).

In this framework, language is seen as irreducible to biological evolution. The notion of language fitness does not play any major role in language evolution; however, as in most population models, language change is not usage-based, but results from the passing of human generations. More precisely, the source of the change is to be found in the process of language acquisition (Niyogi 2009): «Perfect language acquisition would imply perfect transmission. Children would acquire perfectly the language of their parents, language would be mirrored perfectly in successive generation, and languages would not change in time. […] Therefore, for languages to change with time, children must do something differently from their parents».

This strong hypothesis on the origin of language change shifts the focus to this process of language acquisition. In all models exploring this process, a population is considered, and this population is split into two groups: adults and children. Adults have a fixed I-language, and produce utterances; children listen to these utterances, and infer the most plausible I-language compatible with the set of utterances they have been given.

Children have a limited time to learn language; once this time is elapsed, they definitely adopt an I-language (Kirby and Hurford 2002). The very existence of this time limit implies that children are exposed to a finite number of utterances—which is crucial: it is this finiteness which makes the acquisition process unavoidably imperfect.

On the basis of the utterances they have recorded during their learning period, children infer hypotheses to decide which I-language is the most compatible with these utterances. Such a

statement raises an immediate question: from which set must an I-language be chosen? Rule-based models approaches make the following assumption: a set of possible I-languages innately exists in the mind of children; an I-language is chosen among this set of possibilities; such a set is called the Universal Grammar. Interestingly, this Universal Grammar is sometimes considered to be subject to biological evolution (Nowak et al. 2001).

A child is expected to learn an I-language close enough to the I-language of the adults. Otherwise, an individual would be able to communicate with neither individuals of other generations nor individuals of its own generation, which is obviously not the case. The question is then: what constraints should there be on the set of learnable I-languages, for a child to be able to learn an acceptable language? The answer to this question depends on the particular inferential procedure actually used by children; yet, no matter the procedure, this set of I-languages has to be finite (Nowak et al. 2001).

The rule-based models framework offers many more ways to explore language acquisition. Five possible lines of investigation have thus been proposed: «(1) the type of learning algorithm involved; (2) the distribution of the input data; (3) the presence or absence of noise or extraneous examples; (4) the use of memory; (5) the parameterization of the language space itself (the class of possible grammars/languages)» (Niyogi and Berwick 1996). If lines (1), (4) and (5) concern the cognitive process of inference, line (2) and more especially line (3) partly result from the specific social structure of the population. Those questions are indeed of some importance: do the children only listen to their parents, and so hear utterances produced on the basis of a very limited set of different I-languages? Does child—child communication have any influence on the learning process?

3.3. The Pitfall of Grammaticalization

These different types of models (code models and rule-based models) share some serious limitations if we want to address the question of why and how languages change, and especially, the question of grammaticalization. Indeed, grammaticalization intertwines the structure of the language and the particular role of its lexical elements: code models fail to account for the structure; rule-based models fail to describe the specific history of words and constructions. In the following, we will highlight three main limitations of existing models.

3.3.1. Communication

The first of these limitations concerns the communicative mechanisms at play in these models. Indeed, they are often very basic and we cannot expect them to give birth to interesting pragmatic constraints which would act on language change.

These communicative mechanisms are of two types. The first type is encountered in models such as the naming game. Agents try to find an agreement on the name associated with an object (sign-meaning pairing); they then check that there is no misunderstanding through nonverbal means. This latter mechanism has drastic consequences. It means that the language emerging from these games cannot be adapted to situations where one cannot point to the object referred to in its speech. Two objections immediately arise: first, language is obviously much more complex, and can be used in a much wider range of situations; second, one can wonder why language would have been selected through the evolution process of mankind, if

it had been no more efficient than a simple pointing device, which is found in other apes and even in lemurs (Gómez 2007; Leavens et al. 1996). Models such as the Naming Game are thus unable to describe the very essence of language. They have originally been created in a very specific goal (robotic implementation), for which such a nonverbal process has a meaning and a reality; but they might well fail to be extended to deeper linguistic investigations.

The second type escapes this limitation, since it does not refer to any nonverbal way to ensure communication. Let us consider a given context, sufficiently vague in its definition for an agent to be repeatedly exposed to it in a reasonably short time. At each exposure, the agent of interest is surrounded by agents who produce utterances. It may happen that, when exposed anew to a similar context, the agent of interest will himself produce an occurrence which will be more or less alike to what he has heard during the previous exposures. He may also consider what he has himself said in the past. Such a mechanism is at work in rule-based models (the child mimics what he has heard from adults), but also in a few agent models (as the utterance selection model, based on this very mechanism).

This last communicative mechanism seems to be more realistic. However, it does not explain why people produce utterances (the question of the function of language is thus still an open one), nor does it distinguish a situation in which people understand each other from one in which they just say similar things in similar situations, without even knowing why. And, above all, they do not take into account pragmatic considerations.

Yet, pragmatic constraints do seem to be an essential feature of language change (Croft 2013). Agent-based and population models show that a consensus is possible, and likely to occur; in several models, general consensus is even a stable attractor of the dynamics (Baronchelli et al. 2006; Kirby 2001). The question is then the following: if a society of human beings is able to reach a consensus concerning the conventions of language, why do languages change? Several hypotheses may be invoked: according to the invited inference theory, the repeated use of specific constructions in contexts implying further meaning may lead language users to associate this new meaning to the original meaning of the construction (Traugott and Dasher 2001); a desire to valorize one’s own speech may exhort to use new, unusual constructions (Keller 2005); the subtleties of gallantry impose to constantly renew the courtesy formula, leading to a constant depreciation of past constructions (Keller 1989). In all these hypotheses, what drives change is pragmatics.

Even though simplifications and abstractions are the intrinsic preliminaries of any possible model, some of them can be misleading. The specificity of the linguistic phenomenon has been thoroughly argued for in the past; it is therefore crucial that models be able to account for this specificity. Putting apart pragmatic considerations in the communication rules actually implemented in a model may thus be slightly too bold a simplification.

