Palacký University Olomouc

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

Italian as Second Language in native Czechs and Slovaks

From the development of a learner corpus towards a theoretical investigation

MARCO PETOLICCHIO

Supervisors

DR. FRANCESCO BIANCO, PH.D.
Palacký University · Dept. of Romance Languages, Faculty of Arts

draft : : Friday 31st August, 2018: : 09:39

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 (Clark, 2010; Kuebler and Zinsmeister, 2015; Kurdi, 2016), 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 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 collects into a coherent fashion a set of data that represent some spotted linguistic fact 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.

1.1 Background for the thesis

Corpus Linguistics is a field of approaches developed during the last decades in order to give an empirical support to the investigations on language use and variation. It can offer strong support for analyzing the systematics which underlies the variations among the language use, yielding for empirical and quantitative methods.

> In fact, at one level it can be regarded as primarily a methodological approach:
>
> - it is empirical, analyzing the actual patterns of use in natural texts;
> - it utilizes a large and principled collection of natural texts, known as a "corpus", as the basis for analysis;
> - it makes extensive use of computers for analysis, using both automatic and interactive techniques;
> - it depends on both quantitative and qualitative analytical techniques (Biber et al., 1998)

The main tenets of such a discipline still permit to obtain different level of information starting from the texts and their annotations, to result in a general picture of the language variation. A part of this is due to a widespan documentation which overpasses the recognized linguistic theories - under the *corpus-driven* approach. On the other, the *corpus-based* approach permits to ground the hypothesis on a real actual set of data constitutes by language use in an empirically based way.

1.1.1 Corpus-based approach: motivations for the thesis

While a strong opposition between the way to approach the corpora can be fairly molded during the actual analysis of the data in a softer manner, it can be useful to stand up and recognize those models to threat linguistic data as a two different standpoints to keep in mind for the different purposes they grow on:

- **Corpus-based**
  When a general theory on some linguistic fact is tested against a corpus in order to verify the hypotheses. This kind of approach is more *deductive*, while it goes top-down, proceeding from a general statement (the theory) towards a specific environment (the corpus).

- **Corpus-driven**
  Corpus-driven approach tends to proceed from the analysis of the partial specific pieces (the corpus), in order to result into a general picture (the theory). This method is more *inductive*, going bottom-up.

Different views were proposed to face or embrace the corpora in language studies amongst the scholars. The first one is a well-known citation by Noam Chomsky, which substantially regrets any importance to corpora for a theory-oriented language modeling:

> Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description would be no more than a mere list. (Chomsky 1962, *A transformational approach to syntax* in Tognini-Bonelli, 2001)

On the other hand, Charles Fillmore recognizes a structural place to corpora usage into language reflection:

> I have two main observations to make. The first is that I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore; all that I have seen are inadequate. The second observation is that every corpus that I've had a chance to examine, however small, has taught me facts that I couldn't imagine finding out about in any other way. (Fillmore, 1992)

As in Fillmore's quotation, it appears that the distinction between deductive and inductive method cannot be really disentangled in some part of the research planning, moreover in the case when the one which is developing a corpus is the same that is going to write an analysis based on: a simple scan of the data can yields for a purpose of a general theory which needs to be refined on the real data in a more euristic manner. In this sense, while a *corpus-based* approach aims to generalize a picture *before* than the actual recognition of the data and the dataset takes place, it can be possible to softener a bit this difference amongst these models keeping in mind the perspective of corpus-developing related issues.

In the subsequent parts of the thesis I will try to show how the way to develop a linguistic corpus has a certain degree of influence for the successive part of research activities, and how a purely *corpus-based* method could not be apply if the research is conducted by the same person which started to collect the data.

