Classical Armenian Valency Lexicon

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

Petr Kocharov, Lilit Kharatyan

Julius-Maximilians-Universität Würzburg
{petr.kocharov, lilit.kharatyan}@uni-wuerzburg.de

livestream >>

Objective

CAVaL: Classical Armenian Valency Lexicon is an online tool for the study of Classical Armenian predicate-argument structures. It enables access to corpus-driven evidence for verbs and their arguments/adjuncts attested in a treebank of Classical Armenian texts annotated according to the Universal Dependencies (UD) guidelines (de Marneffe et al. 2021). A flexible search interface of CAVaL allows to specify morphological, syntactic, and lexical properties of verbal dependencies. The project continues an array of online resources for the study of valency such as *The Homeric Dependency Lexicon* (Zanchi 2021) and digital tools for Armenian (Avetisyan & Broneske 2023).

Online interface

The search interface allows to access occurrences of a verb by typing the infinitive or 1 sg. pres. verb form ① (in Armenian or using a latinized input) or its meaning ② into respective search fields, or by selecting it from the list of verbs. By default, the interface lists all the verbs attested in the CAVaL treebank. The list can be modified using the text selector ③ and alphabetic selector ④, or by specifying dependency parameters. One can sort ⑤ any list by alphabet or frequency. The Armenian spelling of the output can be switched to transliteration ⑥.

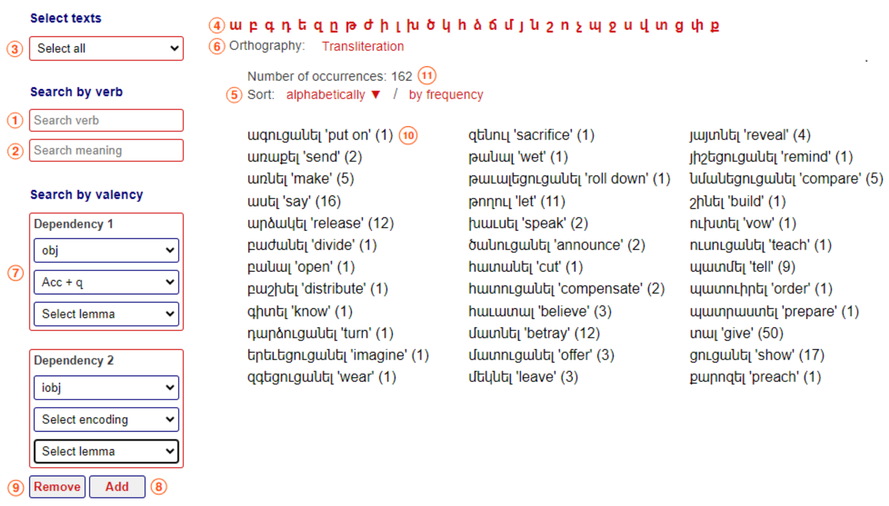

Figure 1: Searching verbs

By default, the interface has a set of parameters ⑦ (relation, encoding, lemma) for one dependency. If at least one of the three parameters is selected for a dependency, a new set can be added ⑧ for another dependency until the verb list is empty; added sets can be removed ⑨. Each selected parameter or its reversal to the default value dynamically updates the verb list, token frequencies ⑩ of individual verbs, the overall number of occurrences ⑪, and the remaining values of parameters in the open set(s).

