Proceedings of the Workshop on Processing Language Variation: Digital Armenian (DigitAm 2022) @LREC2022, pages 28–37
Marseille, 20 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

Eastern Armenian National Corpus: State of the Art and Perspectives

Victoria Khurshudyan, Timofey Arkhangelskiy, Michael Daniel, Dmitri Levonian, Vladimir Plungian, Alex Polyakov, Sergei Rubakov
INALCO/SeDyL/CNRS/IRD, Universität Hamburg, School of Linguistics / Linguistic Convergence Laboratory, HSE University, Corpus Technologies, Russian Academy of Sciences, Corpus Technologies
65 rue des Grands Moulins, 75013 Paris
victoria.khurshudyan@inalco.fr, timofey.arkhangelskiy@uni-hamburg.de, misha.daniel@gmail.com, dlevonian@gmail.com, plungian@gmail.com, pollex@mail.ru, rubakov@gmail.com

Abstract
Eastern Armenian National Corpus (EANC) is a comprehensive corpus of Modern Eastern Armenian with about 110 million tokens, covering written and oral discourses from the mid-19th century to the present. The corpus is provided with morphological, semantic and metatext annotation, as well as English translations. EANC is open access and available at www.eanc.net.

Keywords: Armenian, corpus linguistics, annotation

1. Introduction

Corpus linguistics (McEnery and Wilsonn, 2001; 2012; Stefanowitsch, 2020; i.a.) started to develop actively only from 1990-ies with the evolution of new technologies facilitating the compilation and processing of different types of corpora. Corpus linguistics is based on empirical data and reflects the language reality throughout all forms of language production.

In corpus linguistics a corpus is defined as a set of texts, and a reference corpus as a balanced and representative set of texts (written discourse) and/or transcripts (oral discourse) varied by different parameters (genre, chronology, original and translated literature etc.), provided by various types of annotation (metatextual, morphological, syntactical etc.) and searchable by various linguistic or pragmatic criteria.

Despite being a language with a multisecular written tradition, Armenian[1] is an under-resource language and it lacks significantly digital resources for Natural language Processing (NLP) and linguistic research. Several rare projects for particular Armenian varieties exist, as well as a growing interest in NLP resources is observed.

General purpose untagged Armenian plain texts are represented by a number of open-access online libraries in the Internet offering mainly fiction and press (for a more detailed overview on the existing resources for different Armenian varieties see (Vidal-Gorene et al., 2020)). Often the available data are merely scanned rather than OCRed.

At the time of Eastern Armenian National Corpus (EANC) launching (2006) the availability of Modern Eastern Armenian (MEA) data was quite inadequate with only few e-libraries offering popular fiction with an estimated total volume of about 1 million words. In the available open online resources, non-fiction genres (except press) were often missing. MEA press enjoys better online representation mostly due to online editions of a number of popular Armenian newspapers.

More recently MEA project of Universal Dependencies provides a treebank of about 50K tokens (2502 sentences) with morphological and syntactic annotations in the form of a dependency tree bank (Yavrumyan, 2020; Yavrumyan and Danielyan, 2020).

Currently, several other resources provide MEA plain-text or scanned databases (Armenian Wikipedia and Wikisource (about 50M tokens), Fundamental Scientific Library of the National Academy of Sciences of the Republic of Armenia[2] (considerable number of scanned books of different genres as well as press archives), etc.). Rare tools such as spellcheckers and orthography converters exist for the two modern standards. More recently, some NLP research projects have been conducted to address particular NLP issues, such as named entity recognition (Ghukasyan et al., 2018), word embeddings (Avetisyan and Ghukasyan, 2019) or paraphrase detection for Armenian (Malajyan et al., 2020).

Russian National Corpus[3] provides an aligned sub-corpus of MEA and Russian on the basis of the translated texts existing in EANC. The sub-corpus is

[1] The Armenian language in all its variation encompasses Classical Armenian (5th-10th cen. A.D), preserved exclusively for canonical uses, Middle Armenian (11th-17th cen.), and Modern Armenian (17th cen. – up to present) with its two standards: Modern Eastern Armenian (the official language of the Republic of Armenia, which is also the language of the Armenian communities of Iran and the ex-Soviet republics) and Modern Western Armenian (spoken by traditional Armenian communities in Europe, the Americas and the Middle East originating mainly from the Ottoman Empire), both standardized in the 19th cen. Aside from the two standards, the Armenian language continuum includes various dialects, as well as vernacular forms. All the written varieties of the Armenian language use the unique Armenian alphabet.

[2] https://arar.sci.am/

[3] https://ruscorpora.ru/new/search-para.html?lang=hye

provided with full morphological annotation for both languages and it covers about 2,4M tokens.
In contrast to the written discourse, MEA oral data is rarely available for research. During the last years several projects elaborating MEA Automatic speech recognition (ASR) models[4] came out.
As of today, EANC is the largest Armenian resource.

