Palacký University Olomouc

Ph.D. Thesis in Italian Linguistics
Dept. of Romance Languages, Faculty of Arts

Second Language Acquisition: a corpus based approach for a theoretical investigation

MARCO PETOLICCHIO

Supervisors

DR. LOREM IPSUM
Palacký University
Dept. of Romance Languages, Faculty of Arts

PROF. DOLOREM SIT AMET
University of Lorem
Dept. of Lorem Ipsum, Faculty of Humanities

draft : : Sunday 27th May, 2018: : 15:45

Contents

List of Tables

List of Figures

Preface

Abstract

The main topic of this doctoral dissertation is on the analysis of syntactic structures in language acquisition, specifically in the domain of Czech and Slovak learners which acquire the Italian language. In particular, I will focus on the complex noun phrase subdomain, showing the compositionality of the phrase structure and the hierarchical fashion of this component. The analysis is casted in the Minimalist-oriented framework of the Generative Grammar (Chomsky, 1995, 1998, 2013; Hauser et al., 2002) and its application in the field of the second language acquisition (Rothman and Slabakova, 2017; Slabakova et al., 2014).

The usage of an established computational ground to conduct the work, where the data retrieved by fieldwork is stored in a coherent corpus which easily permits to be queried and interpolated for the research purposes, represents a standpoint for this research in its totality, yielding for a data-based approach to the whole process. The annotation schema of the data is standardized in order to adhere to the major point of discussion into the discipline (Kurdi, 2016; Clark, 2010; Kuebler and Zinsmeister, 2015), representing the plus to furnish a data source which is independent to the merely contingent purposes.

This research aims to offer a way to investigate how second language acquisition can be seen grounding on a coherent set of data in terms of annotation schema: it does insist either on the speculative questions both on computational models involved.

Keywords

- Computational Linguistics
- Syntax
- Second Language Acquisition
- Italian L2
- Corpus Linguistics

1

Introduction

The main question of this thesis yields a twofold mindset that is not a corollary of the research but which represents the process in which the work was conducted: how could I investigate a particular area of the language faculty as language acquisition in a way which can gain from the usage of the digital instruments in order to ground the theoretical analysis on actual data?

The idea under this research moves across the motivation to investigate over an empirically-grounded path the strategies shown by the learners during the acquisition of second languages, using an established coherent digital architecture. My task is twofold: on one side this provides for the developing of a theoretically-grounded framework to research in the fields of Second Language Acquisition (SLA), while on the other this necessitates to develop a linguistic corpus which collect in a coherent fashion a wide set of data that represent some spotted linguistic facts in order to give a transparent documentation of the learning path. The usage of the modern tools in developing a linguistic corpus yields for a fully documentable research path, in which is possible to reconstruct the steps and the choices which underlie its development, the methods used in the analysis, the correctness of the outcomes. This kind of research is intimately multidisciplinar in nature, embracing different approaches and areas of interest: digital humanities, corpus and computational linguistics for the development of the linguistic corpus, general and theoretical linguistics, studies on SLA and interlanguage for the theoretical analysis.

This introductive chapter collects a preliminary way to represent the main areas of the research, the methods involved in the analysis and the possible outcomes of such a way to conduct the work.

The idea under this research moves across the motivation to investigate over an empirically-grounded path the strategies shown by the learners during the acquisition of second languages, using an established coherent digital architecture. My task is twofold: on one side this provides for the developing of a theoretically-grounded framework to research in the fields of Second Language Acquisition (SLA), while on the other this necessitates to develop a linguistic corpus which collect in a coherent fashion a wide set of data that represent some spotted linguistic facts in order to give a transparent documentation of the learning path. The usage of the modern tools in developing a linguistic corpus yields for a fully documentable research path, in which is possible to reconstruct the steps and the choices which underlie its development, the

methods used in the analysis, the correctness of the outcomes. This kind of research is intimately multidisciplinar in nature, embracing different approaches and areas of interest: digital humanities, corpus and computational linguistics for the development of the linguistic corpus, general and theoretical linguistics, studies on SLA and interlanguage for the theoretical analysis.

1.1 Background for the thesis

1.1.1 Data-based research: definition

1.1.2 Data-based research: motivations for the thesis

1.2 Objectives of the thesis

The three main objectives of this thesis are methodological, empirical and theoretical.

1. Methodological objectives
   To address the decision and the methods raised by the compilation, the storage and the design of a learner based corpus, exploring the effective procedures for retrieving the relevant features for the analysis;
2. Empirical objectives
   To explore the previous generalizations of the acquisitional path in SLA literature comparing with the amount of linguistic productions given by different learners;
3. Theoretical objectives
   To describe the features which are relevant for characterise the language variety effect and the place of interlanguage.

1.2.1 Methodological objectives

1.2.2 Empirical objectives

The role of the native language (L1) can conditionates deeply the way in which the target language (L2) is acquired during the learning path: an emblematic case is the *transfer* of the knowledge about the structures of the L1 to the target, revealing the interlanguage.

