International Journal of Scientific Advances
ISSN: 2708-7972
Volume: 4 | Issue: 2 | Mar – Apr 2023 Available Online: www.ijscia.com
DOI: 10.51542/ijscia.v4i2.19

ArmSpeech-POS: Eastern Armenian Part-of-Speech Tagged Corpus

Varuzhan H. Baghdasaryan

Armenia, Yerevan, Romanos Melikian 6/1
National Polytechnic University of Armenia

*Corresponding author details: Varuzhan H. Baghdasaryan; varuzh2014@gmail.com

ABSTRACT

Text chunking, Part-of-speech (POS) tagging, and named entity recognition (NER) are fundamental tasks in natural language processing (NLP). Part-of-speech (POS) tagging involves assigning grammatical labels to words in a sentence. Research shows that Armenian is a low-resourced language and there are not enough materials for developing higher accurate part-of-speech tagging systems in the Armenian language. This paper presents a fresh dataset for POS tagging in Armenian that follows the naming conventions of both Penn Treebank and Universal Dependencies tagsets, with two versions available. The dataset consists of 6081 sentences that were automatically annotated and then manually verified. The data was sourced from Armenian news websites, focusing on topics such as culture, medicine, and lifestyle, as well as 22 Armenian fairytales. The reason for having two versions of the POS tagset was to ensure compatibility and integration with all-natural language processing tools and models that use these standards. By standardizing the tagset, it becomes easier to compare and evaluate the effectiveness of different POS tagging models. The paper also describes data collection, cleaning, preprocessing, and processing steps. The ISMA translator was used for the annotation of the dataset, which not only performs machine translation but also conducts a syntactic and semantic analysis of the text and assigns a POS tag for each word in the sentence. The final corpus contains 13 groups of part-of-speech tags and a total of 57160 tagged tokens including the distinction between singular and plural parts of speech.

Keywords: NLP; natural language processing; POS; part-of-speech; dataset; automated annotation; POS tag; tagset.

INTRODUCTION

An example of sequence labeling, which is a pattern recognition task, is part-of-speech tagging, whereby each word in a sequence is labeled with a corresponding part-of-speech tag [1]. Part-of-speech tagging is a critical task in NLP that involves assigning a grammatical label to each word in a sentence, such as a noun, verb, adjective, or adverb [2, 3]. POS tag/label can also indicate grammatical categories (tense, number (plural/singular), case, and so on) [2, 3]. This information is essential for many downstream NLP tasks that rely on understanding the syntactic structure of natural language, such as named entity recognition, sentiment analysis, and machine translation [1]. Also, POS tagging can be used in corpus searches (search for examples of grammatical or lexical patterns), text analysis, and automatic text processing tools, as well as, word sense disambiguation, question-answering parsing, and so on. POS tagging has been extensively studied for several languages, such as English, Chinese, and Arabic, but there has been relatively little work on POS tagging for the Armenian language.

Armenian is an Indo-European language spoken primarily in Armenia, neighboring countries, and the Armenian diaspora, with approximately 7-9 (according to various estimates) million speakers worldwide. The Armenian language has a complex morphology, with rich inflectional and derivational morphology, and a complex word order, which makes POS tagging a challenging task. There is a lack of large-scale annotated datasets for Armenian. This article aims to address this issue by presenting a new part-of-speech tagging dataset for Armenian.

This paper introduces a fresh POS tagging dataset for the Armenian language, consisting of over 6081 automatically annotated sentences (57160 tokens tagged with part-of-speech and grammatical number tags) collected from Armenian news websites and fairytales. The dataset was annotated using the ISMA translator, which assigns a POS tag for every word in the sentence. Additionally, the paper outlines the data preprocessing and database development procedures involved in establishing the dataset.

Most basic POS tagging datasets include tags for the most common parts of speech (noun, pronoun, verb, adjective, and so on). Other advanced corpora can also include more grammatical details (distinguishing between singular and plural nouns, tag tenses, causes, and much more) [4, 5, 6]. There are two ways to accomplish the annotation process:

- Manual annotation.
- Automated annotation.