1. EANC Overview

The project of Eastern Armenian National Corpus (www.eanc.net) was launched in 2006 (the current version corresponding to the third release as of 2009) by a group of linguists and it was supported by Corpus Technologies, a Moscow-based NLP development company.
EANC is designed as a comprehensive corpus with about 110 million tokens, covering Modern Eastern Armenian written and oral discourses from the mid-19th century to the present. The texts/transcripts have morphological, semantic and metatext annotation and they are provided by English translations for frequent tokens searchable for making complex lexical morphological queries. EANC is an open access corpus available at www.eanc.net. EANC proposes also an electronic library (scanned and processed entirely by the EANC team) with full-view access for over hundreds of works by classical authors in public domain. The library provides the same morphological analysis and translation as the rest of the corpus (displayed on mouse click). Due to copyright considerations, the search function in the main corpus does not provide access to the texts in their entirety.
The term "national", included in the name of EANC, has a terminological rather than emotional value. After British National Corpus[5], the concept of a "national corpus" has come to designate a comprehensive and representative corpus of a language: cf. Russian National Corpus[6], Czech National Corpus[7], Georgian National Corpus[8], among others. It is in this sense that the Eastern Armenian National Corpus qualifies as a national corpus of a language.

2. EANC Composition

EANC is designed as a comprehensive corpus with the objective to include as many MEA texts as practicable – all literary, scientific and oral texts available to us have been indexed for search. The only exception to this is certain widely-available texts, such as electronic press and legal documents, whose presence has been limited for the sake of balance among different genres.
Due to its comprehensive nature, EANC is inherently different from the high-resource languages' corpora such as Russian National Corpus or British National Corpus which choose their collections selectively. BNC additionally imposes a limit on the number of words per document, truncating longer texts. EANC, on the other hand, includes a great majority of all extant Eastern Armenian literary texts. In this respect, EANC is similar to Czech National Corpus, Slovak National Corpus[9] or Georgian National Corpus.
The vast majority of EANC written texts except press are obtained through scanning and OCRing scanned materials using ABBYY Fine Reader 8.0. Most of the EANC press corpus was downloaded from open electronic archives of the newspapers that provide access to such archives (e.g. www.azg.am, www.aravot.am, www.yerkir.am, www.iravunk.com etc.).

Written discourse	# tokens	% EANC	# of docs
Fiction			
prose: novels	29 729 521	27,1%	366
prose: short stories	5 888 695	5,4%	158
prose: plays	1 411 030	1,3%	55
prose subtotal	**37 029 246**	**33,7%**	**579**
poetry	3 627 119	3,3%	208
Press	47 264 735	43,0%	7858
Non-fiction			
science	13 750 358	12,5%	112
essays, memoirs, official, religious	4 680 539	4,3%	360
Written total	**106 351 997**	**96,8%**	**9 117**
Oral discourse	**# tokens**	**% EANC**	**# of docs**
Oral spontaneous discourse	1 029 646	0,94%	208
Oral public discourse	1 933 899	1,76%	543
Oral task-oriented discourse	70 010	0,06%	22
+ Online communication	442 399	0,40%	1
Oral total	**3 475 954**	**3,2%**	**774**
EANC Total	**109 827 951**	**100%**	**9 891**

Table 1: EANC composition by genre

About 1 million tokens of texts have been downloaded from public electronic collections (www.armenianhouse.org, www.hayeren.hayastan.com etc.).

EANC includes written texts of various genres (over 106M tokens) such as fiction, press, poetry, non-fiction, etc., as well as a diversified corpus of oral speech (about 3,5M tokens) (see Table 1).
EANC includes not only all school reading texts in today's Armenian secondary school program, but the vast majority of MEA classical literature starting from mid-19th century, a large number of scientific texts (including the 13-volume Armenian Soviet Encyclopaedia 1974-1987).

[4] MEA ASR model by Public initiative for national acceleration (https://arm.ican24.net/demoasrv4.html), Mozilla common voice project for MEA (https://pontoon.mozilla.org/hy-AM/common-voice/), MEA ASR model integrated in Google translate (https://translate.google.com/?hl=hy&sl=hy&tl=la&op=translate), Sonix's MEA model https://sonix.ai/languages/transcribe-armenian-audio, https://hindityping.info/speech-to-text/armenian/.
[5] www.natcorp.ox.ac.uk
[6] www.ruscorpora.ru
[7] https://ucnk.ff.cuni.cz
[8] http://gnc.gov.ge
[9] https://korpus.sk

Each of the 9,960 document entries in EANC is labeled by metatext information specifying genre and other bibliographic details (e.g. date of creation/publication, name of the author, etc.).