3.3.2. Language

Another important problem of the existing models is the way language is implemented. We have already mentioned the two major representations encountered in the field: the code, and the set of rules.

In code models, language is viewed as but a set of sign-meaning associations; that is, it is no more complex in its structure than, say, the equivalence between Morse alphabet and Latin alphabet. This vision of language seems to be quite far from reality. It may be a useful simplification to study how a group of agents can reach a consensus, but it says only little on changes of linguistic nature.

Rule-based models offer an interesting alternative. In such models, the linguistic system is more elaborate: it is a set of parameters, or rules, which permit communication. For instance, one of these parameters can be a binary variable which is set to 0 if the language is of SVO type and to 1 if it isn't. The bit-string models (Schulze and Stauffer 2005; Zanette 2008), for instance, consider language as a bit-string, that is, a string of a fixed number of bits, each of them taking two values, 0 and 1. All of these bits indicate a dichotomic property of the language.

However, language, as it is actually used by speakers, is not simply an actuation of these parameters; indeed, such a view does not take individual words into account, words whose specific history seems to be an essential part of language change. One could also argue that all linguistic changes occur through the fortunes of individual words, and so it may seem that, by setting words aside, rule-based models also leaves aside (at least one of) the channels through which languages change.

Furthermore, rule-based models do not offer to test the hypothesis that transformational grammar is a relevant framework to study language evolution, since they are based on this very hypothesis. In other words, they take for granted the validity of their own framework. A more open position—as claimed in (Reali and Griffiths 2010)—would have been to construct a representation of language which may change both through children acquisition, and through other factors, to compare their importance, and to judge of their respective relevance. Thus, the predominance of language acquisition in the process of change could be more convincingly justified.

3.3.3. Change

The third line of discussion is possibly the most essential one, and partly follows from the two previous ones. In all these models, either language evolves toward a stable form, and does not subsequently deviate from it, or it evolves through extrinsic factors.

Agent-based models often fall into the first category[3]: they do not describe a language which is continually changing, but the emergence of a cultural agreement whose shape and modalities grasp some actual features of the language. Once such an agreement is met, then language keeps its final form and does not evolve further. This seems to be a consequence of the fact that language, in these models, is reduced to a pairing between two sets of abstract objects, and does not take into account any pragmatic tensions in the production of utterances.

In the majority of cases, whenever language does evolve, it does not evolve by itself, although it has been claimed elsewhere to be a self-organizing system (Bybee 2003; John and Bennett 1982): the evolution does not result from the regular, everyday use of the language, but is brought about by extrinsic factors (Wedel 2006). In rule-based models, language change results from changes in the population of speakers. The timescale of language change is thus

given by the timescale of the passing of successive human generations—a timescale which does not necessarily coincide, however, with actual timescales of language evolution. In other models, change can come from the mixing of two different populations speaking different languages, or from speakers who cannot, or refuse to, follow the established convention (Komarova and Jameson 2008).

These different types of models thus all fail to give an account of language evolution in which language use actually triggers the changes at hand: language evolution is but an epiphenomenon, resulting from other phenomena occurring in the society of the speakers, and it cannot occur without these external perturbations. We prefer to assume that languages are intrinsically dynamic and unstable: being spoken is a sufficient condition for them to change and evolve continually.

3.4. Draft of a Deontology

It is quite surprising that no model, so far, has provided a satisfying description of grammaticalization. We have tried to highlight the intrinsic limitations of the main lines of modeling, to understand why they all seem to fail to account for grammaticalization. From these observations, a few lessons should be learned. We will try to summarize them, sketching thus the draft of a possible deontology for modeling language change.

3.4.1. Role of Analogy

We will first assume the following principle: language is a phenomenon which has no equivalent in Nature; consequently, language evolution is ruled by specific mechanisms and cannot be considered as a subtype of a more general cultural or biological evolution (Croft 2013). The corollary is that analogy should not be used without great care.

Researchers from ‘hard sciences’ are expected to work with analogies, however: otherwise, their specific competences, tools and methods would be of no use. Analogies are useful in the sense that they guide the scientist’s attempts to formalize observed features of a given phenomenon. For instance, if it is observed that, when speaking, a human being chooses a given way to express himself among several possibilities, and that this choice is not fully conscious but subject to infinitesimal mental variations, impossible to be accounted for in detail, then one can formalize this observation by seeing it as a stochastic process of decision, a well-studied framework already used to describe many other phenomena.

On the other hand, more systematic analogies (of the scale of the studied linguistic phenomenon as a whole) should be avoided. Analogies should be limited to individual features of a linguistic phenomenon; using them to a greater extent would be fallacious.

3.4.2. Interpretation of Model Mechanisms

The use of analogy can be considered reasonable if any element of the model, any mechanism at work, is susceptible of a clear, unambiguous linguistic interpretation. Thus, when one tries to model a linguistic phenomenon characterized by some specific and well-known quantitative features, it is not sufficient to be able to reproduce them with the model: it is also of crucial importance to know which linguistic realities all the mechanisms which have given rise to this result actually correspond to. This claim may seem blatant; yet, we previously saw

that the mechanisms at work in some models may not correspond to any plausible linguistic reality, such as the nonverbal sharing of the intended meaning in naming games. The idea of "language fitness," shared by many population models, also has no clear equivalent in reality.

Furthermore, the mechanisms used in quantitative models are mathematical abstractions. The delicate translation from linguistic mechanisms to their mathematical formulation requires a formalization effort which cannot be the work of 'hard' scientists alone. Very often, hard scientists try to guess from scratch which quantitative mechanisms could be the right ones, while the modeling should proceed along the following steps: the identification, by linguists (possibly led by the insight of the hard scientists who may have an idea of what to look for), of possible linguistic mechanisms, and then, an attempt of formalization to give a mathematical shape to these mechanisms. We give an example of such a way of proceeding in the last section of this work.

