Recycling and Comparing Morphological Annotation Models for Armenian Diachronic-Variational Corpus Processing

Chahan Vidal-Gorène
École Nationale des Chartes-PSL
65 rue de Richelieu
75003 Paris
chahan.vidal-gorene@chartes.psl.eu

Victoria Khurshudyan
SeDyL, UMR8202,
INALCO, CNRS, IRD
65 rue des Grands Moulins
75013 Paris
victoria.khurshudyan@inalco.fr

Anaïd Donabédian-Demopoulos
SeDyL, UMR8202,
INALCO, CNRS, IRD
65 rue des Grands Moulins
75013 Paris
anaid.donabedian@inalco.fr

Abstract

Armenian is a language with significant variation and unevenly distributed NLP resources for different varieties. An attempt is made to process an RNN model for morphological annotation on the basis of different Armenian data (provided or not with morphologically annotated corpora), and to compare the annotation results of RNN and rule-based models. Different tests were carried out to evaluate the reuse of an unspecialized model of lemmatization and POS-tagging for under-resourced language varieties. The research focused on three dialects and further extended to Western Armenian with a mean accuracy of 94,00 % in lemmatization and 97,02% in POS-tagging, as well as a possible reusability of models to cover different other Armenian varieties. Interestingly, the comparison of an RNN model trained on Eastern Armenian with the Eastern Armenian National Corpus rule-based model applied to Western Armenian showed an enhancement of 19% in parsing. This model covers 88,79% of a short heterogeneous dataset in Western Armenian, and could be a baseline for a massive corpus annotation in that standard. It is argued that an RNN-based model can be a valid alternative to a rule-based one giving consideration to such factors as time-consumption, reusability for different varieties of a target language and significant qualitative results in morphological annotation.

1 Introduction

So far rule-based (RB) approaches prevailed in the annotation of the Armenian varieties which proved to show very good results provided that the system is sufficiently complete and refined (see Khurshudyan et al. (2020) for Modern Eastern Armenian [henceforth MEA]), or more modest ones if the system is perturbed by certain factors (see Vidal-Gorène and Kindt (2020) for Classical Armenian). However, RB systems have the drawback of being considerably time-consuming and not sufficiently reusable for other varieties of the target language.

The current research aims at exploring an alternative recurrent neural network (RNN) approach to annotate Armenian varieties favored for its flexibility and application rapidity on linguistically and structurally various datasets, as well as for the possibility of making predictions on unknown tokens (predominantly on very different corpora) and contextual disambiguation (Dereza, 2018).

RNN approach has already been applied to some Armenian varieties [(Vidal-Gorène and Kindt, 2020) for Classical Armenian and (Arakelyan et al., 2018; Yavrumyan, 2019) for MEA], highlighting competitive advantages for tagging Armenian data. Trained on more modest (Universal Dependencies [UD]) or specialized (GRE*g*ORI project) datasets, the results described are equivalent in lemmatization to Eastern Armenian National Corpus (EANC) rule-based approach and more precise in POS-tagging. The experiments (currently limited to POS tagging and lemmatization) are extended to three Armenian dialect varieties and to the two Modern Armenian Standards and the results are compared to EANC rule-based tools.

The article is structured as follows: *Armenian language preliminaries* give a highlight to Armenian variation in diachronic and synchronic perspectives; the chapter on the *Armenian ressources online* make a state of the art of existing Armenian online open-access corpora and databases. The chapter on the *datasets* focuses on the target datasets designed and used for current research experiment, whereas the

Proceedings of the 7th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects, pages 90–101
Barcelona, Spain (Online), December 13, 2020

chapters on *methodology and results, lemmatization results, POS-tagging results* spotlight the lemmatisation and POS-tagging results which are furthermore compared in RNN and RB approaches. Finally, the last chapter on *MWA model* explores the feasibility of MWA tagging with a MEA model.

2 Armenian language preliminaries

Armenian is an Indo-European language with a nominative-accusative alignment, predominantly with an agglutinative nominal system and with a more fusional verbal one. It is a left-branching language with flexible word order (SVO/SOV).

The periodization of the Armenian language includes: Classical Armenian (henceforth CA)[1] (5^{th}-10^{th} cen. A.D), Middle Armenian (11^{th}-17^{th} cen.) and Modern Armenian (17^{th} cen. – up to present). Modern Armenian includes two standards: Modern Eastern Armenian and Modern Western Armenian, both standardized in the 19^{th} century. MEA is the official language of the Republic of Armenia and it is also spoken by the Armenian communities of Iran and ex-Soviet republics. MWA is spoken by traditional Armenian communities in Europe, Americas and Middle East originating mainly from Ottoman Empire. Aside from the two standards the Armenian language continuum includes various dialects as well as vernacular forms (Figure 1). Classical Armenian is preserved exclusively for canonical uses. The variation in Armenian continuum can vary from light to significant with or without mutual intelligibility. In particular, MEA can vary from MWA less than from certain Armenian dialects which sometimes lack mutual intelligibility [for more details on Armenian varieties and variation see Donabedian-Demopoulos (2018) and Donabedian-Demopoulos and Sitaridou (2021)].

Different classifications exist for the Armenian dialects depending on the criteria applied (e.g. areal (Aytənean, 1866), morphological (Adjarian, 1909), phonological (Gharibyan, 1953), typological-statistical (Jahukyan, 1972), etc.). In our research the morphological criteria prevail due to their importance in annotation processing.

One of the main distinguishing morphological features for Armenian dialects is the formation of the present indicative according to which three main groups (*-um*, *kə* and *-l* branches)[2] can be outlined as shown in Figure 1 [for more details on the Armenian dialects see Martirosyan (2018) as well as Greppin and Khachaturian (1986)].