1.1.2 Learners corpora of Italian L2: an overview

In this section I am going to summarize the most representative Italian L2 learner corpora available online, including Czech-IT, which I have co-founded since July, 2017. I will present all the relevant information and discuss the central topics of the project in a dedicate part of the thesis, while for now I list the most evaluable corpora for studying Italian as 2nd language:

- **GranVALICO** and **VALICO** (Barbera, 2003)
  Learner corpora provided by Turin University. They represent the most valuable sources of Italian L2 corpora. They are composed by written texts composed by the students which have the assignment to describe the vignettes provided by the teachers. The corpora are accessible online with an advanced search that permits to filter the data by different parameters (e.g. learners' L1 and education, assignments etc.).

- **MERLIN** (Abel, 2014)
  The MERLIN Corpus represents a wide-range multilingual documented resource which collects 2.286 texts written by learners of Czech, Italian and German. Started in 2012, the main objective is to show the different levels of acquiring languages by the usage of written texts, relying on the CEFR level schema on L2 acquisition. The Italian-L2 subcorpus contains 813 texts.

- **LIPS** (Vedovelli et al., 2006)
  The corpus contains the transcriptions of more than 2000 audio files by CILS - Certificazione di Italiano come Lingua Straniera (CILS) at the Università per Stranieri of Siena between the years 1993–2006. With more than 700k of words divided in *monologues* and *dialogues* between the candidate and the examiner, it represents one of the biggest corpora of Italian L2. The corpus is POS annotated using the tool Treetagger (Schmid, 1994).

- **Czech-IT** (Petolicchio and Bolpagni, 2017)
  The Czech-IT corpus contains chat messages, emails, coversations, surveys and assignments by more than 70 Czech and Slovak learners of Italian language. Started in 2017, it is fully accessible online while the data acquisition continues. The whole dataset is fully interrogable by an interactive interface and released with a Creative Commons license; POS and automatic tagging are in tune.

- **Corpus Italiano scritto L2** (Voghera and Turco, 2010)
  The corpus retains 227 written texts by undergraduate students of different native languages, which study Italian as a foreign language for their courses at the University of Greenwich. Learners' L1 are: albanian, bosniac, chinese, french, greek, english, norwegian, portuguese, spanish, tigrinya.
  The type of texts are: *descriptive*, *narrative* and *argumental*. The texts are syntactically annotated and the tagset is available in xml format.

Table 1.1: Size of Italian L2 Corpora

Corpus	L1	Texts	Tokens	Lemma	Years
GranVALICO	Various	4778	784217	13057	2002–2007
VALICO	Various	2502	382098	6935	
LIPS	Various	2198	> 700000		1993–2006
MERLIN	Various	813			2012–?
Czech-IT	cs,sk	316	17064		2017–present
Corpus Italiano Scritto L2	Various	227	22931		2010?

1.2 Objectives of the thesis

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

1. **Methodological objectives**
   To address the decisions 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

While usually seen as a sussidary tool for linguistic investigations, corpus linguistics can be regarded with a certain degree of indipendence by such aims (Sinclair, 2005; Sinclair and Carter, 2004), and involves highly specialized sectors for what concerns the planning, the mantaining, the design and the scalability of the corpora.

The Czech-IT corpus is composed by different kind of texts in order to exhibit the variation in language use across different communicative situations:

- Email subcorpus for the (quasi-) bureaucratic and academic language;
- SMS and other direct platforms for textual messaging for informal situations;
- Spoken discourse analysis for spontaneous modality;
- Online surveys created for obtaining auto-evaluation by learners about their acquisition: the tests are made by a certain amount of questions and tiny writing samples.

75 are the learners inserted in the corpus. Informations about the learners concern the education level, the age group, the level of their italian knowledge, and other known languages - while their real identities are preserved by the assignment of an alpha-numeric ID.

1.2.2 Empirical objectives

Amongst many scholar the role of the native language (L1) has been raised as a factor of possible conditionation in the way 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 intermediate steps of the acquisitional path defined with the term *interlanguage* (Selinker, 1972), that we can refer as to **Interlanguage Hypothesis** (IlH). Different from this hypothesis –which recognizes a central place to the native language in the acquisitional path– is the **Monitor Model** (Krashen, 1981), a multi-focal perspective on language acquisition where different factors are described as involved in the process and where the L1 could not represent that conditionation.