Manual annotation is frequently utilized for tagging a small corpus to be used as training data to develop a new automatic POS tagger. However, for substantial corpora, manual annotation is not feasible, and automatic tagging is used instead.

Large, modern corpora necessitate automatic annotation, which can achieve an accuracy of up to 98%. Errors are typically caused by misspelled words, unusual usage, or interjections such as "yuppeeee," which may be wrongly labeled.

Typically, automatic POS taggers rely on a limited manually annotated training data set to learn how the language should be tagged. Additionally, taggers for each language may differ based on linguistic features.

Two commonly used syntactic annotation schemes for natural language processing are Universal Dependencies (UD) [6] and Penn Treebank (PTB) [4, 5]. UD is a framework that provides standardized labels for parts-of-speech, morphological features, and syntactic dependencies to achieve cross-linguistically consistent annotation of morphosyntactic structure in natural language sentences. With over 100 languages developed for UD treebanks, it has become widely used in natural language processing tasks such as dependency parsing, named entity recognition, and machine translation [6]. Universal POS tags, which are part-of-speech markers used in UD, are freely available and accessible.

On the other hand, the PTB is a parsed and tagged English language dataset that was developed at the University of Pennsylvania. It includes a large corpus of text from sources like the Wall Street Journal, annotated with syntactic structure using a constituency-based treebank annotation scheme [4, 5]. The PTB has been extensively used in natural language processing research, particularly for tasks such as parsing, language modeling, and part-of-speech tagging.

The main differences between Universal Dependencies (UD) and Penn Treebank (PTB) annotation schemes are as follows:

- Tagset: The UD scheme has a simpler tagset compared to the PTB scheme. UD has 17 coarse-grained tags, while PTB has 36 fine-grained tags.
- Consistency: UD has a more consistent tagset, meaning that the same tag is used for the same part of speech across different languages. In contrast, PTB tag names are often language-specific and can vary depending on the language being tagged.
- POS Categories: The UD scheme has more detailed POS categories, including subcategories for nouns, verbs, adjectives, and adverbs, whereas the PTB scheme has fewer subcategories and more general categories.
- Tokenization: The UD scheme is based on tokenization by whitespace and punctuation, whereas the PTB scheme uses a more complex tokenization process that involves handling contractions and punctuation marks differently.
- Dependency Parsing: The UD scheme includes information about dependency relations between words, whereas the PTB scheme does not include this information.

The main differences between UD and PTB are the tagset, syntax representation, language coverage, annotation goals, and the inclusion of explicit dependency relations in UD. UD is a cross-linguistic framework for annotating morphosyntactic structure, while PTB is a specific dataset of parsed and tagged English sentences. Both are important resources for natural language processing research and are widely used in various applications.

The simplest approach to constructing a POS tagging dataset is to use a pre-existing solution, such as a programming language library or an online tool. The ISMA translator, available at http://www.translator.am/am/index.html, is an excellent option, requiring no technical knowledge or IT skills, and is capable of annotating vast amounts of data. The critical technologies employed are data preprocessing, web scraping, and data cleaning techniques, which can be implemented using the Python programming language.

The ISMA Translator is an online machine translation tool that features a rule-based POS tagging system and can also perform grammar and spelling analysis for Armenian language text. The system has demonstrated remarkable accuracy in processing Armenian text, making it an ideal tool for annotating POS tagging. While ISMA offers other types of grammar analysis, such as stem, gender, and article identification, it is not infallible and may occasionally make mistakes or fail to analyze certain word groups. As a result, the database generated through testing only includes tags related to parts of speech and grammatical numbers.

RELATED WORKS

Large-scale annotated datasets are critical for the development of robust and accurate POS tagging models. There are several publicly available POS tagging datasets for other languages, such as the Penn Treebank for English [4, 5] and the Chinese Treebank for Chinese [7]. However, there has been a lack of such datasets for the Armenian language, which limits the development of POS tagging models for the language.