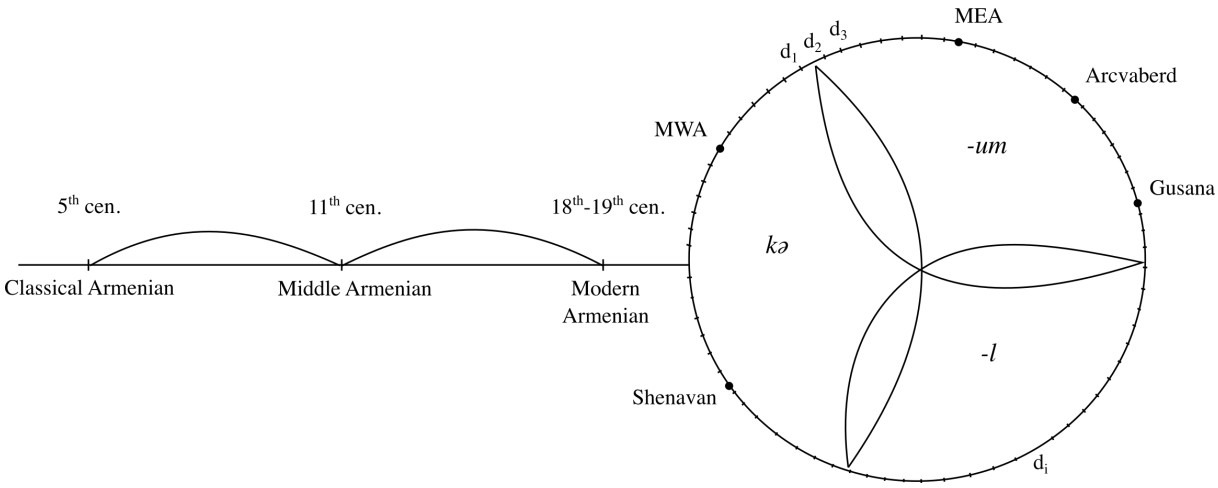

Figure 1: Armenian diachronic and synchronic varieties with d_i corresponding to a dialectal variety

One of the important issues for dialect corpora is how to transcribe the recordings and how to annotate

[1] Classical Armenian traces back to the creation of the Armenian alphabet attributed to the monk Mashtots at the beginning of the 5^{th} century A.D. Currently, the Armenian alphabet is composed of 39 graphemes and it is the only alphabet used for Armenian. At the beginning of the 20^{th} century an orthography reform was carried out in (Eastern/Soviet) Armenia alongside with her sovietization. Currently, two spellings exist for Armenian, since the traditional Armenian diaspora (both Western and Eastern Armenian speakers) continued to preserve the traditional orthographic standard. Technically it requires a conversion system to avoid annotation "noise". Several conversion tools have been developed.

[2] According to this criterion, MEA belongs to the *-um*, whereas MWA belongs to the *kə* branch.

the transcripts. Except for rare specialized dialect corpora, it is usually very difficult (often impossible in case of big corpora) and time-consuming to get a reliable phonetic transcription. An alternative to a phonetic transcription is either a complete standardized (orthographized) transcription (with the condition to have sound alignment) or a semi-standardized (orthographized) one with certain adjustments proper to the target dialect. The most important advantage of the standardized transcription would be the possibility to apply the NLP resources of the standard variety to dialects [for more details on different approaches see Arkhangelskiy and Georgieva (2018) and von Waldenfels et al. (2014)].

3 Armenian resources online

Heterogeneous texts representing various Armenian varieties can be found in a number of online resources which vary in their accessibility, formatting and linguistic background.

1. Classical Armenian. Currently, the most important corpus project with full morphological annotation for Classical Armenian is the Classical Armenian Bible project with parallel King James Version realized by Arak29 foundation[3]. The corpus database contains approximately 630.000 tokens (60.000 unique tokens, 12.000 lexemes) covering a very specific lexicon in Classical Armenian.
GRE*g*ORI project (UCLouvain)[4] is mainly specialized on Hellenophile Classical Armenian texts (6th-7th cen.), thus, Armenian translated texts from Greek. The public database of the project is more modest and it counts 66.812 tokens (16.000 unique tokens) with full morphological annotation and context-disambiguation (Vidal-Gorène and Kindt, 2020).
Several other Classical Armenian text databases exist among which the most significant one is the project of Digital Library of Armenian Literature (American University of Armenia)[5]. The database covers nearly all Classical Armenian texts from the 5th to 18th centuries.
The projects TITUS for Classical Armenian (Johann Wolfgang Goethe University)[6] and the Leiden Armenian Lexical Textbase (University of Leiden)[7] provide searchable databases of the Bible as well as certain historical and hagiographical texts with limited annotation.
The Calfa project[8] is a comprehensive online reference dictionary platform for Classical Armenian with particularly the ongoing project of New Dictionary of the Armenian Language (Awetik‘ean et al., 1836 1837) including 54.000 headwords with 150.000 examples (1.3 million tokens, 190.000 unique tokens) drawn from various Classical Armenian texts provided with full context-disambiguated morphological annotation (Vidal-Gorène et al., 2020).

2. Middle Armenian. No dedicated Middle Armenian corpus or database exists. Certain texts can be randomly found in different databases.

3. a. Modern Western Armenian. No annotated corpus is publicly available for MWA with the exception of a small fully annotated corpus by Nooj (Donabedian-Demopoulos and Boyacioglu, 2007). The Digital Library of Armenian Literature offers the biggest database of MWA texts of the 19th and 20th centuries (1850-2000) with the complete works of 75 authors (about 8.400.000 tokens).