3.4.3. Freedom from Hypotheses

A model cannot be free from hypotheses. The formalization effort, the general form of the model, the identified mechanisms, all result from unavoidable working hypotheses. Yet, a model should allow for the testing of these hypotheses, discriminating between them, even if they happen to differ a lot.

Thus, the formalizing effort on which lies the model should not be too heavily biased toward one specific view of language. Building a quantitative model of language change from one specific linguistic hypothesis does not allow the questioning of this very same hypothesis, nor does it favor the opening of a dialogue with other, concurrent views of language. A good formalization should therefore not tell what language is, but it should be able to express in a single, unified mathematical framework, different possibilities of what language might be.

3.4.4. Stylized Facts

The gathering and use of facts is no less important. Indeed, pretending to submit to a scientific deontology implies to make use of observations: ideas, however right they may seem to be, are but speculations as long as they are not grounded on observed, communicable linguistic facts, which are the exact counterpart of the stylized facts in complex systems sciences.

Concerning language, these facts can hardly be anything else than corpora-based observations (Biber et al. 1994). But what can these observations consist of? Corpus-based studies help the researcher find occurrences, and observe their context of use (Fagard and Combettes 2013; Ogura and Wang 1996). Finding occurrences is the necessary preliminary of the statistical characterization of a phenomenon—this characterization being the sole feature susceptible of a complete modeling. Looking at the context of occurrences is both a linguistic verification which can in no case be eluded, and the only tool we dispose of to try to understand why and how such a phenomenon has taken place in the past history of language. Without this study of occurrences in context, it seems highly dubious to induce the details of the linguistic mechanism at work. In this very sense, corpus-based studies are an unavoidable premise of the formalization which precedes the modeling attempt itself.

Besides, a linguistic study is often full of ambiguities, uncertainties, or ideas which lack a concrete formulation—where by 'concrete,' we mean leading to a possible numerical

implementation. A model cannot be built without filling in those gaps in the understanding of the phenomenon. Hard scientists may have ideas on how to fill in these gaps—using analogies, as we have suggested in Sect. 3.1. These elements of mathematical nature must have their linguistic counterpart—as we have stressed in Sect. 3.2. And so, once a linguistic interpretation has been proposed, it gives a new insight into the phenomenon. Yet, this insight has to be legitimated by corresponding facts; in other terms, one has to find the traces of its corresponding features in corpora.

Models are of little use if they aim only to describe a given, circumscribed phenomenon. Let us imagine for instance that a viable quantitative description has been found to successfully describe a phenomenon. It would be of course all the more interesting if we manage to extend the original use of the model, in order to find out whether a plausible variation in its core mechanisms could lead to the description of new phenomena. If all the elements of the model have been given a correct, unambiguous linguistic interpretation, then a different result would be more than a mathematical curiosity: it would also correspond to a new plausible linguistic phenomenon, which may be thus identified. It would then be possible to track down the recognized statistical features of this phenomenon in corpora to attest its reality. Once again, corpora appear to be the only reservoir of linguistic facts on which to base any scientific understanding of language change.

Corpora are not, however, the only source of facts. Indeed, they leave apart an important aspect of language: cognitive activity—and yet, one can argue that any language activity is cognitive in its very nature. Corpora give us a glimpse of the product of language; but we have to access, one way or another, what produces the language: the human mind. Studies in experimental psychology (Scott-Phillips and Kirby 2010) provide us material to explore the cognitive mechanisms at work in language production and, more generally, in human communication. However, while the raw, unprocessed material found in corpora is free from hypotheses, psychological facts on language are the result of experiments, themselves led according to principles, hypotheses, and models.

3.5. A Formalization Attempt: The Grammatheme Abstraction

To illustrate how such principles could take place in an actual case study, we will now present an attempt of formalization and modeling we have developed on the issue of grammaticalization. Details of the model can be found in Feltgen et al. (2014).

3.5.1. Linguistic Evidences of the Change

Benjamin Fagard and Bernard Combettes (Fagard and Combettes 2013) have led a statistical, extensive study of a specific case of grammaticalization in French, the partial replacement of en “in(to)” by dans “in(to)” as one of the main locative prepositions. This statistical study, performed on the French corpus Frantext (ATILF-CNRS and Université de Lorraine 2014), has highlighted several stylized facts:

- the replacement occurs according to a sigmoid-shaped curve (Kroch 1989);
- the timescale of the change is about 50 years;
- the change is actualized in the prose of different authors of that time.

They have, furthermore, tracked this change in a few morpho-syntactic and semantic contexts of use of the locative proposition. It appears that:

- the change within each specific context also follows a sigmoid shaped curve;
- the timescale of the change within each specific context is shorter than the timescale of the overall change;
- the change within each specific context may end before the overall change has ended;
- the change within each specific context may start after the overall change has started.

These facts suggest the following hypothesis: the replacement of en by dans occurs in each context separately according to the same mechanism; the overall change is the outcome of a spreading of the new gram, dans, in more and more contexts of use of the locative preposition. Such a hypothesis has to be more thoroughly tested by supplementary analyses of this change, preferentially on various corpora.

3.5.2. A Formalization Attempt: The Grammatheme Abstraction

We assume that the above stylized facts are generic of the grammaticalization phenomenon. In order to account for them we propose a theoretical framework, which is quite general, and amenable to further generalizations in order to account for most cases of grammaticalization. The goal of this model is twofold: first, to reproduce qualitatively and, as much as possible, quantitatively, the observed facts; second, to shed some light on the underlying mechanisms, and highlight possible links between different phenomena.

