Proceedings of the Workshop on Processing Language Variation: Digital Armenian (DigitAm 2022) @LREC2022, pages 1–7
Marseille, 20 June 2022

A Free/Open-Source Morphological Transducer for Western Armenian

Hossep Dolatian, Daniel Swanson, Jonathan Washington
Stony Brook University, Indiana University, Swarthmore College
Stony Brook, NY 11794; Bloomington, IN 47405; Swarthmore, PA 19081
hossep.dolatian@alumni.stonybrook.edu, dangswan@iu.edu, jonathan.washington@swarthmore.edu

Abstract

We present a free/open-source morphological transducer for Western Armenian, an endangered and low-resource Indo-European language. The transducer has virtually complete coverage of the language's inflectional morphology. We built the lexicon by scraping online dictionaries. As of submission, the transducer has a lexicon of 75K words. It has over 90% naive coverage on different Western Armenian corpora, and high precision.

Keywords: finite-state morphology, two-level morphology, transducer, computational morphology, low-resource language, Western Armenian

1. Introduction

This paper presents the first known publicly available morphological transducer for Western Armenian (hyw), an endangered Indo-European language currently spoken by an estimated 1 million people (Eberhard et al., 2022).[1]
A morphological transducer is a computational tool that maps between forms and analyses, able to perform both morphological analysis and morphological generation. For example, the form բառէրէն [pʰɑɾeɾen] 'the words' may be analyzed as բառ<n><pl><abl><def>, whereas generation goes the other direction. The morphological transducer reported on in this paper has production-quality coverage and was developed entirely by hand, with some automated support in the form of scraping dictionaries.
Section 2 overviews Western Armenian and positions the present work among other Armenian text processing tools. Section 3 details the implementation of the transducer. Section 4 presents an evaluation of the transducer. Section 5 presents thoughts on future work, and Section 6 focuses on cross-dialectal support. Section 7 concludes.

2. Background on Armenian and language tools

Armenian belongs to an independent branch in the Indo-European family. Armenian is pluricentric with two standard lects (Western and Eastern) and multiple non-standard lects (Adjarian, 1909). The two standard lects share substantial similarities but have many substantial differences in phonology, morphology and syntax (Cowe, 1992; Donabédian, 2018). Both lects are written in the Armenian script. Western Armenian uses a more conservative spelling system than Eastern Armenian (Sanjian, 1996; Dum-Tragut, 2009).
Eastern Armenian is the official language of Armenia, while Western Armenian developed as a koiné lect among ethnic Armenians in the Ottoman Empire (Sayeed and Vaux, 2017). After the Armenian Genocide (1915–1917), Western Armenian became a largely diasporic language that is spoken across communities in the Middle East, Europe, the Americas, and Australia. Western Armenian is classified as an endangered language by UNESCO. Depending on the country, Western Armenian communities have different degrees of language maintenance, language shift, or endangerment (Jebejian, 2007; Al-Bataineh, 2015; Chahinian and Bakalian, 2016).
In terms of pre-existing resources, Armenian is considered a low-resource language with few computational resources (Megerdoomian, 2009). There are more resources for Eastern Armenian than for Western.[2] For example, Eastern Armenian has the EANC corpus (Khurshudian et al., 2009), a spoken corpus (Skopeteas et al., 2015), corpus-processing tools like UniParser (Arkhangelskiy et al., 2012), a treebank (Yavrumyan et al., 2017; Yavrumyan, 2019), and various Deep Learning tools from the YerevaNN[3] research group (Ghukasyan et al., 2018; Arakelyan et al., 2018). Eastern Armenian is also part of the Universal Morphology schema (Kirov et al., 2018; Chiarcos et al., 2018; McCarthy et al., 2020).
In contrast, there are few if any significant resources for Western Armenian. There is report of a two-level finite-state system (Lonsdale and Danielyan, 2004) but it does not appear to be available. There are some small corpora of Western Armenian (Donabédian and Boyacioglu, 2007; Khachatryan, 2012; Khachatryan, 2013; Silberztein, 2016), and a new UD treebank (Yavrumyan, 2019).[4] Complete verbal paradigms are also available (Boyacioglu and Dolatian, 2020). Thus any contribution to computer processing of Western Armenian currently

[1] The source code for the transducer is available at https://github.com/apertium/apertium-hyw, and the transducer may be used online at https://beta.apertium.org/#analysis?aLang=hyx_hyw.