The investigation was conducted on POS datasets of Armenian, as well as those of other frequently employed languages, to investigate and recognize all the features and structural principles present in these datasets.

There have been few studies on POS tagging for the Armenian language. Several pieces of research have been conducted for creating Armenian language resources, such as corpora and lexicons. The Universal Dependencies framework includes 3 freely available corpora for the Armenian language, with 2 designed for Eastern Armenian and 1 for Western Armenian.

The UD_Armenian-ArmTDP treebank for Eastern Armenian was developed in collaboration with Marat M. Yavrumyan and the ArmTDP team at Yerevan State University as part of the Universal Dependencies project [8, 9, 10]. It is based on the ArmTDP V2.0 dataset, which is a broad collection of Modern Standard Armenian texts covering various genres, and includes around 2500 sentences, 52220 tokens, and 52585 syntactic words [8, 9, 10]. The treebank was manually annotated with Universal POS tags. Currently, it consists of 17 POS tags and follows the Universal Dependencies annotation scheme, a cross-linguistic standard for annotating grammatical structures. The treebank was used to create a dependency parser for Eastern Armenian and is currently the largest verified corpus of Eastern Armenian according to the source [8, 9, 10].

Another research effort is the development of the UD_Armenian-BSUT treebank, which was carried out under the auspices of the "HayLingvoTech" excellence center program, implemented by V. Brusov State University with funding from the Competitive Innovation Fund of Armenia [10]. This corpus includes 2300 sentences, 41492 tokens, and 41805 syntactic words, and is annotated with 17 tags [10]. The treebank was created by Marat M. Yavrumyan, Rima R. Grigoryan, Anna S. Danielyan, and Setrag H. M. Hovsepian [10].

For further information, please visit https://universaldependencies.org/.

Eastern Armenian National Corpus (EANC) is a corpus of Modern Eastern Armenian and contains 110 million tokens [11]. According to the paper, the corpus contains written and oral discourses from the mid-19th century to the present [11]. EANC is publicly available at www.eanc.net.

In the majority of these studies, the dataset development process is founded on data obtained from various sources, such as literature, journalism, and spoken language.

As previously noted, non-Armenian datasets were also examined.

The AnCora Corpus is a large dataset of Spanish text annotated with POS tags. It consists of over 500,000 words of text from a variety of sources, including newspapers, literature, and scientific articles [12].

The TIGER Corpus is a dataset of German text annotated with POS tags, syntactic structures, and other linguistic features [13]. It consists of over 50,000 sentences (900,000 tokens) from a variety of genres, including news articles, fiction, and academic texts [13].

The French Treebank is a dataset of French text annotated with POS tags and other linguistic features. It consists of over 21.550 sentences (appx. 664.500 tokens) from the newspaper Le Monde (1990-1993) [14]. According to the paper [14], the corpus has been annotated by software developed specifically for this set of tasks (Clément 2001) and then systematically corrected by hand.

The datasets mentioned above mainly originate from fiction, as well as non-fiction sources such as newspapers, magazines, academic journals, literature, and scientific articles [12, 13, 14]. Most of these datasets follow either Universal Dependencies or Penn Treebank rules to provide tag names for part-of-speech tokens. Below are some of them [4, 5, 6]:

TABLE 1: Some tags Used in Universal Dependencies and Penn Treebank.

Universal Dependencies	Penn Treebank
NOUN - Noun	NN - Noun: singular or mass
VERB - Verb	VB - Verb: base form
ADJ - Adjective	JJ - Adjective: general
ADV - Adverb	RB - Adverb: general
ADP - Adposition	IN - Preposition or subordinating conjunction
PRON - Pronoun	PRP - Personal pronoun
CCONJ - Coordinating conjunction	CC - Coordinating conjunction
DET - Determiner	DT - Determiner
NUM - Numeral	CD - Cardinal number
AUX - Auxiliary verb	VBG - Verb: gerund or present participle

For example, in the sentence "It is hard to breathe." the words, according to the Penn Treebank tagset [4, 5], would be labelled as " It/PRP, is/VBZ, hard/JJ, to/TO, breathe/VB" where "PRP" indicates a personal pronoun, "VBZ" indicates a verb in the present tense (3rd person, singular), JJ indicates Adjective, and so on.