3.5.2.1. Zetemes

The basis of our framework is a new formal representation of language, largely inspired from the Utterance Selection Model (Baxter et al. 2006). We first establish a clear distinction between lexical and grammatical semes (for instance, negation could be a grammatical seme, though quite large and unspecified). It does not imply, however, that words themselves are purely lexical or purely grammatical: they may carry semes of both natures. This sharp distinction relies on the distinction between conceptual and procedural information in relevance theory (Sperber and Wilson 1996), a point more precisely argued in Nicolle (1998). Furthermore, since, as pointed out by Ellis (2008): «Words mean things in the context of other words», we consider that grammatical semes are meaningful only in the context of their use (and may thus convey slightly different meanings in different contexts of use). By ‘context of use’, we mean something very general, to be understood as the conjunction of a phonetic context, a morpho-syntactic context, a semantic context, and even a pragmatic context. We will thus consider only such a conjunction of a seme and a context of use of this seme.

From a set of grammatical semes S, and a set of contexts C (both chosen according to criteria developed below), we define a more general set O(S, C), encoding all possible conjunctions[4] between the grammatical semes of S and the contexts C (see Fig. 3.1), such conjunctions being called zetemes[5] (informative units). The set O(S, C) is a mere collection of sites, each of them associated to a zeteme (seme-context pairing). It plays a role similar to that of a «meaning space» (Kirby 2001), and provides a framework to represent the linguistic phenomena of interest.

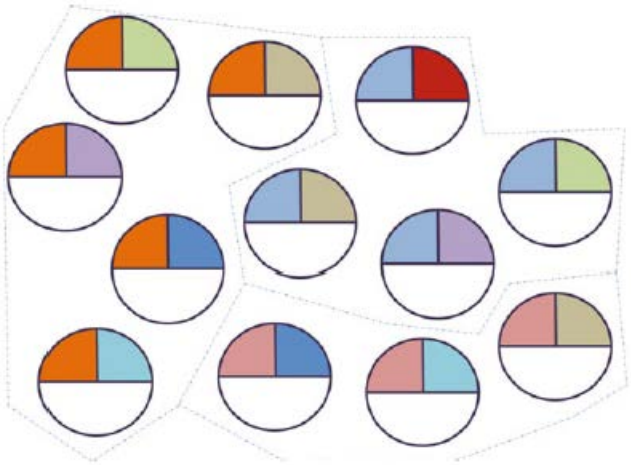

Fig. 3.1: A specific set O(S, C). Each zeteme is the conjunction of a grammatical seme (upper left color) and its context of use (upper right color). Not all possible conjunctions have to be accounted for. In this example, the set of grammatical semes S is made of three different semes, while the set of contexts C is made of seven different contexts. The borders between different semes have been materialized by a blue dotted line

Since these mathematical entities provide a tool of representation rather than a concept whose every detail would have a unique counterpart in reality susceptible to be tracked and found, we can be quite loose with the specific definition of context; pragmatically, contexts have to be wide and imprecise enough to cover a satisfying number of utterances, and, at the same time, specific enough to offer a rich and accurate description of the situation of interest. It may be considered that the previous remark is also applicable (perhaps with greater caution) to the concept of seme.

Once the degree of accuracy in the definition of semes and contexts has been chosen, the borders of the sets have to be specified. Indeed, including all possibly imaginable contexts and semes from all times and all languages (postulating thus the existence of a universal set of zetemes) is a vain ambition, since it would lead to an unclear mathematical specification of the problem. If one wants to represent a specific situation, then one has to choose accordingly the semes and contexts pertaining to this situation. But we need not deal with these questions when exploring generic and theoretical mechanisms without any reference to a particular, historical, linguistic event.

3.5.2.2. Grammatheme

This framework will help us represent, in a structured and organized way, the past linguistic experience of an individual. This representation will be named a grammatheme, in reference to the thematic organization of the Byzantine Empire.[6]

For an individual, different zetemes may be linked to each other, these links expressing a conceptual proximity between the two sites in the mind of the individual. For a given seme, all contexts of use are likely to be linked, but not necessarily; different semes, if used in similar contexts, could also be linked to each other (see the schematic illustration on Fig. 3.2). Those links are furthermore weighted according to the conceptual distance between the two sites they link up (the weight of the link increases as the two sites are conceptually closer).

The grammatheme is thus a network, encoding the conceptual links an individual establishes between different grammatical semes, in different contexts of use.

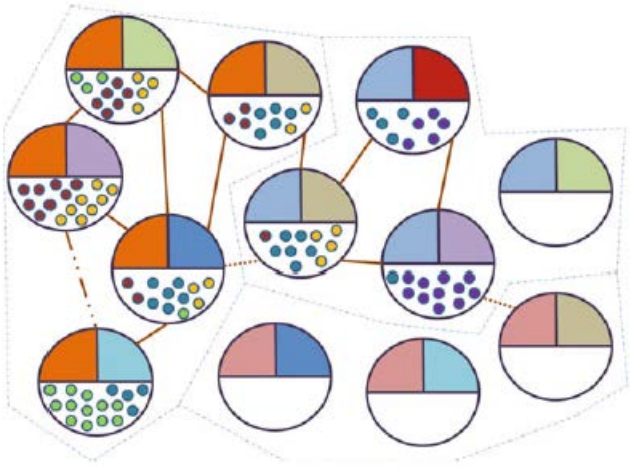

Fig. 3.2: Schematic view of a grammatheme. Zetemes are represented in the same way as in Fig. 3.1. Conceptual links between different zetemes are represented by brown lines. Dotted lines are links about to be created. Half plain, half dotted lines are links about to be deleted. Each occurrence of a gram is represented by a small colored circle. Different colors stand for different grams. Each time an individual hears or produces an utterance implying a given zeteme, the occurrence of the gram expressing this zeteme is recorded in the corresponding site of this individual's grammatheme

What actually encodes an individual's grammar (in the sense of (Bybee 2006): «grammar is the cognitive organization of one's experience with language») are the past occurrences of the linguistic elements having expressed a given grammatical seme in a given context of use (i.e., having expressed a given zeteme), linguistic elements we will refer to as grams, following (Bybee and Dahl 1989). We describe this past linguistic history of the individual by populating each site of the grammatheme with occurrences of the grams (see Fig. 3.2), used to convey the corresponding zeteme. Within a given site, each gram is associated with as many occurrences as it has been heard and produced in the past history of the individual.