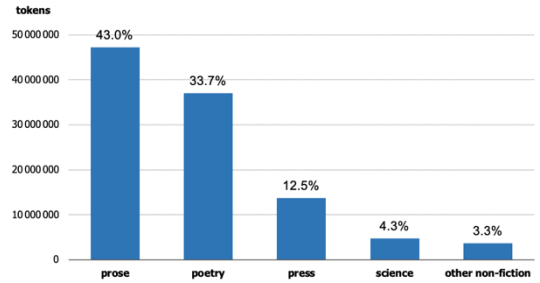

Figure 1: EANC written discourse composition

Written discourse sub-corpus in EANC includes over 106 million tokens covering over 500 authors, a sizeable collection of press, scientific and other non-fiction texts, as well as some 130 translated texts.
Various genres of EANC texts are distributed unevenly over time. The 19th and 20th centuries are mostly represented by literary texts, prose and poetry. Some older press has been added to the corpus in a joined project by EANC and the Armenian National Library to render the press sub-corpus more balanced chronologically. The main bulk of the press sub-corpus, however, was acquired by downloading texts from open newspaper archives and thus represents the modern (from 2000 on) language of internet news resources of the Republic of Armenia. This makes the ratio between press and fiction texts for the last decade very different from the same ratio for the rest of the corpus.
Oral discourse sub-corpus (about 3,5M tokens) is an important part of EANC represented by spontaneous dialogs, polylogs, task-oriented interviews, TV talk shows, movies, and other recordings, all transcribed by EANC.
The oral discourse sub-corpus of the EANC being a linguistic and corpus project on its own, it is impertinent to implement any balance restrictions controlling the proportion of written vs. oral discourse (unlike within the written corpus where a reasonable balance is required between various genres and types of texts).

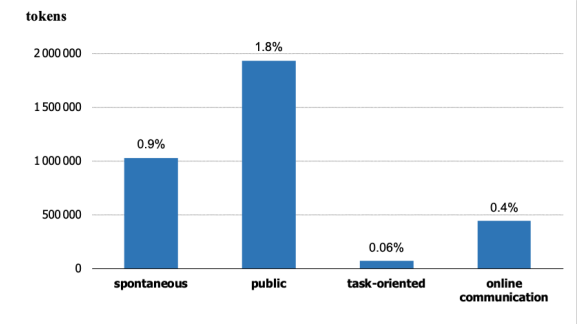

Figure 2: EANC oral discourse composition

Oral discourse in EANC is represented by the Yerevan standard which is justified by the fact that it is the closest spoken dialect to Modern Eastern Armenian, the language of the written sub-corpus and which historically served as a spoken prototype for the MEA literary tradition. The entire oral discourse corpus has been recorded and transcribed within the framework of the EANC project.
Oral public discourse (about 2M tokens) is originally recorded in video format and includes various recordings of TV programs, talk shows, public debates, interviews, etc. broadcasted by Armenian TV stations. Audio data are then extracted and stored as audio files. Oral spontaneous discourse and task-oriented discourse are recorded in audio format (.mp3 or .wav). The respondents are speakers of the Yerevan standard and are selected in an attempt to obtain a balanced mix of age, gender, and social status. The corpus of oral spontaneous discourse (over 1M tokens) includes spontaneous polylogues, dialogues and diverse narratives. The corpus of task-oriented discourse (about 70,000 tokens) covers favorite film narratives and cartoon narratives.
Currently, EANC oral discourse corpus uses a *plain* transcription which basically follows traditional Armenian orthography and punctuation standards. Only few additional special tags are used: == for falsestarts, = for fragmented words, among other tags.

3. Annotation and Grammatical Wordlist EANC Composition

All the annotation information enhances the EANC search capability by allowing the user to build and search sub-corpora and to sort the search results. Three major layers of markup are implemented in EANC:
1. Metatext (bibliographic) markup is assigned to each text unit and includes such metatext information as author, title, year of creation, and genre (genres) etc.
2. Token markup includes lexical and morphological markup assigned to over 90% of tokens as well as English translations for about 85% of tokens. Every token (wordform) in EANC is supplied with a set of lexical morphological tags (labels). These tags cover grammatical categories applicable to MEA (part of speech, case, number, determination, tense-aspect-mood, polarity, inflection type etc.). EANC tagging system follows the Leipzig Glossing Rules (as of 2015)[10] as closely as possible (see Annexe 1: EANC grammatical tags). Few solutions have been made that may appear controversial, but these are mainly connected to controversial or understudied phenomena in Armenian grammar itself (such as interpretating dative / genitive and destinative / dative infinitive syncretism or the morphological composition of relational forms of nouns, such as *սեղանիիդ*).
3. The third markup layer covers *punctuation, sentence boundaries, and auxiliary markup options*.
Linguistic background of the project comprises two main components – the wordlist and a morphological model (inflectional classification).

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

The EANC wordlist is based on a combination of the wordlist of E. Galstian's Armenian-Russian dictionary (1985) and part of E. Aghaian's dictionary of Modern Armenian (1976) (over 70,000 entries), the wordlist of H. Grgearian and N. Harutyunian's dictionary of geographic names (1987-1989) (about 4,000 entries), a list of common first names and family names (about 1000 entries), abbreviation wordlist from D. Gyurdjinian and N. Hekekian's dictionary of acronyms used in Armenian (2007) (about 2000 entries). Additionally, the EANC wordlist includes a limited number of lexemes, such as neologisms, that occur in EANC but are missing from the sources above. Such lexemes were added manually on the basis of the list of non-annotated words filtered by their frequency in EANC.

To make lemmatization possible a morphological model with a formal and exhaustive classification of MEA inflection types for both nominal and verbal categories was worked out. Each inflectable lexeme in the EANC wordlist was then assigned a specific tag corresponding to the relevant inflection type (e.g. N11, N12, V11, V12 etc.).