[2] There are likewise recent resources for Classical Armenian (Vidal-Gorène and Decours-Perez, 2020; Vidal-Gorène and Kindt, 2020), which have been recently applied to the modern lects (Vidal-Gorène et al., 2020): https://calfa.fr/

[3] http://yerevann.com/

[4] https://universaldependencies.org/treebanks/hyw_armtdp/index.html.

(

has the potential to make a large impact.

Note that Vidal-Gorène et al. (2020) develop a quite workable model of Eastern and Western Armenian using Deep Learning. However, this paper sees how far we can go with a rule-based system for the following reasons. First, rule-based methods are more interpretable than neural-based methods, so the designer of the analyzer can directly control the behavior of the analyzer. Second, interpretability allows linguists to directly analyze the analyzer to further their own pen-and-paper analyses (Karttunen, 2006); this is quite important for under-studied langauges. Third, rule-based and neural-based methods aren't in true competition with each other because they have different practical uses. Thus, the rule-based analyzer described here can hypothetically integrate with a neural-based analyzer to cover any gaps (cf. finite-state covering grammar in text normalization: Zhang et al. (2019)).

3. Methodology and implementation

3.1. Software

This transducer was written for use with HFST (Lindén et al., 2011) using the two-level framework (Koskenniemi, 1984; Beesley and Karttunen, 2003; Roark and Sproat, 2007).

The lexicon and morphotactics (combinatorial patterns of morphology) were implemented using `lexd` (Swanson and Howell, 2021), which differs from other formalisms in that it is designed to support non-suffixational patterns, like prefixes. The morphophonology (phonological/orthographic alternations) was implemented using `twolc`. The two separate transducers (morphotactic and morphophonological) are compose-intersected to create both a generator and an analyzer. The bulk of the work was done between October 2020 and January 2021.

3.2. Paradigms

In terms of morphology, Western Armenian is largely agglutinative and it is primarily suffixing. There are some inflectional and derivational prefixes. Verb inflection is primarily agglutinative and synthetic with different suffixes for tense, aspect, agreement, mood, and valency. Verbs are divided into different conjugation classes based on suffix allomorphy, root allomorphy, and other irregularities (Boyacioglu, 2010). For these reasons, we chose to use the "infinitive" forms of verbs as the lemmas, instead of the morphological stems. Similarly, noun inflection is primarily agglutinative with different suffixes for number, case, definiteness, and possession (Hagopian, 2005). To illustrate, we present two morphological forms of a verb in (1) and (2), showing orthographic form, IPA pronunciation, a morpheme-by-morpheme breakdown and gloss,[5] an English translation of the form, and the analysis returned by the transducer.

(1) սիրել [siɾel]
sir -e -l
like TH INF
'to like'
սիրել<v><tv><ger>

(2) սիրեցին [siɾet͡sin]
sir -e -t͡s -i -n
like TH PFV PST 3PL
'they liked'
սիրել<v><tv><past><pret><p3><pl><indc>

The analyses returned by a transducer differ from traditional linguistic analyses in that morpheme breaks are not provided; tags are used instead of abbreviations; word categories (or parts of speech), here VERB or <v>, are annotated; and subcategories of words, here TRANSITIVE or <tv>, are annotated. This particular transducer also differs in that the infinitive is used as the lemma of a verb instead of the morphological stem, and some grammatical labels are different. The tagset used is that provided by Apertium.[6]

To construct this transducer, morphological paradigms were gathered via a combination of pre-existing teaching grammars of Western Armenian (Boyacioglu, 2010; Hagopian, 2005), using cognates from Eastern Armenian grammars (Dum-Tragut, 2009), and native intuition. All paradigms were manually coded into the `lexd` format.

For an irregular word like ճամբայ [d͡ʒampʰa] 'road', the analyzer analyses both standard irregular forms like ճամբու [d͡ʒampʰ-u] (genitive), but also colloquial regularized forms like ճամբայի [d͡ʒampʰaj-i]. However, the generator only produces the standard form.

We added rules to generate some productive derivational processes as well, such as causativization, passivization, and some productive word-forming suffixes like the suffix -որէն *-oɾen* (forms adverbs from adjectives, roughly equivalent to the English suffix *-ly*).