We can formulate the hypothesis, grounded on biological plausibility, that the memory size for each site is actually finite. Therefore, once the memory is saturated, recording a new occurrence implies the deletion of a past memory—i.e., of a past recording of an occurrence. Several ways of implementing this deletion are possible: erasing the oldest recorded memory, or randomly choosing a memory to be erased—the probability distribution function governing this choice being possibly sensitive to the age of the memory.

Not all sites of the grammatheme of an individual are populated by occurrences. The fact that a site is populated reflects his knowledge of the associated zeteme, or at least the (possibly subconscious) awareness of the presence of this pairing in his linguistic environment. Thus, some zetemes are of no relevance to understand and describe his grammatheme. Yet, some unpopulated sites can be related to populated sites by a conceptual link, and so become populated. Thus, an individual may happen to acquire a new grammatical seme (Carlier 2007) which he was unaware of in the past, or to use a seme in a new context.

When a majority of recorded occurrences within a site are occurrences of the same gram, we consider that this gram dominates the site. We can then identify the function or meaning of a given gram—in the structuralist sense—with the dominion of this gram in the individual grammatheme.

Finally, the conceptual links between sites are susceptible of change. They can be reweighted, and even added or removed: similarly populated sites (i.e., sites populated by the same grams) will be more likely to be linked, and conversely, if two sites are populated by different grams, the conceptual distance between them can be revised to become larger than it was.

3.5.2.3. The Grammatheme of a Given Language

How should we characterize a human language within this framework? Like species in biological evolution, languages do not exist as a concrete, directly observable entity. Yet, the ability of individuals to viably reproduce with each other, and the sharing of common traits, allows a regrouping of these individuals into species. Quite similarly, one can regroup idiolects into languages (with the crucial difference that an individual can speak several languages): languages, like species, may be seen as 'existing abstractions;' yet it is not easy to characterize them.

A simple, yet limited way of doing this is by averaging over the idiolects of the individuals recognized to speak the same language, and consider this average to be a satisfying picture of this language. There is an obvious flaw in this characterization: it presupposes that we can identify the sharing of a common language, before characterizing it. It is thus theoretically inadequate. However, it is practically of some relevance, since the identification of a common language is relatively straightforward: interintelligibility can for instance serve as a first, rough criterion to consider that two individuals speak the same language.

Despite these difficulties, we assume this approximation to be acceptable. Indeed, corpus-based studies rarely take into account regional or community differences, and assume the language to define an entity susceptible of being observed. Thus, we can consider the grammatheme of a given language as the grammatheme of a hypothetic, prototypical individual, changing as he 'speaks to itself.' Indeed, we may consider that this prototypical individual produces occurrences on the basis of his past history, and records these same occurrences he has produced. This approximation is in the spirit of the "mean field approximation" in physics and the "representative agent" description in economics—with the same types of qualities and drawbacks.

It should be kept in mind that the mechanisms of producing and recording occurrences (and of redefining the conceptual links of the grammatheme network) at work in the prototypical individual, may be different from the mechanisms characterizing a regular individual. Only a careful statistical derivation could give a satisfying idea of how these self-changing mechanisms implemented in the prototypical individual can be built from and depend on actual, realistic communicative mechanisms between the individuals. For now, the grammatheme leaves this important matter as an open question.

In the following, we will consider a representative agent, embodied with minimal communicative mechanisms we believe to be of relevance for an individual: the ability to

produce utterances, to understand them correctly, a desire to be expressive, and the possibility to conventionalize an invited implicature.

3.5.3. Modeling Grammaticalization: Testing Hypotheses Within the Grammatheme Framework

We try here to answer the following question: how a gram, so far absent from a grammatical paradigm, and entering it at a given time through a given context of use, becomes fixed (in the same sense of fixed as in population genetics) in this paradigm. To this purpose, we will consider a scheme very similar to the utterance selection model (Baxter et al. 2006). The prototypical individual tries to express a given grammatical seme in a context of use (thus corresponding to one zeteme). To do this, he has to choose a gram, and to produce an occurrence of this gram, occurrence which will be subsequently recorded in the memory of this site. The choice itself is driven by the specific memory (i.e., the set of recorded occurrences) associated with the site, thus setting up a Markovian process for the production of occurrences (a process in which, at each time, the next stage depends on the state of a given object at that time; in this case, memory). The future of a gram is thus determined by this stochastic, Markovian process.

3.5.3.1. Criteria of Choice

We consider that, when the prototypical individual tries to express a given zeteme, it chooses a gram according to two conflicting considerations: it may want to reduce both its cognitive effort and the cognitive effort of its interlocutors, using a word which is as frequent as possible (Diessel 2007); or it may try to use a gram which may be less frequent, but with greater expressivity.

Expressivity is not a property of the grammatheme, but of the grams themselves. Obviously, expressivity should vary as the gram becomes more frequent (through phonetic erosion and semantic bleaching, for instance). We assume that the timescale governing the change for the expressivity of grams is longer than the timescale associated to the fixation event within the context of interest. Thus, expressivity can be considered constant in the study of this fixation.

The frequency of grams is a property of the memories encoded in the grammatheme. To compute the actual frequency of a gram in a site, we sum up all the occurrences of this gram in this context, with all the occurrences of this gram in neighboring sites, weighted by the conceptual proximity between sites, and we divide the result by the sum of all occurrences in this site and in neighboring sites, taking into account their respective weights.

3.5.3.2. Origin of the Grams

Where does a new gram come from? We may suggest that some individuals of the population suggest (most often inadvertently) a new way to express a given zeteme, to attract the attention of their interlocutors, to highlight the content of what they have to say, or to fix an insufficiency in the conventional way of expressing it. In the prototypical individual's grammatheme, these individual attempts will be represented by the recording of an occurrence of this new gram.

