A Valency Lexicon for Universal Dependencies

Petr Kocharov
University of Würzburg
petr.kocharov@uni-wuerzburg.de

Lilit Kharatyan
University of Würzburg
lilit.kharatyan@uni-wuerzburg.de

Abstract

The paper presents a prototype of a web-app designed to automatically generate verb valency lexica based on the Universal Dependencies (UD) treebanks. It offers an overview of the structure of the app, its core functionality, and functional extensions designed to handle treebank-specific features. Besides, the paper highlights the limitations of the prototype and the potential of its further development.

1 Introduction

A prototype of the web-app "UDVaL: Valency Lexicon for Universal Dependencies" (see supplementary demo files[1]) is designed to build valency lexica of languages provided with treebanks, the morphosyntactic annotation of which is stored in the CoNNL-U format and follows the guidelines of the Universal Dependency (UD) project[2]. UDVaL is instrumental for research on a wide range of topics related to verbal morphosyntax, since it provides corpus data on valency and valency alternations of individual verbs and verb classes of a given language within a standardized, typologically oriented annotation framework of UD (De Marneffe et al., 2021).

The app is a derivative of the "CAVaL: Classical Armenian Valency Lexicon"[3], which in turn closely follows the structure and functionality of other corpus-driven valency lexica, in particular, "IT-VaLex" for Latin (Passarotti et al., 2016) and "HoDeL: Homeric Dependency Lexicon" for Ancient Greek (Zanchi, 2021), which in turn rely on the model of "PDT-Vallex" for Czech (Hajic et al., 2003).

UDVaL offers a search engine that supports flexible queries on verb frames based on the combinations of morphological, syntactic and lexical properties of verbs and their dependencies, as specified in section 2. This functionality significantly facilitates research on a wide array of topical linguistic issues, such as valency classes of verbs, valency alternations, alignment, coding of verbal dependencies, etc. (see applications in section 2.3), and contributes to the inventory of digital tools for the linguistic analysis of the rich comparative data accumulated within the UD project (the latest UD release v2.17 includes 339 treebanks of 186 languages).

Unlike general-purpose online tools designed to query dependency treebanks such as PML-TQ[4] and TüNDRA[5], UDVaL is customised for queries on verb frames, and does not require knowledge of formal query languages. The presented prototype version of UDVaL offers a package that can be used for deriving, with more or less adaptations indicated below, a valency lexicon of any language included in the UD database.

The app automatically retrieves verb frames for all verbs of a treebank. A verb frame consists of a verb together with its dependents carrying the grammatical relations of core arguments (subject, direct object, indirect object) and oblique nominals (all other modifiers), and their clausal equivalents (see section 2.1). In line with the UD principles, these types of relations aim at grasping cross-linguistic similarities in abstraction from language-specific overt coding properties of verb frames. With that, a wide-spread distinction between oblique arguments (obligatory or selected by the predicate) and adjuncts (facultative) is abandoned, both types being covered by the same tag "obl", since it is difficult to consistently apply it to the annotation of one language and across languages (see Haspelmath, 2014 on the controversies of this distinction in a cross-linguistic perspective). In some treebanks,

[1] https://github.com/caval-project/ud_val
[2] https://universaldependencies.org
[3] https://github.com/caval-project
[4] https://lindat.mff.cuni.cz/services/pmltq/#!/treebanks
[5] https://weblicht.sfs.uni-tuebingen.de/Tundra

Proceedings of the 8th Workshop on Research in Computational Linguistic Typology and Multilingual NLP, pages 12–17

the distinction is partly formalized by the use of specialized tags "obl:arg" (oblique nominal) and "obl:agent" (oblique agent in passive construction), corresponding to arguments, not adjuncts. Regarding UDVaL, the aforementioned annotation principles have the advantage of providing corpus-driven data, less prone to annotators' bias on whether or not a given oblique dependent is selected by the predicate, while clearly indicating core arguments, which constitute the backbone of verb frames. The distinction between arguments and adjuncts can then be assessed within different theoretical frameworks based on the relative frequency of dependents of specific verbs insofar as the size of a treebank allows it.

