Introducing PaVeDa – Pavia Verbs Database: Valency Patterns and Pattern Comparison in Ancient Indo-European Languages

Silvia Luraghi, Alessio Palmero Aprosio, Chiara Zanchi, Martina Giuliani
University of Pavia, Bruno Kessler Foundation Trento, University of Pavia, Universities of Pavia and Bergamo
luraghi@unipv.it, aprosio@fbk.eu, chiara.zanchi@unipv.it, martina.giuliani@unibg.it

Abstract

The paper introduces PaVeDa (Pavia Verbs Database), a resource that builds on the ValPaL database of verbs' valency patterns and alternations by adding a number of ancient languages (completely absent from ValPaL) and a number of new features that enable direct comparison, both diachronic and synchronic. For each verb, ValPaL contains the basic frame and ideally all possible valency alternations allowed by the verb (e.g. passive, causative, reflexive etc.). In order to enable comparison among alternations, an additional level has been added, the alternation class, that overcomes the issue of comparing language specific alternations which were added by individual contributors of ValPaL. The ValPaL had as its main aim typological comparison, and data collection was variously carried out using questionnaires, secondary sources and largely drawing on native speaker intuition by contributors. Working with ancient languages entails a methodological change, as the data is extracted from corpora. This has led to re-thinking the notion of valency as a usage-based feature of verbs and to planning future addition of corpus data to modern languages in the database. It further shows the impact of ancient languages on theoretical reflection.

Keywords: verbal valency, valency patterns, alternations.

1. Introduction

In this paper we introduce a newly created resource, PaVeDa (the Pavia Verbs Database, https://paveda.unipv.it/), which expands on an existing one, the ValPaL database. The latter is a typological database, intended to document valency patterns and alternations in a variety of languages of different areal and genealogical affiliation, in which data from each language can be visualized in isolation. Our new database builds on the existing resource to include ancient Indo-European languages and adds features that allow visualizing direct comparison among languages (Zanchi, Luraghi and Combei 2022).

The paper is organized as follows. In Section 2 we briefly introduce the original ValPaL database and focus on some issues in data collection and presentation that affect cross-linguistic comparability. In Section 3 we describe the new features of PaVeDa and show its possible uses for synchronic and diachronic language comparison. In Section 4 we outline our plans for further extension of PaVeDa. Section 5 contains the conclusion.

2. Background: the ValPaL database

The ValPaL (Valency Patterns Leipzig Online Database) available at https://valpal.info/ is one of the main results of the Leipzig Valency Classes Project, carried out from 2009 to 2013, aimed at a large-scale cross-linguistic comparison of valency classes. The ValPaL project follows up on Levin's (1993) intuitions of providing a semantic classification of verbs based on their syntactic behavior. Valency classes are conceived as groups of verbs sharing morphosyntactic properties, i.e. coding patterns and valency alternations.

The ValPaL database stores information regarding the basic valency patterns and alternations of a selection of verb meanings for 36 languages belonging to 23 language families. The ValPaL verb selection singled out 80 core verb meanings, based on two criteria: a) representativeness of the entire verbal lexicon; b) known instantiations of distinctive grammatical behavior according to previous studies. These verb meanings denote a variety of events with different numbers of participants. They include two-place changes-of-state verbs (e.g. BREAK, KILL), three-place verbs of transfer (e.g. GIVE, BRING) and cognitive transfer (e.g. TELL, TEACH), perception verbs (e.g. SEE, SMELL), verbs of cognitions (e.g. THINK), emotions (e.g. LIKE) and bodily sensations (e.g. BE HUNGRY), activities (e.g. RUN, LAUGH), and weather verbs (e.g. RAIN), which are cross-linguistically zero-place verbs. Each verb meaning is paired with the semantically most fitting basic verb in each project language. Additional verb meanings were occasionally included for specific languages, up to the total of 162 verb meanings currently represented in the database. Only the 80 core verb meanings are represented in the database for each project language, resulting in a partial coverage of the newly added meanings. For example, the basic verb for WINK is stored for three languages, whereas the core meaning BLINK is covered by 35 languages; the meaning ASSASSINATE is available only for Italian.

The basic valency pattern and possible valency alternations available as 'coding frames' are stored along with each verb (cf. Section 2.1). Alternations are classified as 'coded' if morphologically marked on the verb, or as 'uncoded' if unmarked. As its original aim is not diachronic analysis, the ValPaL does not include data from ancient languages.

The ValPaL database has paved the way for the subsequent creation of similar typological databases, e.g. BivalTyp database (which can be found at https://www.bivaltyp.info/), which stores bivalent verbs and their encoding frames for 124 languages (Say 2014). Partly inspired by ValPaL is also the Multilingual Verb Valence Lexicon, which offers verb valency information in a uniform format for four languages: Norwegian, Spanish, Ga and Bulgarian (Hellan et al. 2014). The ValPaL database is

LT4HALA 2024@LREC-COLING-2024, pages 79–88

commonly considered a valuable tool for synchronic cross-linguistic investigations of valency patterns (Malchukov and Comrie 2015), and several studies that relied on its data have achieved important results (e.g. Aldai and Wichmann 2018). In this framework, the fact that no diachronic research is supported by the data stored in ValPaL is an important shortcoming, and ultimately also affects typological comparison. Moreover, even though single valency-related phenomena and certain argument structure constructions are well-studied topics for some ancient Indo-European languages, even these languages generally lack comprehensive overviews of their valency classes and alternations. A partial fill of this gap can be found in the valency lexica automatically induced from treebanks, i.e. morpho-syntactically annotated corpora in which dependency structures are stored as syntactic trees. Currently, such valency lexica are available for a limited set of ancient languages, notably Latin and Ancient Greek (McGillivray et al. 2009, McGillivray and Passarotti 2015, McGillivray and Vatri 2015, Passarotti et al. 2016, Zanchi et al. 2018, Zanchi 2021), and for the ancient languages included in the PROIEL project (available at http://dev.syntacticus.org/proiel.html), i.e. Latin, Ancient Greek, Old Russian, Old Church Slavic, Gothic, Old English, Classical Armenian and Old French. The new PROIEL treebank browser, Syntacticus, allows for visualization of the so-called "valency table", in which argument structure constructions with relative frequencies are given for verbs. These valency tables, too, are automatically generated from the syntactic annotation in the treebanks of the PROIEL project. A valency lexicon of Classical Armenian is currently under construction at the University of Wüzburg in the framework of the project CAVAL – The Classical Armenian Valency Lexicon (see https://www.phil.uni-wuerzburg.de/en/vgsp/research/projects/). As discussed at length by Zanchi et al. (2018) and Zanchi (2021), valency lexica of this type are useful resources if employed with caution: they reflect the classification system for arguments and adjuncts indicated in the annotation guidelines, may contain annotation errors inherited from treebanks, and do not account for null referential arguments, widespread in ancient Indo-European languages and not annotated in treebanks (see Luraghi 2003, Keydana and Luraghi 2012, Haug 2012, Sausa and Zanchi 2015).

