Chapter 1

Introduction

1.1 Introduction to Complexity

Complexity theory is a major interdisciplinary paradigm which unifies natural and social sciences through a combination of quantitative and qualitative methods applied at various phases of the research, from observations and data analysis to modeling, simulation, and interpretation of complex phenomena (Anderson, 1972; Ross and Arkin, 2009). In this framework, cognitive and nonlinear stochastic models and effective representation methods such as complex networks and fractal geometry, represent a part of the standard toolbox. In fact, in complexity theory, phenomena emerge dynamically from hierarchical, multi-modular systems, produced by bundles of possibly stochastic interactions and causalities rather than from correlative determinism.

As far as language and linguistic analysis are concerned, complexity itself can be understood from at least two different perspectives. On the one hand, there is 'constitutional complexity', or 'bit complexity', that is, complexity due to inventories of functional units or structural features, such as phonemes, morphemes, and lexical stems. On the other hand, there is '(socio-)interactional complexity', or, in other words, 'communal complexity', involving intricate modules of units and features, or networks of interactive individuals and aggregates. These different aspects all find their natural description in the *multiplex* paradigm, that is, through a model system composed of a set of interacting, overlapping networks. The unification of the two aforementioned dimensions of complexity represents a major challenge and is a focus point of this book.

For more than a decade, a growing interdisciplinary community has applied the tools of complex systems theory and statistical mechanics to the study of problems that traditionally belong to the field of linguistics. Nowadays, language dynamics represents a relevant branch of complexity theory. The modeling of language dynamics has mainly addressed three fundamental dimensions of language complexity.

(i) **Language spread and competition** (the dynamics of language use in multilingual communities).

(ii) **Language evolution** (how the structure of language evolves).
(iii) **Language cognition** (the way the human brain processes linguistic knowledge).

While these three dimensions of language complexity closely interact with each other and should all be taken into account for an exhaustive description of language complexity, it is useful, for clarity, to consider them as separate aspects. In the present book, we mainly address the first two dimensions, discussing language spread and competition models and considering language evolution models. However, we will also use socio-cognitive models of linguistic and cultural change.

A fundamental characteristic of complex systems theory is its high interdisciplinarity: the methods and the models used share many relevant features with other disciplines. This is simply due to the fact that, by definition, the paradigms of complex systems theory apply to all complex systems, independently of their specific nature. In the following, we present a short list of common points between the complex systems approach to language dynamics and other disciplines.

First, language dynamics is clearly connected to problems related to social interactions. Such a direct connection with social sciences and social dynamics follows from the mere fact that linguistic features represent cultural traits of a specific nature, namely, semiotic traits, distributed on two levels of organization: as morphemes, words, and phrases on the one hand; distinctive features and the so-called auto-segmental traits on the other hand. The propagation and evolution of these complex, double-sided items (words versus traits) can in principle be modeled through dynamics analogous to those of cultural spread and opinion dynamics. However, one should also consider that the conditions of their spreading in time and space may turn out to be more intricate, as they resort more to indexical traits than mere artifacts, opinions, or ideas. In other words, their liability is higher than for artifacts. Moreover, the complexity of diffusion and equilibrium between competing languages or dialects differs strongly according to the patterns and properties shared among languages: the more akin languages are, the more they mingle through contact; whereas, genetically and typologically distant languages undergo stiffer processes.

Many models considered in the book are models of competition between cultural traits and are usually referred to as ‘language competition models’. Even when language change and evolution processes are ignored, language competition models can offer the invaluable opportunity to understand the mechanisms regulating the size evolution of linguistic communities, representing an important part of the information that needs to be taken into account, for example, in the design of appropriate linguistic policies. Language competition is a central concept in this book and reveals another relevant analogy, namely between language dynamics and the dynamics of ecological systems. This analogy can be particularly useful for representing in a simple way the more complex landscape of interactions between the individuals of two different-speaking communities. Interestingly, such an analogy seems to suggest that single languages (or linguistic traits) are the actual entities competing with each other for resources, while speakers are the ‘mere niches’ where selection and evolution eventually take place. The promotion of languages to the role of main characters recalls the definition of ‘meme’ introduced by Dawkins (1976), in this case, with memes representing languages or linguistic traits competing with each other for resources in the minds of speakers,

analogously to the way genes compete for resources at the genetic level. This type of framework is certainly useful when considering minimal models of language competition, where the relevant variables are the sizes of the various interacting communities, as in a Lotka–Volterra modeling framework. However, even in the perspective where languages interact as competing species, one should bear in mind that a competition process between two languages differs in many respects from competition between species.

