Claudia Marzi

Modelling the morphological lexicon

A computational approach to mono- and bilingual learning and processing of verb inflection

FrancoAngeli
Edizioni La passione per le conoscenze

Informazioni per il lettore

Questo file PDF è una versione gratuita di sole 20 pagine ed è leggibile con **Adobe Acrobat Reader**

La versione completa dell'e-book (a pagamento) è leggibile **con Adobe Digital Editions**.

Per tutte le informazioni sulle condizioni dei nostri e-book (con quali dispositivi leggerli e quali funzioni sono consentite) consulta **cliccando qui** le nostre F.A.Q.

Materiali Linguistici. Collana a cura del Dipartimento di Studi Umanistici – Sezione di Linguistica Teorica e Applicata – dell'Università di Pavia

Materiali Linguistici è una collana fondata nel 1990, che pubblica studi su vari aspetti delle lingue naturali e della loro evoluzione storica. La collana riflette la molteplicità di punti di vista teorici e metodologici che la linguistica oggi comprende, con riferimento in particolare ai seguenti temi: descrizione dei sistemi linguistici, tipologia, sociolinguistica, semantica, pragmatica e apprendimento linguistico.
Ogni volume è sottoposto a processo di peer review.

Materiali Linguistici. *Book Series edited by the Department of Humanities – Section of Theoretical and Applied Linguistics – University of Pavia*

Materiali Linguistici *is a peer-reviewed series founded in 1990. It publishes studies on various aspects of natural languages and their historical evolution. The series reflects the wide range of methodological and theoretical approaches of contemporary linguistics, with particular reference to the following themes: description of linguistic systems, typology, sociolinguistics, semantics, pragmatics and language learning.*

Segreteria – Editorial Assistant
Ilaria Fiorentini *(Università di Pavia)*
Dipartimento di Studi umanistici – Sezione di Linguistica
C.so Carlo Alberto 5, I-27100 Pavia; tel. 0382984484.
Per maggiori informazioni: *www.lettere.unipv.it/diplinguistica*

Claudia Marzi

MODELLING THE MORPHOLOGICAL LEXICON

A computational approach to mono- and bilingual learning and processing of verb inflection

FRANCOANGELI

Questo volume è stato pubblicato grazie al contributo dell'Istituto di Linguistica Computazionale "A. Zampolli", Consiglio Nazionale delle Ricerche, nell'ambito del progetto "Modelli (bio-)computazionali dell'uso linguistico" DUS.AD016.075.004.

Contents

Abstract

This work aims at defining an explanatory model of the morphological lexicon as a dynamic system of word learning and processing in both mono- and bilingual contexts. The main focus is on exploring some relevant aspects of the paradigmatic organisation of the mental lexicon in language learning, based on a dynamic analysis of mono- and bilingual contexts.

The proposed interdisciplinary approach to lexical acquisition combines theoretically-motivated accounts, psycho-cognitive evidence and methodologies, and machine learning technologies. In particular, I will take into account those basic psychological and cognitive mechanisms that are considered as crucial in language acquisition: (i) the ability to perceive recurrent morphological structures (invariances) in varying temporal contexts, (ii) the capability to access/activate time series of symbols in the short term memory and to selectively integrate them with long term memory expectations, (iii) the attitude towards building novel forms through analogical extension of intra- and inter-paradigmatic relations (generalisation).

This investigation is pursued through a computational model based on a recurrent Self-Organising Map, with Hebbian connections defined over a temporal layer (*Temporal Self-Organising Map, TSOM*), providing a principled algorithmic account of effects of lexical acquisition, processing and access. The computational simulation of a biologically inspired neural architecture of the mental lexicon offers the possibility to reproduce a wide range of conditions of mono- and bi-lingual input exposure, and to illustrate the dynamic of word acquisition and the emergence of morphological organisation.