Comprising a wordlist and providing an internationally-compatible inventory of morphological categories was mainly a technical task. The main challenge has been to work out a formal morphological model of Modern Eastern Armenian inflection that would be comprehensive enough to cover most of the corpus tokens. In other words, each lexeme that inflects had to be provided with information about its paradigmatic type (or types, in case of inflectional variance) that predicts its forms. This challenge may seem unexpected, provided a long tradition of Armenian studies. However, the conventional grammars of Eastern Armenian proved not to be formal enough for an automatic analysis (lemmatization) of EANC electronic library, which is quite justifiable because conventional grammars serve a purpose other than automatic processing (mostly educational).

By way of example, the current inflectional classification of MEA nouns used in EANC includes 45 types, nine of which could be considered as subtypes and grouped into nine larger classes, which roughly correspond to conventional declensions. Some types are different from others by vowel reduction or, for nouns, plural formation, which cross-cuts the whole system of MEA nominal inflection; some classes are not real declensions, being limited to few lexemes only. The full list of types is available at the project site and covers all or, strictly speaking, all we are currently aware of, types of paradigms that have at least one position that distinguishes it from all other types of paradigms. Similar classifications have been elaborated for pronouns and verbs.

The inflectional classification of MEA applied is based on orthography, and, thus, more of an applied linguistic than a purely linguistic project (although a speech-oriented linguistic classification may be obtained relatively easily).

Figure 3:

One of the challenges has been analysing orthographic variants[11] widespread in MEA texts, including old writings or Western Armenian inserts. The markup was designed to allow to find regular and deviant orthographic variants in one same query, as well as tokens using non-standard orthography.

Supplemented with part-of-speech and inflectional information, the EANC wordlist became a grammatical e-

dictionary, similar to those used by Internet search engines for other morphologically rich languages.

4. Software

EANC database software consists of four major parts: parser, indexer, server and user interface and client.

The collection of raw electronic texts is first processed by EANC *Parser* (a PERL program), which adds XML-compliant or tab-delimited metatext and token markup. Next, the resulting files are processed by the *Indexer* to create the corpus database structure. *Server* implements search and sorting algorithms in the corpus database. Finally, *User interface* and *Client* provide web access to the EANC database and its search functionality.

The EANC Parser assigns token markup tags to each wordform, provided that the respective lexeme is present in the EANC grammatical wordlist. Overall, 92,5% of all tokens are recognized and annotated with 72,6% analyzed unambiguously, 17% ambiguously, and 7,5% not recognized. Parsing success rate varies depending on a genre. The highest percentage of unrecognized tokens occurs, unsurprisingly, in oral discourse.

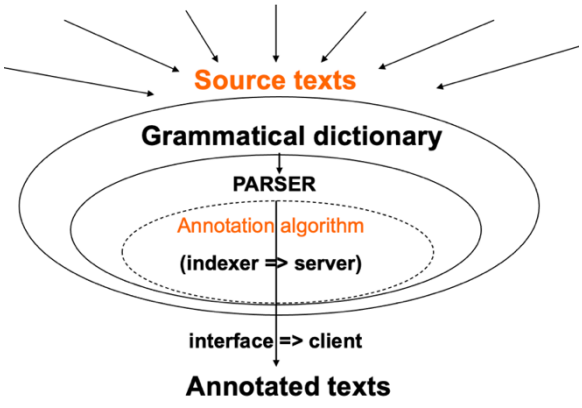

Figure 3: EANC database software

Some wordforms have multiple analysis. For example, the forms for infinitive and perfective converb in MEA are regularly homonymous for the *-l̄* (-e) conjugation

[11] Up to the 1920s both MEA and Modern Western Armenian (MWA) together with Classical Armenian had a common spelling. Once Armenia was sovietised an orthography reform was made with the objective of simplification and rendering it more phonetic, though political reasons were certainly not of last importance either. Currently, classical spelling is used for MWA and by the Armenian community in Iran for MEA, whereas reformed spelling is applied in Armenia and other Eastern Armenian communities.

type (*գրել grel* 'to write'). An example of an occasional homonymy is *հարգի hargi* 'respectable': it is analyzed both as an adjective and as a subjunctive, 3rd person, present of the verb *հարգել hargel* 'to respect'. This lexical morphological homonymy, both regular and coincidental, is quite common in MEA, the overall percentage of tokens with multiple analysis being as high as about 12%. Currently, EANC parser deals exclusively with the wordform, completely ignoring their context. The noise level can be cut down by adding specific constraints to the query, e.g. by introducing another wordform that is supposed to co-occur with the relevant reading.

Indexer is a PHP+MySQL program that extracts address information for each token and each markup element from the XML output provided by the *EANC Parser*. The output of *Indexer* is a set of hash tables that establish a pointer connection between each unique lexeme, wordform and grammatical attribute occurring in EANC, and their respective positions (addresses) in the corpus data files. The corpus data files represent a non-relational database consisting of binary address arrays. Sorting keys for each token are also stored in the data files. This allows sorting output contexts by specific key criteria, such as alphabetically, by period/genre, etc.

Server is a C++ program which implements core search algorithms over the corpus data files via the ISAM method. Search algorithms are designed to minimize response time for most common queries. Given the size of EANC (well over 100M tokens), response time may exceed the standard 0.5-0.8 second threshold for some contextual queries such as searching for complex collocation sequences of frequent gram attributes.

