On the syntax of DP in Italian as non-native language

A case study on Czech and Slovak learners

Marco Petolicchio

This study aims to investigate over the internal syntax of Noun phrases in non-native Italian by Czech and Slovak learners. The typological difference in those languages are used to reconnect with the generative models on language acquisition, while the linguistic data is threated quantitatively to show the gap between L2 production and the standard behaviour in mono-lingual corpora.

Keywords: Corpus linguistics, Determiner phrase, Italian L2, Second language acquisition, Syntax

Contents

1 Introduction

Czech (ces) and Slovak (slk) are languages of the Slavic branch in the Indo-european family. Alongside a certain morphological complexity in noun declension systems, these languages –except for Bulgarian (3b) and Macedonian (Dryer, 2013)– don't show an overt realization of the Determiner position inside the noun phrase (1) (Harkins, 1953). Conversly, Italian (ita) and the other romance languages explicit that position as a default behaviour, usually with a free morpheme preceding the noun (2) or by cliticization of the definite article (3a):

(1) Articleless

a. ces (Veselovská, 2014, 14)
Chlapec/Marie/Ona/každý miluje ryby/ {své rodiče}.
Boy/Marie/She/Everyone Love.3sg Fish/ {POSS parent}
"SUBJ loves [the] fish/ his parents"

b. slk (Kamenárová, 2007, 113)
Večer čítam knihy, píšem referáty...
Evening Read.1SG Book.PL Write.1SG Paper.PL
"In the evening I read [the] books, I write (school) papers ..."

(2) Proclitic

a. ita (Bianco, 2017, 60)
Il terremoto ha distrutto
ART.DEF Earthquake AUX.3sg Destroy.PTCP.PST
la città.
ART.DEF City
"The earthquake destroyed the city"

b. fro (Dufournet and Lecoy, 2008, 3261)
La dame estoit devant la sale.
ART.DEF Girl Be.3sg ADV ART.DEF Room
"The *dame* was in front to the room"

(3) Enclitic

a. ron (Cojocaru, 2003, 45)
Prieten=ul meu este aici.
Friend=ART.DEF POSS Be.PRES.3sg Here
"My friend is here"

b. bul (Leafgren, 2011, 37)
Къде е книга=та ми?
Where Be.3SG Book=ART.DEF POSS.1SG
"Where is my book?"

The general idea of this paper is to address the question of how linguistic structures which are not overtly marked in L1 can be accessed during the acquisition of a target language which show them. While doing this can be either both purely speculative and meanwhile grounded on actual data, I will show how the usage of a targeted linguistic corpora which can be useful to concentrate the main hypotheses into narrower facts. The language under observation are manifold: on one side Czech (ces) and Slovak (slk) as native languages–with no position for the articles–with Italian (ita) on the other side as the target language.

The section 2 provides a theoretical discussion on the top of different theories inside the generative framework (Chomsky, 1995) on the status of DP and NP. The section 3 is twofold: firstly I present the methods used into the current study in terms of *reproducibility* of research and an analysis of the expected results; while the second subsection is built upon a case study made test hypotheses about the categorial differences of DPs during the acquisition of ita by ces and slk native speakers involved in the test. A summary conclusion (Section 4) closes the paper.

2 Theoretical background

2.1 The generative view on language

From a generative-oriented point of view, the human language is a computational procedure which relies on a hierarchical organization of structures, and language variations are connected to a parametrizing of choice among those structures (Adger, 2013; Chomsky, 1995, 1998, 2013, 2015; Rizzi, 2013):

> 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)

From this perspective, the possibility of comparation is offered either by different languages or among different states of language acquisition: structures can be compared and analysed into a coherent grid in order to perform further analyses and reveal similarities and differences in the parametrizing of syntax.

2.2 The role and the study of interlanguage

For many scholars the role of the native language (L1) carries possible conditioning for the way which the target language (L2) is acquired during the path to learning: 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, which is the hypothesis of *interlanguage* addressed by Selinker (Selinker, 1972). One of the main areas of research in Generative Studies on Second Language Acquisition (GenSLA) regards investigating how linguistic structures can be accessed in L2 and how the transitional stages of acquisition work into the learning *continuum* (Rothman and Slabakova, 2017).