For complex verbs like causatives and passives, we adopted a dual approach to lemmatization and analysis. If the dictionary listed a passive verb like ձգուիլ [t͡səkʰ-v-i-l] 'to be left', then that means that this verb likely developed some opaque semantics when compared to the active form ձգել [t͡səkʰ-e-l] 'to let'. We treated such listed passives as their own lemmas. But for most verbs like երգել [jɛɾkʰ-e-l] 'to sing', most dictionaries don't list the passive երգուիլ [jɛɾkʰə-v-i-l] 'to be sung' because the morphology and semantics are predictable. For such unlisted passives, we derive them at run-time from the lemma of the active. Similar annotation and strategies are used for causatives.

3.3. Lexicon

The lexicon was at first compiled by scraping an Armenian-English dictionary (Kouyoumdjian, 1970) from Nayiri.[7] The dictionary contained at least 60k words.

[5] Glossing conventions and abbreviations are based on Leipzig standards: `https://www.eva.mpg.de/lingua/resources/glossing-rules.php`

[6] `https://wiki.apertium.org/wiki/Symbols`

[7] `http://nayiri.com/`

The dictionary items were catalogued into the right conjugation or declension class. A sample of common Armenian names was gathered from lists of names on different websites.[8] Table 1 provides a breakdown of the lexicon.[9]

	category	entries	tag
Core POS	Noun	39006	`<n>`
	Adjective	18617	`<adj>`
	Verb	7441	`<v>`
	Adverb	1895	`<adv>`
Names	Given name	4848	`<np><ant>`
	Surname	2052	`<np><cog>`
	Location name	1183	`<np><top>`
	Other name	22	`<np><al>`
Function, other	Pronoun	415	`<prn>`
	Adposition	130	`<pr>`, `<post>`
	Abbreviation	81	`<abbr>`
	Conjunction	48	`<conj>`
	Interjection	49	`<ij>`
	Numeral	41	`<num>`
	Particle	9	`<particle>`
	Total	75837	

Table 1: Current lexicon by part-of-speech

3.4. Morpho-phonology

Some morpho-phonological processes are reflected in the orthography. These were implemented through use of special symbols in the morphological side of the morphotactic transducer (`lexd`). Such symbols encode allomorphy and other morphophonological processes. These diacritics were then used in the morphophonological transducer (`twol`) to trigger the appropriate processes.

As an example, the definite suffix is [ə] after consonants (3a) and [n] after vowels (3b). However, with stems ending in the glide letter յ ‹j› (a consonant), the pattern is slightly different: monosyllabic nouns of this sort (3c) behave as expected: the glide is pronounced and the definite suffix is [ə]. But in multisyllabic stems ending in յ ‹j› (3d), the glide letter is silent when not before a vowel, and is not represented orthographically when before a consonant. Hence, in the definite form, the glide letter is not used, and the suffix [n] is added.

(3) Allomorphy of the definite suffix

a.	բառ	‹paṙ›	[pʰɑɾ]	‘word’
	բառը	‹paṙə›	[pʰɑɾ-ə]	‘the word’
b.	կատու	‹gadu›	[gɑdu]	‘cat’
	կատուն	‹gadun›	[gɑdu-n]	‘the cat’
c.	խոյ	‹xoj›	[χoj]	‘ram’
	խոյը	‹xojə›	[χoj-ə]	‘the ram’
d.	ծառայ	‹d͡zaṙaj›	[d͡zɑɾɑ]	‘servant’
	ծառան	‹d͡zaṙan›	[d͡zɑɾɑ-n]	‘the servant’

In our code, the definite suffix was generated in the `lexd` file as the symbol `{dfU}`. The mapping of this to the correct output symbol was conditioned using rules in the `twol` file.

3.5. Infixed punctuation

For punctuation, some punctuation elements are placed outside of words, but others are placed inside words on the stressed vowel. For example, the word [pʰáɾ] ‘word’ when unquestioned is spelled բառ ‹paṙ›. When this word is questioned, the interrogative marker is added on top of the stressed letter: բա՞ռ ‹pa$^{?}$ṙ›. Stress is generally predictable in the language as being word-final while ignoring schwas. Some function words have idiosyncratic stress placement.

To handle word-internal punctuation, we specified a final punctuation marker for every word in the lexicon (`lexd` file). In another transducer built to handle infixed punctuation, also written in the `lexd` formalism, we defined ‘metathesis’ rules to move these final punctuation symbols into the correct word-internal location.