3. b. Modern Eastern Armenian. The largest resource for MEA is the open-access Eastern Armenian National Corpus[9]. EANC is designed as a comprehensive corpus with about 110 million tokens, covering MEA written and oral discourses from the mid-19th century to the present. The texts/transcripts have full morphological, semantic and metatext annotation and they are provided by English translations[10] for frequent tokens searchable for making complex lexical morphological queries. Besides the corpus, EANC proposes also an electronic library with full-view access for over hundreds of works by classical

[3] https://www.arak29.am/bible_28E/index.htm
[4] https://gregoriproject.com
[5] http://digilib.aua.am
[6] http://titus.uni-frankfurt.de/indexe.htm
[7] http://sd-editions.com/LALT/index.html
[8] https://calfa.fr
[9] http://eanc.net
[10] EANC allows search by an English lexeme (for about 85%) with the same functionality as for an Armenian lexeme, e.g. grammatical and lexical features, sentence position, punctuation etc.

authors in public domain. The library provides the same morphological analysis and translation as the rest of the corpus. The EANC annotation relies on a rule-based approach, combining a wordlist (about 80.000 lexemes composed of a combination of different dictionaries (Galstyan, 1985; Aġayan, 1976; Grgearyan and Harutyunian, 1987 1989; Gyurdjinyan and Hekekyan, 2007) and a morphological model. Overall, 92,5% of all tokens are recognized and annotated with 72,6% analyzed unambiguously, 17% ambiguously, and 7,5% not recognized[11] [for more details on EANC see Khurshudyan et al. (2020)].

The Universal Dependencies project includes MEA[12] (Yavrumyan, 2019) providing 2.502 manually annotated sentences in MEA (about 53.000 tokens [v. 2.5]) with morphological and syntactic annotations in the form of a complete dependency tree bank.
Several other databases (not always searchable) provide MEA and MWA texts: Armenian Wikisource project, Fundamental Scientific Library of the National Academy of Sciences of the Republic of Armenia, etc.

3. c. Armenian Dialects. Except for some rare scanned books containing dialectal texts[13] with various types of linguistic accuracy and transcription approaches (not always reusable for linguistic research), no dedicated dialectal resources exist online. The Armenian dialectology started developing from the mid-19th century and has recorded important advances during the 20th century with a number of dedicated dialect descriptions, important attempts to collect dialectal data with a systematic approach throughout fieldworks as well as various types of researches carried out by the Institute of Language, National Academy of Sciences of Armenia. After the collapse of the Soviet Union the scientific thrust was significantly stopped. An attempt to set up a dialectal corpus was made in the framework of EANC research grant project during 2008-2009. Three dialects were chosen (1. Arcvaberd dialect (Shamshadin, Tavush region), 2. Shenavan dialect (Aparan, Aragatsotn region), and 3. Gusana dialect (Maralik, Shirak region) for each of which about 15 hours of recordings were made and transcribed entirely by the grantees[14] (about 100.000 tokens for each dialect corpus, see *infra* Datasets). For each corpus a list of unique wordforms was processed and the grantees provided full morphological annotation manually. The pilot version of the three dialectal corpora is available online[15].

Project	Tokens	Variety	Contextual annotation	Annotation type	Accessibility
Arak29	630.000	CA	no	full	OD
GRE*g*ORI	66.812	CA	yes	lemma, pos	O
Calfa	1,3 million	CA	yes	full	O
EANC	110 million	MEA	no	full	O
UD	53.000	MEA	yes	full	OD
EANC	300.000	dialects	no	full	O

Table 1: Target annotated corpora used in datasets (O = open access, D = downloadable)

Besides the lack of disambiguation for certain target corpora used in our datasets (e.g. in Arak only interlexical homonymy is disambiguated), different projects rely on various tagging systems for POS and morphological annotation (see annotation differences in Table 2 from the examples (1) and (2) for CA and MEA respectively) and sometimes on a various level of lemma annotation (e.g. GRE*g*ORI and Calfa consider *mardoyn* as a polylexical wordform because of the definite article). The datasets were automatically standardized, however, the lack of interoperability can have an impact on the results (see *infra*).

[11] EANC original analyzer was updated by Timofey Arkhangelskiy and Aleksei Fedorenko and current open source version is available at https://bitbucket.org/timarkh/uniparser-grammar-eastern-armenian/.

[12] https://universaldependencies.org/hy/

[13] E.g. one can find the scanned 18 volumes of Armenian folk tales (1959-2016) available at the site of the Fundamental Scientific Library of the National Academy of Sciences of the Republic of Armenia http://serials.flib.sci.am/.

[14] Shushan Asilbekian (Institute of Linguistics, Armenian Academy of Sciences); Garik Mkrtchian (Yerevan State University); Susanna Davtian (Yerevan State University).

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

(1) Mk 7:20

or	inč‘	i	**mardoyn**	elan-ē
which	what	PREP	**man.ABL.SG.DEF**	go.out-3SG

”That which cometh out of the man ...”

(2) UD

ergel	em	Irlandiay-um
sing-PFV	be.AUX.1SG	Ireland-LOC

”I have sung in Ireland.”

	Variety	Wordform	Annotation
Arak29	CA	*mardoyn*	*mard* noun.gen.dat.abl.sg.def
GRE*g*ORI	CA	*mardoyn*	*mard@n* N+Com:Âs
Calfa	CA	*mardoyn*	*mard@n* 1. NOUN:abl.sg@DEF 2. N+COM:Âs@DEF
UD	MEA	*ergel*	*ergel* Aspect=Perf, Polarity=Pos, VerbForm=Part, Voice=Act
EANC	MEA	*ergel*	1. *ergel* (V,intr/tr) cvb, pfv ‘sing’ 2. *ergel* (V,intr/tr) inf ‘sing’

Table 2: Target corpora annotation samples

4 Datasets

Five datasets were set up to conduct experiments (see Figure 2 and Table 3): three dialect variety and two Modern Armenian standard datasets. Besides, three mixed datasets were constituted to assess the potential advantages of mixed data drawn from EANC database. All the datasets have full token morphological analysis (lemma, POS and morphological features).