During the last 20 years, a considerable part of linguistic activity has been involved in developing some sort of models to describe how the faculty of language works, through its biological (Hauser et al., 2002), computational (Fodor, 2001) and cognitive components in a highly interdisciplinary environment. SLA is a fertile field, which relies on the comparative and contrastive analyses of linguistic phenomena, either both from an applied view (Ellis, 1994) or by theoretically grounded perspective focused on GenSLA (Guasti, 2002; Hawkins, 2001; Rothman and Slabakova, 2017; Sorace, 2011).

2.3 The position of DP and NP

There are striking differences between languages that display an overt D position and those that do not in respect to the syntactic behaviour of NP, as such as Left Branch Extraction allowing, scrambling or adjective extraction. Those properties are summarized in the table below (in Salzmann, 2018, from Bošković (2009)):

Table 1: Typology of Overt D vs. Covert D languages

	Overt D	Covert D
allow adj extraction from NP	no	yes
allow LBE	no	yes
allow Neg-raising	yes	no
allow scrambling	no	yes
allow the majority superlative reading	yes	no
allow trans. nominals with 2 non-lex. genitives	yes	no
can be polysynthetic	no	yes
island sensitivity in head-internal relatives	no	yes
superiority effect in wh-mvt	yes	no

Since Abney's seminal work (Abney, 1987) there two hypotheses have been established to represent this structure: (i) NP-over-DP, for which the DP is at the edge of NP as specifier; (ii) DP-over-NP, where the DP dominates the NP:

Figure 1: Phrase structure in NP-over-DP vs. DP-over-NP Hypotheses

Following this approach (Abney, 1987), in generative studies it is assumed that nouns project a higher functional category–Determiners–with their phrase DP, wherein a restricted class of items can be (articles, quantifiers etc.). The *DP Hypothesis* aims to reconnect the internal geometry of the Determiner Phrase to the

generalized phrase structure of complex elements (Bernstein, 2008; Zamparelli, 1995) for which, independent of the actual presence of DP elements on a morphological level, the functional category of DP is still in the derivation.

3 The analysis of data

3.1 The datasets

For the analysis of `ita` as a non-native language for `ces` and `slk` learners, the 3 corpora below have been subsetted and collected into a dataframe (henceforth "*collection*"):

- **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 vignettes provided by the teachers. The corpora are accessible online with an advanced search that permits filtering the data along different parameters (e.g. learners' L1 and education, assignments etc.).

- **MERLIN** (Abel, 2014)
  The MERLIN Corpus is a wide-ranging 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 language acquisition using written texts, relying on the CEFR level schema on L2 acquisition. The Italian-L2 subcorpus contains 813 texts.

- **Czech-IT** (Petolicchio and Bolpagni, 2017)
  The Czech-IT corpus contains chat messages, emails, conversations, surveys and assignments by more than 70 Czech and Slovak learners of Italian language. Started in 2017, it is fully accessible online.

Additionally, two monolingual L1 corpora have been used for `ita` and `ces`:

	Texts		Tokens	
	ces	slk	ces	slk
Czech-IT	212	74	11129	4440
Merlin	1	0	256	0
Valico	107	17	16250	3316

Table 2: Structure of data in the collection

- **Google nGram Viewer Italian** (Michel et al., 2011)
  With more than 40 billions words with an estimated accuracy rate of 95.6% for POS-tagging and 80.0% for dependency parsing (Lin et al., 2012), the Italian corpus represents a wide collection of data to study monolingual `ita` in written form. Developed at Google, the nGram Viewer presents an interface to deal with those corpora in a standalone way.
- **Syn2010** (Křen et al., 2010)
  Part of the a documentation project of the Czech National Corpus (CNK, *Český Národní Korpus*), SYN2010 is a representive corpus of contemporary Czech writing containing more than 100 million words, which includes texts of fiction (40%), journal articles (27%), and professional literature (33%).