Many queries may correspond to a large number of matches in EANC; however, only up to 10,000 matches are displayed to the user. These 10,000 are drawn from various parts of the Corpus proportionally to the way *all* matches are distributed throughout EANC, so as to form a representative sample (if a sub-corpus has been defined, the same distribution sampling is performed over the sub-corpus).

EANC user interface is a PHP/HTML program that provides user access to the full search functionality of the server. Visually, the user interface is a collection of browser windows, including: Search form appearing on the right side of the EANC web page, gram selection form, sub-corpus selection form, display options form, search output area and a number of auxiliary windows such as virtual Armenian keyboard.

The main search form is the central element of the EANC user interface. It is used to build various types of queries (e.g. for a lexeme or a wordform, gram attributes, punctuation, case-sensitivity etc.).

When the user defines a search query, the user interface transmits that query to *Client*. *Client* is a PHP program that pre-processes user input in the User interface, builds and sends a query to *Server*, and then receives and post-processes the search output. *Client* is also responsible for more advanced interface operations, such as displaying token markup or transliterating the output. The grammatical wordlist of MEA is used by the parser, EANC corpus software that ascribes each token a lexical morphological analysis.

Apart from the parser EANC software is designed as a scalable and a language-independent software platform for corpus studies. The system is built in a way that corpora of structurally different languages can be indexed and made available for search provided that such corpora follow the specific XML markup standards developed by Corpus Technologies (cf. in 2011-2017 EANC software was used for Albanian, Ossetic, Buryat, Mongolian, Kazakh corpora[12]).

Morphological analysis in general can be either rule-based or statistical. In case of statistical analysis certain amount of training data (100,000-1,000,000 words) is annotated manually on which a smart algorithm is trained which finally learns and provides the rules to annotate texts. One of the advantages of this method is the possibility to analyze previously unseen words, thus no dictionary is required. This mode of analysis is popular for large languages and the more fine-grained the tagset, the larger the training dataset is needed.

Figure 4: EANC annotated example

Current EANC morphological analysis is rule-based with manually compiled dictionary and morphological rules that the analyzer applies to the text. Such analysis results in ambiguous analyses since words are analyzed regardless of context and out-of-vocabulary words are not recognized. Rule-based analysis is advantageous for adding dictionary lexical information (e.g. translations, animacy, diathesis etc.) and it does not require training data. However, the description format is not really transparent, as it only provides grammatical tags rather than glossing, which is a standard in typology and many other linguistic subdisciplines.

Tagging + translation	**կփորձեն** attempt, feel, try (V,tr) cond,prs,pl,3
Glossing	կ-փորձ-են k-p'orj-en COND-attempt-SBJV.PRS.3PL

Figure 5: Example with standard typological glossing

By the initiative of Timofey Arkhangelskiy and Aleksei Fedorenko the existing analyzer was improved and updated. The rules of the analyzer were rewritten in a format allowing glossing (Uniparser); the vocabulary was converted automatically, whereas

[12] http://web-corpora.net/

the inflection was rewritten manually. Certain procedures were applied to prepare stem glosses. Importantly, the analyzer[13] is now open source (MIT license).

The analyzer was tested on about 10 million tokens from EANC. The test dataset included 19th and 20th century fiction, press, scientific literature, as well as oral discourse. The test proved 93% coverage (not including tokens in non-Armenian script) and 1,25 ambiguity analysis per analyzed word. The updated test dataset was published through *tsakorpus*[14] corpus platform. The objective is to move entire EANC to *tsakorpus*.

5. Search Functionality and Display Options

EANC was designed first as an instrument of linguistic analysis and thus has to provide efficient tools of looking for linguistic information.

EANC allows to make token queries by wordforms (e.g. *մարդու mardu 'man.SG.GEN'*), lexemes (e.g. *մարդ mard 'man.SG.NOM'*, *մարդու **mardu** **'man.SG.GEN'***, *մարդիկ mardik 'man.PL.NOM'* and so on for the lexeme *մարդ* mard) or English translation (e.g. *man*) or queries based on a specific grammatical attribute or a combination of attributes (e.g. passive imperfective converbs or searching for *տուն tun 'house'* in singular definite yields such forms as nominative *տունը* tunə and *տուն* tunn, dative *տանը* tanə etc.).

Additional search criteria and options, such as case-sensitivity (e.g. capitalized tokens only), adjacent punctuation (only tokens preceding a comma) or position in the sentence (e.g. only tokens neither in the beginning nor in the end of a sentence) can be applied as well.

The most fascinating (and, in terms of software support, the most challenging) query option is a context query, a combination of several token queries. Using a context query, the use of the corpus may look for co-occurrences of tokens defined in each token query included in the context query in the same context. Co-occurrence is subject to distance limitations which may require that tokens occur next to each other (default option), at a distance between two values specified, simply within the same sentence, or in different sentences in the document.

Examples of context queries include, for instance, searching for a noun preceded by a genitive and an adjective, perfective converbs followed by any wordforms of the stative verb *է* with not more than one other wordform between them*, or co-occurrence of two negative verb forms in two adjacent sentences.