A shared characteristic of these datasets is that they are partitioned into training, validation, and testing sets, with each section containing a collection of labeled sentences where each word is assigned a corresponding part-of-speech tag. The token and its corresponding tag are separated by a forward slash (/), vertical bar (|), or tab, and each sentence is separated by a blank line.

DATA COLLECTION AND PROCESSING

Creating a database necessitates a substantial amount of data. However, since there is no extensive collection available for the Armenian language, the required data was sourced from the internet, specifically from websites that offer information in Armenian. The focus was placed on articles about culture, sports, politics, medicine, lifestyle, and other subjects, as they encompass a broad range of vocabulary and sentence structures. The final dataset consists of 117550 sentences and also includes 22 Armenian fairytales.

The process began with web scraping technology [15, 16] to collect data from various Armenian websites. The collected data was then normalized, which involved removing all incomplete sentences, sentences with typos, non-Armenian symbols, and punctuation marks that are not used in Armenian. The resulting dataset contains only grammatically correct and error-free sentences, saved as a text file with each sentence on a separate line. Subsequently, each token in the dataset was assigned a corresponding part-of-speech tag and grammatical number (if applicable) using the ISMA translator.

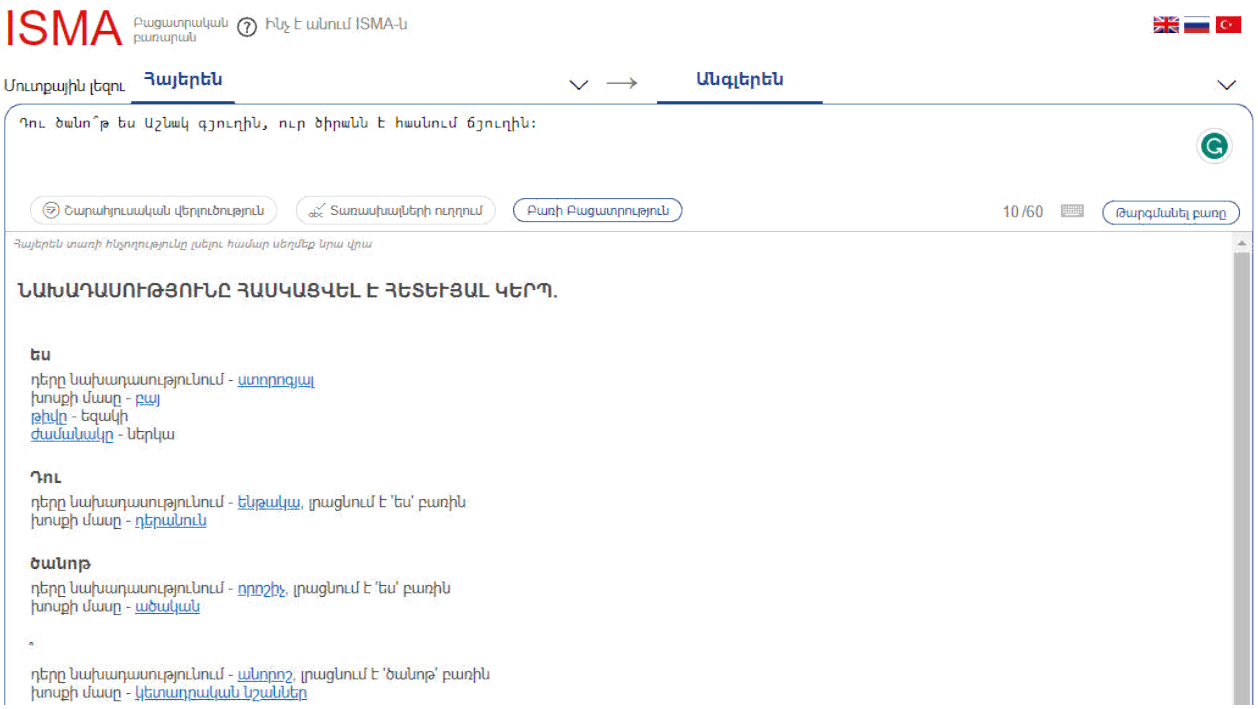

FIGURE 1: Example Of ISMA Translator's Morphological Analysis.

