Palacký University Olomouc

Ph.D. Research in Italian LInguistics
Dept of Romance Languages, Faculty of Arts

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

MARCO PETOLICCHIO

Supervisors

SUPERVISOR 1 Dept. of Romance Languages
Univerzita Palackeho v Olomouci

SUPERVISOR 2 Dept. of Language and Cultures
University of Xyz

2018 : : DRAFT

Contents

1

Preface

1.1 Abstract

The aim of the present PhD research is to analyse in a coherent way the learning path displayed by Czech learners acquiring the Italian Language, basing on the evidences which result from the independent linguistic corpus Czech-IT!. This study grows up from the researches in Second Language Acquisition and wishes to retrieve data from applied fields turning them into theoretic and formal questions in general linguistics.

1.2 Keywords

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

2

Introduction

The main idea under this PhD proposal moves across the idea of an empirically-grounded research to investigate the strategies of the learners during the acquisition of the second languages. My task is twofold: on one side this provide for the developing of a theoretically-grounded framework to research in the fields of Second Language Acquisition (SLA), while on the other this necessitate to realize a linguistic corpus which collect in a coherent fashion a wide set of data and facts in order to give a transparent documentation of the learning path. At the current date, the corpus is actively mantained online under the name "Czech-IT!" and counts 220 items by more than 50 learners (Petolicchio and Bolpagni, 2017). In this order of ideas, the project is focused around the analysis of Czech and Slovak learners which acquire the Italian language as L2.

Czech (CZ) is a West Slavic language of the Indoeuropean (IE) family (Beekes and Cor de Vaan, Hammarström et al. (2017)), widespreadly known for its morphological complexity with a very rich inflectional and derivational system in the word formation. Italian (IT) is a Romance language of the IE family, strictly related to Latin[1] , which exhibits a wide range of variation in dialects, regional languages and specialistic styles. Commercial and cultural links between the Czech Republic and Italy are effective and deep, and the studying of Italian language amongst native Czech learners is a remarkable fact. This kind of research is intimately twofold in nature and embraces differences approaches and discipline: the developing of a linguistic corpus (digital humanities, corpus and computational linguistics) and a theoretically-oriented analysis (general and theoretical linguistics, studies on SLA and interlanguage).

2.1 The empirical ground

By a strictly linguistic analysis, CZ and IT know a set of phenomena which diverges a while and permits to afford for a comparative investigation, focused in the errors displayed during the acquisitional path: the absence of the Determiner phrase (DP) and the rich morphological declension in the CZ noun syntax, where IT does not exhibit this kind of morphological complexity and does not permit the delection of the Determiners in such contexts (Bianchi, 1992, Longobardi (1994)), which gives examples of omission or ipercorrected forms or examples due to the L1 habits. As a framework-free corpus with no theoretical issues, Czech-IT! aims to be a resource either for speculative, data-based studies, as well than for empirically based L2 acquisition teaching processes. The

[1] For the sake of clarity: I am going to refer to Italian in its standard variety. Cfr. (Berruto, 2012, D'Achille (2003)) for a general overview of the contemporary IT.

project and the datasets are licensed under a Creative Commons Attribution 4.0 International License, for which it represents an open source and an open data project, in the universe of the *open knowledge* works. This represent also a tempt to gain indipendence from data to the analysis of the data itself, creating a linguistic corpus threated in a computational manner (Abney, 1997, Kuebler and Zinsmeister (2015), Schmid (1994), Bird et al. (2009), Kurdi, Clark), in line with other wll established learner corpora on Italian L2 (Barbera, 003 , Vedovelli et al. (2006)).

2.1.1 Data

In order to define a wider range of linguistic situations, there are different kind of linguistic productions in the corpus:

- an 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;
- some online surveys created for obtaining auto valutation by learners about their acquisition: the tests are made by a certain amount of questions and tiny writing samples.