The proposed model provides an adaptive multifactorial account of morphology acquisition affected by a variety of input factors, such as word frequency distributions, paradigm regularity and wordlikeness, whereby lexical perception and organisation are grounded in memory-based processing

strategies. In addition, it suggests a processing-based notion of morpheme, as a by-product of processing dynamics, with paradigms emerging as specialised surface relations between inflected forms.

Il presente lavoro si propone di definire un modello esplicativo del lessico morfologico come sistema dinamico di acquisizione ed elaborazione in contesti mono- e bi-lingui.

L'attenzione è rivolta all'esplorazione di alcuni degli aspetti principali dell'organizzazione paradigmatica nell'acquisizione evolutiva del lessico mentale, attraverso un'analisi dinamica delle tappe nelle quali essa si articola in contesti mono- e bilingui, partendo dai contributi più recenti della riflessione teorica contemporanea sul lessico mentale, dell'evidenza psico-linguistica sperimentale, e della modellistica computazionale su base neurale. In particolare, si approfondiscono alcuni dei meccanismi psico-cognitivi di base ritenuti determinanti in questo processo: (i) la capacità di percepire strutture invarianti che ricorrono in diversi contesti temporali, (ii) la capacità di attivare sequenze simboliche nella memoria a breve termine e di integrarle selettivamente con le aspettative della memoria a lungo termine, (iii) la capacità di produrre forme nuove attraverso processi di estensione analogica intra- ed inter-paradigmatica (generalizzazione).

La ricerca si avvale del contributo di mappe ricorrenti auto-organizzanti dotate di un livello di connettività hebbiana (TSOM, mappe temporali auto-organizzanti), che offrono una modellizzazione algoritmica di effetti di acquisizione, elaborazione e accesso lessicali. La simulazione computazionale di architetture neuro-biologicamente ispirate del lessico mentale consente di riprodurre artificialmente un ampio ventaglio di condizioni di esposizione a input mono- e bi-lingue, e di analizzare la dinamica dell'acquisizione lessicale e l'insorgenza di una organizzazione morfologica basata sui paradigmi.

Il modello si propone di illustrare una dinamica multifattoriale del processo di acquisizione morfologica, risultante dall'interazione tra fattori di frequenza, regolarità paradigmatica e familiarità lessicale, e meccanismi di percezione e organizzazione lessicali basati su strategie di elaborazione e memorizzazione. Nella prospettiva suggerita, la nozione di morfema acquisisce una identità funzionale in relazione alle dinamiche di percezione delle relazioni superficiali tra forme pienamente flesse, e la nozione di paradigma emerge come risultato dell'interazione di principi di auto-organizzazione emergente.

List of abbreviations

ANN	Artificial Neural Networks
BMU	Best Matching Unit
CxG	Construction Grammar
fMRI	functional Magnetic Resonance Imaging
GAM	Generalised Additive Model
L1	First Language
L2	Second Language
MB	Morpheme Boundary
ms	millisecond
SLA	Second Language Acquisition
SOM	Self-Organising Map
STG	Superior Temporal Gyrus
TSOM	Temporal Self-Organising Map
UG	Universal Grammar
WM	Working Memory

This book would not have been possible without the precious advice of Pierluigi Cuzzolin (University of Bergamo) and Vito Pirrelli (Institute for Computational Linguistics, CNR Pisa). I am grateful to them for their insightful comments. I want to acknowledge the Institute for Computational Linguistics of the Italian National Research Council, the General and Editorial Boards of the book series *Materiali Linguistici* for giving me the opportunity of publishing this work.

1. Introduction

1.1 Objectives and methodology

The main goal of the present work is to propose an explanatory model of the morphological lexicon as a dynamic system of word learning and processing in both mono- and multi-lingual contexts, based on the interplay between two important research areas in the current debate on morphological competence: the psycholinguistic modelling of the mental lexicon (Aitchison 1994; Bybee 1985, 1995; Baayen *et al*. 1997; among others) and word-based theories (lexicalist theories) on word structure and morphological processes (Jackendoff 1975; Aronoff 1976, 1994; Scalise 1994; among others).

