Italian as non-native language in Czech and Slovak learners

From the development of a learner corpus towards a theoretical investigation

MARCO PETOLICCHIO

Supervisors

DR. FRANCESCO BIANCO, PH.D.
Palacký University Olomouc
Department of Romance Languages

PROF. ANTONIO FÁBREGAS
UiT The Artic University of Norway Tromsø
Department of Language and Culture

draft : : Monday 23rd May, 2022: : 00:45

Contents

IV. Conclusion   31

10. Conclusive remarks   33

11. Bibliography   35

List of Tables

List of Figures

Part I.

Outline

Preface

Declaration

Annotation

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.

Anotace

xvi

Acknowledgements

1. Introduction

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 their possible outcomes.

1.1. Corpus–based approach: motivations 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 method-
ological 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 analyt-
  ical 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 re-
sult 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 con-
stitutes by language use in an empirically based way.

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 man-
ner, it can be useful to stand up and recognize those models to threat lin-
guistic 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 lan-
guage 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.

1.2. 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.

r toString(czechit_datasetCountLearners) 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 alphanumeric 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:
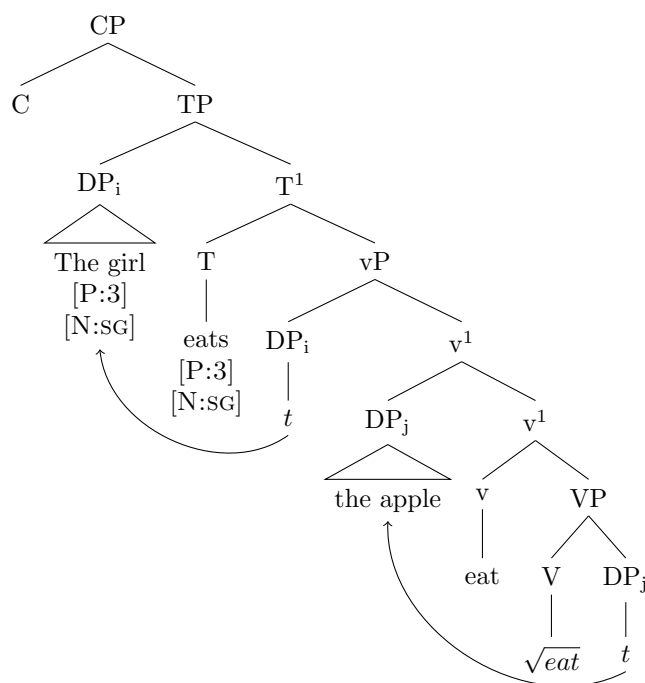
Figure 1.1.: Structural representation of a simple sentence

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.

190   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.

195 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 200 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 …

205   **Chapter 3** introduces …

      Chapter 4 introduces …

      Chapter 5 introduces …

Part II.

Background

2. Evidences and theories in a linguistic research

Pántōn gàr hósa pleíō mérē ékhei kaì mè̄ éstin hoîon sōròs tò pān.

⁵ *The totality is not, as it were, a mere heap, but the whole is something besides the parts.*

– Aristotle, *Metaphysics, Book VIII, 1045a.8–10*

It seems that a certain grade of analysis, theories and empirical collect of data do not follow the same path, while they strike one against the other:
¹⁰ on one side the theoretical generalizations can involve or not a verification of the hypotheses on the actual data that the researcher can handle with, on the other the work around the data collection can still confined without a well-grounded theoretical approach. Nevertheless, while it does not imply *per sé* that a theoretical approach can be regarded as the primary goal
¹⁵ for a scientific approach, also the opposite knows some problems. Defining a neutral way in which data should be collected is not an easy matter, and someone could certainly ask if there is at all some sort of *rawness* or *neutrality* in data itself[1].

---

[1]Cfr. the *Observer Paradox* as stated by William Labov: "To obtain data on the most systematic form of language (the vernacular), we must observe how people speak when they are not being observed" (Labov, 1973, xvii).

2.1. Inductivism and deductivism in linguistics

20 The inductivist approach to research begins with observations in forms of singular events: they borrow a singular context of the place, the time, and the particular situation in which each observation is made, while the analysis of the similarities between such events yields for generalizations.

In order to attempt a logic basis for a research method, Aristotle dis-
25 tinguishes the *induction* (*epagōgē*) as the way which preceeds from the particular to the universal, and the *deduction* (*syllogismos*):

> L'osservazione della somiglianza [...] è utile, poiché siamo con-
> vinti di suscitar l'universale attraverso l'induzione sui casi sin-
> goli, che risultano simili: non è invero facile indurre, quando
30 > non si conoscono le somiglianze degli oggetti. [...] quanto
> si applica eventualmente ad uno degli oggetti simili, si ap-
> plicherà allo stesso modo anche ai rimanenti. Di conseguenza
> [...] quanto si applica eventualmente ad essi si applichi allo
> stesso modo anche all'oggetto della discussione (Aristotele,
35 > 2003, *Topici* 1.18.108b).