(i) First, language is a **cognitive system** or, as Noam Chomsky put it (Chomsky, 1986), a cognitive device, based on universal constraints (principles of universal grammar, for example, markedness versus default, topicality and comment constituent structure of predication, lexicon versus functional heads, etc.) and accommodated to local parameters (for example, accusative versus ergative alignment, head marking versus dependent marking, etc.).

(ii) Second, **social time can evolve much more rapidly than biological time**—and language, as a matter of fact, is embedded in social periods and social time, interconnected with a complex set of cultural and political constraints; see Dixon (1997) for a punctuated model of language evolution and shift.

(iii) Third, language is ontologically a **rather freely articulated system** of rules and constraints, with a very high sensitivity to setting (that is, status of the speakers involved in social intercourse and to the context of communication), plasticity being the rule, rather than the exception, in colloquial interaction.

The first of the factors listed here accounts for the potential complexity of repertoires. The second accounts for complex patterns of emergence, while the third factor for powerful patterns of miscegenation and self-organization. The emergence of creoles provides a good example of the influence of these factors in the evolution of new means of expression and linguistic patterns leading to totally new languages. The first factor accounts for universal trends such as low markedness and primarization of complex structures, and local parameters, such as right-branching determination (as in Haitian Creole) or left-branching determination (as in Seselwa or other Mauritian Creoles). The second factor powerfully accelerated the emergence of creoles in just a few generations, in spite of the appalling social violence enforced or enslavement of human aggregates. The third factor expands the complexity and the plasticity of the repertoire, from basilect to acrolect (see further for a definition of these terms).

Before proceeding further, it would be good to clarify that the complex systems approach developed in this volume differs conspicuously from the so-called computational linguistics or automated text analysis, whose purpose is to process linguistic data or corpora in order to implement grammatical models through computing automata, or to widen the empirical range of analysis with powerful tools for processing and parsing data as done by Clark et al. (2010) and Grishman (1986). Computational linguistics and automated text analysis are concerned with algorithmic complexity, as is complexity theory, but the main goal of those approaches is data extraction and a more accurate and comprehensive description and modeling of the fabric of the lexicon, of grammar or discourse. For example, typical research in computational linguistics may attempt to account for the complexity of verb inflection in Bulgarian or Macedonian, or to contrive more accurate and efficient multilingual automatic translators or devices for speech synthesis, etc.

1.2 Aim of the Book

Many of the canonical models that will be discussed in the present book have been developed over more than a decade by physicists, who have applied the tools of statistical mechanics and complex systems to study the problems that traditionally belong to the field of linguistics. The interest among physicists and the complex systems community in modeling language competition was originally raised by the work of Abrams and Strogatz (2003) (see also Wichmann [2008a,b] for a review). However, research on the dynamical modeling of the interaction between linguistic communities had already begun more than a decade earlier with the papers written by Baggs and Freedman (1990, 1993). These papers address the problem from two different sides that are still central in the current research in language dynamics:

(i) A mathematical formulation based on a close analogy with the ecology of competing biological species: thus, language dynamics (or at least the study of language competition) can be considered a natural branching of the mathematical modeling of ecological processes.

(ii) The concern for the problem of the disappearance of language and cultural diversity, similar to its ecological counterpart of the disappearance of species diversity.

In our opinion, it is now the right moment to look back on what has been done so far by the growing interdisciplinary community of researchers working in language dynamics, and to draw some conclusions on the picture obtained by the various studies. This introspection will also allow us to ponder over the most relevant and reasonable research directions to be taken in the near future by applying complexity theory to language dynamics. These questions have not been considered systematically so far. The scientific goal of the present book is to fill this gap, which affects different disciplines, and to provide a reference point for the highly interdisciplinary studies concerning the mathematical modeling of language competition, evolution, and spread. On the social side, as long as the defense of language diversity is a concern, the models studied can be used for planning optimal strategies to defend endangered languages or at least limit what in most cases seems to be an unavoidable imminent cultural catastrophe.

The treatment of mathematical modeling of (socio)linguistic complexity is developed in the present book around the following two lines:

- First, the exploration of a theoretical or virtual range of facts and phenomena that can be applied to language in space and time, from a logical and hypothetico-deductive standpoint.
- Second, the verification of the match between the models and their corresponding results obtained through simulations, on the one hand, and with the corresponding (socio)linguistic data, on the other hand.

The aforementioned two points are closely related to each other. The first point resorts especially to analytical calculations and numerical simulations in complex system modeling as main investigation tools; the second one aims at challenging and validating theoretical models with empirical facts. Various interesting case studies, such as that of the Mazatec dialects or the Basque language, will be employed throughout the text to present applications

to real problems that illustrate concepts and methods: language ecology, linguistic evolution, structural complexity, and the various computing tools employed.