D-Ab: Arcvaberd dialect (Shamshadin, Tavush region) dataset includes the transcripts of about 15 hours of recordings (16 informants) and 120.258 manually annotated wordforms (14.405 unique). This is the most important and yet the least varied dialect dataset with only 4.120 unique lemmata. **D-Ab** has a significant number of ambiguous forms due to free-form (vs. context-based) annotation. Arcvaberd dialect belongs to the *-um* branch, like the dialect of Gusana, and is considered to be a blend of two *-um* type dialects (Ararat and Karabakh).

D-Ga: Gusana dialect (Maralik, Shirak region) dataset is composed of the transcripts of about 15 hours of recordings (26 informants) and 100.352 manually annotated wordforms (20.647 unique). Although it is equivalent to **D-Ab** by its volume, it is much more varied with 9.087 unique lemmata. As a consequence, much more unknown tokens are found in the associated test set which makes **D-Ga** an interesting benchmark for the evaluation of predictions on unknown tokens. Although the main population of Gusana originates from Kars, Van and partly Mush (immigrated at the beginning of the 19th century) and the village is areally situated in a *kə* branch region, the dialect is of *-um* type (like Arcvaberd dialect) with certain mixed features.

D-Shn: Shenavan dialect (Aparan, Aragatsotn region) includes the transcripts of about 15 hours of recordings (18 informants) and 89.632 manually annotated wordforms (17.940 unique). Proportionally, this is the most varied dataset with 7.568 unique lemmata, thus, with many ambiguous and unknown

forms. Shenavan population immigrated from Mush at the beginning of the 19th century and the dialect originates from Western *kə* branch Mush dialect. However, being in contact with *-um* branch varieties for almost two centuries certain contact-induced changes are present.

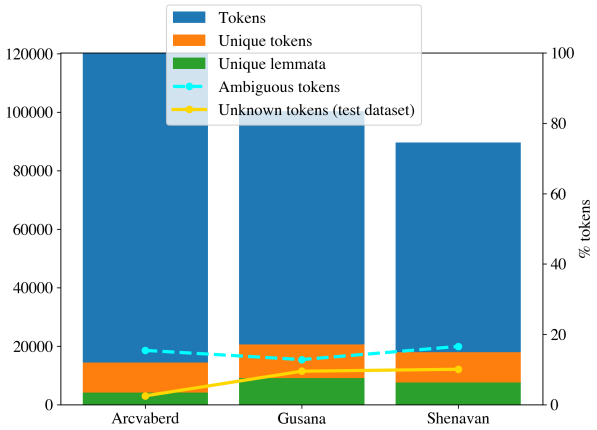

Figure 2: Dialect datasets composition

D-MEA: This dataset is the reference representative dataset for MEA, a subset drawn from EANC database. It counts 5.111.614 tokens (201.710 unique). The sentences are taken from heterogeneous sources: press (2.037.629 tokens), fiction (1.453.894 tokens) and non-fiction (2.031.055 tokens).

D-MWA: This dataset is an experimental dataset for MWA designed for assessing the reuse and reproducibility relevance of the models. The database counts 3.531 tokens (1.788 unique). The sentences are drawn from miscellaneous data such as press, fiction, non-fiction and the Bible.

	D-MEA	**D-Ab**	**D-Ga**	**D-Shn**	**D-MWA**
Wordforms	5.111.614	120.258	100.352	89.632	3.531
Unique tokens	201.710	14.405	20.647	17.940	1.788
Unique lemmata	-	4.120	9.087	7.568	1.311
Ambiguous tokens	-	18.584	12.883	14.844	250
Unknown tokens (test dataset)	13.145	364	1.968	1.810	1.080

Table 3: Dedicated dataset composition

Three more mixed datasets were created: **D-Ab+MEA**, **D-Ga+MEA** and **D-Shn+MEA** composed of the target dialect dataset and increased by a third of its volume with varied data from **D-MEA**.

Finally, three other external datasets were considered for the creation of the models: Universal Dependencies (**D-UD**) for MEA, and GREgORI (**D-CA1**) and Arak29 (**D-CA2**) for Classical Armenian (see *supra* Armenian resources online).

Unlike **D-UD**, **D-CA1** and **D-CA2**, the annotations for **D-Ab**, **D-GA** and **D-Shn** are not context-based (matching between the list of wordforms and the corpus) and its influence on the annotation is discussed below. The evaluation of the different models (see *infra*) showed less considerable results, since **D-Ab**, **D-GA** and **D-Shn** have a phonetic transcription based on the Armenian alphabet and figures[16]. Finally, **D-Ab**, **D-Ga**, **D-Shn**, **D-MEA**, **D-UD** use reformed orthography, while **D-MWA**, **D-CA1** and **D-CA2** use the traditional one. Orthography conversion with some limitations has been performed to allow the evaluation[17].

[16]E.g. հըվընդաւցաւվ həvəndäc^{h}äv '[he/she] fell ill' (vs. the MEA wordform հիվանդացավ hivandac^{h}av).