Thus, a valency database compiled by humans and storing valency frames of verbs from ancient languages is certainly a *desideratum*. Notably, some work in this direction has also been done in the framework of the LiLa project (see https://lila-erc.eu/#page-top), whereby valency frames are added to the verbal synsets contained in the Latin WordNet (Mambrini et al. 2021). Similar endeavors with Sanskrit and Ancient Greek are described in Biagetti et al. (2023a, 2023b).

2.1 The data available in ValPaL

For each meaning stored in ValPaL one can find and visualize data related to individual languages, the geographical distribution, and the list of alternations available across languages, as shown in Figure 1.

Figure 1. Instantiations of the meaning LOAD

All verb meanings stored in ValPaL are cross-referenced with Concepticon, a resource that links concept labels from different concept lists to concept sets, which are given a unique identifier, a unique label and a human-readable definition (see https://concepticon.clld.org). Languages stored in ValPaL are paired with their Glottocodes, i.e. unique and stable identifiers that allow ValPaL to be cross-referenced with Glottolog (see https://glottolog.org/glottolog/language).
All coding frames and alternations are illustrated with examples including grammatical glosses and translations; verb-specific microroles and information about word order and argument type are also featured in the coding frame. As an example, let us consider the Italian verb *caricare* 'load' in Figure 2.

caricare

Simplex verb

Verb meaning: LOAD [load]

Examples: see at the bottom

Basic coding frame

Schema: 1 > V.subj[1] > 2 (su+3)

#	Microrole	Coding set	Argument type
1	loader	V.subj	A
2	loaded thing	Ø	P
3	loading place	su+NP	X

Figure 2: The coding frame of Italian *caricare*

In the coding frame the symbol > indicates word order, [...] indicates agreement and (...) optionality. The coding set refers to the morphological marking of the arguments, while possible argument types are A (transitive verb subject), P (direct object), S (intransitive verb subject), I (instrument), L (locative) and X (other).

Verb meaning FEAR [fear]

Meaning list: Core list

Typical context: The man feared the bear.

Role frame: E fears M

Microroles: fearer, fear stimulus, fear causer

Figure 3: role frame and microroles for FEAR

Along with each verb meaning, a list of microroles is provided. Figure 3 shows the microroles associated with FEAR. From Figure 2, one can see that each microrole is assigned a number, for example #1 indicates the "loader", #2 "the loading thing" and #3 "the loading place". This numbering is crucial to

interpreting basic and derived coding frames and to understanding how microroles are mapped to different basic and alternating argument structure constructions. For example, as shown in Figure 2, "1 > V.subj[1] > 2 (su+3)" is the basic coding frame of Italian LOAD *caricare*. The same assignment of numbers to microroles is kept in the derived coding frame of the passive alternation: "2 > passV'.subj[2] (su+3) (daParteDi+1)", from which one understands that microrole #2 "loaded thing" is passivized and microrole #1 "loader" becomes non obligatory, as expected for passive agents.

In the ValPaL database there are a number of inconsistencies concerning microrole labels: for example the first argument of EAT is labelled as *eater*, whereas the same argument of DRINK is tagged as *drinking person.* These inconsistencies are partly related to the addition of new meanings discussed above: EAT is one of the core 80 verb meanings, whereas DRINK has been added to ValPaL at a later stage.

In a separate section of the database, microroles (but not all of them, see Section 2.2) are grouped into roles. The latter are also employed in the role frame provided for each verb meaning (see, in Figure 3, the role frame for FEAR). Such roles partly overlap with the set of argument types, partly add to them. A list of roles can be found in a footnote in the guidelines downloadable from the ValPaL webpage, which leaves space for possible additions, and possibly modifications: "*We often use letters that can be thought of as standing mnemonically for particular roles (A: agent, P: patient, S: single central argument of intransitive verb, T: theme (of ditransitive verb), R: recipient (of ditransitive verb), L: location (including goal), I: instrument, E: experiencer, M: stimulus, X, Y, Z: other). No claims are associated with the use of these letters, and they could be replaced by other arbitrary variable symbols*." (Database Questionnaire Manual, fn. 6, https://valpal.info/database).

2.2 Some issues related to the ValPaL data

Some inconsistencies related to the new meaning addition and microroles have been discussed in Sections 2 and 2.1 above. In this section, we add to this, by elaborating on issues regarding data collection, alternation storage and labelling, microrole grouping and derived coding frame collection.

The data collected for the database has been elicited by contributors in different ways. Often, contributors were also native speakers of the language for which they were responsible, and heavily relied on their intuition for data collection. In other cases, they relied on their own fieldwork, or on data from previous works by themselves or by other authors. Only occasionally the data was collected from corpora.

The number of alternations listed varies widely across languages, ranging from 42 for English to 5 for Besta. This makes comparison complicated, as it may indicate that contributors stored alternations based on different levels of granularity. In addition, there is no consensus across contributions on how the same alternation is labelled: for example, the same alternation occurring with the meaning FILL whereby an instrumental adjunct is promoted to subject (as in *Water filled the tub*) is labelled 'Instrumental subject' in English, 'Instrument to subject alternation' in German and Russian, and 'Oblique subject' in Italian, making cross-linguistic comparison complicated.

Each verb meaning is assigned a role frame, with semantic roles covering a number of more fine-grained microroles (see Figure 3 in Section 2.1 and cf. Haspelmath & Hartmann 42-43; Malchukov 2015: 74). For example, the role frame for the meaning BRING is "A brings T to R" (see Section 2.1 for the role labels), possible microroles are *bringer, brought thing, bringing recipient, bring causer, bringing instrument*. Microroles have been added by contributors without specifying under which role label they should be grouped. So while in the case of BRING one finds *bringer* A, *brought thing* T*, bringing recipient* R*,* the remaining two microroles, *bring causer* and *bringing instrument* are not further specified (see the data in https://valpal.info/microroles).