For words with irregular stress, the main lexicon file contained a diacritic to mark this irregular stressed location. For example, the word ‘how much’ has irregular stress on the first vowel: [**vóɾ**kʰɑn]. The question marker is added on the first syllable: ո՞րքան ‹oʾrkʼan›. The lexicon represents this word as `ո{՞}րքան` with a diacritic question mark. Upon intersection with the punctuation transducer, the value of the question marker is changed, moved, or deleted as needed.

4. Evaluation

4.1. Corpora

To perform evaluation, we prepared several corpora.[10] The **Bible corpus** is the contents of a Western Armenian translation of the Bible, available from an Armenian church website.[11] The **News corpus** consists of the the contents of the Kantsasar Armenian News website from Syria.[12] Content was scraped in early November, 2021, using a web spider written using Scrapy.[13] The **Wikipedia corpus** consists of the pages and articles dump of the Western Armenian Wikipedia[14] from January 1, 2022. Text files were extracted from the XML dump.[15] We likewise tested our Western transducer over the **UD Treebank**

[8] The source URLs for these websites are listed as comments in the `.lexd` files for names. Some names were taken from Eastern Armenian sources or were written in the non-conservative orthography. These were manually adapted to Western Armenian spelling conventions.

[9] These numbers reflect the state of the transducer as of mid-January, 2022.

[10] All evaluation was performed on revision `a2ad591`, from mid-January, 2022.

[11] `https://hycatholic.ru/biblia/` The name of the translated edition is not specified, but the translation is stated as being from 1994.

[12] `http://www.kantsasar.com/news/`

[13] `https://scrapy.org/`

[14] `https://hyw.wikipedia.org/`

[15] `https://wiki.apertium.org/wiki/Wikipedia_Extractor`

for Western Armenian (in UD v2.9) (Yavrumyan et al., 2021b). The treebank included a training set, development set, and test set.

4.2. Naive coverage

Naive coverage is the number of forms in a corpus for which the analyzer returns an analysis, regardless of whether the analysis is correct or not. Ambiguity is the average number of analyses returned by the analyzer per analyzed form. Table 2 shows the naive coverage and ambiguity of the Western Armenian transducer on the corpora described in §4.1.

corpus	tokens	coverage	ambiguity
Bible	744K	99.33%	1.54
News	1.78M	95.00%	1.56
Wikipedia	3.56M	90.67%	1.37
UD training	70K	95.33%	1.44
UD dev	9.6K	96.35%	1.48
UD test	10K	96.72%	1.46

Table 2: Naive coverage on Western Armenian

Naive coverage is above 90% for all corpora, and at or above 95% for most. This level of coverage is very high, and should be considered sufficient for many tasks. Many of the top unanalyzed forms are in fact forms from other languages which should not be analyzed, especially in the Wikipedia corpus. Actual missing content in the transducer mostly consists of proper nouns and some rarely occurring stems which are not found in Armenian-English dictionaries.[16] Some tokens are also words from other Armenian dialects, such as Classical Armenian and Eastern Armenian (whether in the traditional or reformed spelling).

Ambiguity is around 1.5, meaning that there are approximately 3 analyses returned for every 2 analyzed tokens. Disambiguation is a task for future work.

4.3. Accuracy

We evaluated the precision and recall of our transducer over a random sample of words. We first retrieved 1300 random tokens from the News corpus. We then cleaned the sample by removing words that were typos, foreign words, words from other dialects or spelling systems, or were words that were so low-frequency that we couldn't find them in any modern dictionary. In all, 1225 tokens were hand-annotated. The results are shown in Table 3.

Tokens	Precision	Recall
1225	90.58%	74.82%

Table 3: Precision and recall measurements

Precision measures how many of the transducer-provided analyses for the tokens were correct. Recall measures how many of the correct analyses were retrieved from the transducer. Although our precision was high at nearly 90%, our recall rate was around 75%. This was because the transducer currently accepts more forms for a given analysis than is correct. This "overanalysis" is due to complications in the variable application of some phonological rules that are reflected in the orthography (vowel reduction), and semantically-induced variation in plural marking (§5.2). Future work would remedy this issue.

[16] A future step would be incorporate digitized Armenian-Armenian dictionaries which can have as many as 100K lemmas.