3.2 Methods

The data from the three corpora have been subsett for analysis, including only the data which present `ita` as the target language by `ces` and `slk` learners and merged into a collection which consists of 411 texts and 35391 tokens. The texts in the collection are computationally processed in sequential steps in order to retrieve a comparable basis for data analyses. In the first step, only the relevant pieces were extracted from their original dataset and then they were processed towards the use of the library UDPIPE (Straka and Straková, 2017) in R. The corpora were cleaned by the deletion of non-informative structures (e.g. punctuation marks), and merged (Table 2):

Additionally, mono-lingual data were been analysed for comparison. For the Czech language, the analysis relies on the work of

Veselovská (Veselovská, 2014) based on SYN2010 (Křen et al., 2010). The statistics on the Italian corpus were been provided by the submission of syntactic queries against the Google NGram API (Michel et al., 2011) on Google Books ITA (1500-).

3.3 Results

The data in the collection was computationally processed in order to retrieve quantitative information about the overall distribution of the syntactic phrase, specifically elicited in the environments that present a Noun element. These clusters were been analysed by their condition in the environment, giving the possibility to compare the distribution of single tags in the antecedent position of a noun or in the subsequent position. A general POS tagging pipeline was established with the usage of the free library UDPIPE for R. While those tools reach far beyond 90% accuracy in POS-tagging for mono-lingual corpora, it was also determined that learner-based corpora posit a challenge for automated tasks.

The chart below (Figure 2) presents the occurrence of bigrams clustering with N, extracted from the collection of corpora.

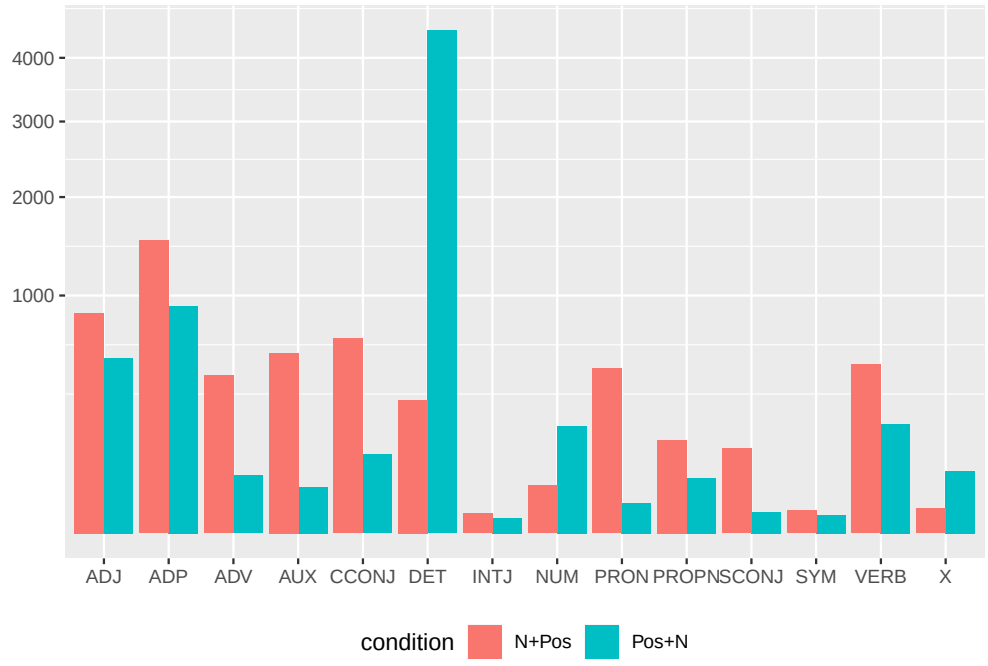
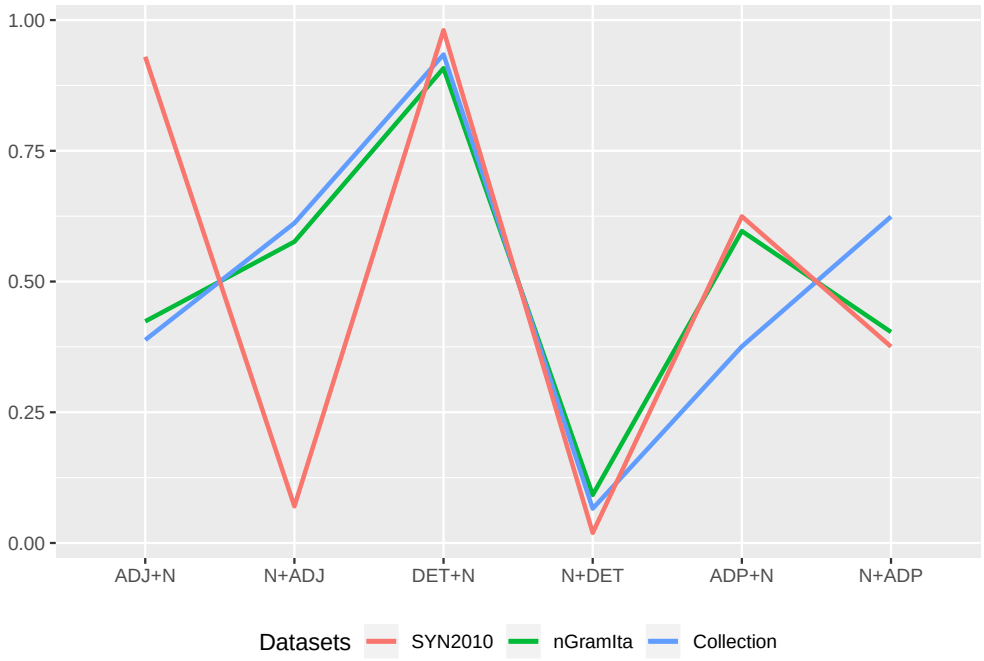