Another origin of the grams can be proposed, following Steve Nicolle (Nicolle 1998). According to him, all grammaticalization processes start with the conventionalization of an implicature. We can translate this hypothesis into the grammatheme framework, by considering that the conventionalization of an implicature is the appearance of a unilateral conceptual link between one of the grammatheme sites and an external site, associated with a lexical seme. Due to this new link, occurrences recorded in the external site will be taken into account when computing the actual frequency of the grams in the linked up grammatheme site. The external site being populated by occurrences of a new gram (the one responsible for the implicature), this new gram will thus be able to 'invade' the grammatheme.

3.5.3.3. Hypotheses and Results

To answer our question (how does a new gram become fixed?), we have considered (Feltgen et al. 2014) a given site of the grammatheme, where a new gram is supposed to appear. We tested the two following hypotheses, sketched in Fig. 3.3: 1—new grams are able to become fixed because they have a greater expressivity; 2—new grams are able to become fixed because they are the result of the conventionalization of an implicature.

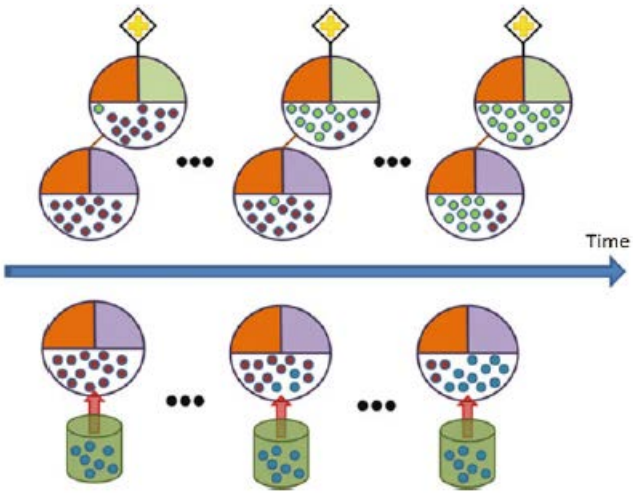

Fig. 3.3 : Two scenarios for the fixation of a new gram. Zetemes, grams and conceptual links are represented as in Fig. 3.2. In the first scenario (first row), one context of use is sensitive to expressivity considerations (green context, with sensitivity represented by a yellow plus sign), while the other is not (purple context). The red gram is frequent, yet unexpressive, and the green one is newly arrived in the paradigm, but very expressive. In the green context, this new, expressive gram will be favored over the red one; in the purple context, it will not. Yet, once the green gram has dominated the site with the green context, it is able to invade the other site, and finally prevail. In the second scenario (second row), an implicature is conventionalized (red arrow). A new gram (the blue one) will now arrive from outside the grammatheme (green cylinder) to the site of interest, already populated by a gram (the red one). If the implicature is strong enough, the blue gram will eventually prevail over the red one.

Hypothesis (1) fails as it is: indeed, a new gram never becomes fixed, unless the criterion of choice is dominated by expressivity considerations. This seems highly unrealistic: the repetition of past occurrences seems to be the dominant mechanism in actual communication; otherwise, there would be little understanding between individuals only caring to endlessly innovate. Yet, this hypothesis may be saved by considering that the criterion of choice

depends on contexts of use: in some very specific contexts, it is plausible that expressivity considerations are favored over repetition. Those (highly hypothetic) contexts would then make the fixation of a new gram possible, and this gram will then be able (though this next step is in no way automatic) to spread over the rest of the grammatheme (see Fig. 3.4a).

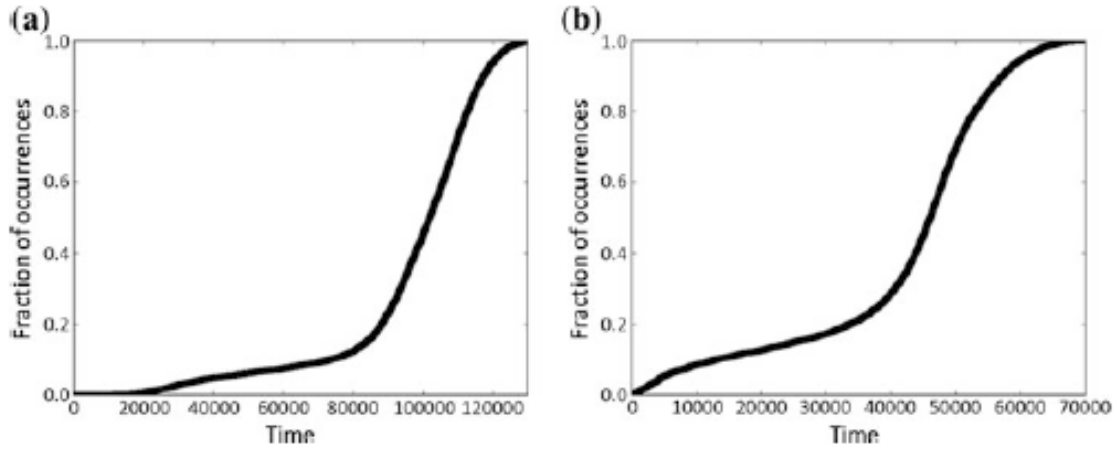

Fig. 3.4: Time evolution of the fraction of occurrences of the new gram (a) in the frequency-oriented zeteme (orange-purple zeteme in Fig. 3.3), following the first scenario and (b) in the entering zeteme, following the second scenario. Time is given by numerical time, i.e., the number of iterations of the model. In the first scenario (a), the new gram cannot enter the zeteme at first (0–2000), since grams are chosen to be uttered according to their frequency. Later (2000–8000), as the new gram dominates the expressivity-oriented zeteme, it starts to be used in the frequency-oriented zeteme. The new gram will eventually prevail in both zetemes. In the second scenario (2), the new gram is immediately used, and replaces the older gram according to a sigmoid-shaped curve.