4.4. Compilation speed

One current weakness of the `lexd` compiler is compilation speed and memory use. As of revision `41b8555`, the transducer took 2 minutes 56 seconds and peak memory usage of 4.29GB to compile using a single core of an Intel i9-9900X CPU (3.50GHz). We were able to optimise many of the definitions by factoring out common subpatterns (revision `49a7487`). After this, compilation on the same system took only 48 seconds with peak memory usage of 387MB. This constitutes a nearly four-fold decrease in speed and an over 11 times decrease in memory usage.

5. Future work

This section briefly outlines our thoughts on how this transducer could be improved through increasing coverage (5.1) and handling overgeneration (5.2). Expansions to handle additional dialects which is a quite complicated problem, postponed to (6).

5.1. Increasing coverage

As stated, our lexicon was based off of a published dictionary that had at least 60k lemmas. Both the original dictionary and its digitized content had a few errors in terms of spelling or part-of-speech assignment. We tried to find as many errors as possible. Future work should go through the entire dictionary more carefully to weed out other errors. We can also cross-reference our dictionary with another dictionary in order to help find other errors or increase coverage. We are currently trying to do so with additional digitized dictionaries from Nayiri.

5.2. Handling overgeneration

One complication for our generator comes from compounds. Compounds are formed by concatenating two stems with a vowel ա /ɑ/ intervening. Compounds are listed as single orthographic words in the dictionary. For inflecting a compounds, knowing the right plural suffix depends on knowing the word's semantics (Donabédian, 2004; Dolatian, 2021). Such information cannot be easily determined from the dictionary, so without further work our generator overgenerates. To fix this issue, a possible future step is to use the lemma list of the EANC, which provides this semantic information.

6. Cross-dialectal support

It would be ideal if the current Western Armenian transducer can interface with a transducer for Eastern Armenian, cf. strategies in Vidal-Gorène et al. (2020). The two

dialects share large portions of their morphology and orthography, and code switching can be found within large corpora.

6.1. Differences between dialects

Eastern Armenian is the official language and dialect of Armenia. It has many morphological differences from Western Armenian, which are reflected in the orthography. Thus a morphological transducer for Western Armenian is not expected to work perfectly for Eastern Armenian, even when orthographic differences are accounted for.

In terms of orthography, up until the mid 20th century, Eastern Armenian in Armenia was written in the Classical Orthography system (Sanjian, 1996). This is the system that is still in use for Western Armenian. But during the Soviet era, various spelling reforms were applied to Eastern Armenian as spoken within the Soviet Union. The current spelling system is called the Reformed Orthographic system. This system applies to Eastern Armenian as spoken in Armenia and most of the Eastern Armenian diaspora. The exception is the Eastern Armenian community in Iran which still uses the Classical Orthography. Some Eastern liturgical literature is still published in the Classical Orthography.

To illustrate, in Table 4, we show the pronunciation and spelling of a passive verb 'to be gathered' for Western and Eastern Armenian. The main morphological difference is that Western Armenian uses a theme vowel ի /-i-/ for passives, while Eastern Armenian uses a theme vowel ե /-e-/. The classical spelling of the passive suffix /-v-/ is ու ‹ow›, while the reformed spelling is վ ‹v›.

		Spelling	
	Pronunciation	Traditional	Reformed
W	[kʰɑʁ-v-i-l]	քաղուիլ	—
	'gather-PASS-TH-INF'	‹kʻaɣowil›	
E	[kʰɑʁ-v-e-l]	քաղուել	քաղվել
	'gather-PASS-TH-INF'	‹kʻaɣowel›	‹kʻaɣvel›

Table 4: Example of orthographic and morphological differences between Western (W) and Eastern (E) Armenian for the form `քաղ<v><iv><pass><inf>`.