[17]The orthography conversion was processed by the converter designed by Arak29 foundation available at

5 Methodology and results

A number of tests were carried out to develop an RNN annotation model for three dialects (with manually annotated corpora available) and MWA (no annotated corpus available). Three sets of neural networks have been trained and evaluated:

1. univariational targeted variety RNN model;
2. mixed model (2/3 dialect model +1/3 MEA);
3. univariational non-targeted variety RNN model.

The RNN relies on Pie (Manjavacas et al., 2019), which offers a highly modular architecture particularly designed to process historical (cf. Classical Armenian) and non-standard languages (cf. Armenian dialects). The RNN model adopted in this research was successfully tested on Classical Armenian [for more details on the RNN model used see Vidal-Gorène and Kindt (2020)]. Generally, Pie learning ability exploits fully sentence context to increase the lemmatization accuracy and POS-tagging tasks, particularly in case of ambiguous tokens (Eger et al., 2016; Sprugnoli et al., 2020). However, in our experiments Pie learning ability has been limited because of the unresolved ambiguity of the annotations in **D-Ab**, **D-Ga** and **D-Shn**. Consequently, the RNN preserves either all possible categories or only the most probable one. Although the linear decoder showed better results for Classical Armenian POS-tagging, it was compared with the CRF decoder provided by MarMoT and LEMMING (Mueller et al., 2013; Müller et al., 2015), which obtained convincing results on equivalent datasets at the last Evalatin Evaluation Campaign (Sprugnoli et al., 2020; Stoeckel et al., 2020). The model of the lemmatizer and POS-tagger has been trained jointly using a single multitask architecture.

Finally, the relevance of the architecture was evaluated for a standard language (MEA) on the basis of two datasets (**D-MEA** and **D-UD**). COMBO (Rybak and Wróblewska, 2018) trained with **D-UD** (v. 2.3) is at 88.05% for lemmatization and 85.07% for POS-tagging (Arakelyan et al., 2018; Yavrumyan, 2019). The present architecture (**m-UD**) trained with **D-UD** (v. 2.5) obtains 91.56% in lemmatization (74.35% for the ambiguous tokens and 61,85% for the unknown tokens) and 92.54% in POS-tagging (87,81% ambiguous tokens and 83,56% unknown tokens).

5.1 Lemmatization results

Arcvaberd and Gusana being morphologically of *-um* branch, thus, closer to each other, as well as to MEA, a working hypothesis could be to have more positive annotation overlapping between these two dialect data. On the contrary, Shenavan belonging to the *kə* branch would be morphologically more distinct from the two other dialects and MEA, and closer to MWA (see Figure 1), thus the two other dialect and MEA models could be expected to be less relevant for the annotation of its data.

Specialized models: The results of lemmatization of all the dialect tokens (known and unknown) vary between 92.05% and 97.69% (see Table 4) with greater discrepancy for unknown tokens (from 46.52% to 66.87%) (see Table 5). The **m-Ab** model (trained with **D-Ab** which is the biggest dialect dataset) turns out to be the best performer in the general task (97.69%), but the lack of token and lemma variety leads to poor predictions on unknown tokens, whereas **m-Ga** and **m-Shn** prove to be more robust. The confusion matrix shows that **m-Ab** mostly fails on the verbal forms with no phonetic particularity in the transcript (i.e. formally similar to the standard language forms). **m-Ga** and **m-Shn** generate much more false forms for the same token, in addition to being penalized by the wide variety of phonetic transcriptions reproduced in the corpora. Despite being very robust, **M-MEA** suffers from the ambiguity of the data. The model proves to be more efficient for generating all the possible analyses rather than just one (unlike **m-UD** described previously). It processes successfully 94.34% of **D-UD**.

Mixed models: Adding data from **D-MEA** to **D-Ab**, **D-Ga** and **D-Shn** for training mixed models is relatively advantageous for the Arcvaberd dialect (+ 0.64% in accuracy and + 4.4% in precision and

https://www.arak29.am/template/_msconv.php/

recall), including the prediction on unknown tokens ranging from 46.52% to 51.10%. On the other hand, this disadvantages Gusana and Shenavan dialects (see *infra* Non-specialized models for a possible explanation).

Non-specialized models: Similar to the models trained on Classical Armenian (**m-CA1** and **m-CA2**) a model strictly trained on MEA (**m-MEA** and **m-UD**) does not currently allow dialect lemmatization. However, these results should be nuanced, since **D-MEA**, **D-Ab**, **D-Ga** and **D-Shn** are transcribed very differently and in reformed spelling. **D-CA1** and **D-CA2** also have differences at lemma description level (e.g. lemmas in -em and not in -el for verbs) which results in a large number of false negatives despite automatic smoothing. Arcvaberd and Gusana dialects being linguistically close to each other (see *supra*) show right lemmatization for less than 50% (49.47% and 46.38% respectively), which, nevertheless, is a better result than **D-Ab** annotation by **m-Shn** (42.90%). **M-Shn** correctly annotates **D-Ga** at 58.45%, while **D-Shn** annotated by **m-Ga** is at 52.32%. Mixed models provide better results (see Table 4 and Table 5).