Moreover, a number of derived coding frames are missing. For example, according to the ValPaL data, the Italian verb *caricare* regularly features 10 alternations: the so-called *Object omission, Passive, Reflexive passive, Locative alternation, Anticausative (coded), Indirect/dative reflexive, Impersonal reflexive, Causative, Impersonal of Reflexives, Impersonal passive.* Among these, only eight alternations are paired with their derived coding frame; for example, "2 > passV'.subj[2] (su+3) (daParteDi+1)" is the derived coding frame of the *Passive* alternation, as we discussed in Section 2.1. In the cases of the *Indirect/dative reflexive* and of the *Impersonal of reflexive* alternations, this piece of information is missing, which makes it hard for database users to understand the coding details of certain alternations.

3. New features in PaVeDa

The aim of PaVeDa is twofold. In the first place, more languages have been and will be added, starting with, but not limited to ancient Indo-European languages that have a modern counterpart already stored in ValPaL. This enables diachronic comparison and offers evidence for changes in valency patterns and alternations (Section 3.1). In the second place, an intermediate level of annotation to the original ValPal has been added, called "alternation class", which categorizes language-specific alternations into four cross-linguistic types. Because comparison is an essential part of our research, we added a dedicated tool to compare basic frames and alternations across all languages and between individual languages (Section 3.2). PaVeDa also aims to add the missing role labels to all microroles and correcting some discrepancies discussed above (Section 3.3).

3.1 Adding a diachronic dimension

To date, ancient Indo-European languages added to PaVeDa are Old Latin, Ionic-Attic Ancient Greek, Gothic, Old English, Classical Armenian and Old High German. Apart from Gothic, that does not have any modern descendent, four other languages have their modern counterpart already stored in ValPaL: Italian,

English, Eastern Armenian and German. Because Ionic-Attic Ancient Greek did not have its modern counterpart already available, we also added Modern Greek to the database. The information on basic valency patterns and alternations included for ancient languages relies on corpus data. Old Latin is based on the Plautus' corpus, whereas a corpus of Classical Greek prose comprising orators, historians and Plato has been scrutinized for Ionic-Attic Ancient Greek[1]. The reference corpus for Gothic is the fourth-century translation of the Bible, traditionally attributed to the Gothic bishop Wulfila (see Zanchi & Tarsi 2021: 31-34)[2]. The corpus for Old English consists of both prose (e.g. Ælfric's Catholic Homilies and Bede's History of the English Church; see Taylor et al. 2003) and poetry (e.g. the Beowulf and the Anglo-Saxon Elegies; see Pintzuk & Plug 2002), and includes texts differing in period, genres, and dialect. For Classical Armenian the New Testament has been scrutinized. Finally, data for Old High German is based on the REA corpus (Krause and Zeldes 2016), limited to Old High German texts[3].

The corpora used for such languages differ in terms of corpus-size and genre; these differences are due to the fact that, even though these languages all qualify as corpus languages, the available corpora that survived up to the present time are very different, which makes corpus harmonization virtually impossible. Concerning data extraction, PaVeDa contributors adopt different methodologies. In the case of languages with small and close corpora such as Gothic and Old Latin, all the occurrences of verb lemmas selected have been analyzed. In case of large-corpus languages such as Ionic-Attic Ancient Greek all the occurrences of verb lemmas whose frequency in the reference corpus is lower than 100 occurrences have been analyzed, whereas, for verb lemmas with frequency higher than 100, a stratified random sample of 100 occurrences has been extracted. These 100 occurrences are assumed to contain instantiations of all alternations featured by a certain verb. Notably, this assumption has always been double-checked against reference dictionaries and grammars. All added ancient languages are cross-referenced to Glottolog. For this reason, we tried to adhere to Glottolog language names as close as possible, as in the case of Old Latin and Ionic-Attic Ancient Greek (Glottolog does not feature a generic label Ancient Greek, while the label Latin only refers to Late, Vulgar and Medieval Latin, cf. https://glottolog.org/resource/languoid/id/lati1261).

3.2 Issues brought about by the addition of ancient languages

Elicitation of data for ancient languages brings about a number of theoretical issues that have a more general scope. The most challenging issue is of course the impossibility to rely upon native speakers' judgments to rate the basicness of competing verbs for any given core meaning, let alone alternations. Following the methodology laid out in Zanchi & Tarsi (2021), we used a combination of morphological and frequency criteria to overcome this issue as detailed below.

Verb lemmas that are morphologically underived or that exhibit the simplest morphological structure are regarded as more basic (e.g. in Old Latin the verb *eō* is preferred over the preverbed *ad-eō* 'approach' for the meaning GO). If a verb is underived but is scarcely attested in the reference corpus, a derived verb is selected instead, provided that its number of occurrences is significantly higher. For example, for the meaning LIKE the derived Gothic verb *ga-leikan* (attested 20 times in the Gothic corpus) has been selected instead of *leikan* (one occurrence) because of its higher frequency. Frequency also drives the choice between verbal lemmas with comparable degrees of morphological complexity (e.g. in Old Latin for the meaning FEAR the verb *metuō*, 154 occurrences in the reference corpus, is preferred over *timeō*, 35 occurrences). In cases in which neither of these criteria is applicable, we decided to take into account the historical developments of the candidate lemmas, and possibly select more than one verb. For example, for the meaning EAT both the Gothic verbs *matjan* and *itan* were included in the database, as the latter, despite being less-frequent than the former in the Gothic corpus, continues in several modern Germanic languages (e.g. English *eat*, German *essen*).

Frequency is also disfavored in cases in which the more frequent verb for a given meaning is polysemous. Take as an example the two competing Old Latin lemmas *petō* and *poscō* for the meaning ASK FOR. Despite its lower frequency, *poscō* has

[1] Corpora for Old Latin and Ionic-Attic Ancient Greek have been scrutinized with the Perseus Digital Library (https://www.perseus.tufts.edu/hopper/).

[2] The Gothic Gospels are available at the PROIEL project and Wulfila project websites (PROIEL Project: http://foni.uio.no:3000/sources/11; Wulfila project: http://www.wulfila.be.