2 The UDVaL app

2.1 Backend

UDVaL has a multi-tier architecture. Its backend employs a MariaDB relational database, which is automatically populated with data extracted from a UD treebank of a selected language stored in the CoNNL-U format using parameterised SQL queries and a Python code for dynamic data fetching. The application layer is developed using the Python-based Flask micro web framework to handle application logic and query processing. The frontend layer utilises HTML, CSS, and JavaScript to support a responsive and user-friendly interface. Advanced database indexing and query optimisation techniques have been implemented to facilitate complex queries and ensure high performance.

The interface supports search queries to verb frames and generates complete lists of their occurrences in the database.

The core of the interface functionality is to query constructions with a verbal head and a subset of its immediate syntactic dependents as well as second-order coordinated dependents linked by the "conj" relation.

The presented app prototype only processes predicative constructions headed by verbs (UPOS tag "VERB"). The backend database indexes and stores all verbs of a treebank together with their dependents, carrying the following grammatical relations (and their subtypes):

- nominals: "nsubj" (nominal subject), "obj" (direct object), "iobj" (indirect object), "obl" (oblique modifier), as well as "nmod" (nominal modifier) in the case of dependents of nominalised heads;
- clausal dependents fulfilling the function of core arguments: "csubj" (clausal subject), "ccomp" (clausal complement without an obligatorily controlled subject), "xcomp" (clausal complement with an obligatorily controlled subject);
- auxiliaries of analytic verb forms carrying the relation tag "aux".

Another feature of the app, determined by the UD annotation guidelines, concerns the morphological expression of nominal dependents. To support the cross-linguistically consistent comparison of grammatical functions of nominals, all adpositions are considered as part of case marking along with inflectional case morphology in UD (see Haspelmath, 2019 on the comparable grammatical status of case markers and adpositions). To account for this annotation feature, nominals are stored in the UDVaL database together with adpositions. This allows making queries to the morphological expression of nominal dependents in terms of attested combinations of the inflectional case (morphological tag "Case") and adpositions linked to the nominals by a grammatical relation "case". In case of multiword adpositions (such as English "according to", cf. the UD English-GUM treebank of the UD database under fn. 2), the constituents linked by the "fixed" relation are included in the encoding pattern.

2.2 Frontend

Some key functionalities of the search engine and the user interface are summarised below.

By default, the interface lists all the verbs which are attested in a treebank. The verbs are linked to pages with complete sets of their occurrences in the treebank. By modifying search parameters, the user restricts the list of verbs and occurrences. Besides selecting a verb from the list, it can be found via a free input field "Search by verb". A string of characters in the utf-8 format matching a verb lemma (or several homonymous lemmas) restricts the verb list. The default and restricted lists can be arranged in alphabetical order or by token frequency (ascending and descending).

A valency frame can be configured by morphological, syntactic and lexical properties of an open number of dependents of a verbal head insofar as they co-occur in the treebank. These search parameters can be applied to a list of verbs as well as to a specific verb.

By default, the verb frame query contains a set of the following three parameters for one dependent (see Figure 1):

- "Select relation": the grammatical relation tag from the list specified in section 2.1;
- "Select encoding": the inflectional and/or adpositional case marking;
- "Select lemma": a lemma from the list of lemmas filling the dependent in a given treebank.

In the second of these parameters, encoding patterns are given as a closed list of values, automatically generated from the treebank data, in the format "Case + Adp(s)", regardless of the linear order of constituents, to facilitate the value selection. For clausal dependents, the value of this parameter is conventionally set to "Clause" for tokens with the relations "csubj" and "ccomp" and to "Nominal" for the tokens with the relation "xcomp".

Each selected parameter or its reversal to the default value dynamically updates the verb list and lists of occurrences. When at least one of the parameters is selected, a new set can be added for the next dependent. Sets can be added or removed and their parameters specified until the output list of verbs or occurrences is empty.

The flexibility of interface allows to configure queries of increasing complexity suitable for specific research tasks. Disentangling the morphological, syntactic, and lexical tiers enables queries, in which, for example, a syntactic feature is specified for one dependent and a morphological one for another dependent. The dynamic update of the output with frequency data allows to instantly determine the availability and relative frequency of frames or their specific features in the underlying treebank.

A page with occurrences of a verb includes a complete list of its attestations in the treebank. Every occurrence consists of a sentence, its reference id in the treebank (based on the compulsory "text" and "sent_id" comment fields of the CoNNL-U format), and a morphosyntactic BRAT-based[6] visualisation of the verb frame. The visualisation includes part-of-speech attributes of all tokens and syntactic relations that constitute a verb frame specified in a search query. Besides, the visualisation is provided with the mouse-over glossing of all words in the sentence. The UDVaL engine supports conversion of the UD annotation tags to the Leipzig glossing conventions[7].