Hypothesis (2) gives promising results (see Fig. 3.4b). New grams may or may not become fixed in their entering site, depending on the degree of conventionalization of the implicature (formalized as the weight of the link between the external site and the grammatheme site). This is in line with the claim that conventionalization of implicature is indeed the starting point of any grammaticalization (Nicolle 1998).

3.6. Conclusion

We have proposed a new framework, the grammatheme, which offers a convenient mathematical representation of various hypotheses and linguistic concepts. The model can be developed in many directions. Future works will explore how to implement other important features of grammaticalization, such as phonetic erosion and semantic bleaching, but also further pragmatic considerations. It would also be interesting to incorporate some information theory elements, a path already explored in (Fortuny and Corominas-Murtra 2013).

Yet, for now, it is already an accomplishment to be able to compare different scenarios, relying on different hypotheses, of the very beginning of the grammaticalization process. This led us to point out the importance of conventionalization of implicatures. Linguistically, it is far from being new, and (Bybee 2006) provides a comprehensive study on this matter. However, our attempt is one of the first to account for the viability of this process in a quantitative fashion.

In this task, we believe to have fulfilled what is expected from a modeling approach. A modeling approach cannot definitely rule out a possible explanation—one can always think of a mathematical variant, or add a new mechanism which would change everything. But a model can affirm that an explanation actually works—it can attest the efficiency of this explanation. We do not rule out the importance of expressivity considerations: we just prove that conventionalization of implicatures is an effective mechanism, susceptible of implementation, and that it yields the results it had been elsewhere claimed it would be able to.

And this leads to new questions. We have assumed that the conventionalization just happens, without exploring how, or why. We did not even raise the question of what could cause the weight of the conceptual link symbolizing this conventionalization to vary. To answer this question, we could extend the grammatheme representation to the whole language, accounting for purely lexical aspects of language, as well as the important distinction between the two poles of the lexico-grammatical continuum. We would also need to carefully check that they do differ, that they are not blurred into each other in an inaccurate representation. Our approach thus offers an opening towards the investigation of new questions, that it has helped to identify. We therefore, consider the grammatheme as a promising modeling tool to explore the pitfall of grammaticalization.

References

ATILF-CNRS, Université de Lorraine. (2014). Base textuelle FRANTEXT [WWW Document]. URL http://www.frantext.fr .

Baronchelli, A., Felici, M., Loreto, V., Caglioti, E., Steels, L. (2006). Sharp transition towards shared vocabularies in multi-agent systems. Journal of Statistical Mechanics: Theory and Experiment, 2006, P06014.

Baronchelli, A., Loreto, V., & Steels, L. (2008). In-depth analysis of the naming game dynamics: The homogeneous mixing case. International Journal of Modern Physics C, 19, 785–812.

Baxter, G. J., Blythe, R. A., Croft, W., & McKane, A. J. (2006). Utterance selection model of language change. Physical Review E, 73, 046118.

Biber, D., Conrad, S., & Reppen, R. (1994). Corpus-based approaches to issues in applied linguistics. Applied Linguistics, 15, 169–189.

Bybee, J. L. (2003). Phonology and language use. Cambridge: Cambridge University Press.

Bybee, J. L. (2006). From usage to grammar: The mind's response to repetition. Language, 82, 711–733.

Bybee, J. L., & Dahl, O. (1989). The creation of tense and aspect systems in the languages of the world. Studies in Language, 13, 51–103.

Carlier, A. (2007). From preposition to article: The grammaticalization of the French partitive. Studies in Language, 31, 1–49.

Castellano, C., Fortunato, S., & Loreto, V. (2009). Statistical physics of social dynamics. Reviews of Modern Physics, 81, 591–646.

Castelló, X., Baronchelli, A., & Loreto, V. (2009). Consensus and ordering in language dynamics. European Physical Journal B: Condensed Matter and Complex Systems, 71, 557–564.

Chomsky, N. (1981). Lectures on government and binding. Dordrecht: Foris Publications.

Christiansen, M. H., & Kirby, S. (2003). Language evolution: Consensus and controversies. Trends in Cognitive Sciences, 7, 300–307.

Croft, W. (2013). Evolution: Language use and the evolution of languages. In P.-M. Binder & K. Smith (Eds.), The language phenomenon, The Frontiers Collection (pp. 93–120). Berlin: Springer.

Cucker, F., Smale, S., & Zhou, D.-X. (2004). Modeling language evolution. Foundations of Computational Mathematics, 4, 315–343.

Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. New Ideas Psychology Modern Approaches to Language, 25, 108–127.

Ellis, N. C. (2008). The periphery and the heart of language. In S. Granger & F. Meunier (Eds.), Phraseology: An interdisciplinary perspective (pp. 1–13).

Fagard, B., & Combettes, B. (2013). De en à dans, un simple remplacement? Une étude diachronique. Langue Française, 178, 93.

Fitch, W. T. (2005). The evolution of language: A comparative review. Biology and Philosophy, 20, 193–203.

Fortuny, J., & Corominas-Murtra, B. (2013). On the origin of ambiguity in efficient communication. Journal of Logic, Language and Information, 22, 249–267.

Gómez, J.-C. (2007). Pointing behaviors in apes and human infants: A balanced interpretation. Child Development, 78, 729–734.

John, T., & Bennett, A. (1982). Language as a self-organizing system. Cybernetics and System, 13, 201–212.

Keller, R. (1989). Invisible-hand theory and language evolution. Lingua, 77, 113–127.

Keller, R. (2005). On language change: The invisible hand in language. New York: Routledge.

Kirby, S. (2001). Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. IEEE Transactions on Evolutionary Computation, 5, 102–110.

Kirby, S., & Hurford, J. R. (2002). The Emergence of Linguistic Structure: An Overview of the Iterated Learning Model. In A. C. Laurea & D. P. L. Laurea (Eds.), Simulating the Evolution of Language (pp. 121–147). London: Springer.