[3] For the REA corpus see https://www.deutschdiachrondigital.de/rea/ and http://dsh.oxfordjournals.org/content/31/1/118 available at https://korpling.german.hu-berlin.de/annis/ddd. Notably, the REA corpus also contains texts in Old Saxon and Old Low Franconian, which have been left out from our account. This has been easily done, as texts can be selected individually in REA. Old Latin data was collected by Martina Giuliani (University of Pavia / University of Bergamo); Chiara Zanchi (University of Pavia) and Guglielmo Inglese (University of Turin) are responsible for Ionic-Attic Ancient Greek; Matteo Tarsi (Uppsala University) and Chiara Zanchi added the Gothic data. The Old English data was collected by Martina Giarda (University of Pavia / University of Bergamo), and the Old High German one by Giacomo Bucci (Ghent University). Petr Kocharov took care of the Classical Armenian section of the database. The addition of corpus data for Modern Russian was carried out by Erica Pinelli, Irina Parshina and Maria Bocharova. Lucrezia Carnesale collaborated in the creation of the database.

been selected instead of *petō,* because its semantics better fits the meaning ASK FOR. The verb *petō* is highly polysemous and is frequently used with the meanings 'assault, attack' and 'go, travel toward', along with expressing requests. As argued by Inglese (2021: 142) "verbs that are primarily associated with a given meaning are preferred over those that express that meaning only secondarily and/or metaphorically". Selecting *poscō* would have forced us to analyze all the occurrences of the lemma to look for those instantiating the meaning relevant for the database.

Of course, especially with languages such as Gothic for which only a limited corpus is available, missing attestation of some verb meanings or constructions does not necessarily reflect a gap in a language's lexicon or grammar but it may reflect a gap in the corpus. The same is true for Old Latin whose reference corpus is the collection of Plautus' comedies (see Section 3.1). For verb meanings not sufficiently represented because of corpus selection, additional corpora (e.g. Terence's corpus for Old Latin) and lexicographic resources have also been checked.

In spite of these challenges, using corpus data has an undoubted advantage over relying on the intuition of individual speakers, as corpora contain more than what is evident to speakers' intuition, provide real usage-based occurrences and also data about their actual frequency. As Fillmore's (1992: 35) puts it: "[…] every corpus I have had the chance to examine, however small, has taught me facts I couldn't imagine finding out any other way". We will return on this important point in Section 4.

In languages that do not rely on a large enough corpus of attestations it may be the case that some of the ValPaL verb meanings are not retrievable. In such cases, other verb meanings have been selected, to partly compensate for the gaps in coverage, that can reasonably be expected to elicit verbs with a comparable syntactic behavior to those which are not attested. For example, the Gothic section of the database does not comprise lemmas for the ValPaL core meanings BE A HUNTER, BLINK, BOIL, COUGH, FEEL COLD, HUG, PLAY and SMELL. In order to partially compensate these gaps, new meanings have been added, i.e. CRY, DIG, DRINK, FALL, GRIND and LIGHTEN. All new meanings are cross-referenced to Concepticon.

Corpus-based approaches also challenge the assumption that ValPaL core meanings are representative of the entire verbal lexicon, as some argument structure constructions are under-represented due to verb meanings selection. An example is the domain of experience in Old Latin. ValPaL core meanings fail to account for a group of Latin experiential verbs denoting negative emotions (e.g. *pudet* 'be ashamed'), which show a peculiar argument structure construction (see Fedriani 2014 among others). These verbs are constructed impersonally: they are inflected in the third person singular active (rarely passive) form, without a fully-fledged syntactic subject in the nominative, and take two arguments: an accusative experiencer and a genitive stimulus. To also included Latin verbs featuring this construction in the database, five new verb meanings have been added to PaVeDa: BE ANNOYED, BE ASHAMED (cf. (1), (2)), DISPLEASE, HAVE PITY and REGRET[4].

(1) Verb meaning: BE ASHAMED
Old Latin verb: *pudet*
Microroles:
1. ashamed person
2. ashaming thing
Basic coding frame: 1-acc 2-gen V.3SG

(2) Example of the basic coding frame:

quoius	*me*	*nunc*
REL.GEN.SG	1SG.ACC	now
facti	*pudet*	
deed:GEN.SG	be_ashamed:PRS.3SG	

'a deed which I am now ashamed of.' (Plaut. *Bacch*. 1016)

As the addition of new meanings leads to the addition of new microroles and, ideally, should be extended to all languages in the database, such additions are discussed with the project coordinators and managed by them. All newly added meanings will also be externally cross-referenced with Concepticon.

The role of frequency in corpora cannot be underestimated, and has brought us to reconsider the way in which the data stored in ValPaL have been elicited and, more in general, how one should elicit data for modern languages and how the valency of a verb should be established. For this reason, we plan to add corpus data to languages already stored in ValPaL, following a usage-based notion of valency (see Section 4).

3.3 Alternation classes

In order to make cross-linguistic comparison easier, we added an intermediate level of alternations that we have called "alternation class". Following Malchukov (2015: 96-103 and references therein) language-specific alternations have been classified into four coarse-grained groups: (i) Argument-decreasing; (ii) Argument-increasing; (iii) Argument-rearranging; and (iv) Argument identifying. Alternations affecting the number of verbs' arguments have been marked either as Argument-decreasing or -increasing. As argument-decreasing strategy see the generic argument omission in Ionic-Attic Ancient Greek, as in (3) and (4).

(3) Verb meaning: EAT
Ionic-Attic Ancient Greek verb: ***esthíō***
Basic coding frame:
1-nom V.act.subj[1] 2-acc

[4] All examples used in this paper are from the PaVeDa database. In case the new language employs a script different from the Latin one, the original text is provided, along with its transliteration, glosses and translation. When available, we plan to add links to external language resources indicating to the loci of the added examples.

Derived coding frame: 1-nom V.act.subj[1]

(4) Example of the generic argument omission alternation in Ionic-Attic Ancient Greek:
ὅτι ἀηδῶς ἐσθίοι
hóti aēdôs esthíoi
that unpleasantly eat.PRS.OPT.3SG
'That he eats unpleasantly.' (Xen. Mem. 3.13.2.1)

An argument-augmenting strategy is the cognate/ kindred argument alternation in Old English, shown in (5) and (6).

(5) Verb meaning: LIVE
Old English verb: *lifian*
Basic coding frame: 1-nom V.subj[1] (in 2-dat)
Derived coding frame: 1-nom V.subj[1] 4-acc-cognate (in 2-dat)

(6) Example of the cognate/kindred argument alternation in Old English:
Lifd se
live.IND.PRET.3SG DET.NOM.SG.M
mon his
man(M).NOM.SG POSS.3SG.M
liif in micelre
life(N).ACC.SG in great.DAT.SG.F
forhæfdnisse
abstinence(F).DAT.SG
'The man lived a life of great abstinence.'
(Bede_4:26.350.6.3521_ID)

Alternations implying a change in the encoding of verbs' arguments but not in their number are Argument-rearranging. An example is the partitive alternation attested in Ionic-Attic Ancient Greek in (7) and (8).