In particular, through a computational model of monolingual and bilingual lexical self-organisation, the focus is devoted to suggesting possible answers to some main basic questions, such as: what general cognitive mechanisms are involved in the speaker's lexical competence? What strategies are at work when the speaker shows the ability to get access to both unknown and already memorised words? How does a speaker perceive the internal structure of morphological complex words? And, in case of bilingual speakers, what cognitive strategies may optimise contextual learning and access of/to more than one lexical system?

To locate this contribution more clearly on the vast theoretical map of word-based approaches, I intend to heavily rely on a specific perspective on morphological theorising that identifies word forms as the basic units for morphological competence, and their recurrent parts as abstractions over full forms. This type of perspective has been defined by Blevins (2006) as *abstractive*, in contrast with a *constructive* approach to word structure assuming a redundancy-free lexicon based on morphemes as the basic building blocks of morphological competence.

The constructive idea I want to oppose, namely that morphological competence comprises a repository of basic building blocks (be they constituents on a par with each other, or base forms modified by rules), and a set of rules for combining morphemes into full words, has deep roots in theoretical linguistics, dating back to the European and American *Structuralism* (for a detailed overview and reanalysis of grammatical theories that have dominated linguistics in the last century, see Matthews 1993).

Since Hockett (1954), considerable disagreement has been built on (i) how morphological rules operate and access the morphological lexicon, and (ii) what the ultimate content of the morphological lexicon should be.

For example, *Item-and-Process* approaches have classically assumed that the morphological lexicon contains only lexical roots and stems, all other formatives (e.g. affixes) being the by-product of morphological rules (e.g. Anderson 1992; Aronoff 1994). On the other hand, according to *Item-and-Arrangement* theories, which are particularly associated with the American Structuralism, all sublexical constituents are distinct entries in the morphological lexicon, each associated with a specific content (see Lieber 1992, among others). Fully spelled-out word forms are just the result of combining morphological entries through rules.

What both approaches have in common is the idea that the content of the morphological lexicon is governed by a principle of economy of storage, since the role of rules is precisely that of simplifying the lexicon by reducing the number of stored units to a small set of primitives. In other words, both approaches assume that surface word-forms are constructed by assembling smaller units.

However formally well-structured, both constructive views have been seriously challenged by a vast psycholinguistic literature on the topic, in favour of the idea of a redundant lexicon based on fully stored word-forms. According to this literature (see Aitchison 1987; Bybee 1985, 1988; Derwing 1990; for early but insightful overviews), which represents a grounding contribution for my present approach, there is no reason to think that the mental morphological lexicon should be redundancy-free. Speakers of a language know more words than they imagine (about 250,000 entries for a typical educated English adult speaker, Diller 1978), and the addition of more other languages does not seem to represent a problem for human memory. Hence, the role of morphological competence in a morphologically redundant brain cannot be that of storage compression.

Rather than being organised to minimise storage or to respond to logical necessity, it can be suggested that the structure of the morphological lexicon

must be maximally functional to lexical access and retrieval, as normal processes of speech production and comprehension must operate within time constraints of the order of hundreds of milliseconds (200 ms; Marslen-Wilson & Tyler 1980).

Jackendoff (1975) alludes to this aspect when he claims that morphological rules are formal tools that mutually relate fully-specified lexical entries. Rules are not intended to make such entries smaller[1]. In a similar vein, Matthews (1991: 204) provocatively questions why a language should have rules obscuring identity and function of its minimal elements, suggesting that words "are not wholes composed of simple parts, but are themselves the parts within a complex whole".

Probably, the first overtly clear connection between redundancy and structure is pointed out by Vennemann (1974), who suggests that the function of morphological competence is to help organise the lexicon, to give it structure, and to make words easier to store, which is something that is independently needed anyway, whenever efficient access or retrieval is an issue.