A well known example given by Bertrand Russell points out how the inductivism approach can be a fallacy, making an expectation over similar past events and applying these categories on the future:

> Domestic animals expect food when they see the person who
40 > feeds them. We know that all these rather crude expectations
> of uniformity are liable to be misleading. The man who has
> fed the chicken every day throughout its life at last wrings
> its neck instead, showing that more refined views as to the
> uniformity of nature would have been useful to the chicken
45 > (Russell, 2008).

2.1.1. Induction and empiricism

The inductive method proceeds bottom-up from a particular event to a generalization of similar events into an uniform class of items, commoned up by the property to display some analogies into their core components.
50 In this sense, the inductive way focuses on single, individual phenomena as the starting point, collecting these into subsets of similarities in order to attempt a rationale hypothesis which can explain these similarities.

Basing from evidences as the starting point for critical investigation, an inductive reasoning proceeds towards the elaboration of a general rule that can expain the behaviour of different events in a similar class. For achieving such kind of generalization, inductive hypotheses rely on the principle which falls under the definition "uniformity of nature"

2.1.2. Deduction

2.1.3. The role of empirical data

2.2. A theoretic framework to analyze the data

3. An overview of the second language acquisition

3.1. L1 vs. non-native language

3.2. Outline of different approaches about SLA

3.2.1. Contrastive approach

3.2.2. Interlanguage Hypothesis

3.2.3. Generative SLA

4. The generative framework and SLA

4.1. Narrow syntax and interfaces

4.1.1. Nature and culture in language faculty

5 ### 4.1.2. The modular architecture of the linguistic information

4.2. Core operations in syntax

4.2.1. The computational ability of mind

4.2.2. Phrasehood and workspace

4.2.3. Two types of Merge

10 ### 4.2.4. Agree, Labeling, Locality

4.3. Parameters and learnability

4.3.1. Language variation and uniformity

4.3.2. Acquisition and language change

5. The corpora

5.1. A comparison across the corpora of Italian L2

5.1.1. Description and size

5.1.2. Accessibility

5.2. Mono-lingual corpora

5.2.1. Czech and Slovak corpora

5.2.2. Italian corpora

5.3. Limits and needs

5.4. The relevance of corpora

5.4.1. Corpora for the study of language

5.4.2. Corpora in machine-aided tasks

Part III.

Application

6. The dataset(s)

6.1. The corpus "Czech–IT"

6.1.1. Description and size

6.1.2. Planning the corpus

6.1.2.1. Type of texts and data

6.1.2.2. Why a learner–based corpus?

6.1.2.3. Data–retrieving

6.1.3. Interacting with the corpus

6.1.3.1. The user interface

6.1.3.2. The computational tools

6.1.3.3. NLP tasks and examples

6.2. Additional materials

6.2.1. Online surveys and questionaires

6.2.2. Auto–corrective forms

7. Case study I: The syntax domain of DPs

7.1. The theoretical status of DP and NP

7.1.1. The DP–over–NP and NP–over–DP debate

7.1.1.1. The early–days view

7.1.1.2. Abney's proposal

7.1.2. Consequencies for the theory

7.1.2.1. Parameters and locality effects

7.2. Determiners in Ces/Slk and Ita

7.2.1. Typological differences

7.2.2. Theoretical consequencies

7.2.3. Learning consequencies

7.3. The acquisition of Determiners

7.3.1. Patterns and regularities

7.3.1.1. Errors

7.3.1.2. Choices

7.3.2. The acquisitional path

7.3.2.1. Language level and acquisition

7.3.2.2. Other important facts

7.3.3. Determiners with PPs

8. Case Study II: Adverbs

8.1. Adverbs in syntax

8.1.1. The role of Adverbs

8.1.2. Adverbs and functional projections

8.2. Adverbs and information structure

8.2.1. Prominence

8.2.2. Prosodic effects in Italian

8.3. Acquisition

8.3.1. Fronting phenomena

8.3.2. Scrambling

8.3.3. Learning strategies

8.3.4. Teaching strategies

8.4. Conclusion

9. Case study III: Verbs

9.1. Aspect

9.1.1. Verbal aspect in CES/SLK

9.1.2. Verbal aspect in ITA

9.1.3. Consequencies and data analysis

9.2. Auxiliary

9.2.1. Typology of Auxiliaries

9.2.2. AUX selection

9.3. Conditional clauses

9.3.1. Choice of TAM

9.4. Conclusion

Part IV.

Conclusion

10. Conclusive remarks

11. Bibliography

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

Aristotele (2003). *Organon / Aristotele ; Traduzione di Giorgio Colli.* Adelphi (Series). Adelphi.

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.

Labov, W. (1973). *Language in the Inner City: Studies in the Black English Vernacular*. Conduct and Communication. University of Pennsylvania Press.

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

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.

Russell, B. (2008). *The problems of philosophy*. Arc Manor.

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.

60   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.

65   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.