(7) Verb meaning: CUT
Ionic-Attic Ancient Greek verb: *témnō*
Basic coding frame: 1-nom V.act.subj[1] 2-acc (3-dat)
Derived coding frame: 1-nom V.act.subj[1] 2-gen

(8) Example of the partitive alternation in Ionic-Attic Ancient Greek:
τῆς ὕλης
tês húlēs
ART.GEN.F wood(F).GEN
τέμνοντα
témnonta
cut.AOR.PTCP.ACC
'Having cut wood' (Xen. Cyneg. 2.9.3)

Finally, the class Argument-identifying has been assigned to reflexive and reciprocal alternations, see e.g. the direct reflexive alternation in Old Latin shown in (9) and (10).

(9) Verb meaning: COVER
Latin verb: *tegō*
Basic coding frame: 1-nom 2-acc V.subj[1]
Derived coding frame: 1=2-nom 1=2-acc-refl V.subj[1=2]

(10) Example of the direct reflexive alternation in Old Latin:
capite se
top(N):ABL.SG REFL.ACC
totum tegit
entire(N):ACC.SG cover:IND.PRS.3SG
'He covers himself entirely with his top' (Plaut. Trin. 851)

Having added this level, which does not exist in the original ValPaL, we now have new options for comparison. In order to compare ancient languages with their modern counterpart, we have added it not only in the new languages stored in PaVeDa but also to some of the languages already stored in ValPaL and imported into PaVeDa, i.e. English, German, Italian and East Armenian. We can now look for all alternations belonging to one of the four groups in the relevant languages, or all alternations, again divided into the four groups under each verb meaning.
In addition, we implemented the option of directly comparing a verbal meaning, with basic frames and alternations in two given languages.
Let us take the verb meaning BREAK. Presently, ValPaL offers the option of visualizing the basic frames occurring in all languages.

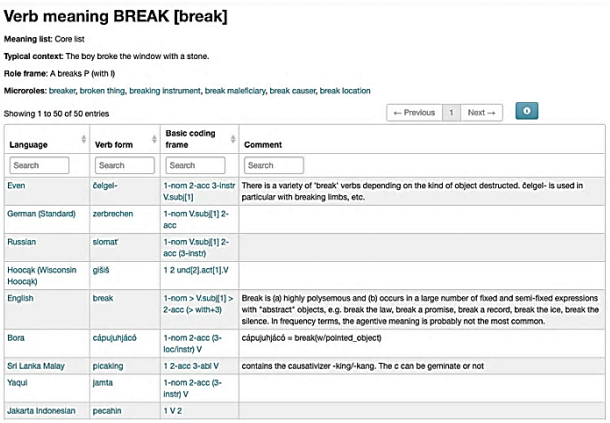

Figure 4: Basic frames of BREAK

To this, we added the option of visualizing all attested alternations (for BREAK they are 250), or to select those belonging to one of the four groups at the intermediate level. In Figure 5 we show all argument -decreasing alternation contained in the database for the meaning BREAK.

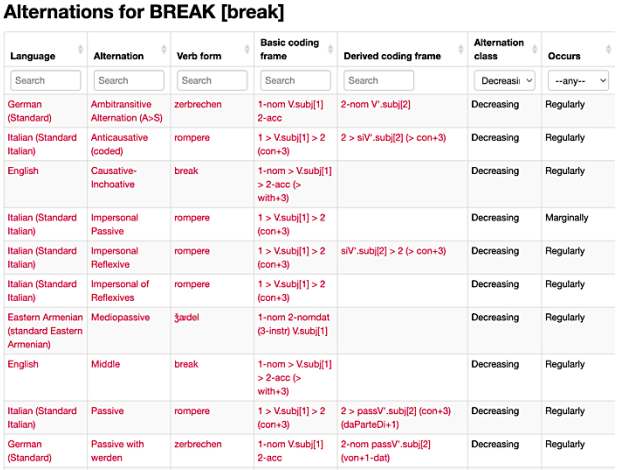

Figure 5: Argument-decreasing alternations for BREAK

Comparison between two languages allows visualizing the basic frame and all the alternations that occur in those two languages (see Figure 6).

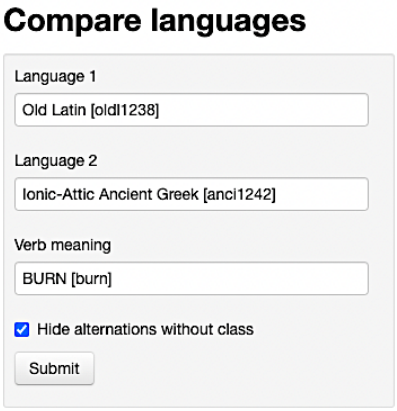

Figure 6: BREAK in Old Latin and Ionic-Attic Ancient Greek

Comparison between Old Latin [oldl1238] and Ionic-Attic Ancient Greek [anci1242] on BREAK [break]

Basic frames | Alternations

Language	Alternation	Verb form	Basic coding frame	Derived coding frame	Alternation class	Occurs	
Old Latin	r-passive (P)	frango	1-nom 2-acc V.subj[1]	2-nom (a/ab 1-abl) passV'.subj[2]	Decreasing	Regularly	Details
Old Latin	anticausative with r-passive	frango	1-nom 2-acc V.subj[1]	2-nom passV'.subj[2]	Decreasing	No data	Details
Ionic-Attic Ancient Greek	voice alternation - anticausative	hrḗgnumi	1-nom V.act.subj[1] 2-acc	2-nom V.mid.subj[2]	Decreasing	Regularly	Details
Ionic-Attic Ancient Greek	generic argument omission	hrḗgnumi	1-nom V.act.subj[1] 2-acc	1-nom V.act.subj[1]	Decreasing	Marginally	Details

Figure 7: BREAK in Old Latin and Ionic-Attic Ancient Greek

In Figure 7 we compare the alternations of Ionic-Attic Ancient Greek *hrḗgnumi* and Old Latin *frangō*. We can remark the mediopassive voice encodes the anticausative alternation in both languages, but it encodes the passive only in Latin.

