PaVeDa – Pavia Verbs Database: Challenges and Perspectives

Chiara Zanchi
University of Pavia
chiara.zanchi01@unipv.it

Silvia Luraghi
University of Pavia
silvia.luraghi@unipv.it

Claudia Roberta Combei
University of Pavia
claudiaroberta.combei@unipv.it

Abstract

This paper describes an ongoing endeavor to construct Pavia Verbs Database (PaVeDa) – an open-access typological resource that builds upon previous work on verb argument structure, and in particular the Valency Patterns Leipzig (ValPaL) project (Hartmann et al., 2013). The PaVeDa database features four major innovations as compared to the ValPaL database: (i) it includes data from ancient languages enabling diachronic research; (ii) it expands the language sample to language families that are not represented in the ValPaL; (iii) it is linked to external corpora that are used as sources of usage-based examples of stored patterns; (iv) it introduces a new cross-linguistic layer of annotation for valency patterns which allows for contrastive data visualization.

1 Introduction

In this paper, we introduce a new typological resource, the Pavia Verbs Database (PaVeDa)[1] which is modeled upon the Valency Patterns Leipzig (ValPaL)[2] database by Hartmann et al. (2013), while introducing a number of innovations detailed below.

PaVeDa has been designed at the University of Pavia and is currently being constructed. The final version of the resource will allow to contrastively and simultaneously display valency patterns and alternations for ancient and modern languages.

2 State of the art and current challenges

In the last decades, researchers have observed that cross-linguistically verb classes show similar patterns as to their valency patterns and possible alternations.

This observation led scholars to study the extent of possible variation across verb classes emerging from languages of different genetic and areal affiliation in order to discover general tendencies. Additionally, typologists have striven to design and build foundational toolkits, data-sets, and other resources useful to ease and systematize research on verb classes.

The open-access ValPaL database which is the output of the 2009-2013 project “Valency classes in the languages of the world” carried out at Leipzig University represented a ground-breaking tool for the field. The rationale behind its construction is fully documented in Malchukov and Comrie (2015). The ValPaL contains data for 80 verb meanings (and occasionally additional others) from 36 languages. Data includes translational equivalents for these verb meanings together with their associated participants called “microroles” (see example (1)a), the basic valency pattern (see example (1)b) and the valency alternations (see example (1)c), all represented through “coding frames”.

(1) Verb meaning: LOAD
  a. Italian equivalent: *caricare*
    1. loader
    2. loaded thing
    3. loading place
  b. Basic coding frame:
    1 > V.subj[1] > 2 (su+3)
  c. Alternations: Locative alternation:
    1 > V’.subj[1] > 3 > di+2

Examples of basic (2) and non-basic usages and alternations (3) of stored verbs are also provided.

(2) *I venditori caricano i giornali e i libri sulla loro macchina.*
i venditor-i carica-no i giornal-i e i libr-i su-lla loro macchin-a
ART.DEF.M.PL seller-M.PL load-PRS.PL ART.DEF.M.PL newspaper-M.PL and ART.DEF.M.PL book-M.PL on-ART.DEF.F.SG their car-F.SG
‘The sellers load the newspapers and the books into their car.’

[1] https://hodel.unipv.it/paveda
[2] https://ValPaL.info

(3) *I venditori caricano la loro macchina di giornali e libri.*
i venditor-i carica-no la loro macchin-a di giornal-i e libr-i
ART.DEF.M.PL seller-M.PL load-PRS.PL ART.DEF.F.SG their car-F.SG of newspaper-M.PL and book-M.PL
'The sellers load newspapers and books onto their car.'

The ValPaL paved the way both for investigating transitivity scales and construction alternations Aldai and Wichmann (2018) while further verbal databases, such as the BivalTyp database[3] and the MultiVal lexicons[4] have also been created. In spite of these advances, the ValPaL database can be further improved, in terms of architecture, languages, and data coverage.

In the first place, (i) the language sample is unbalanced from the point of view of representability, as several language families are not included. Moreover, (ii) it does not contain data from ancient languages: this reflects a more general issues, as no systematic comparative study on diachronic developments across languages is available (see Luraghi and Roma, 2021a, Luraghi and Roma, 2021b). Additionally, (iii) the examples stored in the database are only occasionally extracted from corpora; their elicitation relies mainly on the native speakers' intuition of contributors (or elicited from speakers by contributors) or on reference handbooks and dictionaries. Finally, (iv) the current interface does not support comparative visualization of constructions and alternations, and comparison is further complicated by the use of language specific labels that make it virtually impossible to retrieve functionally similar alternations across languages.