Since the last 20 years, a considerable part of linguistic activity is involved in developing some sort of models to describe 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; Hawkins, 2001; Rothman and Slabakova, 2017; Sorace, 2011). In this sense appears that the adoption of a general picture in which analysing the variation in grammar into a *parametric* model (Chomsky, 1995) can be suitable for long-standing researches on SLA and interlanguage.

The dataset used in this thesis aims to display either the different linguistic outcomes in a wide range of communicative situations by the same learner, both than a sociolinguistic grained analysis where the variety of educational or age range can show different linguistic behaviors in the range of learners' variety.

1.2.3 Theoretical objectives

From a theoretical viewpoint, the research is inserted in the current theories that rely on the hierarchical functioning of the language faculty, for which the variation among languages are reconducted to a parametrizing of choice amongst the languages (Adger, 2013; Chomsky, 1995, 1998, 2013, 2015; Rizzi, 2013), which are structurally constant, despite of the wideness of the linguistic variation:

> We are concerned, then, with states of the language faculty, which we understand to be some array of cognitive traits and capacities, a particular component of the human mind/brain. The language faculty has an initial state, genetically determined; in the normal course of development it passes through a series of states in early childhood, reaching a relatively stable steady state that undergoes little subsequent change, apart from the lexicon. To a good first approximation, the initial state appears to be uniform for the species. (Chomsky, 1995)

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 (Kayne, 1994; Moro, 2000). In this perspective is generally assumed that the hierarchical phrase structure plays a central role in syntactic computation, while the *flattering* of such structures into a mono-dimensional workspace is a matter of externalization constraints and interface conditions (e.g. the need to give an ordered array where every item of the sentence is present at one time in order to be spelled out). I will summarize this in a representational way with the usual tree-diagram in Fig.1.1.

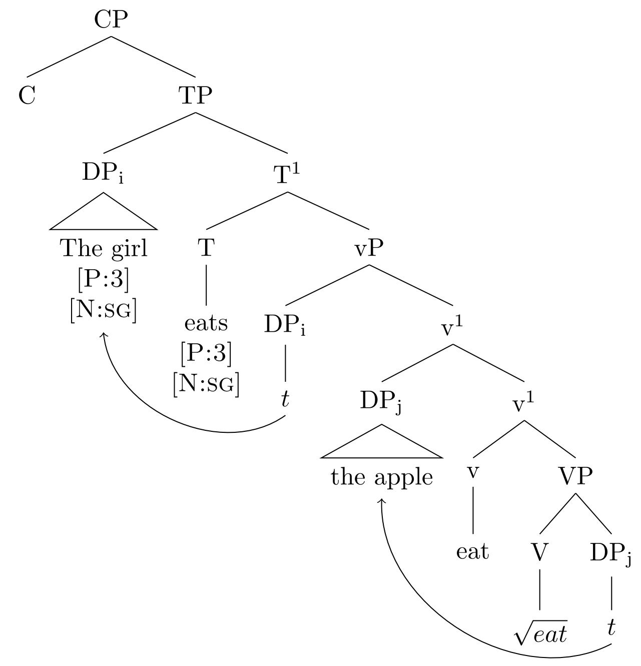


Figure 1.1: Structural representation of a simple sentence

Given this way to proceed, that assures a coherent framework to compare languages in a parametric way, the main theoretical question addressed here concerns the relevance and the potential usage of this perspective in the analysis of a dynamic system as during the acquisitional path and the strategies raised up by learners during the various steps in the interlanguage.

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

1 Colophon

This document is typeset with LaTeX using a custom template based on KOMA-script SCRBOOK class. The layout is based on a standard A4 paper (210 x 297mm), with 40mm margins and 10mm of binding offset.

The typesetting software used the XeTeX engine and the text is set in the open source IBM Plex font family – in Serif, Sans Serif and Monospace variants.