Comparing an ancient language with its modern counterpart also leads to interesting remarks. In Figure 8 we compare the alternations of the Ionic-Attic Ancient Greek verb *kaíō* Modern Greek *kéo* 'burn'. We can see that the main function of the mediopassive voice remains the encoding of the anticausative alternation, while encoding of the passive voice remains marginal at both languages stages (Luraghi and Mertyris 2021).

Comparison between Ionic-Attic Ancient Greek [anci1242] and Modern Greek [mode1248] on BURN [burn]

Basic frames | Alternations

Language	Alternation	Verb form	Basic coding frame	Derived coding frame	Alternation class	Occurs	
Modern Greek	reflexive lability	kéo	1-nom V.act.subj[1] 2-acc	1=2-nom V.act.subj[1=2]	Rearranging	Marginally	Details
Modern Greek	voice alternation - anticausative	kéo	1-nom V.act.subj[1] 2-acc	2-nom V.nonact.subj[2]	Decreasing	Regularly	Details
Modern Greek	voice alternation - passive	kéo	1-nom V.act.subj[1] 2-acc	2-nom V.pass.subj[2] (1-apó+acc)	Decreasing	Marginally	Details
Ionic-Attic Ancient Greek	reflexive lability	kaíō	1-nom V.act.subj[1] 2-acc (3-dat)	1=2-nom V.act.subj[1=2]	Identifying	Marginally	Details

Figure 8: BURN in Ancient and Modern Greek

In Figure 9 we compare the verb meaning EAT in Old High German and Modern Standard German.

Comparison between Old High German [oldh1241] and German (Standard) [stan1295] on EAT [eat]

Basic frames | Alternations

Language	Alternation	Verb form	Basic coding frame	Derived coding frame	Alternation class	Occurs	
German (Standard)	Object Omission Alternation	essen	1-nom V.subj[1] 2-acc	1-nom V'.subj[1]	Decreasing	Regularly	Details
German (Standard)	Passive with werden	essen	1-nom V.subj[1] 2-acc	2-nom passV'.subj[2] (von+1-dat)	Decreasing	Regularly	Details
German (Standard)	Impersonal Passive	essen	1-nom V.subj[1] 2-acc		Decreasing	Marginally	Details

Figure 9: EAT in Old High German and Modern Standard German

In Old High German the partitive alternation occurs, which has disappeared in Modern German. Indeed, this particular Argument-rearranging alternation, which involves the partitive genitive (or the ablative case in Classical Armenian) as direct object case is typical of ancient, as opposed to modern Indo-European languages, coherently with the data in Figure 10 (see Luraghi and Kittilä 2014).

Alternations

Showing 1 to 5 of 5 entries (filtered from 709 total entries)

Language	Alternation	Alternation class	Type
Search	partitive	--any--	--any--
Ionic-Attic Ancient Greek	partitive alternation	Rearranging	Uncoded
Classical Armenian	partitive	Rearranging	Uncoded
Gothic	partitive alternation	Rearranging	Uncoded
Old English	partitive	Rearranging	Uncoded
Old High German	Partitive	Rearranging	Uncoded

Figure 10: The partitive alternation

3.4 Semantic roles and microroles

As we said in Section 2.2 labels for semantic roles are introduced in a footnote of the guidelines, and it is explicitly stated that they are arbitrary. When one looks at the classification of microroles according to their correspondence to a role, one can see a number of discrepancies. Some of them are connected with the use of the label S, defined as *single central argument of intransitive verb*. Indeed, this definition is problematic because it refers to a syntactic, rather than semantic property. In particular, experiential verbs are often monovalent, so their subjects should be labelled S, but as the label E *experiencer* is also in the list, in the database they are variously labelled S (as in the case of FEEL COLD) or E (as in the case of BE HUNGRY and BE SAD, whose role frame also contains a single argument). Similarly, the single argument of motion verbs is usually variously assigned the role S (e.g. the verb meaning GO) or A (the verb meanings RUN and JUMP). Other discrepancies are shown by the use of the label R (recipient of a ditransitive verb). While the third participant of verb meanings such as GIVE and

BRING (*bringing recipient*, *giving recipient*, respectively) is assigned to R (cf. Section 2.2), the third participant of the meaning SEND is instead assigned to X: other.
These discrepancies have not allowed us to implement a further level of comparison among semantic roles yet. This comparative level will allow users to visualize how a certain semantic role is encoded in the project languages. To reach this goal, we are presently trying to unify role assignment to microroles.

4. PaVeDa in the (near) future

As for the diachronic dimension, we plan to add other ancient Indo-European languages to PaVeDa. Recruited project members have already started working on Old Italian, Old Church Slavonic, Old Icelandic, Sanskrit, Old Irish and Hittite.

Thus, besides including ancient Indo-European languages that already have a modern counterpart in ValPaL (e.g. Old Icelandic - Icelandic) we will also include languages for which no modern descendent is stored in ValPaL, as is the case of Sanskrit or Old Irish. In such cases, we plan to also add modern counterparts and have already recruited contributors for Hindi and Modern Irish. We aim to have all sub-branches of the Indo-European language family stored in the database in order to allow employing the data for syntactic reconstruction, and reconstruct valency patterns and alternations for the proto-language (for previous efforts in this direction, see e.g. Barðdal and Smitherman 2013, Barðdal and Eythorsson 2016).

In regard to data coverage, we plan to include languages from families that are currently not stored in ValPaL, in particular Uralic and Turkic: our contributors are currently working on Finnish, Hungarian, Turkish and Chuvash, as well as from language families that are currently underrepresented, such as Afro-Asiatic (only Modern Standard Arabic is included in ValPaL). Increasing the number of Afro-Asiatic languages will also enable us to expand diachronic research outside the Indo-European languages: our contributors are currently working on Modern and Biblical Hebrew, and we have plans to further include diachronically diverse Arabic varieties.

Moreover, we are currently working on revising the data of some modern languages in light of corpus-based evidence provided by reference corpora. It is important to stress that our decision to add corpus data both to the modern languages that we have started adding (such as Modern Greek) and to the languages already stored in ValPaL has been prompted by our work with ancient languages. As we remarked in Section 3.2, working with a closed corpus may have limitations, but it also provides real data from language usage rather than data specifically elicited by a linguist from his/her native speaker intuition. Hence, work with ancient languages has had an impact on our view on how modern spoken languages should be investigated with concrete consequences on our methodology.