Komarova, N. L., & Jameson, K. A. (2008). Population heterogeneity and color stimulus heterogeneity in agent-based color categorization. Journal of Theoretical Biology, 253, 680–700.

Kroch, A. S. (1989). Reflexes of grammar in patterns of language change. Language Variation and Change, 1, 199–244.

Leavens, D. A., Hopkins, W. D., & Bard, K. A. (1996). Indexical and referential pointing in chimpanzees (Pan troglodytes). Journal of Comparative Psychology, 110, 346–353.

Loreto, V., Baronchelli, A., Mukherjee, A., Puglisi, A., & Tria, F. (2011). Statistical physics of language dynamics. Journal of Statistical Mechanics: Theory and Experiment, 2011, P04006.

Meillet, A. (1912). L'évolution des formes grammaticales, Zanichelli.

Mukherjee, A., Tria, F., Baronchelli, A., Puglisi, A., Loreto, V. (2011). Aging in language dynamics. PLoS ONE, 6, e16677.

Nicolle, S. (1998). A relevance theory perspective on grammaticalization. Cognitive Linguistics, 9, 1–36.

Niyogi, P. (2009). The computational nature of language learning and evolution. Cambridge, Massachusetts: MIT Press.

Niyogi, P., & Berwick, R. C. (1996). A language learning model for finite parameter spaces. Cognition, Compositional Language Acquisition, 61, 161–193.

Nowak, M. A., Komarova, N. L., & Niyogi, P. (2001). Evolution of universal grammar. Science, 291, 114–118.

Nowak, M. A., & Krakauer, D. C. (1999). The evolution of language. Proceedings of National Academy of Sciences, 96, 8028–8033.

Ogura, M., & Wang, W. S.-Y. (1996). Snowball Effect in Lexical Diffusion, English Historical Linguistics 1994: Papers from the 8th International Conference on English Historical Linguistic (8. ICEHL, Edinburgh, 19–23 September 1994).

Omodei, E., & Fagard, B. (2013). Cases, Prepositions, and In-Betweens: Sketching a Model of Grammatical Evolution. In European Conference on Complex Systems (ECCS, Barcelona, 16–20 September 2013).

Pawlowitsch, C. (2007). Finite populations choose an optimal language. Journal of Theoretical Biology, 249, 606–616.

Pawlowitsch, C., Mertikopoulos, P., & Ritt, N. (2011). Neutral stability, drift, and the diversification of languages. Journal of Theoretical Biology, 287, 1–12.

Peyraube, A. (2002). L'évolution des structures grammaticales. Langages, 1, 46–58.

Puglisi, A., Baronchelli, A., & Loreto, V. (2008). Cultural route to the emergence of linguistic categories. Proceedings of National Academy of Sciences, 105, 7936–7940.

Reali, F., & Griffiths, T. L. (2010). Words as alleles: connecting language evolution with Bayesian learners to models of genetic drift. Proceedings of the Royal Society of London B: Biological Sciences, 277, 429–436.

Schulze, C., & Stauffer, D. (2005). Monte Carlo simulation and the rise and fall of languages. International Journal of Modern Physics C, 16, 781–787.

Scott-Phillips, T. C., & Kirby, S. (2010). Language evolution in the laboratory. Trends in Cognitive Sciences, 14, 411–417.

Sperber, D., & Wilson, D. (1996). Relevance: Communication and Cognition, (2 ed.). Oxford ; Cambridge, MA: Wiley-Blackwell.

Steels, L. (1995). A self-organizing spatial vocabulary. Artificial Life, 2, 319–332.

Steels, L. (1999). The talking heads experiment, words and meanings.

Steels, L. (2011). Modeling the cultural evolution of language. Physics of Life Reviews, 8(339–356), 4.

Steels, L., & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. Behavioral and Brain Sciences, 28, 469–529.

Traugott, E. C., & Dasher, R. B. (2001). Regularity in semantic change. Cambridge: Cambridge University Press.

Victorri, B., & Fuchs, C. (1996). La polysémie - construction dynamique du sens, Langue, raisonnement, calcul. Paris: Hermès Science Publications.

Vogt, P., & De Boer, B. (2010). Language evolution: Computer models for empirical data. Adaptive Behavior: Animals, Animats, Software Agents, Robots, Adaptive Systems, 18, 5–11.

Wedel, A. B. (2006). Exemplar models, evolution and language change. The Linguistic Review, 23, 247–274.

Zanette, D. H. (2008). Analytical approaches to bit-string models of language evolution. International Journal of Modern Physics C, 19, 569–581.

[1] Note that some models are hard to classify; for instance, the Iterative Learning Model (Kirby and Hurford 2002), while closer to transformational grammar in its approach, falls into the first category.

[2] See however (Victorri and Fuchs 1996) on how such links may evolve and generate polysemy.

[3] With a few exceptions: the most notable one, the category game, yields a glassy behavior which blocks the language in an ever though slowly changing, non-stable state (Mukherjee et al. 2011); the Fagard and Omodei model (Omodei and Fagard 2013) implements an ad hoc mechanism to ensure that language always changes (a form which is well established comes to carry new meanings); another model (Nadal and Pierrehumbert, unpublished) leads to a state where language yields an unchanging hierarchical structure, but where words constantly exchange their respective roles within this hierarchy.

[4] All pairings are not relevant. Some of them may be excluded when representing a situation of interest.

[5] From the Greek root “zêtê” meaning “information”.

[6] As the themata were both the territorial districts of the Empire, and the armies occupying them, the grammatheme deals both with zetemes (the ‘territory’ of grammar), and with occurrences ‘populating’ the associated sites; as the number of the themata and the borders of the Empire were fluctuating, depending on whether the Byzantine army was keeping them or not, the condition of being expressed determine the belonging of a given zeteme to the grammatheme, the ‘borders’ of the latter being subject to change.