6.2. Evaluating the analyzer on Eastern Armenian

For exploratory purposes, we tested our Western transducer on Eastern corpora. We found two Eastern **Bibles**. One Eastern Bible was written with the traditional orthography,[17] and one with the reformed orthography.[18] Besides orthographic differences, the two Bibles are non-identical translations, both against each other and against the Western Bible. For example, the traditional Eastern Bible used more archaic syntactic constructions, obsolete function words, and more footnotes. We also tested the transducer on pages and articles from the Eastern Armenian **Wikipedia**, from January 1 2022.[19] We likewise tested our transducer over the **UD Treebank** for Eastern Armenian (v2.9) (Yavrumyan et al., 2021a), which uses the reformed orthography. In Table 5, we report naive coverage of our Western Armenian transducer on these Eastern Armenian corpora.

corpus	spelling type	tokens	coverage
Bible	traditional	832k	93.61%
Bible	reformed	775k	79.96%
Wikipedia	reformed	62M	67.92%
UD training	reformed	42K	74.65%
UD dev	reformed	5.3K	72.44%
UD test	reformed	5.3K	74.76%
UD BSUT	reformed	3.1K	74.69%

Table 5: Naive coverage on Eastern Armenian corpora

6.2.1. High coverage on the traditional orthography

For Eastern Armenian corpora with traditional spelling, our transducer works quite well: 93% for the Eastern Bible, while 99% for the Western Bible. The high coverage rate is not surprising because the two dialects share the bulk of the same lexicon and derivational/inflectional morphology. They differ significantly in their phonology and pronunciations, but the orthography doesn't show these differences.

The fact that the two dialects have unequal naive coverage is because some inflectional suffixes are present in Eastern but not Western Armenian. Some high-frequency words likewise have different orthographic representations across the two lects. For example, the most common 'unknown' word in the traditional Eastern Bible is 'he said' at 3812 tokens. This word is [ɑsɑt͡sʰ] աuաց ‹asat͡s'› in Eastern Armenian, but [əsɑv] ըսաւ ‹əsav› in Western.

6.2.2. Low coverage on the reformed orthography

The coverage of the Western Armenian transducer over Eastern corpora with the reformed spelling is drastically lower, anywhere between 67% to 79% percent. This difference is likely because of rampant spelling differences across the two spelling systems. For example, the most common 'unknown' word over the reformed Eastern Bible is the word [jev] 'and' at 4026 tokens. This word is spelled as եւ ‹ew› in the traditional system (in both Western and Eastern Armenian) but եվ or և ‹ev› in the reformed system. The reformed Bible that we used almost always used the եվ form.

6.3. Combining the dialects in one analyzer

There are several ways that the transducer could be expanded to support multiple dialects. We have already be-

[17] `http://ter-hambardzum.net/armenia-bible-online/`

[18] `https://hycatholic.ru/biblio/wuuվwծw2nւնչ/`

[19] The Wikipedia (`https://hy.wikipedia.org/`) is primarily written in Eastern with the reformed orthography, but there are some articles in Western or in the traditional orthography.

gun expanding the transducer source code and compilation instructions in one such way. When not the same across dialects, stems and inflectional morphology may be specified on a per-dialect level. This allows the compilation of separate analyzers, separate generators, and a combined analyzer.

7. Conclusions

This paper overviewed the development of a free/open-source morphological analyzer and generator for Western Armenian. In terms of naive coverage, it performs quite well over various Western Armenian corpora. It has high precision and okay recall. It likewise has some coverage over other dialects, thus paving the way for creating a pan-dialectal transducer.

8. Bibliographical References

Adjarian, H. (1909). *Classification des dialectes arméniens*. Librairie Honoré Champion, Paris.

Al-Bataineh, A. (2015). *Cent ans après: Politiques scolaires et la vitalité des langues en danger le cas de l'arménien occidental*. Ph.D. thesis, Sorbonne Paris Cité.

Arakelyan, G., Hambardzumyan, K., and Khachatrian, H. (2018). Towards JointUD: Part-of-speech tagging and lemmatization using recurrent neural networks. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 180–186, Brussels, Belgium, October. Association for Computational Linguistics.

Arkhangelskiy, T., Belyaev, O., and Vydrin, A. (2012). The creation of large-scale annotated corpora of minority languages using UniParser and the EANC platform. In *Proceedings of COLING 2012: Posters*, pages 83–92.

Beesley, K. and Karttunen, L. (2003). *Finite-state morphology: Xerox tools and techniques*. CSLI Publications, Stanford, CA.

Boyacioglu, N. (2010). *Hay-Pay: Les Verbs de l'arménien occidental*. L'Asiatheque, Paris.

Chahinian, T. and Bakalian, A. (2016). Language in Armenian American communities: Western Armenian and efforts for preservation. *International Journal of the Sociology of Language*, 2016(237):37–57.