Models	Lem. **D-MEA**	POS-t. **D-MEA**	Lem. **D-Ab**	POS-t. **D-Ab**	Lem. **D-Ga**	POS-t. **D-Ga**	Lem. **D-Shn**	POS-t. **D-Shn**
m-MEA	A: 0.9870 P: 0.8859 R: 0.8774	0.9974 P: 0.9989 R: 0.9976	A: 0.3576	-	0.3219	-	0.4264	-
m-Ab	-	-	0.9769 P: 0.7795 R: 0.7796	0.9894 P: 0.992 R: 0.9893	0.4947	0.6288	0.4695	0.6487
m-Ga	-	-	0.4638	0.6627	0.9228 P: 0.6332 R: 0.6229)	0.9645 P: 0.74 R: 0.7019	0.5232	0.7553
m-Shn	-	-	0.4290	0.7452	0.5845	0.8173	0.9205 P: 0.6509 R: 0.6398	0.9569 P: 0.8384 R: 0.8218
m-Ab+MEA	-	-	0.9833 P: 0.8235 R: 0.8219	0.9912 P: 0.9959 R: 0.9899	0.5010	0.6744	0.5010	0.6579
m-Ga+MEA	-	-	0.4856	0.6833	0.9151 P: 0.6205 R: 0.6064	0.9700 P: 0.7867 R: 0.7793	0.5460	0.7534
m-Shn+MEA	-	-	0.4337	0.7684	0.5650	0.8222	0.9166 P: 0.6396 R: 0.6246	0.9645 P: 0.7479 R: 0.7268
m-UD	0.6508	-	0.2035	-	0.2117	-	0.2513	-
m-CA1	0.2616	-	0.1364	-	0.1607	-	0.1787	-
m-CA2	0.2930	-	0.2104	-	0.2435	-	0.2397	-

Table 4: Lemmatization and POS-tagging evaluation of the models on **D-MEA**, **D-Ab**, **D-Ga**, and **D-Shn** for all tokens (A = accuracy, P = precision and R = recall)

5.2 POS-tagging results

Taking into account the limits exposed for the lemmatization task described above, the results in POS-tagging are much more regular. All the POS-tagging evaluations could not be performed due to the excessive conventional variation in the annotation of the corpora (morphological and lexical tagging, formatting, transcription etc.). The conventional discrepancies existing in different projects is an important issue for conducting additional experiments.

Specialized models: POS-tagging models provided significant results (> 95%) for all the dialects including unknown tokens. **M-Ab** is the least efficient because of its diversity. More than two thirds of the errors are caused by the confusion between noun and adjective, which can be explained by the absence of context and especially by the potential homonymy between these two categories, since in Armenian adjectives (and other parts of speech functioning as an attribute) can be easily nominalized, thus having formal endings similar to nouns. In EANC nominalized adjectives are tagged as A, NMLZ which facilitates the search of the target matches.

Mixed models: Adding MEA data to the dialects improves significantly the results for POS-tagging annotation, in particular for unknown tokens (see Table 3).

Non-specialized models: The reuse of models between dialects prove viable in particular for Gusana and Shenavan. The two target dialects do not belong to the same linguistic branch (see *supra*), and yet **m-Shn+MEA** provides 82.22% for **D-Ga**. Moreover, even though **m-Ga** covers only 66.27% for **D-Ab** it may provide a considerable basis for faster annotation of dialect corpora.

Models	Lem. **D-MEA**	POS-t. **D-MEA**	Lem. **D-Ab**	POS-t. **D-Ab**	Lem. **D-Ga**	POS-t. **D-Ga**	Lem. **D-Shn**	POS-t. **D-Shn**
m-MEA	A: 0.9239 P: 0.8487 R: 0.8455)	0.9614 P: 0.9108 R: 0.9497	A: 0.1950	-	0.1996	-	0.1937	-
m-Ab	-	-	0.4652 P: 0.2984 R: 0.2897	0.7406 P: 0.3299 R: 0.3284	0.3758	0.5474	0.3856	0.5280
m-Ga	-	-	0.2504	0.5278	0.6687 P: 0.4861 R: 0.468	0.8318 P: 0.4862 R: 0.3992	0.4067	0.6729
m-Shn	-	-	0.2929	0.5933	0.5468	0.7456	0.6547 P: 0.4700 R: 0.4547	0.8083 P: 0.3369 R: 0.3292
m-Ab+MEA	-	-	0.5110 P: 0.3170 R: 0.3064	0.7634 P: 0.4016 R: 0.4020	0.3803	0.5853	0.3799	0.5097
m-Ga+MEA	-	-	0.3241	0.5629	0.6660 P: 0.4826 R: 0.4639	0.8507 P: 0.3850 R: 0.4074	0.4135	0.6609
m-Shn+MEA	-	-	0.2986	0.6365	0.5043	0.7508	0.6459 P: 0.4553 R: 0.4347	0.8358 P: 0.4726 R: 0.3994
m-UD	0.5545	-	0.0916	-	0.1162	-	0.1301	-
m-CA1	0.1991	-	0.1161	-	0.1599	-	0.1553	-
m-CA2	0.2371	-	0.2072	-	0.2240	-	0.2165	-

Table 5: Lemmatization and POS-tagging evaluation of the models on **D-MEA**, **D-Ab**, **D-Ga**, and **D-Shn** for the unknown tokens (A = accuracy, P = precision and R = recall)

5.3 MWA Model

D-MWA was first processed by **m-MEA** model after which the predicted lemmata were checked and manually corrected. The results were compared to the rule-based EANC parser predictions (see Table 4).