For example, if one searches for the "locative" type among the "All alternations" variable in the database, the online query interface returns 28 entries, i.e., all alternations of all 36 languages available that contain the word "locative" in their name such as "Locative Alternation" (Standard Italian, see (1)c and (3) above), "Locative applicative o-" (Ainu), and "Locative alternation (argument rearranging)" and "Locative alternation (oblique to object)" (Balinese); however, it does not return the Russian "Prefixal Goal-Instrumental alternation". As shown in examples (4) and (5) taken from Malchukov (2015) and made available on the ValPaL database, this Russian alternation is functionally similar to the Italian "Locative alternation" in example (3), although in Russian it is coded on the verb through prefixation, whereas in Italian it is not.

(4) *Он нагрузил сено на телегу.*
On nagruzil seno na telegu.
on na-gruzil seno na teleg-u
he PFV-loaded hay.ACC on cart-ACC
'He loaded the hay onto the cart.'

(5) *Он загрузил телегу сеном.*
On zagruzil telegu senom.
on za-gruzil teleg-u sen-om
he PFV-loaded cart-ACC hay-INS
'He loaded the cart with hay.'

Instead, Ainu "Locative applicative o-" and Balinese "Locative alternation (oblique to object)" point to a valency increasing alternation usually called "applicative" (see Peterson, 2007).

Therefore, the results of this simple query seem both inhomogeneous and incomplete, which is a byproduct of the alternations being language-specific and at the same time being included in the database by separate contributors. Impossible as collecting these data would otherwise be, we find the current architecture problematic in different respects: one may not simultaneously visualize how alternations are encoded cross-linguistically (e.g., comparing Italian and Russian data) nor how similar alternations are encoded in the same language (e.g., comparing all Balinese locative alternations).

3 PaVeDa: where it stands and the road ahead

To overcome the issues discussed above, a research team based at the University of Pavia developed the PaVeDa database – the output of the PaVeDa project which received funding in 2021 from the University of Pavia[5]. Besides the local team, several international partners have agreed to take part in this project, in an attempt to build a more insightful typological resource that also allows for diachronic research.

The PaVeDa resource complies to the up-to-date standards regarding cross-linguistic data formats, as proposed by Forkel et al. (2018) in the *Cross-Linguistic Data Formats initiative (CLDF)*[6].

[3] https://www.bivaltyp.info

[4] https://typecraft.org/tc2wiki/Multilingual_Verb_Valence_Lexicon

[5] https://hodel.unipv.it/paveda

[6] https://cldf.clld.org

Particular attention has been paid to preserving compatibility with other linguistic resources and to avoiding multiplying information by referencing existing data. More specifically, we plan to enhance the database with the four main features described below.

(i) Concerning data coverage, we plan to include languages from families that are currently not represented on ValPaL (Uralic and Turkic) or are underrepresented (Afro-Asiatic).

(ii) To allow diachronic research, we also plan to add ancient Indo-European and Afro-Asiatic languages. For part of such languages the relevant data-sets have already been extracted and will soon be uploaded into the database (Early Latin, Ancient Greek, Gothic, Old Irish, Old English, and Classical Armenian, see Giuliani, 2021, Inglese and Zanchi, 2022 forthc., Zanchi and Tarsi, 2021, Roma, 2021) and they will be made available to the research community.

(iii) The work we have conducted so far on ancient languages has prompted us to redesign our methodology, as no native speakers are available for these languages they may only be studied by means of data retrieved from corpora. In order to assess which coding frames are basic, we plan to give a greater weight to the frequencies of attested patterns and to implement this type of usage-based methodology for modern languages as well, linking the data on constructional patterns to existing corpora and to other machine-readable resources. This raises further challenges: while for some modern languages reference corpora are indeed available and represent both the written and the spoken varieties (see e.g., the Russian National Corpus[7]), some of the languages already in the database or that we plan to include are low-resourced in terms of reference corpora for either oral varieties or written varieties and in same cases for both. For example, in the case of Italian, we will use CORIS[8]– the reference corpus for written Italian consisting of a balanced, representative, and up-to-date reference sample of 165 million tokens. As far as the oral variety of Italian is concerned, in the absence of a reference corpus, we will supplement the PaVeDa database with data from two spoken corpora, KIParla[9] and RadioCast-it[10]. We will adopt similar strategies of documentation for all low-resourced languages included in the PaVeDa database.

(iv) Regarding the database structure, PaVeDa makes contrastive visualizations possible and at the same time ensures interoperability with the ValPaL resource, due to a new layer of annotation that contains non-language-specific alternations. This layer enables generalizations over language-specific patterns, as it is mapped onto the current ValPaL set of alternations. This solution still allows including a language-specific level of alternations, which is crucial to account for the coding aspects and distributional restrictions of alternations.