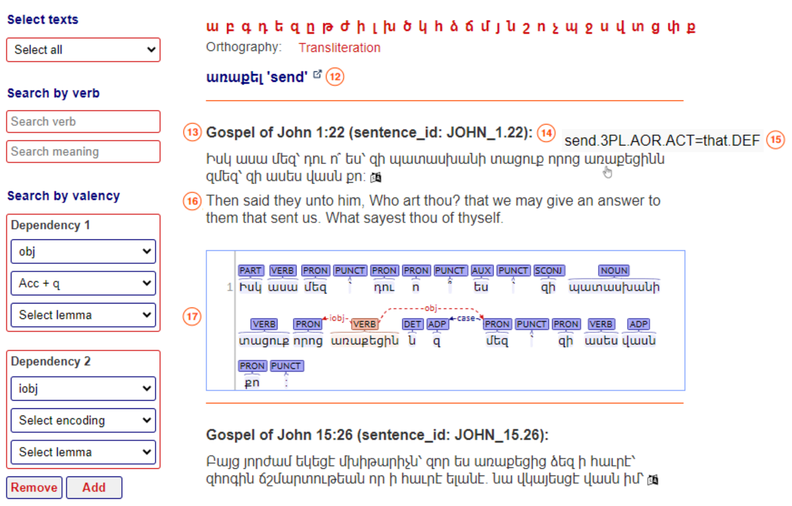

Figure 2: Browsing occurrences

Selecting a verb leads to a page with its occurrence in the CAVaL treebank. The verb is provided with a link to the online Calfa dictionary of Classical Armenian ⑫. Every occurrence contains a citation source ⑬ and sentence ID ⑭ in the UD treebank. Sentences have mouseover grammatical glossing ⑮ and English translations ⑯. The brat-based visualization ⑰ of sentences highlights the selected verb with its nominal, clausal, auxiliary dependencies.

Dataset

CAVaL retrieves data from the latest version of the UD Classical Armenian-CAVaL treebank. The treebank of the UD release 2.15 (15.11.2024; ca. 87.500 tokens) includes the Classical Armenian translation of the Gospels and the first ten chapters of the *History of Armenians* by Movses Khorenatsi.

Database structure

The backend of CAVaL is a relational database, which is automatically generated from the UD treebank using parameterized SQL queries and dynamic data fetching. The database indexes all personal verb forms of the treebank based on the ”VerbForm” attribute in the morphological layer of the UD annotation and dependent auxiliaries. Verbs are indexed together with their nominal and clausal dependencies tagged by relations nsubj, csubj, obj, ccomp, iobj, obl, and their subtypes. In line with the UD framework, the encoding of arguments/adjuncts includes case forms of nominals and accompanying adpositions. The syncretic case forms are disambiguated in the CAVaL treebank, which allows to describe encoding patterns by one of the seven cases (Nom, Acc, Dat, Gen, Abl, Loc, or Ins) with a possible addition of adposition(s). These morphosyntactic parameters are used as filters in the online search interface.

The advantage of the CAVaL search interface is that it disentangles the syntactic, morphological and lexical properties of verbal dependencies, and allows to apply combinations of these parameters to configure predicative structures of increasing precision, tailored to specific research needs, insofar as they are attested in texts.

Applications

CAVaL is suitable for many research purposes relevant for the grammar of Classical Armenian, historical and typological linguistics. In particular, it provides corpus data on default valency and its alternations for individual verbs, see the Classical Armenian datasets for the *World Atlas of Transitivity Pairs* (Kocharov 2023) and the *Pavia Verbs Database* (Kocharov, https://paveda.unipv.it/contributions/clas1256#tverbs, access date: 19.11.2024). It also helps to assess such major morphosyntactic parameters as the alignment of different types of predicates, as examplified below.

The **Classical Armenian alignment** involves the split flagging of S, A, P, T, and R.

Split S/A: S and A are flagged by Nom, except for the analytic perfect tenses, where A is typically flagged by Gen (Kölligan 2013; Meyer 2023). Occasionally S is also flagged by Gen, which leads to a question whether the flagging is conditioned by syntactic transitivity or lexical properties of the first argument such as agentivity or affectedness.

→ CAVaL: Dep1 = relation: nsubj, encoding: Nom; Dep2 = relation: aux — 189x
Dep1 = relation: obl:arg, encoding: Gen; Dep2 = relation: aux — 36x

Split P/T: Referentially prominent P is flagged by a proclitic particle *z* added to Acc, whereas non-referential P is typically coded by bare Acc (Müth 2014; Scala 2022).