Chiarcos, C., Donandt, K., Ionov, M., Rind-Pawlowski, M., Sargsian, H., Wichers Schreur, J., Abromeit, F., and Fäth, C. (2018). Universal Morphologies for the Caucasus region. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Cowe, S. P. (1992). Amēn teł hay kay: Armenian as a pluricentric language. In Michael Clyne, editor, *Pluricentric Languages: Differing Norms in Different Nations*, pages 325–346. De Gruyter Mouton.

Dolatian, H. (2021). The role of heads and cyclicity in bracketing paradoxes in Armenian compounds. *Morphology*, 31(1):1–43.

Donabédian, A. and Boyacioglu, N. (2007). La lemmatisation de l'arménien occidental avec nooj. In Svetla Koeva, et al., editors, *Formaliser les langues avec l'ordinateur: De INTEX à NooJ*, Cahiers de la MSH Ledoux, pages 55–76. Presses Universitaires de Franche-Comté, France.

Donabédian, A. (2004). Le nom composé en arménien. In Pierre J.L. Arnaud, editor, *Le nom composé: Données sur seize langues*, pages 3–20. Presses Universitaires de Lyon, Lyon.

Donabédian, A. (2018). Middle east and beyond - Western Armenian at the crossroads: A sociolinguistic and typological sketch. In Christiane Bulut, editor, *Linguistic minorities in Turkey and Turkic-speaking minorities of the periphery*, pages 89–148. Harrazowitz Verlag, Wiesbaden.

Dum-Tragut, J. (2009). *Armenian: Modern Eastern Armenian*. Number 14 in London Oriental and African Language Library. John Benjamins Publishing Company, Amsterdam/Philadelphia.

David M. Eberhard, et al., editors. (2022). *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, online, twenty-fifth edition.

Ghukasyan, T., Davtyan, G., Avetisyan, K., and Andrianov, I. (2018). pioNER: Datasets and baselines for Armenian named entity recognition. In *2018 Ivannikov Ispras Open Conference (ISPRAS)*, pages 56–61. IEEE.

Hagopian, G. (2005). *Armenian for everyone: Western and Eastern Armenian in parallel lessons*. Caravan Books, Ann Arbor, MI.

Jebejian, A. (2007). *Changing ideologies and extralinguistic determinants in language maintenance and shift among ethnic diaspora Armenians in Beirut*. Ph.D. thesis, University of Leicester.

Karttunen, L. (2006). The insufficiency of paper-and-pencil linguistics: The case of Finnish prosody. In Miriam Butt, et al., editors, *Intelligent linguistic architectures: Variations on themes by Ronald M. Kaplan*, number 179 in CSLI Lecture Notes, pages 287–300. CSLI, Stanford, CA.

Khachatryan, L. (2012). Formalization of proper names in the Western Armenian press. In Kristina Vučković, et al., editors, *Formalising Natural Languages with NooJ: Selected Papers from the NooJ 2011 International Conference (Dubrovnik, Croatia)*, pages 75–85. Cambridge Scholars Publishing, Newcastle, UK.

Khachatryan, L. (2013). An Armenian grammar for proper names. In Anaïd Donabédian, et al., editors, *Formalising Natural Languages with NooJ: Selected Papers from the NooJ 2012 International Conference (Paris, France)*, pages 233–234. Cambridge Scholars Publishing, Newcastle, UK.

Khurshudian, V. G., Daniel, M. A., Levonian, D. V., Plungian, V. A., Polyakov, A. E., and Rubakov, S. A. (2009). Eastern Armenian National Corpus. In *Computational Linguistics and Intellectual Technologies (Papers from the Annual International Conference*

"Dialogue 2009", volume 8, pages 509–518, Moscow. RGGU.

Kirov, C., Cotterell, R., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Mielke, S. J., McCarthy, A., Kübler, S., Yarowsky, D., Eisner, J., and Hulden, M. (2018). UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Koskenniemi, K. (1984). A general computational model for word-form recognition and production. In *Proceedings of the 10th international conference on Computational Linguistics*, pages 178–181. Association for Computational Linguistics.

Kouyoumdjian, M. G. (1970). *A comprehensive dictionary, Armenian-English*. Atlas Press, Beirut.

Lindén, K., Axelson, E., Hardwick, S., Pirinen, T. A., and Silfverberg, M. (2011). Hfst—framework for compiling and applying morphologies. In *International Workshop on Systems and Frameworks for Computational Morphology*, pages 67–85. Springer.