ID	Original	Converted	Lemma			POS		
			GT	RNN	RB-EANC	GT	RNN	RB-EANC
1	*ałǰikə*	*ałǰikə*	*ałǰik*	*ałǰik*	*ałǰik*	N	N	N
2	*grkac*	*grkac*	*grkel*	*grkel*	*grkel*	V	V	V
3	*ēr*	*ēr*	*ē*	*ē*	*ē*	V	V	V
4	*etewi*	*etewi*	*etev*	*etev*	*etew*	N	V	N
5	*kołmēn*	*kołmen*	*kołm*	*kołmel*	*kołmel*	N	N	V
6	*t'ewerə*	*t'ewerə*	*t'ev*	*t'ev*	*t'ew*	N	N	N
7	*anut'nerēn*	*anut'neren*	anut'	*anut'*	*anut'nerel*	N	N	V
8	*anc'uc'ac*	*anc'uc'ac*	*ancənel*	*anc'uc'el*	*anc'uc'el*	V	V	V
9	*u*	*u*	*u*	*u*	*u*	CONJ	CONJ	CONJ
10	*ir*	*ir*	*ink'ə*	*ir*	*ink'ə*	PRON	N	PRON
11	*k'it'ə*	*k'it'ə*	*k'it'*	*k'it'*	*k'it'*	N	N	N
12	*hetzhetē*	*hetzhete*	*hetzhete*	*hetzhete*	*hetzhete*	ADV	ADV	ADV
13	*aweli*	*aveli*	*aveli*	*aveli*	*aveli*	A	V	ADV
14	*ke mxrčuēr*	*kmxrčver*	*mxrčvil*	*mxrčvel*	*kmxrčvel*	V	V	V
15	*anor*	*anor*	*an*	*aner*	*aner*	PRON	N	N
16	*akanǰin*	*akanǰin*	*akanǰ*	*akanǰ*	*akanǰ*	N	N	N
17	*etew*	*etew*	*etev*	*etev*	*etew*	N	N	N
18	*zmayleli*	*zmayleli*	*zmayleli*	*zmayleli*	*zmayleli*	A	A	A
19	*anušahotut'eamb*	*anušahotut'yamb*	*anušahotut'yun*	*anušahotut'yun*	*anušahotut'yun*	N	N	N
20	*mə*	*mə*	*mə*	*mə*	*mə*	INDEF	NUM	N
21	*glxē*	*glxe*	*glux*	*glxel*		N	A	
22	*elac*	*elac*	*elnel*	*elnel*	*elnel*	V	V	V
23	:	:	:	:	:	PUNCT	PUNCT	

Table 6: **m-MEA** and EANC rule-based parser comparison of lemmatization and POS-Tagging for **D-MWA** data

The **m-MEA** model provides 88.79% correct lemmatization and 87.33% correct POS-tagging on the **D-MWA**. EANC parser (RB-EANC) obtains 74.09% and 68.57% respectively for the same dataset. As

shown in Table 6, the original spelling of the text was converted in order to make **m-MEA** operational. The errors are focused on radically different lemmas between MWA and MEA (e.g. ըllաl əllal/ լինել linel "to be", certain pronouns (#10 and #15 in Table 6), indefinite article (#20 in Table 6), etc.) which are easily identifiable and could be corrected in the future. **M-MEA** processes successfully lemmas from unknown declined forms (e.g. ablative forms, #7 in Table 6).

The **m-MEA** model can allow rapid corpus processing. Manual correction of such a corpus can provide a specialized MWA model which is of utmost importance for MWA documentation, an endangered language with crucially decreasing native speakers (Donabedian-Demopoulos, 2000; Donabedian-Demopoulos and Al-Bataineh, 2014). **D-MWA** was first processed by **m-MEA** model and predicted lemmata were further checked and manually corrected. The results were compared to the rule-based EANC parser predictions (see Table 6).

6 Conclusion

Different experiments were carried out to illustrate for the first time the automatic morphological annotation of Armenian dialect varieties, and the possible reuse of non-specialized models for rapid corpus processing. The first results are more than relevant with very accurate models specialized on the dialects and MEA showing more than 92% accuracy in lemmatization and 95% in POS-tagging.

The mixed and non-specialized models prove to be insignificant for the annotation rate improvement. These models lack interoperability between the target databases which is detrimental to the models and results in producing many false negatives.

The experiments show considerable relevance in model reuse for Armenian diachronic and variational data, as illustrated more particularly by MWA target test (88.79% in lemmatization and 87.33% in POS-tagging).

The standardization and harmonization of annotation conventions is one of the further challenges. The upcoming experiments will be extended to context-based and full morphological annotation. The experimental data show that parallel to the rule-based approaches RNN models can be a sound alternative to process Armenian diachronic and variational corpora.

No precise estimations are available for the models trained on RB tagged corpora to assess RNN time-consumption. However, current tagging results for MWA and CA allow to take into account an iterative and mixed approach which can partially cover the annotation of non-specialized models reducing the time-consumption for new corpora design.

Language varieties have usually fragile vitality when lacking the "standard" status (cf. dialects) and/or natural regenerating native speakers' rotation and evolution (cf. MWA). The first is true for the Armenian dialects which have always "secondary" status as compared to the standard language with which they coexist. MWA is a standard language and yet it has become mainly a diaspora/heritage language for more than a century. Therefore, the documentation of these Armenian varieties as well as the processing of the documented data is of foremost importance not only in NLP but also and especially in linguistic, anthropological and social perspectives.

References

Hratchia Adjarian. 1909. *Classification des dialectes arméniens*. H. Champion, Paris, France.

Gor Arakelyan, Karen Hambardzumyan, and Hrant Khachatrian. 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. Association for Computational Linguistics.

Timofey Arkhangelskiy and Ekaterina Georgieva. 2018. Sound-aligned corpus of Udmurt dialectal texts. In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, pages 26–38, Helsinki, Finland. Association for Computational Linguistics.

Gabriēl Awetik‘ean, Xač‘atur Siwrmēlean, and Mkrtič‘ Awgerean. 1836–1837. *New Dictionary of the Armenian Language*. Tparan i Srboyn Łazaru, Venice, Italia.

Arsēn Aytənean. 1866. *K‘nnakan k‘erakanowt‘iwn ašxarhabar kam ardi hayerēn lezowi [Critical Grammar of the Vernacular or Modern Armenian Language]*. Vienna, Austria.