Up to now, Russian data has been partly revised by one of our project members based on data from the *Russian National Corpus* (available at https://ruscorpora.ru/en/), and discrepancies have indeed emerged from ValPaL examples coming from native speaker intuition and what is actually contained in corpora. For example for the verb ***slomat'*** 'break' not all alternations listed in ValPaL have been found in the *Russian National Corpus*; conversely, for the verb meaning LOOK AT the verb ***smotret'*** is given with no alternations, but in the corpus our contributors found the reflexive passive, as in (11).

(11) *Fil'm* *smotritsja*
film.SG.NOM watch.PRES.3SG.REFL
očen' legko.
very easily
'The movie is very easy to watch.'

In addition, while for various verbs possible alternations are listed that involve different verbal prefixes, e.g. under the meaning LOAD ***nagruzit'*** is used for the basic coding frame, but for the 'Prefixal Goal-Instrumental alternation' the verb ***zagruzit'*** is used. Following the same approach, for ***smotret'*** one could also add the participial passive alternation, which is documented in the corpus again in connection with a different prefix, ***osmotret'***, but the contributors of ValPaL failed to do so.

To enhance the comparative possibilities offered by our database, we will further group languages specific alternations in a more fine-grained layer of 'comparative concepts' (Haspelmath 2010) describing alternation types such as 'passive', 'antipassive', 'applicative', and so forth. For coded alternations, we will build on the taxonomy proposed in Haspelmath (2022), while for uncoded alternations we will try to identify and correct the inconsistencies of the type described in Section 2.2.

So far, we have described implemented and planned comparative visualization for verb meanings and for functional units, such as alternations and semantic roles. Our last goal for the near future of PaVeDa is to introduce a lemma-based comparative visualization option, which will allow tracking whether and how cognate verbs change their valency patterns and alternations over time. This is possible as our contributors for ancient Indo-European languages have been asked to indicate cognates of the basic verbs they choose to include in the database.

5. Conclusion

This paper documents the work that has been done so far to create a new resource, PaVeDa, which is specifically designed for cross-linguistic and diachronic comparison of verb valency classes and alternations. Building on the ValPaL database, we implemented modifications regarding language coverage, data elicitation and database structure.

As for language coverage, to date we have added six ancient and one modern Indo-European languages, for a total of nine new meanings, 46 new microroles, 211 new coding frames. Two new options for

searching the database have been implemented, one that allows to visualize simultaneously all alternations stored in the database for each verb meaning across all languages, and a second one that allows direct comparison of the alternations found in two given languages for each verb meaning.

Further plans concern the addition of other, both ancient and modern languages, as well as corpus data for all languages, including those stored in ValPaL, in order to have real, usage-based data on valency patterns and alternations, and minimize the impact of constructed data, based on native speaker intuition of individual researchers.

Finally, we are planning to implement a comparison option based on etymological information (that has been annotated for ancient languages but not yet uploaded into the database) to make possible tracking changes in valency patterns and alternations over time.

Our research shows how working with ancient languages may also bring about a change of perspective on the methodology adopted for research on modern languages, as in the case of favoring corpora over native speaker intuition as source for data elicitation.

Concerning the relation between PaVeDa and ValPaL, while the main goal is language comparison for both databases, we view diachronic comparison as equally important as typological comparison. In this regard, PaVeDa should not simply be viewed as an enhanced version of ValPaL, but as a new and independent resource in its own right, and a completely new resource for what concerns ancient languages.

6. Bibliographical References

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

Barðdal, J. and Smitherman, T. (2013). The quest for cognates: a reconstruction of oblique subject constructions in proto-Indo-European. *Language dynamics and change*, 3(1):28–67.

Biagetti, E., Brigada Villa, L., Zanchi, C., and Luraghi S. (2023a). Enhancing the semantic and conceptual description of Ancient Greek verbs in WordNet with VerbNet and FrameNet: a treebank-based study. In *Papers from the Annual International Conference "Dialogue"*, Vol. 22 (Supplementary volume), pages 1009–1020.

Biagetti, E., Zanchi, C., and Luraghi, S. (2023b). Linking the Sanskrit WordNet to the Vedic Dependency Treebank: a pilot study. In *Proceedings of the 12th Global Wordnet Conference*, pages 77–83, University of the Basque Country, Donostia – San Sebastian, Basque Country. Global Wordnet Association.

Eythórsson, Th. and Barðdal, J. (2016). Syntactic reconstruction in Indo-European: State of the art. *Veleia*, 33:83–102.

Fedriani, C. (2014). *Experiential predicates in Latin.* Brill, Leiden.

Haspelmath, M. (2010). Comparative concepts and descriptive categories in cross-linguistic studies. *Languages*, 86(3):663–687.

Haspelmath, M. (2022). Valency and voice constructions (https://lingbuzz.net/lingbuzz/005941).

Haspelmath, M. and Hartmann, I. (2015). Comparing verbal valency across languages. In A. Malchukov & B. Comrie (Eds.), *Valency classes in the world's languages*. Berlin & New York: Mouton de Gruyter, pp. 41–72.

Haug, D. T.T. (2012). Syntactic conditions on null arguments in the Indo-European Bible translations. *Acta Linguistica Hafniensia*, 44(2):129–141.

Inglese G. (2021). Anticausativization and basic valency orientation in Latin. In S. Luraghi & E. Roma (Eds.), *Valency over Time: Diachronic Perspectives on Valency Patterns and Valency Orientation*. Berlin & New York: Mouton de Gruyter, pp. 133–168.

Keydana, G. and Luraghi, S. (2012). Definite referential null objects in Vedic Sanskrit and Ancient Greek. *Acta Linguistica Hafniensia*, 44(2):116–128.

Levin, B. (1993). *English verb classes and alternations*. University of Chicago Press, Chicago.

Luraghi, S. (2003). Definite referential null objects in Ancient Greek. *Indogermanische Forschungen* 108:167–194.

Luraghi, S. and Kittilä, S. (2014). Typology and diachrony of partitive case markers. In S. Luraghi & T. Huumo (Eds.), *Partitive cases and related categories*. Berlin & New York: Mouton de Gruyter, pp. 17–62.

Luraghi, S. and Mertyris, D. (2021). Basic valency in diachrony: from Ancient to Modern Greek. In S. Luraghi & E. Roma (Eds.), *Valency over Time: Diachronic Perspectives on Valency Patterns and Valency Orientation*. Berlin & New York: De Gruyter, pp. 169–208.