Our proposed set of "general" alternations is conceived with the following aims in mind: avoid multiplying terminology, be as functional and conceptual as possible, and maximize language coverage. For instance, the Italian "Locative alternation", the Balinese "Locative alternation (argument rearranging)", and the Russian "Prefixal Goal-Instrumental alternation" are mapped onto the "Locative alternation (argument rearranging)", whereas Ainu "Locative applicative o-" and Balinese "Locative alternation (oblique to object)" are mapped onto "Locative alternation (oblique to object)". "Locative alternation (oblique to object)" is preferable over "Applicative", as the latter usually implies explicit encoding on the verb. In contrast, the former also accounts for marginal transitive occurrences of *abitare* 'inhabit' in Italian, shown in example (7) as compared to the basic coding frame in example (6) in which the locative microrole is indicated by the preposition *in* – both examples are retrieved from Cennamo and Fabrizio (2013).

(6) 1 > V.subj[1] > LOC 2
Mario abita in campagna.
Mario abit-a in campagn-a
Mario live-PRS.SG in countryside-F.SG
'Mario lives in the countryside.'

(7) 1 > V.subj[1] > LOC 2
La famiglia abita una villa abbandonata.
la famigli-a abit-a un-a vill-a abbandonat-a
ART.DEF.F.SG family-F.SG live-PRS.SG ART.INDF-F.SG country_house-F.SG abandoned-F.SG
'The family lives in an abandoned country house'

The database will be made freely available to the scientific community on an open-source basis through a dedicated web platform. This will promote the collaboration and reproducible research in

[7] https://ruscorpora.ru/old/en/index.html
[8] https://corpora.ficlit.unibo.it/TCORIS/
[9] http://kiparla.it/
[10] https://site.unibo.it/radiocast/it

linguistics. Last but not least, besides contributing to the scholar research on verbal constructions, the contrastive studies promoted by the PaVeDa resource will also enable applications with an impact on first language acquisition and on teaching and learning typologically diverse languages.

References

Gontzal Aldai and Søren Wichmann. 2018. Statistical observations on hierarchies of transitivity. *Folia Linguistica*, 52(2):249–281.

Michela Cennamo and Claudia Fabrizio. 2013. Italian. In Iren Hartmann, Martin Haspelmath, and Bradley Taylor, editors, *Valency Patterns Leipzig*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Michael Cysouw Sebastian Bank, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5:1–10.

Martina Giuliani. 2021. Valency patterns in latin. Master's thesis, University of Pavia, Italy.

Iren Hartmann, Martin Haspelmath, and Bradley Taylor. 2013. *The Valency Patterns Leipzig online database*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Guglielmo Inglese and Chiara Zanchi. 2022 forthc. Ancient greek valency patterns and alternations. In *The 3rd International Colloquium on Ancient Greek Linguistics*, Universidad Autónoma de Madrid, Spain, 16-18 June 2022.

Silvia Luraghi and Elisa Roma. 2021a. Valency and transitivity over time: An introduction. In Silvia Luraghi and Elisa Roma, editors, *Valency over Time: Diachronic Perspectives on Valency Patterns and Valency Orientation*, pages 1–12. De Gruyter Mouton, Berlin.

Silvia Luraghi and Elisa Roma, editors. 2021b. *Valency over Time: Diachronic Perspectives on Valency Patterns and Valency Orientation*. De Gruyter Mouton, Berlin.

Andrej Malchukov. 2015. Valency classes and alternations: parameters of variation. In Andrej Malchukov and Bernard Comrie, editors, *Valency Classes in the World's Languages. Volume 1: Introducing the Framework, and Case Studies from Africa and Eurasia*, pages 73–130. De Gruyter Mouton, Berlin.

Andrej Malchukov and Bernard Comrie, editors. 2015. *Valency Classes in the World's Languages. Volume 1: Introducing the Framework, and Case Studies from Africa and Eurasia*. De Gruyter Mouton, Berlin.

David A. Peterson. 2007. *Applicative constructions*. Oxford University Press, Oxford.

Elisa Roma. 2021. Valency patterns of old irish verbs: finite and non-finite syntax. In Silvia Luraghi and Elisa Roma, editors, *Valency over Time: Diachronic Perspectives on Valency Patterns and Valency Orientation*, pages 89–132. De Gruyter Mouton, Berlin.

Chiara Zanchi and Matteo Tarsi. 2021. Valency patterns and alternations in gothic. In Silvia Luraghi and Elisa Roma, editors, *Valency over Time: Diachronic Perspectives on Valency Patterns and Valency Orientation*, pages 31–88. De Gruyter Mouton, Berlin.