By default, the user interface utilises an alphabetic selector, which is automatically generated based on the initial characters of verb lemmas. This functionality only applies to treebanks of languages with alphabetic or syllabic writing systems, the inventory of initial characters of which is manageable within the selector field of the user interface and can therefore be seamlessly integrated into the app.

2.3 Applications

UDVaL inherits all key functionalities of the online CAVaL app[8], has been approbated while assembling corpus data for the ongoing research on Classical Armenian morphosyntax, in particular, as part of the "PaVeDa: Pavia Verbs Database" project[9].

In particular, it provides corpus data on the encoding of core arguments and alignment of Classical Armenian, which involves the split marking of all core arguments (Kölligan, 2013; Müth, 2014). For example, it provides instant access to all the occurrences of perfect tenses with the genitive subject (33x, incl. 16x in a transitive construction) and the nominative subject (189x, incl. 3x in a transitive construction) in the Gospels. The emerging mismatches between transitivity and genitive flagging contribute to the ongoing discussion on the role of animacy and affectedness of the subject in this morphosyntactic split.

Similarly, the app provides corpus data on the split marking of direct object by a bare accusative and a prepositional phrase built with proclitic *z* plus the accusative, with 1289 and 3124 occurrences in the Gospels, respectively. This data suggests that a more frequent marking pattern, associated with a referentially prominent direct object, is morphosyntactically more complex, again a cross-linguistic generalization suggested in (Haspelmath, 2021). Within the same subcorpus, the ditransitive construction occurs 181 times with the prepositional marking of direct object, and 299 times with an accusative marking. This data points to the tripartite ditransitive alignment: P (*z* plus Acc), T (Acc), and R (Dat) are expressed by different majority encoding types. Such ditransitive is typologically rare (Haspelmath, 2005).

[6] https://brat.nlplab.org/embed.html

[7] https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf

[8] https://caval.dch.phil-fak.uni-koeln.de

[9] https://paveda.unipv.it/contributions/clas1256

Figure 1: Adaptation of the UDVaL app for the Arabic-PADT UD treebank (v2.17).

2.4 Functional extensions

The core functionality of UDVaL, summarised above, can be supplemented by secondary functions, which were developed for the UD treebank of Classical Armenian and can be applied with due modifications to subsets of UD treebanks. By default, these extensions are included in the supplemented demo codes as customised for Classical Armenian. Adjustments of the app to other treebanks with respective features are facilitated by internal comments of the demo codes.

Latinized input: In order to facilitate the search of verbs via the free input field "Search by verb", the code allows setting up latinized input conventions. This functionality requires extending the code with a chart of equations between original and input characters or groups of characters.

Transliteration: Based on the "Translit" and "LTranslit" (wordform and lemma transliteration, respectively) attributes of the MISC field for miscellaneous information of the CoNNL-U format, the UDVaL search engine supports representation of verb lists and occurrences in the transliterated form. The implementation of this feature requires adding a chart of correspondences between the original and transliterated characters, and integrating these correspondences with the alphabetic selector.

Translation: The UDVaL engine allows searching verbs by meaning provided that a treebank has the "Gloss" attribute in the MISC field; the attribute contains an approximate translation of the word form or lemma to another language, typically but not exclusively English.

A translation of sentences can be switched on/off in the lists of occurrences for treebanks with sentential translations. In the demo code, this feature is based on the "# translated_text" comment field of the CoNNL-U format, and can be adjusted to the metadata properties of a given treebank.

External lexicographic sources: Given the necessarily approximate lexicographic information stored in the "Gloss" attribute (when available), the UDVaL engine allows linking verb lemmas to external online lexicographic sources. The code enriches the database with external links based on a txt-file with a list of verb lemmas and corresponding URLs of the external lexeme entries.