Figure 2: Distribution of 2-grams with N in Collection

Additionally, SYN2010 and Google NGram data were compared to the collection.
While the analysis of the SYN2010 corpus relied on the study by Veselovská (Veselovská, 2014) except for the statistics of ADP+N/N+ADP clusters, results yielded by the diachronic analysis on Google nGram were processed by their central tendency, calculated by the arithmetic mean (AM). With a data set containing the values `a_1,a_2,…,a_n` thus the arithmetic mean is defined by the formula:

$$AM = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + ... + a_n}{n} \tag{1}$$

The amounts of the clusters in the dataset were weighted by their absolute distribution and refactored to a value equal to 1. The plot below (Figure **??**) shows the close comparison of these clusters in the dataset[1].



The presented data does not deal directly with the problem of evaluating the automated tasks involved in the process, and had to be cleared out indicating that there can be some weaknesses

[1]The category of DET in SYN2010 corresponds to DEM, Q, PRON (not POSS) (Veselovská, 2014, 20).

in the usage of mono-lingual trained processors in learner corpora. Moreover, while these can be implemented towards an application in multilingual contexts, it appears that an overall application of this process cannot be done flawlessly since the complexity of the language acquisition path. Also, the differing sizes of the corpora involved in the analysis, which display different magnitude of breadth (nGram is over 40 billion of tokens, the collection 39 thousand), and plays a role in the analysis of the data itself.

In this sense it appears that a certain tendency can be seen in the examples discussed above. While it shows a certain uniformity in the absolute value of DET-N clusters in the dataset, a major shift arises in the position of the adjectives, which can be traced by the typological differences among the languages under analysis:

(4) a. `ces`
To *červené* *auto.*
DEM.NT ADJ.NT NOUN.NT
"That red car"
b. `ita`
Quella *macchina* *rossa.*
DEM.FEM NOUN.FEM ADJ.FEM
"That red car"

Conversely, the position of the *collection* in respect to ADP-N clusters can be due to either some lexical choices present in the texts of that dataset, which can contribute to complexity in the noun phrases as well to some inconsistencies due to the application of these tools, more than to effective syntactic difference in such cases.