Further important search option is limiting the search domain to a subset of the texts of the corpus. The criteria might be the time of book creation, genres or types of texts, author or authors or the title of the book. The user may thus choose to look only for the matches occurring in Raffi's Samvel, in all Raffi's novels, in all novels of the 19-th century, or in all texts dating from the 19th century in general. This is extremely useful when e.g. trying to investigate semantic or other diachronical processes in the language, e.g. comparing the contexts using the verb *պրծացնել prcac'nel* 'finish' in the 19th and 20th centuries.

The display options allow to have the output in Armenian characters or in transliteration, to choose the layout (full (by default), light, glossed or KWIC), as well as extending the matches per page from 10 to 50 and the sentences in the context from 1 to 3.

The user can also choose the way in which the contexts matching the query are sorted, i.e. in what order they appear on the screen. Important sorting options are sorting alphabetically by the lexeme or wordform matching the query, by year of creation or by the name of the author (note that sorting criteria may be combined). Thus, sorting by the year of creation provides a convenient tool of observing the change of meaning of a word or use of a form over time.

6. Objective and Target Audience

Current state of Armenian studies requires new approaches and linguistic tools to validate key empirical hypotheses and findings as well as to expand the field of research. Corpus-based approach will allow revisiting the aspects of the traditional grammar that have not been sufficiently studied and will facilitate developing new descriptive and theoretical concepts.

EANC provides linguists with a searchable annotated database of MEA. EANC includes empirical linguistic data ranging from classical standard Eastern Armenian literature to Yerevan street talk recorded and transcribed in 2008.

EANC also provides a researcher with an option to build a user-defined sub-corpus, such as a single author sub-corpus, or a sub-corpus containing specific genres and/or periods.

Since EANC provides samples of actual MEA usage across periods, genres, and discourse formats, it can also be used as a powerful educational resource. English translations are provided for about 85 percent of the tokens, facilitating the use of the corpus by non-native speakers, e.g. Armenian language learners. EANC can also be used in various fields such as literature and culture studies, journalism, history, and others.

Importantly, EANC is as much about corpus linguistics as it is about Armenian studies. The EANC team aimed to build a modern flexible linguistic database that can be used as a platform for creating corpora of other languages, exploring statistical approaches to language description, as well as applying natural language processing methods.

7. Problems and Perspectives

A major problem of the EANC is the presence of numerous mistakes in optical character recognition. Wrong or impossible spellings result in losing hits and/or returning wrong hits. A number of procedures have been implemented to increase the accuracy,

[13] https://bitbucket.org/timarkh/uniparser-grammar-eastern-armenian

[14] https://github.com/timarkh/tsakorpus

including human-assisted proofreading of the most important texts.

As mentioned above, most of the press corpus has been downloaded from the open electronic archives, which means that these periodicals are extremely over represented in EANC.

An important problem is the absence of syntactic and morphosyntactic markup. MEA is rich in periphrastic constructions in verbal morphology which are ignored by the parser. One of the perspectives of the project could be the implementation of basic collocation markup, including markup of auxiliary verb constructions. Now, querying these constructions is only possible indirectly (such as submitting context queries for converbs plus the verb *'to be'*, although these queries are obviously not enough restrictive).

Ignoring the context also leads to significant number of ambiguous cases in parsing results, which, for some queries, is a strong 'noise' factor. One of the solutions is human-assisted ambiguity removal.

In some cases, the two (or more) grammatical analyses of a wordform are by far not equally probable. It is possible to decrease the probability rank for less probable analyses depending on the context. Applying statistical procedures may be used to decrease the rank of morphosyntactic interpretations that are impossible or improbable in some types of contexts. For selected highly frequent cases of an extremely improbable homonymy, second readings have already been eliminated (e.g. the locative *asum* from the noun *as*).

Another useful development prospective would be allowing for context output provided with morphological glossing, more convenient for users coming from the field of linguistic typology and ready-to-use in typological publications which is already integrated in the updated version.

Providing the wordlist with phonetic tags indicating orthographically unpredictable phenomenon such as devoicing vs. non-devoicing after sonorants or between vowels or shwa insertion, orthographically would be a useful addon. Ultimately, that will provide a tool to show phonetic transcription of the word and wordform.

More detailed oral discourse transcription which requires serious theoretic background would also be a precious extension for the oral sub-corpus. Discourse transcription segments discourse into units with time synchronization for each unit; designates pauses, both silent (i.e. complete absence of verbal expression) and filled pauses (cf. English *'um', 'uh'* etc.); and tracks other phenomena peculiar to oral discourse, e.g. parceling, embeddings, discourse markers, etc. The transcripts should also be synchronized with light versions of audio files so that the user may not only read the transcript but also listen to the original audio.

An attempt of dialect corpus was made in the framework of EANC research grant project during 2008-2009. Interviews and narratives in three dialects of Armenian (1. Arcvaberd dialect (Shamshadin, Tavush region), 2. Shenavan dialect (Aparan, Aragatsotn region), 3. Gusana dialect (Maralik, Shirak region)) were collected and transcribed by three postgraduate grantee students in Yerevan. The target size of each corpus is 15 hours of recordings or about 100,000 tokens. The data is lemmatized and is available for online search similar to EANC[15].