Eduard Aġayan. 1976. *Ardi hayereni bac‘atrakan baŕaran [Explanatory dictionary of Modern Armenian language]*. Yerevan, Armenia.

Oksana Dereza, 2018. *Lemmatization for Ancient Languages: Rules or Neural Networks?*, pages 35–47. Springer International Publishing, Cham.

Anaïd Donabedian-Demopoulos and Anke Al-Bataineh. 2014. L’arménien occidental en France : dynamiques actuelles. Research report, SeDyL UMR8202 (Inalco, CNRS, IRD).

Anaïd Donabedian-Demopoulos and Nisan Boyacioglu. 2007. La lemmatisation de l’arménien occidental avec NooJ. In S. Koeva, D. Maurel, and M. Silberztein, editors, *Formaliser les langues avec l’ordinateur, de INTEX à NooJ*, pages 55–75. Presses Universitaires de Franche Comté.

Anaïd Donabedian-Demopoulos and Ioanna Sitaridou. 2021. Anatolia. In E. Adamou and Y. Matras, editors, *The Routledge Handbook of Language Contact*, pages 404–433. Routledge, London, England.

Anaïd Donabedian-Demopoulos. 2000. Langues de diaspora, langues en danger : le cas de l’arménien occidental. In *Les langues en danger*, Mémoires de la Société de Linguistique de Paris, Nouvelle Série, t. VIII, pages 137–156. Société de Linguistique de Paris, Paris, France.

Anaïd Donabedian-Demopoulos. 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*, volume 111 of *Turcologica*, pages 89–148. Harrazowitz Verlag, Wiesbaden, Allemagne.

Steffen Eger, Rüdiger Gleim, and Alexander Mehler. 2016. Lemmatization and Morphological Tagging in German and Latin: A comparison and a survey of the state-of-the-art. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 1507–1513, Portorož, Slovenia. European Language Resources Association (ELRA).

E. Galstyan. 1985. *Hay-ŕuseren baŕaran [Armenian-Russian dictionary]*. Yerevan, Armenia.

Ararat Gharibyan. 1953. *Hay barbaŕagitut‘yun [Armenian Dialectology]*. Yerevan, Armenia.

John A. C. Greppin and Amalya Khachaturian. 1986. *A handbook of Armenian dialectology*. Caravan Books, Delmar N.Y., USA.

Hagop Grgearyan and Nora Harutyunian. 1987-1989. *Ašxarhagrakan anunneri baŕaran [Dictionary of Geographic Names]*. Yerevan.

Davit Gyurdjinyan and Narine Hekekyan. 2007. *Hayerenum gorcacvoł taŕayin hapavumneri baŕaran [Dictionary of acronyms used in Armenian]*. Yerevan, Armenia.

Gevorg Jahukyan. 1972. *Hay barbaŕagitut‘yan neracut‘yun [Introduction to Armenian Dialectology]*. Yerevan, Armenia.

Victoria Khurshudyan, Timofey Arkhangelskiy, Michael Daniel, Dmitri Levonian, Vladimir Plungian, Alex Polyakov, and Sergey Rubakov. 2020. Introduction to Eastern Armenian National Corpus: www.eanc.net. *Études arméniennes contemporaines*. submitted.

Enrique Manjavacas, Kádár Ákos, and Kestemont Mike. 2019. Improving Lemmatization of Non-Standard Languages with Joint Learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1493–1503, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Hrach Martirosyan. 2018. The Armenian Dialects. In Geoffrey Haig and Geoffrey Khan, editors, *The languages and linguistics of Western Asia: an areal perspective*, The world of linguistics, 6, pages 46–105. De Gruyter Mouton, Berlin, Boston, USA.

Thomas Mueller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient Higher-Order CRFs for Morphological Tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA. Association for Computational Linguistics.

Thomas Müller, Ryan Cotterell, Alexander Fraser, and Hinrich Schütze. 2015. Joint Lemmatization and Morphological Tagging with Lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2268–2274, Lisbon, Portugal. Association for Computational Linguistics.

Piotr Rybak and Alina Wróblewska. 2018. Semi-Supervised Neural System for Tagging, Parsing and Lematization. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 45–54, Brussels, Belgium. Association for Computational Linguistics.

Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, and Matteo Pellegrini. 2020. Overview of the EvaLatin 2020 Evaluation Campaign. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 105–110, Marseille, France. European Language Resources Association (ELRA).

Manuel Stoeckel, Alexander Henlein, Wahed Hemati, and Alexander Mehler. 2020. Voting for POS tagging of Latin texts: Using the flair of FLAIR to better Ensemble Classifiers by Example of Latin. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 130–135, Marseille, France. European Language Resources Association (ELRA).

Chahan Vidal-Gorène and Bastien Kindt. 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. European Language Resources Association (ELRA).

Chahan Vidal-Gorène, Aliénor Decours-Perez, Baptiste Queuche, Agnès Ouzounian, and Thomas Riccioli. 2020. Digitalization and Enrichment of the Nor Baṙgirk‘ Haykazean Lezui: Work in Progress for Armenian Lexicography. *Journal of the Society of Armenian Studies*, 27. submitted.

Ruprecht von Waldenfels, Michael Daniel, and Nina Dobrushina. 2014. Why standard orthography? Building the Ustya River Basin corpus, an online corpus of a Russian dialect. In *Компьютерная лингвистика и интеллектуальные технологии*, pages 720–728.

Marat Yavrumyan. 2019. Tek‘sti mek‘enakan hatuyt‘avorumə arewelahayereni šarahyusakan caṙeri UD Armenian-ArmTDP bankum [Tokenization and Word Segmentation in the UD ARMENIAN-ArmTDP Treebank]. *Banber Erewani hamalsarani*, pages 52–65.