Lonsdale, D. and Danielyan, I. (2004). A two-level implementation for western Armenian morphology. *Annual of Armenian linguistics*, 24:35–51.

McCarthy, A. D., Kirov, C., Grella, M., Nidhi, A., Xia, P., Gorman, K., Vylomova, E., Mielke, S. J., Nicolai, G., Silfverberg, M., Arkhangelskiy, T., Krizhanovsky, N., Krizhanovsky, A., Klyachko, E., Sorokin, A., Mansfield, J., Ernštreits, V., Pinter, Y., Jacobs, C. L., Cotterell, R., Hulden, M., and Yarowsky, D. (2020). UniMorph 3.0: Universal Morphology. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France, May. European Language Resources Association.

Megerdoomian, K. (2009). Low-density language strategies for persian and Armenian. In Sergei Nirenburg, editor, *Language Engineering for Lesser-Studied Languages*, pages 291–312. IOS Press, Amsterdam.

Roark, B. and Sproat, R. (2007). *Computational Approaches to Morphology and Syntax*. Oxford University Press, Oxford.

Sanjian, A. K. (1996). The Armenian alphabet. In Peter T. Daniels et al., editors, *The World's Writing Systems*, pages 356–357. Oxford University Press, New York and Oxford.

Sayeed, O. and Vaux, B. (2017). The evolution of Armenian. In Jared S Klein, et al., editors, *Handbook of Comparative and Historical Indo-European Linguistics*, pages 1146–1167. Walter de Gruyter, Berlin/Boston.

Silberztein, M. (2016). *Formalizing Natural Languages: The NooJ Approach*. ISTE LTd and John Wiley & Sons, London and Hoboken, NJ.

Skopeteas, S., Hovhannisyan, H., and Brokmann, C. (2015). Eastern Armenian spoken corpus.

Swanson, D. and Howell, N. (2021). Lexd: A finite-state lexicon compiler for non-suffixational morphologies. In Mika Hämäläinen, et al., editors, *Multilingual Facilitation*, pages 133–146. Helsingin yliopisto.

Vidal-Gorène, C. and Decours-Perez, A. (2020). Languages resources for poorly endowed languages : The case study of Classical Armenian. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3145–3152, Marseille, France, May. European Language Resources Association.

Vidal-Gorène, C. and Kindt, B. (2020). Lemmatization and POS-tagging process by using joint learning approach. experimental results on Classical Armenian, Old Georgian, and Syriac. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 22–27, Marseille, France, May. European Language Resources Association (ELRA).

Vidal-Gorène, C., Khurshudyan, V., and Donabédian-Demopoulos, A. (2020). Recycling and comparing morphological annotation models for Armenian diachronic-variational corpus processing. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 90–101.

Yavrumyan, M. M., Khachatrian, H. H., Danielyan, A. S., and Arakelyan, G. D. (2017). ArmTDP: Eastern Armenian Treebank and Dependency Parser. In *XI International Conference on Armenian Linguistics, Abstracts*, Yerevan.

Yavrumyan, M. M. (2019). Universal dependencies for Armenian. Presented at the International Conference on Digital Armenian, Inalco, Paris, October 3-5.

Zhang, H., Sproat, R., Ng, A. H., Stahlberg, F., Peng, X., Gorman, K., and Roark, B. (2019). Neural models of text normalization for speech applications. *Computational Linguistics*, 45(2):293–337.

9. Language Resource References

Nisan Boyacioglu and Hossep Dolatian. (2020). *Armenian Verbs: Paradigms and verb lists of Western Armenian conjugation classes (v.1.0.0)*. Zenodo.

Marat M. Yavrumyan and Hrant H. Khachatrian and Anna S. Danielyan and Gor D. Arakelyan and Martin S. Mirakyan and Liana G. Minasyan. (2021a). *UD Eastern Armenian ArmTDP*. In *Universal Dependencies 2.9*, ed. Zeman, D., J. Nivre, et al.

Marat M. Yavrumyan and Hrant H. Khachatrian and Anna S. Danielyan and Setrag H.M. Hovsepian and Liana G. Minasyan. (2021b). *UD Western Armenian ArmTDP*. In *Universal Dependencies 2.9*, ed. Zeman, D., J. Nivre, et al.