During the last 20 years, a considerable part of linguistic literature is involved in developing some sort of models to think how the faculty of language can work, in its biological (Hauser et al., 2002), computational (Fodor, 2001) and cognitive components in a highly interdisciplinary environment. Studies on SLA is

a fertile field, which relies on comparative and contrastive analyses of linguistic phenomena, either both from an applied view (Ellis, 1994) than by theoretically grounded perspective focused on Generative framework (GenSLA) (Guasti, 2002, Rothman and Slabakova (2017), Hawkins (2001), Sorace (2011)). In this sense appears that the adoption of a coherent model to analyze the variation in grammar into a parametric model can be suitable for long-standing researches on SLA and interlanguage, also based on independent data-mining initiatives as in the case of the present purposes, which disentangles the data in a form of a linguistic corpus and the linguistic analysis.

1.2.3 Theoretical objectives

From a theoretical viewpoint, the research is inserted in the current theories which rely on the hierarchical functioning of the language faculty, for which the variation among languages are reconducted to a parametrizing of choice among the structures of languages (Chomsky, 1995, Chomsky (1998), Chomsky (2013), Chomsky (2015), Adger (2013), Rizzi (2013)). This view permits on one side to compare the syntactic structures in a coherent and schematic way, while on the other it concentrates moreover on the hierarchical fashion of the language faculty than on the linear order displayed by the utterances.

1.3 Outline of the thesis

The first year is dedicated to the setting-up of the corpus, with the starting operations to acquire the data and elaborate a coherent way to annotate the texts with a standard schema. During the second year the corpus is planned to grow up for reach a significance level of >15000 words in order to provide quantitative analyses. Third and fourth year will be spent in developing the theoretical analyses and refining the informatic architecture of the project, evolving in a user-friendly and interrogable way to dispense the data. The theoretical outcome constitues the main topic of the research.

Chapter 2 introduces ...

Chapter 3 introduces ...

Chapter 4 introduces ...

Chapter 5 introduces ...

Backmatter

Credits

This project is constituted by files written in Markdown syntax and exported either as a standalone website both as printer-ready product. This is due to the awesome work of the people behind different libraries:

- Pandoc
- Bookdown
- RMarkdown and R environment.

As well, for the computational infrastructure, a lot of open source tools have been used:

- NLTK

For the typographic setting, the print-ready file is composed on LaTeX with the usage of SCRBOOK class and some custom component. I am aware I can't thank everyone on the web about that. By the way, thank you!

About the author

I am a Graduate Researcher involved in a Ph.D. Program in Italian Linguistics at the Department of Romance Studies in the Faculty of Phylosophy at Palacky University in Olomouc, Czech Republic.

My interests span across digital humanities, syntax theories and computational linguistics.

Feel free to write me at marco.petolicchio@gmail.com or visit marcopetolicchio.com for the detailed contact list.

Bibliography

Adger, D. (2013). *A Syntax of Substance*. Linguistic inquiry monographs. MIT Press.

Chomsky, N. (1995). *The Minimalist Program*. Current studies in linguistics series. MIT Press.

Chomsky, N. (1998). *Minimalist Inquiries: The Framework*. MIT occasional papers in linguistics. MIT Working Papers in Linguistics, MIT, Department of Linguistics.

Chomsky, N. (2013). Problems of projection. *Lingua*, 130:33 – 49. SI: Syntax and cognition: core ideas and results in syntax.

Chomsky, N. (2015). *Problems of projection: Extensions*, volume 223 of *Linguistic Aktuell*, pages 1–16.

Clark, A. (2010). *The Handbook of Computational Linguistics and Natural Language Processing*, volume 1.

Ellis, R. (1994). *The Study of Second Language Acquisition*. Oxford applied linguistics. Oxford University Press.

Fodor, J. (2001). *The Mind Doesn't Work that Way: The Scope and Limits of Computational Psychology*. Bradford book. MIT Press.

Guasti, M. T. (2002). *Language Acquisition: The Growth of Grammar*. The MIT Press.

Hauser, M. D., Chomsky, N., and Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298:1569–1579.

Hawkins, R. (2001). *Second language syntax: A generative introduction*. Wiley-Blackwell.

Kuebler, S. and Zinsmeister, H. (2015). *Corpus Linguistics and Linguistically Annotated Corpora*. Bloomsbury Academic, annotated edition edition.

Kurdi, M. Z. (2016). *Natural Language Processing and Computational Linguistics*.

Rizzi, L. (2013). Introduction: Core computational principles in natural language syntax. *Lingua*, 130(Supplement C):1 – 13. SI: Syntax and cognition: core ideas and results in syntax.

Rothman, J. and Slabakova, R. (2017). The generative approach to sla and its place in modern second language studies. *Studies in Second Language Acquisition*, page 1–26.

Slabakova, R., Leal, T. L., and Liskin-Gasparro, J. (2014). We have moved on: Current concepts and positions in generative sla. *Applied Linguistics*, 35(5):601–606.

Sorace, A. (2011). Pinning down the concept of "interface" in bilingualism. *Linguistic approaches to bilingualism*, 1(1):1–33.