One of the most important developments of Armenian corpus processing is to have a multivariational with all the diachronical stages of the Armenian language on the one hand (Classical Armenian, Middle Armenian, Modern Armenian), and the language varieties of Modern Armenian continuum (Modern Western Armenian, Armenian dialects, oral standards).

To address the existing drawbacks and outlined perspectives mentioned above, the project Digitizing Armenian Linguistic Heritage: Armenian Multivariational Corpus and Data Processing (DALiH)[16] was designed. The project aims at building for the first time an open-access and open-source unified digital linguistic platform for the whole spectrum of Armenian language variation. Each language variety will be represented by a comprehensive corpus which will be provided with full morphological annotation. More particularly, DALiH will be the first to design six new annotated corpora for 1) Classical Armenian; 2) Modern Western Armenian; 3) a pilot corpus of Middle Armenian; 4) three pilot corpora of dialects, and 5) one updated Modern Eastern Armenian corpus on the basis of EANC.

More particularly, the following updates will be proposed for MEA:

a. EANC database will be completed by compilation of new texts (10M tokens of various genres, about 50M tokens coming from Wikipedia and Wikisource, about 200M tokens from general Google database);

b. EANC rule-based annotation model will be accompanied by RNN, transformer-based and hybrid models in order to attune the ambiguity and to provide context-based (hence future syntactic) annotation;

c. EANC grammatical dictionary will be updated with new lexemes compiled from the most frequent unrecognized tokens of the corpus;

d. golden standard annotated written and oral corpora will be provided;

e. EANC oral sub-corpus will be sound-aligned;

f. ASR model will be elaborated on the basis of the aligned oral corpus.

DALiH started in April 2021 and the project will be launched in 2025.

8. Bibliographical References:

Aġayan, E. (1976). Արդի հայերենի բացատրական բառարան (Explanatory dictionary of Modern Armenian language). V. 1-2. Yerevan.

Avetisyan, K. and Ghukasyan, T. (2019). Word Embeddings for the Armenian Language: Intrinsic and Extrinsic Evaluation. arXiv:1906.03134 [cs].

Donabedian-Demopoulos, A. and Boyacioglu, N. (2007). La lemmatisation de l'arménien occidental

[15] http://web-corpora.net/EANC_dialects/search/

[16] http://www.inalco.fr/actualite/projet-prc-dalih-digitizing-armenian-linguistic-heritage-laureat-aapg-2021-anr.

The project DALiH is funded by French National Research Agency ANR-21-CE38-0006.

avec NooJ. S. Koeva, D. Maurel, M. Silberztein. *Formaliser les langues avec l'ordinateur, de INTEX à NooJ*, Presses Universitaires de Franche Comté, pp. 55-75.
Galstyan, E. (1985). Հայ-ռուսերեն բառարան (Armenian-Russian Dictionary). Yerevan.
Ghukasyan, T., Davtyan, G., Avetisyan, K. and Andrianov, I. (2018). pioNER: Datasets and baselines for Armenian named entity recognition. In *Proceedings of Ivannikov Ispras Open Conference (ISPRAS)*, pp. 56–61. IEEE.
Grgearian, A. and Harutyunian, N. (1987-1989). Աշխարհագրական անունների բառարան (Dictionary of geographic names). Yerevan.
Gyurdjinian, D. (2005). Անուն խոսքի մասերի թվի կարգը արդի հայերենում. Քերականական բառարան-տեղեկատու (The category of number of nominals in Modern Armenian). Yerevan.
Gyurdjinian, D. and Hekekian, N. (2007). Հայերենում գործածվող տառային հապավումների բառարան (Dictionary of acronyms used in Armenian). Yerevan.
Malajyan, A., Avetisyan, K., Ghukasyan, T. (2020). ARPA: Armenian Paraphrase Detection Corpus and Models. In *Proceedings of Ivannikov Memorial Workshop (IVMEM)*, pp. 35-39.
McEnery, T. and Hardie, A. (2012). Corpus linguistics: method, theory and practice. Cambridge & New York, Cambridge University Press.
McEnery, T. and Wilson, A. (2001). Corpus linguistics: an introduction. Edinburgh, Edinburgh University Press.
Stefanowitsch, A. (2020). Corpus linguistics: A guide to the methodology (Textbooks in Language Sciences 7). Berlin: Language Science Press.
Vidal-Gorène, C., Khurshudyan, V. and Donabédian-Demopoulos, A. (2020). Recycling and Comparing Morphological Annotation Models for Armenian Diachronic-Variational Corpus Processing. *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, 90-101.
Yavrumyan, M. and Danielyan A. (2020). Համընդհանուր կախվածություններ և հայերենի ծառադարանը (Universal Dependencies and Armenian Tree-Bank). *Lraber hasarakakan gitut'yunneri*, 231-244.
Yavrumyan, M. (2019). Տեքստի մեքենական հատույթավորումը արևելահայերենի շարահյուսական ծառերի UD_ARMENIAN-ArmTDP բանկում (Tokenization and Word Segmentation in the UD_ARMENIAN-ArmTDP treebank). *Banber Yerevani hamalsarani. Philology*, 2019 № 3 (30). Yerevan. Pp. 52-65.