Subcorpus selector: If relevant for a given treebank, a selector of subcorpora or individual texts can be integrated into the search interface to facilitate research of morphosyntactic variation across texts. This feature has proven to be useful for the study of Classical Armenian verb frames in the CAVaL implementation of the app. For example,

while 14,5% of verbs are attested in the Gospels with a dative argument, a post-classical text, "History of the Armenians" by Movses Khorenatsi, the precise dating of which is debated, has 18,2% of such verbs. This data is relevant for the study of the gradual increase of dative marking towards Modern Armenian (Daniel and Khurshudian, 2015). In the demo code, this feature is based on the "# sent_id" comment field of the CoNNL-U format, and can be adjusted to the metadata of a given treebank.

Verb features: For treebanks with annotated morphological features, the verb frame query can be further extended with verbal features insofar as they are attested in the treebank for a specified verb frame. This functionality requires adjusting the code to integrate the language-specific list of verbal features into the search interface.

Limitations

Insofar as the CAVaL implementation allows to judge, the retrieval accuracy of the UDVaL engine, based on deterministic database queries, is exact (cf. the frequency data on the co-occurrence of tags in the Classical Armenian treebank of the release UD v2.17 and in the online app under the link in fn. 3). Any perceived errors must be attributed to the quality of the underlying treebank annotations and evaluated as such.

The presented prototype of the UDVaL app requires manual adjustments of the code to tailor functional extensions, mentioned in section 2.4, to available features of a given treebank. Subsequent iterations of the app envisage automated detection of treebank features and integration of respective search functions into the user interface.

The current version of the app is limited to the queries of verb frames with a verbal head and a restricted subset of types of syntactic dependents, listed in section 2.1. This limitation excludes queries to nominal predicates, polipredicative constructions, clausal adverbial modifiers, etc., that are relevant for the study of valency. Thus, the app does not allow the comparison of case frames and their alternations for verbal and nominal predicates.

Being oriented towards typological comparative studies, the app does not allow integrating multiple treebanks into one interface. This limitation points to the potential of further development of the app to implement functionality that would support instant comparison of verb frames across treebanks of different languages within one search interface.

The response time of the app is currently coupled with corpus size: while interaction remains efficient for smaller datasets, integration of larger corpora yields a noticeable increase in latency. Future iterations will address this scalability bottleneck by refactoring the data-access layer around an Object–Relational Mapping (ORM) approach, to reduce response time, especially for complex frame queries, and enable seamless interaction with substantially larger corpora.

Ethical Considerations

The publication complies with the ACL Ethics Policy.[10] In particular, neither part of the presented technology violates the license permissions of reuse for non-commercial purposes.

Acknowledgments

This research is part of the Classical Armenian Valency Lexicon project, funded by the Deutsche Forschungsgemeinschaft (DFG), project number 518003859. We are grateful to two anonymous reviewers for their valuable comments.

References

Michael Daniel and Victoria Khurshudian. 2015. *14. Valency classes in Eastern Armenian*, pages 483–540. De Gruyter Mouton, Berlin, Boston.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Jan Hajic, Jarmila Panevová, Zdenka Urešová, Alevtina Bémová, Veronika Kolárová, and Petr Pajas. 2003. Pdt-vallex: Creating a large-coverage valency lexicon for treebank annotation. In *Proceedings of the second workshop on treebanks and linguistic theories*, volume 9, pages 57–68.

Martin Haspelmath. 2005. Argument marking in ditransitive alignment types. *Linguistic Discovery*, 3:1–21.

Martin Haspelmath. 2014. Arguments and adjuncts as language-particular syntactic categories and as comparative concepts. *Linguistic Discovery*, 12:3–11.

Martin Haspelmath. 2019. Indexing and flagging, and head and dependent marking. volume 62, pages 93–115.

[10] https://www.aclweb.org/portal/content/acl-code-ethics

Martin Haspelmath. 2021. Role-reference associations and the explanation of argument coding splits. *Linguistics*, 59(1):123–174.

Daniel Kölligan. 2013. *Non-canonical subject marking: Genitive subjects in Classical Armenian*, pages 73–90. John Benjamins Publishing Company.

Angelika Müth. 2014. *Indefiniteness, animacy and object marking: A quantitative study based on the Classical Armenian Gospel translation. PhD Thesis*. University of Oslo, Oslo.

Marco Passarotti, Berta González Saavedra, and Christophe Onambele. 2016. Latin vallex. a treebank-based semantic valency lexicon for latin. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2599–2606.

Chiara Zanchi. 2021. The homeric dependency lexicon: What it is and how to use it. *Journal of Greek Linguistics*, 21(2):263–297.