→ CAVaL: Dep1 = relation: obj, encoding: Acc — 992x
Dep1 = relation: obj, encoding: Acc + z — 2085x

Split R: A recipient-like participant can either be coded by Dat or by a prepositional construction *c‘* plus Acc (Kölligan 2016). The later type is found in a few verbs including prototypical ditransitive verbs. It is the majority encoding for the addressee of *asel* ’to say’ (CAVaL: Acc + c‘: 683x vs. Dat: 250x) and is common for the recipient of *tal* ’to give’ (CAVaL: Acc + c‘: 21x vs. Dat: 277x).

→ CAVaL: Dep1 = relation: iobj, encoding: Dat — 970x
Dep1 = relation: obl:arg, encoding: Acc + c‘ — 732x

The alignment is further complicated by the case syncretism: Nom coincides with Acc in the singular of all nominals except the 1st and 2nd singular pronouns, but differs from it in the plural; Dat coincides with Gen in the majority of nominals. CAVaL significantly facilitates evaluation of the relative frequency of competing patterns, and availability of specific alignment patterns for individual verbs.

Monotranslitive alignment: the Split P in conjunction with the Nom/Acc syncretism results in competing accusative and neutral alignment. According to CAVaL, the majority type is accusative. In the perfect, the combination of Split S/A, Split P and the Nom/Acc syncretism yields competing accusative, nominative, tripartite and ergative alignments. According to CAVaL, the tripartite and ergative ones are most common.

Ditransitive alignment: The diversity of ditransitive patterns results from the combinations of Split S/A, Split P/T, Split R, and the sincretism of Nom/Acc and Dat/Gen. If one defines the basic ditransitive pattern with respect to the most frequent flagging of P (Acc + z) of the monotransitive accusative type, the ditransitive tripartite alignment (P = Acc + z : T = Acc : R = Dat), which is typologically rare (Haspelmath 2005), must be postulated as the default one for Classical Armenian.

Outlook

CAVaL is subject to further improvement in terms of the scope of corpus and functionality. Its successive iterations will include a larger set of the 5th century texts, parallel to the extensions of the UD Classical Armenian-CAVaL treebank. The search interface will be extended with a possibility to specify morphological features of the verb such as tense, aspect, and mood based on the morphological data of the UD treebank. The CAVaL techonology can be further applied to any UD treebank to build, with some language-specific adjustments, a valency lexicon of a given language.

References

Avetisyan, H., and Bronseske, D. 2023. Large Language Models and Low-Resource Languages: An Examination of Armenian NLP. In: *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023*: 199–210.

Haspelmath, M. 2005. Argument marking in ditransitive alignment types. *Linguistic Discovery* 3/1: 1-20.

Kölligan, D. 2013. Non-canonical subject marking. Genitive subjects in Classical Armenian. In: I. Seržant, L. Kulikov (eds.), *The Diachronic Typology of Non-Canonical Subjects.* Amsterdam: Benjamins: 73–89.

Kölligan, D. 2016. Valenzalternation im Klassisch-Armenischen: Dativkomplemente vs. Präpositionalphrasen. *Österreichische Linguistiktagung.* DOI: 10.5281/zenodo.11176018.

Kocharov, P. 2023. Anticausatives in Classical Armenian. *Journal of Historical Linguistics* 13/2: 255–294.

De Marneffe, M.-C., et al. 2021. Universal Dependencies. *Computational Linguistics* 47/2: 255–308.

Meyer, R. 2023. *Iranian Syntax in Classical Armenian: The Armenian Perfect and Other Cases of Pattern Replication*. Oxford University Press.

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

Scala, A. 2022. Sulla marcatura differenziale dell’oggetto in armeno classico: una questione di definitezza? *REArm* 41: 15–29.

Zanchi, Ch. 2021. The Homeric Dependency Lexicon: What it is and how to use it. *Journal of Greek Linguistics* 21/2: 263-297.

Acknowledgements

The *CAVaL: Classical Armenian Valency Lexicon* is funded by the Deutsche Forschungsgemeinschaft. DFG