Annexe 1: EANC grammatical tags

#	EANC Tag	Description	Traditional Armenian Label	Example
	Parts of Speech		**Խոսքի մասեր**	
1	N	Noun	Գոյական	*սեղան*
2	A	Adjective	Ածական	*գեղեցիկ*
3	V	Verb	Բայ	*կարդալ*
4	ADV	Adverb	Մակբայ	*արագ*
5	NUM	Numeral	Թվական	*երեք*
6	PRON	Pronoun	Դերանուն	*ես*
7	PREP	Preposition	Կապ	*առանց*
8	POST	Postposition	Կապ	*մեջ*
9	CONJ	Conjunction	Շաղկապ	*և*
10	PART	Particle	Եղանակավորող բառեր	*թերևս*
11	INTJ	Interjection	Ձայնարկություն	*վա՛յ*
	Parts of Speech: lexical subcategories			
12	S	Independent pronouns	Անկախ դերանուն	*նա*
13	Dem	Demonstrative pronoun	Ցուցական դերանուն	*այդ*
14	Intrg	Interrogative pronoun	Հարցական դերանուն	*ինչ*
15	Hum	Human noun or pronoun	Անձի առում	*մարդ*
16	Anim	Animate noun or pronoun	Շնչավոր	*գայլ*
17	Inanim	Inanimate noun or pronoun	Անշունչ	*սեղան*
18	Coll	Collective noun	Հավաքական գոյական	*խումբ*
19	Topn	Toponym	Տեղանուն	*Հայաստան*
20	Persn	First name	Անձնանուն	*Արմեն*
21	Famn	Family name	Ազգանուն	*Պետրոսյան*
22	Abbr	Abbreviation	Հապավում	*ԱՊՀ*
23	Card	Cardinal numeral	Քանակական թվական	*երեք*
24	Tr	Transitive verb	Անցողական բայ	*տալ*
25	Intr	Intransitive verb	Անանցողական բայ	*վազել*
	Nominalization			
26	Inf	Infinitive	Անորոշ դերբայ	*կարդալ*
27	Rel	Relational noun	-	*սեղանիկը*
28	Nmlz	Nominalized attribute (adjective, participle, genitive)	Գոյականացված (ածական, դերբայ, սեռական)	*գեղեցիկը, կարդացածը, սեղանիկը*
	Case		**Հոլով**	
29	Nom	Nominative	Ուղղական	*քաղաք*
30	Gen	Genitive	Սեռական	*քաղաքի*
31	Dat	Dative	Տրական	*քաղաքին*
32	Abl	Ablative	Բացառական	*քաղաքից*
33	Ins	Instrumental	Գործիական	*քաղաքով*
34	Loc	Locative	Ներգոյական	*քաղաքում*

#	EANC Tag	Description	Traditional Armenian Label	Example
	Number		**Թիվ**	
35	Sg	Singular (nouns, pronouns or verbs)	Եզակի	*քաղաք*
36	Pl	Plural (nouns, pronouns or verbs)	Հոգնակի	*քաղաքներ*
37	Apl	Associative plural (nouns and pronouns)	հավաքական անեզական հոգնակի	*Վարդանանց*
	Determination/Possession		**Առում**	
38	Def	Definite form of a noun	Որոշյալ	*քաղաքը*
39	Poss1	First person possessed noun	Ստացական հոդ 1	*քաղաքս*
40	Poss2	Second person possessed noun	Ստացական հոդ 2	*քաղաքդ*
	Degree of Comparison		**Համեմատական աստիճան**	
41	Sup	Superlative	Գերադրական աստիճան	*ամենագեղեցիկ*
	Converb			
42	Cvb	Converb	Դերբայ	
43	Sim	Simultaneous converb	Անկատար դերբայ II	*կարդալիս*
44	Ipfv	Imperfective converb	Անկատար դերբայ I	*կարդում*
45	Pfv	Perfective converb	Վաղակատար դերբայ	*կարդացել*
46	Des	Destinative (future converb)	Ապառնի I	*կարդալու*
47	Conneg	Connegative converb	Ժխտական դերբայ	*կարդա*
	Participle			
48	Ptcp	Participle	Դերբայ	
49	Sbj	Subject participle	Ենթակայական դերբայ	*կարդացող*
50	Res	Resultative participle	Հարակատար դերբայ	*կարդացած*
	Valency Changing			
51	Caus	Causative (morphological)	Պատճառական	*վախեցնել*
52	Med	Medial (passive)	Կրավորական	*կառուցվել*
	Tense-Aspect-Mood		**Ժամանակ-Կերպ-Եղանակ**	
53	Pres	Present	Ներկա	*է*
54	Past	Past	Անցյալ	*էր*
55	Aor	Aorist	Անցյալ կատարյալ	*կարդաց*
56	Sbjv	Subjunctive	Ըղձական	*կարդա*
57	Cond	Conditional	Պայմանական	*կկարդա*
58	Imp	Imperative	Հրամայական	*կարդա՛*
	Polarity			
59	Neg	Negative form of a verb	Ժխտական	*չկարդաց*
	Person			
60	1	1st person category	Առաջին դեմք	*եմ*
61	2	2nd person category	Երկրորդ դեմք	*ես*
62	3	3rd person category	Երրորդ դեմք	*է*