The process of generating the tagset involved several steps. The followings are the steps and descriptions for generating the tagset:

- Parallelly take one sentence from scraped data.
- Generate temporary sentence from taken data, where all punctuation marks are replaced by space or nothing (depending on punctuation mark type).
- Pass the complete sentence as input to the ISMA translator for context-based analysis, as a major challenge with POS tagging is that some words can have multiple labels depending on their role in the sentence, such as "pay," which can be used as both a verb (base form VB and present form VBP) and a noun (NN).
- Obtain the resulting grammatical properties (part-of-speech tag and grammatical number) using the Selenium browser-based regression automation toolkit [17] and assign them to the corresponding tokens in the sentence. Exclude those sentences, which contain tokens for which ISMA could not do synthetic analysis and provide a result.

It is important to mention the fact that ISMA is not perfect and has certain shortcomings. The ISMA translator has some limitations when it comes to the grammatical analysis of sentences, including:

- Failure to label auxiliary verbs.
- Inability to parse many words, particularly those that are rarely used (or dialect words).
- Omission of grammatical number labelling for certain parts of speech.
- Tagging only the last occurrence of a repeated word in a sentence, while skipping the previous occurrences. For instance, in the sentence "Դու ասացիր, բայց ես չլսեցի։" (You said, but I didn't hear.), the word "ես" (I) is used twice, but ISMA only tags the last occurrence and assigns the "pronoun" tag, omitting the auxiliary verb "ես" (I) used earlier.

It was imperative to address the shortcomings of the ISMA translator during the project. While generating token-tag pairs, sentences that had identical repeated words and sentences in which ISMA failed to tag any tokens were automatically excluded. Subsequently, all auxiliary verbs and pronouns were labelled as pronouns and then manually corrected.

RESULTS

To elaborate further, the resulting database not only includes the 12 parts of speech of the Armenian language and punctuation marks (see table 2), but it also follows the standard naming conventions of the Penn Treebank [4, 5] and Universal Dependencies[6] tag names list. Out of the 117550 sentences collected from Armenian news websites and 22 Armenian fairytales, only 6081 sentences were validated for processing. The total number of tokens is 57160. In addition to standard part-of-speech tags, the tagset also includes specific tags indicating whether a token is singular or plural.

TABLE 2: Part-Of-Speech, And Punctuation Mark Count In The Corpus.

Tag	Count	Tag	Count
Noun (Armenian: գոյական)	18536	Hyphen (Armenian: միության գծիկ)	521
Verb (Armenian: բայ)	7132	Refering (Armenian: վերաբերական)	503

Tag	Count	Tag	Count
Full stop (Armenian: վերջակետ)	6077	Question mark (Armenian: հարցական նշան)	284
Adjective (Armenian: ածական)	5103	Connection (preposition) (Armenian: կապ (նախդիր))	170
Pronoun (Armenian: դերանուն)	4234	Connection (postposition) (Armenian: կապ (հետդիր))	163
Auxiliary verb (Armenian: օժանդակ բայ)	3991	Colon (Armenian: միջակետ)	58
Comma (Armenian: ստորակետ)	2533	Emphasis (Armenian: շեշտ)	45
Conjunction (Armenian: շաղկապ)	2102	Bracket (Armenian: փակագիծ)	43
Numeral (Armenian: թվական)	1658	Interjection (Armenian: ձայնարկություն)	41
Quotation mark (Armenian: չակերտ)	1640	Exclamation mark (Armenian: բացականչական նշան)	11
Adverb (Armenian: մակբայ)	1571	dash (Armenian: անջատման գիծ)	7
Punctuation mark (Armenian: բութ)	736	Armenian hyphen (Armenian: ենթամնա)	1

An organized format was used to store the database, with each sentence represented as a series of word-POS tag pairs separated by a vertical bar (|), and sentences separated by blank lines. This format facilitates easy access and use for training and evaluating POS tagging models.