More recently, these lines of argument have developed into a view of the lexicon as a dynamic memory system (Elman 1995, 2004, 2009; Li 2009). Contrary to the *list-hypothesis*[2], which looks at the lexicon as a repository of stored items that are represented and accessed independently, an approach to the mental lexicon as a *dynamic system* argues that what is called lexical information is the resulting combination of various dynamic properties, such as the relation of a particular acquired stimulus to other co-occurring stimuli in a certain time window (syntagmatic relation), or the possible relation of a stimulus (i.e. a fully inflected or derived word form) to other competing stimuli in complementary distribution (paradigmatic relation). The idea of syntagmatic and paradigmatic relations between words has been suggested by Saussure: "*[...] words acquire relations based on the linear nature of language because they are chained together. [...] The co-ordinations formed outside discourse*

[1] Later on, Jackendoff (2002) focusses on these issues more clearly by including the necessity of redundancy in the lexicon and the rule system: where redundancy allows regularities to be captured and lexical entries to be evaluated in the lexicon. For Jackendoff (*ibidem* 189) "[...] rule learning is accomplished by the same process as word learning – because both are types of lexical item. […] stored items and rules are of the same formal character […]. The only innovation necessary in the learning theory is a way to learn variables form instances – which is needed in any event".

[2] The *full-listing* models assume an explicit representation of all complex word forms in the lexicon, suggesting that morphological effects would be an emergent by-product of the system (Butterworth 1983; Rueckl *et al*. 1997, among others).

differ strikingly from those formed inside discourse.[...] The syntagmatic relations is in praesentia. *It is based on two or more terms that occur in an effective series. Against this, the associative relation unites terms* in absentia *in a potential mnemonic series. [...] A syntagm immediately suggests an order of succession and a fixed number of elements, terms in an associative family occur neither in fixed numbers nor in a definite order. [...] A particular word is like the center of a constellation; it is the point of convergence of an indefinite number of co-ordinated terms*"[3].

From this perspective, the dynamic interaction between input stimuli and the emergence of structure in the mental lexicon is key to understanding lexical representations and lexical organisation.

Over the past three decades, in fact, considerable evidence has been supporting a more dynamic view of the lexicon than classical *structuralism* or *generative* theories of a redundancy-free lexicon are ready to acknowledge. Theoretical approaches and psycholinguistic studies of language processing have increasingly shared issues, empirical findings, experimental analysis, and explanatory models, as witnessed by the increasing importance played, of late, by morphological families (e.g. inflectional paradigms, derivational families, etc.) in recent psycholinguistic experimental protocols (e.g. de Jong *et al*. 2000; Moscoso *et al*. 2004b; Milin *et al*. 2009; Mulder *et al*. 2014).

It is nonetheless clear that a thorough understanding of the lexicon as a dynamic memory system goes well beyond the limits of human intuition and the reach of *box-and-arrow* approaches to cognition, as the interaction through time of even a small number of processing factors may lead to very diverse patterns of lexical organisation. *Box-and-arrow* models are static, consisting of a set of boxes interpreting stored representations, which are connected by arrows representing the processes that map one representation onto another. Moreover, these models make explicit the hypothesized information-processing activities carried out in a particular cognitive system (such as language), in a manner analogous to computer flow charts that depict the processes and decisions carried out by a computer program. While these types of description may suffice for capturing general characteristics of cognitive functions, they are unreliable to account for more detailed and dynamic phenomena (for a discussion, see, for example, Westermann & Plunkett 2007; Norris 2005).

[3] Ferdinand de Suassure (1922). *Cours de linguistique générale*. Paris, Edition Payot. Translated by W. Baskin, *Course in General Linguistics*. Meisel P. & H. Saussy (eds.). 2011, Columbia University Press, 123-126.

In this perspective, there is growing awareness that computer models can contribute considerably to a more precise characterisation and assessment of alternative functional models of the mental lexicon, since they provide experimental ways of testing the behaviour of a complex system as a function of different settings of configuration parameters and input conditions (Pirrelli 2007).