Particular attention will be devoted to the (in)adequacy of abstract models of language diffusion in space and time, or language competition in contexts of diglossia (subordinate bilingualism or bidialectalism) or language planning, highlighting the relevance of (socio-)linguistic facts from other fields not strictly (socio-)linguistic.

The goal of these types of studies been clear in the complex systems community for a while. The need to make social dynamics, in general, a more 'sound discipline grounded on empirical evidence' has been well explained, for example, by Castellano et al. (2009) and Taagepera (2008). In other words, we are looking, in the field of language study, for the standard procedure common to any scientific research, namely, the development of a set of models and laws which are applicable within well-defined ranges of validity. In fact, considering the possible applications of this method, the approach advocated in this book has a strong predictive potential on the future trends of linguistic communities and the evaluation of the state of endangered languages, or for shaping the choice of effective language policies to protect minority languages.

Furthermore, the models developed can also be reapplied to interpret facts with models and algorithms, that is, to analyze new data sets in light of the phenomenological models validated, in order to test historical hypotheses and the actual mechanisms of cultural spread. In questions of a historical character, predictions based on an initial or an abstract state need to be checked by facts; conjuncture should sustain conjecture.

Our approach also aims at filling a gap between advanced research in language and complex systems and routines already deeply rooted in both fields of expertise. On the one hand, in hard sciences, one often relies too much on abstract models, providing results considered as irrelevant by linguists. On the other hand, linguists observe intricate situations of multilingualism and linguistic diffusion, without being conscious of the weight and consequences of micro-trends and parameters. Mathematical modeling can help to conceptualize and deliberate on empirical facts and situations better than from a mere ethnographic standpoint, while the pragmatic interpretation of facts and figures through sociolinguistics may considerably foster mathematical modeling in its endeavor to predict and represent reality from as broad a scope as possible. Even if we do not fill this gap, we hope that, by putting the different approaches side by side and using them to study some real problems, we are making further steps toward a constructive integration of the different methods and viewpoints.

1.3 The Readership

While the topics and goals of this book can be easily situated within the wide and new research field of complex systems theory, it may seem that this is another instance of physics trying to invade other disciplines. Physics has been undergoing deep, symbiotic interdisciplinary developments that address many problems of biology (biophysics), economics (econophysics),

sociology (socio-physics), ecology, and now linguistics as well, resulting in interdisciplinary approaches that are sometimes criticized, as in the case of econophysics.

However, the case of linguistics is different for many reasons. In fact, besides the physicists interested in linguistic problems, there is also a growing part of the linguistic community that has recently become aware of the possibilities offered by complex systems theory. Therefore, the target audience of this book includes those physicists willing to discover the ongoing problems in linguistics that one could address using the theory of complex systems and linguists willing to learn the basics of complex systems theory in order to communicate and collaborate with the former. In order to favor productive interaction between these two communities in the future, it is important to first learn the other's language, that is, the terminology, as well as the corresponding methods and specific ways to approach a problem. For these reasons, the book is written in a way that also makes it accessible to non-specialists, for example, graduate and undergraduate students of both fields in general. Thus, this book may be used in academic courses on complex systems, social dynamics, social sciences, ecological linguistics, quantitative linguistics, mathematical modeling, etc.

1.4 Structure of the Book

The book is organized into two parts. The first part may look more 'linguistic', being concerned about the analysis of linguistic databases, while the second part may look more 'mathematical', being more focused on the modeling of language dynamics. In fact, both the parts have a deeply linguistic side and require the formalization of some quantitative method.

Part One concerns the problem of interpreting linguistic databases to reveal the presence of possibly related languages and determine their main features. This is a relevant side of the approach presented in this book since linguistic databases provide the empirical comparison terms for validating quantitative methods and mathematical models. Part One also contains basic notions about language change and how to quantify the degree of relation between languages, about complex networks, and about techniques based on Levenshtein distance. Their use is illustrated with the help of some working examples, including the Mazatec, Basque, Tseltal, and the Numic languages. Some of these examples come from our own research, started in 2008 as an informal, interdisciplinary study by a group of mathematicians, physicists, and linguists who met at a symposium in Tartu.

Part Two deals with the modeling of linguistic systems, that is, with the actual mathematical models developed so far. Various types of models are discussed, ranging from microscopic individual-based models to macroscopic continuous models of language evolution, competition, and diffusion. The overview is neither exhaustive nor systematic—many interesting models are not discussed—but tries to provide an outline of the field through a selection of some minimal models.