Also, the database is accessible through SQL queries and can be converted into a tab-separated file format using a provided script, which can be easily loaded into various machine-learning libraries. FIGURE 2 shows the distribution of the Armenian part-of-speech, and punctuation marks in the corpus.

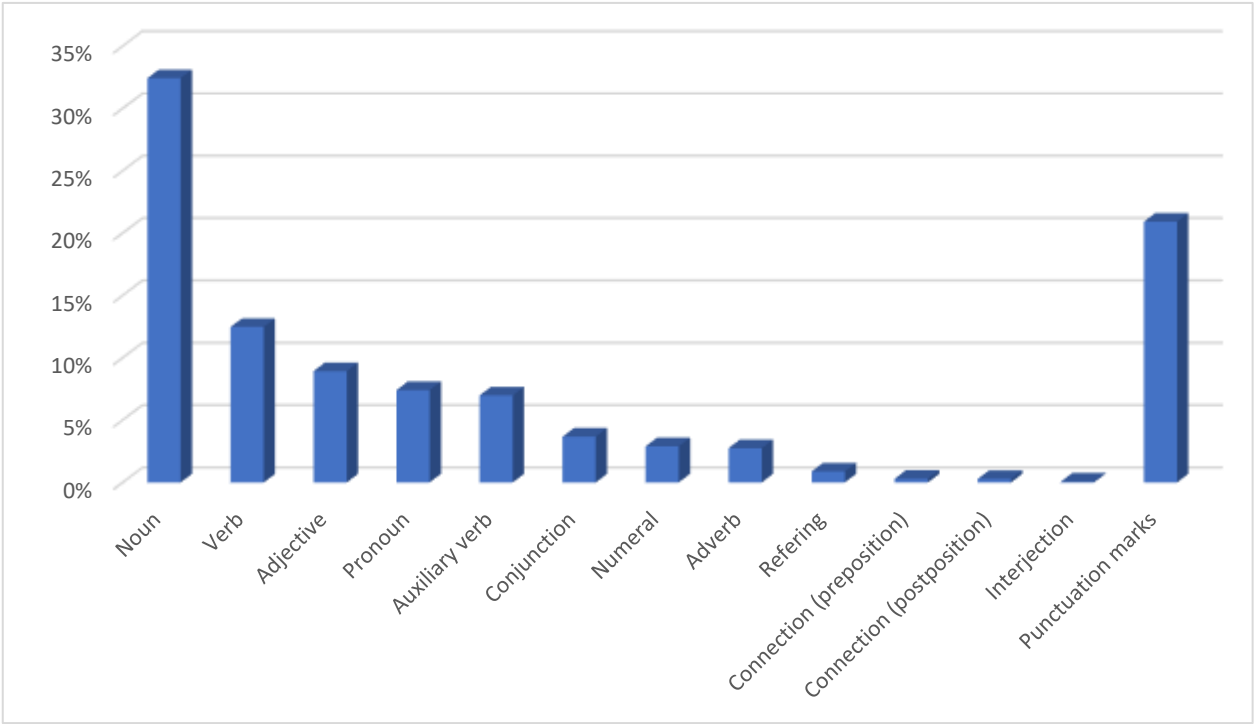

FIGURE 2: Distribution Of The Armenian Part-Of-Speech, And Punctuation Marks In The Corpus.

CONCLUSIONS

This paper presents a novel POS tagging dataset for Armenian, comprising more than 6081 sentences collected from Armenian news websites covering various topics such as culture, medicine, and lifestyle, as well as 22 Armenian fairytales. The dataset adheres to both the naming conventions of the Penn Treebank and Universal Dependencies tagsets (with two versions) and is automatically annotated. An important consideration for having two versions of the POS tagset was to ensure seamless integration and compatibility with all natural language processing tools and models that rely on this standard. The use of a standardized tagset also simplifies the comparison and evaluation of different POS tagging models.