The proliferation of connectionist models for language learning and processing, particularly in connection with what has been commonly referred to as the "past-tense debate" (Pinker & Ullman 2002a; McClelland & Patterson 2002a), bears witness to a general expectation that computational models of language processing – i.e. knowledge of *how* we master our language – cannot be decoupled from linguistic data structures – i.e. knowledge of *what* we know when we know a language. A deeper understanding of some fundamental issues of the language architecture goes through better models of how it functions.

The present work is cast into such an interdisciplinary approach to the acquisition of the lexicon as a dynamic system. I believe that the combination of psycho-cognitive approaches, theoretically-grounded descriptive methodologies, and machine learning technologies, that in recent years have provided innovative ways to put *box-and-arrow* models of language architecture to the challenging test of a computer implementation, will offer better prospects to defining explanatory models of morphological competence.

By paraphrasing Jackendoff (2002: 34), the present attempt at understanding the nature of morphological competence can be said to lie at the crossroad of three lines of inquiry: (i) linguistic theory, providing a formal characterisation of the data structures that are stored and represented in the mind of a speaker; (ii) psycholinguistic models, providing the functional characterisation of the use of these data structures in the course of language perception and production; (iii) computational models, quantitatively describing how the data structures and the processes that store and assemble them are instantiated algorithmically.

Accordingly, psycholinguistic evidence and a cognitive perspective put functional constraints on abstract representations, while computational models explain how this comes about by implementing both representations and functional constraints.

In addition, the possibility to observe and examine these computational models at work can give an important contribution to our understanding of the nature of lexical knowledge, and can clarify the interplay between general principles of brain architecture and connectivity, computational issues of

language processing, and functional and descriptive models of contemporary linguistic theories. Finally, the main challenge of the proposed approach is to concurrently use the most recent contributions of theoretical approaches to the mental lexicon, the substantial body of experimental evidence offered by psycholinguistic literature, together with biologically-inspired computer models of language acquisition and processing, based on artificial neural-networks for a two-fold objective: (i) an analytical study of the developmental and self-organising processes governing the acquisition of the morphological lexicon in different languages; (ii) an overall assessment of how these models can shed light on the complex dynamics triggered by the acquisition of more than one language – either concurrently or deferred in time.

1.2 Outline of the book

The present work firstly outlines the fundamental theoretical reference frame (Chapter 2) and the most recent psycholinguistic experimental evidence (Chapter 3), supporting the idea of the morphological lexicon as the result of a dynamic self-organisation process. Secondly, it describes methodologies and computational architectures that underlie the machine-learning approach (Chapter 4). The most experimental part (Chapter 5) investigates the dynamics of morphological acquisition by exploring some ontogenetic aspects of lexical organisation in the domain of inflection (i.e. by focussing on *word families* constituted by *inflectional paradigms*) in language acquisition, in both mono- and bi-lingual contexts. Extensive computer simulations cover German, Italian, English, Spanish, Russian and Arabic, and differential dynamics/outcomes of morphological organisation due to different conditions of input exposure and different strategies of acquisition[4].

[4] The choice of these languages is made on grounds of data availability and graded inflectional complexity. In comparing the developmental stages in the acquisition of inflection in nearly two dozen languages (mainly in the Indo-European, Ugro-Finnic and Semitic families), Bittner and colleagues (2003) argue that the abstraction of morphological patterns from rote lexical storage by the child is conditioned by typological factors such as richness, uniformity and transparency of the inflectional paradigms. In an ideal graded scale where languages are arranged left-to-right from the more inflecting-fusional types to the more isolating types, Bittner and colleagues locate Greek and Russian to the left end, English and French to the right end, and Italian, Spanish, German in between. Arabic provides an example of language with a non-concatenative morphology. As I am interested in inspecting paradigm effects on global organisation of the morphological lexicon, this range of languages offers an interesting sample of more and less complex and more and less transparent/predictable inflectional processes.