4 Conclusion

This study examined the possibility of a data-based comparison across mono-lingual corpora and learner corpora which yielded quantitative information useful for understanding of second language acquisition, specifically in the syntactic domain of the noun phrases in the Italian grammar by Czech and Slovak learners. On one side it reconnects to a generative framework and deals with the problem of the understanding phrase structure in the nominal domain (Abney, 1987, Bernstein (2008), Zamparelli

(1995)) and its place in the study of non-native language acquisition (Rothman and Slabakova, 2017). A computational method was established to deal with different linguistic datasets (Sinclair, 2005) in order to obtain absolute values of the distribution of the select elements, and to identify some tendencies in the linguistic productions of non-native speakers.

In this sense, a diachronic study on such types of learner-based research can shed a light on more fine-grained analyses, specifically to spot forms of *analogy* or *overcorrection* during the learning path of those construction, and it appears an encouraging perspective to follow in the subsequent steps, aware of the necessary interplay of quantitative and qualitative processes in such interdisciplinary models.

Abbreviations

Languages are indicated by the abbreviations provided in the ISO 639-3 format (SIL International, 2009). Morphological gloss styles adher to the widely recognized *Leipzig Glossing Rules* (Comrie et al., 2008), while other abbreviations respect (Boeckx, 2012).

Financial coverage

This work was supported by the grant IGA_FF2018_015 (*Románské literatury a jazyky: tradice, současné tendence a nové perspektivy*) financed by the Ministry of Education, Czech Republic.

References

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.

Abney, S. P. (1987). *The English Noun Phrase in Its Sentential Aspect*. PhD thesis.

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

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

Bernstein, J. B. (2008). *The DP Hypothesis: Identifying Clausal Properties in the Nominal Domain*, chapter 17, pages 536–561. John Wiley and Sons, Ltd.

Bianco, F. (2017). *Breve guida alla sintassi italiana*. Pillole. Linguistica. Cesati.

Boeckx, C. (2012). List of abbreviations and symbols. In Boeckx, C., editor, *The Oxford Handbook of Linguistic Minimalism*, pages xv–xx. Oxford University Press.

Bošković, Z. (2009). More on the no-dp analysis of article-less languages. *Studia Linguistica*, 63:187–203.

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.

Cojocaru, D. (2003). *Romanian Grammar*. Slavic and East European Language Research Center (SEELRC), Duke University.

Comrie, B., Haspelmath, M., and Bickel, B. (2008). The leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses.

Dryer, M. S. (2013). Definite articles. In Dryer, M. S. and Haspelmath, M., editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Dufournet, J. and Lecoy, F. (2008). *Le roman de la rose ou de Guillaume de Dole par Jean Renart*. Champion Classiques. Champion.

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.

Harkins, W. E. (1953). *A Modern Czech Grammar*. King's Crown Press, New York.

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.

Kamenárová, R. (2007). *Krížom-krážom: Slovenčina A1*. Number sv. 1 in Studia Academica Slovaca. Univerzita Komenského.

Křen, M., Bartoň, T., CvrČek, V., Hnátková, M., Jelínek, T., Kocek, J., Novotná, R., PetkeviČ, V., Procházka, P., Schmiedtová, V., and Skoumalová, H. (2010). SYN2010: Balanced corpus of written Czech. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Leafgren, J. (2011). *A Concise Bulgarian Grammar*. SEELRC.

Lin, Y., Michel, J.-B., Aiden, E. L., Orwant, J., Brockman, W., and Petrov, S. (2012). Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 System Demonstrations*, ACL '12, pages 169–174, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., and Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

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.

Salzmann, M. (2018). Revisiting the np vs. dp debate.

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

SIL International (2009). Iso 639-3 - codes for the representation of names of languages.

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.

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

Straka, M. and Straková, J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Veselovská, L. (2014). Universal dp-analysis in articleless language: A case study in czech. In Veselovská, L. and Janebová, M., editors, *Nominal Structures: All in Complex DPs*, Olomouc modern language monographs, pages 12–28. Palacký University, 1 edition.

Zamparelli, R. (1995). *Layers in the Determiner Phrase*. University of Rochester. Department of Linguistics.