The dataset was annotated using the ISMA translator, which is an online translator system and has a built-in rule-based POS tagging feature. The annotations were checked by native Armenian speakers who have expertise in Armenian grammar. The final tagset contains 57160 tagged tokens grouped into 13 groups (tags) and includes singular and plural parts of speech distinction.

Baseline results on this dataset demonstrate the challenges of POS tagging in Armenian, due to the language's complex morphology and limited language resources. The dataset introduced in this paper provides a valuable resource for future research on Armenian POS tagging and NLP in general. There is a hope that this work will encourage further research in this area and contribute to the development of better NLP tools for the Armenian language.

REFERENCES

[1] Helmut Schmid. (24.10.1994). Part-of-speech tagging with neural networks. arXiv preprint arXiv:cmp-lg/9410018.

[2] Chowdhary, KR1442, and K. R. Chowdhary. (2020) Natural language processing. Fundamentals of artificial intelligence, 603-649.

[3] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. (2011). Natural language processing: an introduction. Journal of the American Medical Informatics Association, 18 (5), 544-551.

[4] Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. (1993). Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics, 19 (2), 313-330.

[5] Ann Taylor, Mitchell Marcus & Beatrice Santorini. (2003). The Penn treebank: an overview. Treebanks: Building and using parsed corpora, 5-22. DOI: 10.1007/978-94-010-0201-1_1.

[6] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. (2021). Universal Dependencies v2.8. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic. ISBN 978-80-88280-42-2.

[7] Nianwen Xue, Fei Xia, Fu-Dong Chiou, Martha Palmer. (01.06.2005). The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. Natural Language Engineering, 11 (2), 207-238. DOI: https://doi.org/10.1017/S135132490400364X.

[8] Marat M. Yavrumyan, Hrant H. Khachatrian, Anna S. Danielyan, Gor D. Arakelyan. (2017). ArmTDP: Eastern Armenian Treebank and Dependency Parser. XI International Conference on Armenian Linguistics, Abstracts, Yerevan.

[9] Marat M. Yavrumyan. Universal Dependencies for Armenian. (2019). International Conference on Digital Armenian, Abstracts. Inalco, Paris.

[10] Marat M. Yavrumyan, Anna S. Danielyan. (2020). Universal Dependencies and the Armenian Treebank. Herald of the Social Sciences (2), 231-244, Armenian.

[11] Victoria Khurshudyan, Timofey Arkhangelskiy, Michael Daniel, Dmitri Levonian, Vladimir Plungian, Alex Polyakov, Sergei Rubakov. (20.06.2022) Eastern Armenian National Corpus: State of the Art and Perspectives, Proceedings of the Workshop on Processing Language Variation: Digital Armenian (DigitAm 2022) @LREC2022, 28–37, Marseille.

[12] Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco. European Language Resources Association (ELRA).

[13] Brants, Sabine, Stefanie Dipper, Peter Eisenberg, Silvia Hansen, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. (2004). TIGER: Linguistic Interpretation of a German Corpus. Journal of Language and Computation, (2), 597-620.

[14] Abeillé A., L. Clément, and F. Toussenel. (2003). Building a treebank for French, in A. Abeillé (ed) Treebanks , Kluwer, Dordrecht.

[15] Lunn S., Zhu J., & Ross M. (2020). Utilizing web scraping and natural language processing to better inform pedagogical practice. In 2020 IEEE Frontiers in Education Conference (FIE), 1-9. IEEE.

[16] Zhao B. (2017). Web scraping. Encyclopedia of big data, 1-3.

[17] P. Molyneux and D. Ferguson. (2017). Automated Web Scraping with Selenium. In proceedings of the 2017 IEEE International Conference on Software Maintenance and Evolution (ICSME), 643-647. DOI: 10.1109/ICSME.2017.73.