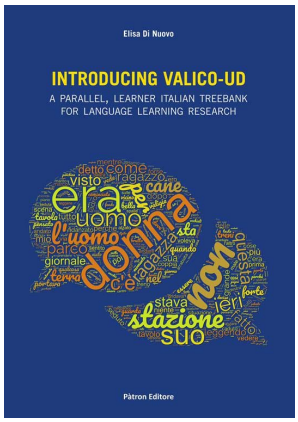

Elisa Di Nuovo,
Introducing Valico-DU.
A Parallel, Learner Italian Treebank
for Language Learning Research
Bologna, Pàtron, 2023, 180 pp.
ISBN: 9788855536110

Bianca Maria De Paolis

Introducing Valico-UD is a monograph that contains the presentation and core results of Elisa Di Nuovo's doctoral project. The work falls within the perspective of computational analysis of learner corpora, which, as indicated by the author herself in the introduction, constitutes a complementary branch to SLA (second language acquisition) studies. The aim of the work is to propose and refine a methodology, along with a set of tools and materials, that integrates and complement studies conducted using typical Second Language Acquisition (SLA) methods (often focusing on psycholinguistic and experimental linguistic analyses, and primarily on oral productions). Additionally, the goal is to incorporate a computational and corpus-driven approach, testing the applicability of tools and formalisms developed for standard varieties of natural language to interlanguage analysis.

The volume begins with an introduction that effectively argues for the necessity of this work (Chapter 1). Firstly, it highlights the potential of quantitative methodology and large datasets to offer robust insights and serve as a model for future studies; additionally, it emphasizes the significance of focusing on Italian as an L2, which is still relatively understudied. In this first section, the research objectives are stated: assess the reliability of error identification with contextual cues, explore agreement among annotators on error presence and normalization, evaluate the impact of explicit target hypotheses on error annotation, examine the adaptability of Universal Dependencies to L2 Italian, analyze the performance loss of parsers on learner language, and investigate using similarity metrics to gauge language development/proficiency.

Chapter 2 features an extensive literature review, that has both the function of arguing for the need of the upcoming work, and to delineate the scope of the research. The chapter collects and describes the work-related contributions, dividing them systematically into three macro topics, which will then be the three key nodes addressed and developed by the research: learner corpora, error annotation and linguistic annotation applied to learner corpora. The conclusion of the chapter also comments on the results of some tasks previously carried out, in which the limitations that emerged, and consequently the gap that the work intends to focus on, are effectively highlighted.

Chapter 3 provides an overview of the VALICO-UD corpus, the foundation of this work. The section begins with a discussion of the selection criteria for creating the corpus and its composition. The corpus holds significance for two main reasons: firstly, the texts are generated through comic strips, enhancing the reliability of reconstructing Target Hypotheses (THs) as lexical choices and semantic frames are constrained by the comic strip context (p.

39). Secondly, extensive metadata accompanies learners' productions, enabling diverse custom queries and the formation of numerous sub-corpora. Following this, the author delves into the principles guiding the development of the parallel normalized version of Learner Sentences, referred to as target hypotheses, showcasing the author's adeptness in balancing a keen awareness of linguistic and acquisitional nuances with the technical and scientific demands necessitated by computational purposes and NLP.

Chapter 4 focuses on error annotation, detailing the methodology, error taxonomy, and statistics regarding error distributions. Additionally, it explores three inter-annotator agreement experiments aimed at validating the annotations. This chapter serves as a crucial aspect of the work, as its primary objective is to provide experimental evidence and quantitative measures to support the tagging system's validity. Successfully achieving this goal, the chapter reinforces the credibility of the entire toolset, encompassing the corpus, tagset and annotations. Through the presentation of the error annotation system and the inter-annotator agreement experiments, in fact, the author solidifies the adequacy of the tagset's design and its elucidation in the guidelines, establishing a robust foundation for processing and operability.

Chapter 5 focuses on linguistic annotation, structured within the Universal Dependencies framework and applied to VALICO, making VALICO-UD a true parallel treebank. This includes segmentation, tokenization, lemmatization, part-of-speech (PoS) tagging, morphological annotation, and dependency annotation. Within this section, comprehensive assessment procedures, including inter-annotator agreement, are detailed. Statistical analysis of the manually-corrected section of the treebank and an incremental evaluation of the model are provided, leading to the generation of the first automatically annotated draft of the resource.

Chapter 6 discusses the quantitative measures employed to analyze the treebank, aimed at assessing the data's quality and comprehending its potential role in computational linguistics. The results presented in this section demonstrate that, despite the tool's primary focus on exploring interlanguage features and not directly training a parser for the target language, it yields an interesting side effect. As the author states, in fact, "it could be useful for conversational systems, which often have to deal with sentences produced by non-native speakers, or have educational purposes" (p. 147). The experiments yield promising results in this regard.

Chapter 7 concludes with a discussion on future perspectives, strengths, and weaknesses of the work. It emphasizes the hope that the resource will garner widespread collective use and contribution to enhance its dissemination and validity. This collective enrichment, facilitated by the input of numerous contributors, is seen as fundamental for solidifying the resource's credibility and utility: "In a non perfectly repeatable and non-deterministic task, the number of annotators is crucial to avoid errors stemming from manual coding and variability of linguistic features" (p. 159).

The contribution of this work is notable for several reasons. It introduces a methodological approach and addresses a significant gap in research on L2 Italian and its combinations with different source languages. Moreover, the work makes a considerable quantity of annotated material accessible, usable and inter-operable, enriching the field.

What sets this work apart is its linguistic sensitivity, a quality that is not to be taken for granted in computational and quantitative studies. The author, in fact, demonstrates a commendable attention to the complexities of SLA, and particularly regarding the treatment of errors. By addressing and analyzing this risk, the work offers valuable methodological and theoretical insights that could inspire future research. Moreover, Di Nuovo's work represents a significant contribution by providing free access to tools and materials, as well as offering valuable methodological reflections; but most of all the work sets a commendable example of interoperability, openness, documentation, and thorough assessment, qualities often standard in computer sciences but still not common enough

in general linguistics and second language studies.

BIANCA MARIA DE PAOLIS • is a PostDoc in Linguistics at Università di Torino. She completed her PhD with joint work at the Laboratorio di fonetica sperimentale 'Arturo Genre' and the Laboratoire Structures formelles du langage (Université Paris 8). Her research focuses on phonetics, prosody, information structure, and cross-linguistic influence in L2 acquisition. She has published on prosodic and syntactic markers of focus in Italian and French, as well as on the phonetic analysis of diatopic and diamesic variation in Italian speech.

E-MAIL • biancamaria.depaolis@unito.it