A discussion and evaluation of the resulting self-organisation dynamics follow (Chapter 6) to illustrate and shed some light on analyses and mechanism of morphological competence that determine the speaker's capacity to have access and produce both known and unknown word forms, to perceive the internal structure of morphological complex words, and, in the case of multi-lingual speakers, to concurrently master more than one language system.

2. Theoretical background

2.1 Lexicon and rules in morphology

Over the last three decades there have been two dominant views on the architecture of the morphological lexicon and its relation to grammar, which offer the theoretical background against which the present approach is set off.

According to the so-called *dual-route* approach to word structure (Pinker & Prince 1988; Prasada & Pinker 1993; Pinker & Ullman 2002a, 2002b; Pinker 1994; Clahsen 1999; Marcus *et al*. 1995; Marcus 2001) recognition – and production – of a morphologically complex input word involves two interlocked steps: (i) a preliminary full-form access (lexical way), and (ii) an optional morpheme-based access to sub-constituents of the input word, resulting from application of combinatorial rules (grammatical way). The second step is taken only if the first one fails to find any matching access entry, i.e. if the word is not memorised as a full form in the lexicon, or when it is not available for immediate retrieval from the mental lexicon.

Such a view has encountered considerable consensus, for years, in the psycholinguistic and cognitive literature because it rests on the simple hypothesis of a direct correspondence between principles of grammar organisation (lexicon versus rules: Clahsen 2006), processing correlates (storage versus computation: Pinker & Ullman 2002a), and localization of the cortical areas functionally involved in word processing (temporo-parietal cortex versus frontal cortex and basal ganglia structures: Ullman 2004, 2016).

By mainly focusing on English, Pinker and Ullman (2002a, 2002b) foster the view that irregular past-tense forms are stored in the lexicon (supported

by the declarative memory[5]), whereas rules and a procedural system explain the computation of regular forms (i.e. by relying on the procedural memory [6]). Irregular forms are fully acquired and stored in long-term memory, together with grammatical features incorporated into lexical entries. Regular forms, on the other hand, can be productively generated by combinatorial rules that associate morphemes and simple words into complex or inflected words, similarly to phrases and sentences. However, such a view implies the ability to always reduce morphologically complex word-forms into their regular constituents. In addition, it implies that basic regular constituents are always memory-stored, off-line, as well as that regular word-forms are always processed on-line. Conversely, as an irregular form is stored, it blocks the application of a possible overgeneralisation to the regular pattern, and precludes a possible parsing into a stem and an affix that codes grammatical features.

In alternative to *dual-route* thinking, the mainstream *connectionist* answer to word storage and processing assumes a one-route model (Rumelhart & McClelland 1986), a pattern associator, which does not require lexical entries or combinatorial rules, but it defines a direct correspondence relation between a surface representation of the input form (a lexical base, e.g. *walk*) and a representation of the corresponding output form (an inflected form, e.g. *walked*). This approach proposes that both regular and exceptional aspects of verb inflection emerge from a unique integrated mechanism, subject to phonology and semantics (McClelland & Patterson 2002a), which learns stems and past-tense inflection of English words. Language acquisition – as well as other abilities – is the result of gradual adjustment of connections between meaning and context-sensitive processing units (Plaut & Gonnerman 2000).

Morphological structure plays no direct role in this, but it is conceptualised in connectionist models as the epiphenomenal by-product of an identity mapping between invariant portions of input and output patterns, irrespective of degrees of morphological regularity. The connectionist perspective suggests that "*decomposition is not an all-or-none phenomenon and that behavioural effects should be graded, reflecting the degree of convergence among semantics, phonological, and orthographic codes*" (Plaut & Gonnerman 2000: 452).

[5] The declarative or explicit memory refers to the human ability to retain, for long time, factual information, past experiences, including word-specific knowledge such as word meanings and sounds.

[6] Procedural memory, as a type on non-declarative memory, refers to skills and the ability to learn relations based on complex structures, such as phonology and morphology.