2 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

3 About the author

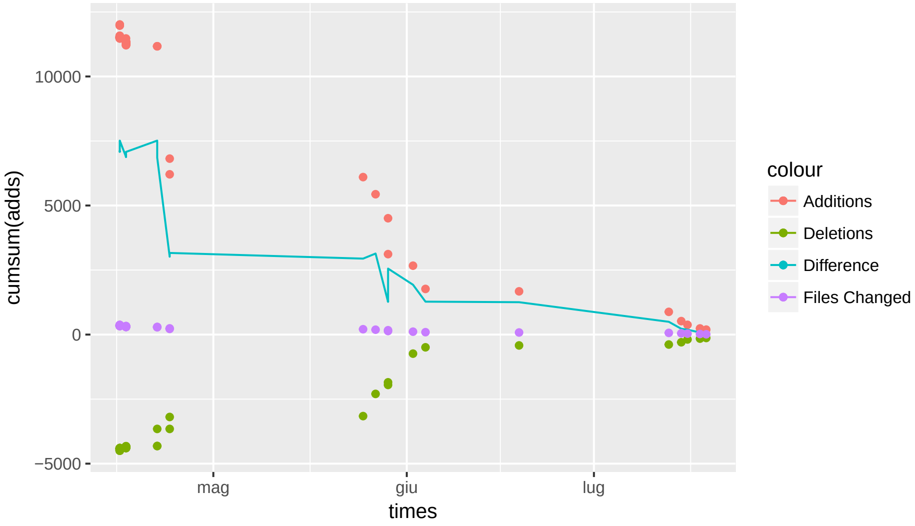
I am a Graduate Researcher involved in a Ph.D. Program in Italian Linguistics at the Department of Romance Studies in the Faculty of Philosophy at Palacky University in Olomouc, Czech Republic.

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.

4 Progress in the repository

This graph represents the addition and deletion amount in the files of the project in function of time.



Bibliography

Abel, A. (2014). A trilingual learner corpus illustrating european reference levels. *RiCOGNIZIONI. Rivista di Lingue e Letterature straniere e Culture moderne*, 1(2):111–126.

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

Barbera, M. e. a. (2003). Valico: Varietà apprendimento lingua italiana corpus online.

Biber, D., Conrad, S., Reppen, R., and University, C. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge Approaches to Linguistics. Cambridge University 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.

Fillmore, C. J. (1992). Corpus linguistics vs. computer-aided armchair linguistics. In *Directions in Corpus Linguistics: Proceedings from a 1991 Nobel Symposium on Corpus Linguistics*, pages 35–66, Stockholm. Mouton de Gruyter, Mouton de Gruyter.

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.

Kayne, R. S. (1994). *The Antisymmetry of Syntax*. Linguistic inquiry monographs. MIT Press.

Krashen, S. (1981). *Second language acquisition and second language learning*. Language teaching methodology series. Pergamon Press.

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

Kurdi, M. (2016). *Natural Language Processing and Computational Linguistics: Speech, Morphology and Syntax*. Wiley-ISTE, 1 edition.

Moro, A. (2000). Dynamic antisymmetry: Movement as a symmetry-breaking phenomenon. *Studia Linguistica*, 51(1):50–76.

Petolicchio, M. and Bolpagni, M. (2017). Czech-IT! - Linguistic corpus of native Czech learners acquiring Italian language.

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.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, 10(1–4):209–232.

Sinclair, J. M. (2005). Corpus and text - basic principles. In Wynne, M., editor, *Developing Linguistic Corpora: a Guide to Good Practice*, chapter 1, pages 1–16. Oxbow Books.

Sinclair, J. M. and Carter, R. (2004). *Trust the Text: Language, Corpus and Discourse*. Taylor & Francis.

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.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Studies in Corpus Linguistics. John Benjamins Publishing Company.

Vedovelli, M., Pallassini, A., Machetti, S., Barni, M., Bagna, C., Pieroni, S., and Gallina, F. (1993-2006). Corpus lips.

Voghera, M. and Turco, G. (2010). *From text to lexicon: the annotation of pre-target structures in an Italian learner corpus*, pages 141–173. Firenze University Press.