Malchukov, A. (2015). Valency classes and alternations: parameters of variation. In A. Malchukov & B. Comrie (Eds.), *Valency classes in the world's languages*. Berlin & New York: Mouton de Gruyter, pp. 73–130.

Malchukov, A. and Comrie, B. (2015). *Valency classes in the world's languages*. Berlin & New York: Mouton de Gruyter.

Mambrini, F., Passarotti, M. C., Litta Modignani Picozzi, E. M. G., and Moretti, G. (2021). Interlinking Valency Frames and WordNet Synsets in the LiLa Knowledge Base of Linguistic Resources for Latin. In *Proceedings of the 17th International Conference on Semantic Systems*, pages 16–28, Amsterdam, The Netherlands.

McGillivray, B., Passarotti, M. C., and Ruffolo, P. (2009). The Index Thomisticus Treebank Project: Annotation, Parsing and Valency Lexicon. *TAL*, 50(2):103–127.

McGillivray, B. and Passarotti, M. C. (2015). Accessing and using a corpus-driven Latin Valency Lexicon. In G. V. M. Haverling (Ed.), *Latin Linguistics in the Early 21st Century. Acts of the 16th International Colloquium on Latin Linguistics*, pages 289–300, Uppsala, Sweden, 6–11 June.

McGillivray, B. and Vatri, A. (2015). Computational valency lexica for Latin and Greek in use: a case study of syntactic ambiguity. *Journal of Latin Linguistics*, 14(1):101–126.

Say, S. (2014). Bivalent Verb Classes in the Languages of Europe: A Quantitative Typological Study. *Language dynamics and change*, 4(1): 116–166.

Sausa, E. and Zanchi, C. (2015). Non-accusative null objects in the Homeric Dependency Treebank. In F. Mambrini, M. Passarotti, & C. Sporleder (Eds.), *Proceedings of the Workshop on Corpus-Based Research in the Humanities* (CRH), pages 107–116, Warsaw, Poland, 10 December.

Zanchi, C., Sausa, E. and Luraghi, S. (2018). HoDeL, a Dependency Lexicon for Homeric Greek: issues and perspectives. In P. Cotticelli & F. Giusfredi (Eds.), *Proceedings of Formal Representation and Digital Humanities*. Cambridge: Cambridge Scholars Publishing, pp. 230–256.

Zanchi, C., Luraghi, S. and Combei, C. R. (2022). PaVeDa – Pavia Verbs Databse: Challenges and perspectives. In *Proceedings of the 4th workshop on research in computational linguistic typology and multilingual NLP*, pages 99-102. Seattle, Washington, Association for Computational Linguistics.

Zanchi, C. and Tarsi, M. (2021). Valency patterns and alternations in gothic. In S. Luraghi & E. Roma (Eds.), *Valency over Time: Diachronic Perspectives on Valency Patterns and Valency Orientation*. Berlin & New York: Mouton de Gruyter, pp. 31–88.

7. Language Resource References

Crane, G. R. (Ed.). *Perseus Digital Library*. Tufts University. Accessed on 22 February 2024.

Dag, T. T. H. and Jøhndal, M. L. (2008). Creating a Parallel Treebank of the Old Indo-European Bible Translations. In C. Sporleder & K. Ribarov (Eds.), *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 27–34. http://dev.syntacticus.org/proiel.html#downloads. Accessed on 22 February 2024.

Hartmann, I., Haspelmath, M., and Taylor, B. (2013). *Valency Patterns Leipzig*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Krause, Th. and Zeldes, A. (2016): ANNIS3: A new architecture for generic corpus query and visualization. In *Digital Scholarship in the Humanities* 31(1):118–139.

List, J. M., Tjuka, A., van Zantwijk, M., Blum, F., Ugarte, C. B., Rzymski, Ch., Greenhill, S., and Forkel, R. (Eds.) (2023). CLLD Concepticon 3.1.0 [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7777629. Accessed 24 February 2024.

Pintzuk, S. and Plug, L. (2002). *The York Helsinki Parsed Corpus of Old English Poetry*. Department of Linguistics, University of York. Oxford Text Archive, first edition; http://www.users.york.ac.uk/~lang18/pcorpus.html. Accessed on 22 February 2024.

Passarotti, M. C., Gonzalez Saavedra, B. and Onambele Manga, C. L. (2016). Latin Vallex. A Treebank-based Semantic Valency Lexicon for Latin. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2599–2606.

RNM = The Russian National Corpus. Плунгян, В. А. Резникова, Т. И., Сичинава Д. В. (2005). Национальный корпус русского языка: общая характеристика. Научно-техническая информация. Сер. 2. № 3. С. 9—13. [Plungjan, V. A., Rezhnikova, T. I., Sichinava D. V. (2005). *Natsionalnyj corpus russkogo jazyka: obshchaja kharakteristika. Nauchno-tekhnicheskaja informats*]

Say, S. (2020). BivalTyp: Typological database of bivalent verbs and their encoding frames. https://www.bivaltyp.info. Accessed on 22 February 2024.

Taylor, A., Warner, A., Pintzuk S., and Beths, F. (2003). *The York Toronto Helsinki Parsed Corpus of Old English Prose (YCOE)*. Department of Linguistics, University of York. Oxford Text Archive, first edition. http://wwwusers.york.ac.uk/~lang22/YcoeHome1.htm. Accessed on 22 February 2024.

Zanchi, C. (2021). The Homeric Dependency Lexicon. What it is and how to use it. Journal of Greek Linguistics 21(2):263–297. https://hodel.unipv.it/hodel-res. Accessed on 22 February 2024.

Zeige, L. E., Schnelle, G., Klotz, M., Donhauser, K. Gippert, J., and Lühr, R. (2022). Deutsch Diachron Digital. Referenzkorpus Altdeutsch. Humboldt-Universität zu Berlin. http://www.deutschdiachrondigital.de/rea/. DOI https://doi.org/10.34644/laudatio-dev-MiXVDnMB7CArCQ9CABmW. Accessed on 22 February 2024.

8. Acknowledgments

Research for this paper and for the creation of PaVeDa has been supported by European Union funding – NextGenerationEU – Missione 4 Istruzione e ricerca - componente 2, investimento 1.1" Fondo per il Programma Nazionale della Ricerca (PNR) e Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN)" progetto 20223XH5XM "Verbs' constructional patterns across languages: a multi-dimensional investigation" CUP F53D23004570006.