The data are inserted at first in textual forms, where are stored the relevant informations about the learner, the date and notes of the revisor, while the textual content of each relevant example is processed towards the usage of automatic machinery, which yields syntactical, morphological and part of speech tagging annotations, relevant for quantitative and statistical outcomes. Currently, a primary dataset which contains the items is linked to other two datasheets, one relative to learners and the other for manual categorization of linguistic phenomena and automatic treatment of the texts, as for tokenization, lemmatization and POS-tagging procedures. Separating the raw data from the annotation scheme seems to be a feasible way to retain data in a wide output directions, e.g. for data-visualization outcomes, and can be effectively implemented towards the successive implementation without the necessity to rethink the overall platform. Also, it permits to data to be independent from contingent purposes and easily accessible and used by the whole community of scholars, researchers, and interested users. It could be usable for data-driven approaches to learning second language and for theoretically-oriented researches on interlanguage, syntactic variation and computational linguistics.

2.1.1.1 The learners

Currently, the number of the learners inserted in the dataset is 51: they are in the most part native-Czech learners but a small part of Slovak is represented. The level of education testimoniates a representative range of different kinds of acquisition paths, as well than the different ages of the learners.

2.1.1.2 The texts

At the present date, there are 220 entries in the corpus, which reveal a large range of different communicative situations, from spontaneous messages as chat, spoken con-

versations, and email towards written homeworks for retrieve targeted hypotheses on the learning way.

2.2 The theoretical framework

2.2.1 Variation in grammar

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 (2011), 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.

2.2.2 Comparative analyses and the role of the interlanguage

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) 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 gramar in 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.

2.3 Models and methods

[compLing]

A similar project aims to show an affordable platform for linguistic data-based researches.

The advantage of a such type of way to proceed is twofold: on one side it permits a clear separation between the data and the investigations of the data itself, while it offers a theoretically-agnostic way to collect the data which can be used in a widespread linguistic researches and model, not confined to some theoretically-oriented approaches. Along this path, such a kind of corpora can be suited either in academic enterprises than for private and corporate initiatives, as well in teaching models in the SLA field, oriented

towards an empirically-grounded perspective on error and interlanguage analyses. The usage of computational and digital architecture (Clark, Kurdi, Kuebler and Zinsmeister (2015)) represents a standpoint in the current path of linguistic studies, resulting in a highly interdisciplinary model to researching. The theoretical model established relies on generative studies to language, applied to a new and insightful field of research as the SLA studies. It permits to obtain an empirically-grounded and theoretically coherent perspective on some pattern displayed during the learning path.

2.4 Structure of the project

2.4.1 Roadmap

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.

2.4.2 Thesis structure

here at last.

3

Backmatter

3.1 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!

3.2 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.

4

Bibliography

Abney, S. (1997). *Part-of-Speech Tagging and Partial Parsing*, pages 118–136. Springer Netherlands, Dordrecht.

Adger, D. (2011). Label and structures.

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

Beekes, R. S. P. and Cor de Vaan, M. A. *Comparative Indo-European Linguistics*, volume 1. John Benjamins Publishing.

Berruto, G. (2012). *Sociolinguistica dell'italiano contemporaneo*. Manuali universitari: Linguistica. Carocci.

Bianchi, V. (1992). Sulla struttura funzionale del sintagma nominale italiano. *Rivista di Grammatica Generativa*, 17:105–127.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

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. *The Handbook of Computational Linguistics and Natural Language Processing*, volume 1.

D'Achille, P. (2003). *L'italiano contemporaneo*. Itinerari: Linguistica. Il Mulino.

Ellis, R. *The Study of Second Language Acquisition*, volume 1.

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.

Hammarström, H., Forkel, R., and Haspelmath, M. (2017). clld/glottolog: Glottolog database 3.0.

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. *Natural Language Processing and Computational Linguistics*.

Longobardi, G. (1994). Reference and proper names: a theory of n-movement in syntax and logical form. *Linguistic Inquiry*, pages 609–665.

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*.

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

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

List of Tables

List of Figures

Timestamp

Printed on: April 24, 2018