AAPG2021	DALiH		PRC
Coordinated by:	Victoria Khurshudyan	42 months	
scientific evaluation committee 8.6 (38)			
Interest for TAP « AquaticPollutants »* □			

Digitizing Armenian Linguistic Heritage (DALiH):
Armenian Multivariational Corpus and Data Processing

Summary table of persons involved in the project:

Partner	Name	First name	Current position	Role & responsibilities in the project (4 lines max)	Involvement (person.month) throughout the project's total duration
INALCO / SEDYL	KHURSHUDYAN	Victoria	Maître de conférences	Coordinator Linguistics, NLP, WP0-6	27,3 p.m (65%)
INALCO / SEDYL	DONABEDIAN	Anaïd	Professeur des Universités	Linguistics, WP1-6 PHD co-supervision	14,7 p.m (35%)
INALCO / SEDYL				PhD student to be hired within the framework of the project, WP1-4	36 p.m
INALCO / SEDYL				Master student intern to be hired within the framework of the project, WP1-3	6 p.m
LIPN	TOMEH	Nadi	Maître de conférences	Partner's scientific leader NLP, WP1-2, WP4-6 PostDoc co-supervision	14,7 p.m (35%)
LIPN	CARTIER	Emmanuel	Maître de conférences	NLP, WP1-2, WP4-6 PostDoc co-supervision	8,4 p.m (20%)
LIPN	CHARNOIS	Thierry	Professeur des Universités	NLP, WP1-2, WP4-6 PostDoc co-supervision	8,4 p.m (20%)
LIPN				Postdoctoral researcher to be hired within the framework of the project, WP2-WP4	24 p.m
ERTIM	NOUVEL	Damien	Maître de conférences	Partner's scientific leader NLP, WP3-6 PHD co-supervision	14,7 p.m (35%)
ERTIM	WANG	Ilaine	Research engineer	NLP, WP3-6	10,5 p.m (25%)
ERTIM				Research engineer to be hired within the framework of the project, WP3-6	18 p.m
Digilib	KIZOGHYAN	Hovhannes	Technical director	Partner's scientific leader Philology, NLP, WP1, 5, 6	14,7 p.m (35%)
RAS	PLUNGIAN	Vladimir	Professor	Partner's scientific leader Linguistics, WP1-WP6	14,7 p.m (35%)
RAS	KOCHAROV	Petr	Associate professor	Partner's scientific leader Linguistics, WP1, WP2, WP4-6	14,7 p.m (35%)

The scientific leader of LIPN Emmanuel Cartier has been replaced by Nadi Tomeh, given his greater specialization on the NLP tasks detailed during the second phase, and his greater availability for the project. Emmanuel Cartier still remains a permanent member of LIPN in the DALiH project.

AAPG2021	DALiH		PRC
Coordinated by:	Victoria Khurshudyan	42 months	
scientific evaluation committee 8.6 (38)			
Interest for TAP « AquaticPollutants »* □			

I. Proposal's context, positioning and objective(s)

The project **Digitizing Armenian Linguistic Heritage: Armenian Multivariational Corpus and Data Processing (DALiH)** aims at building for the first time an open-access and open-source unified digital linguistic platform for the whole spectrum of Armenian language variation. Each language variety will be represented by a comprehensive corpus which will be provided with full morphological annotation. More particularly, DALiH will be the first to design six new annotated corpora for 1) Classical Armenian; 2) Modern Western Armenian; 3) a pilot corpus of Middle Armenian; 4) three pilot corpora of dialects, and 5) one updated Modern Eastern Armenian corpus on the basis of the existing one. The principal stages for a corpus development in the pipeline include source texts'/transcripts' compilation, annotation system/model elaboration with or without a grammatical dictionary compilation, annotating/lemmatizing the texts and publishing the annotated corpus using a search engine.

Research will be conducted in Natural language processing (NLP) and linguistic perspectives in order to provide full grammatical annotation and Automatic speech recognition (ASR) models for the Armenian varieties. Multi-approach deep-learning and rule-based resources will be designed in order to process the written and oral databases and to cross-check their value for further corpus enlargement, in a context of multiparameter language variation for an under-resourced language.

NLP-based linguistic researches, such as language identification and variety distance measuring, lexical and morphological disambiguation, will be carried out to revisit the existing research issues and to introduce new ones backed by the new available processed written and oral data.

Within the framework of DALiH a number of essential open-access and open-source NLP and linguistic resources and tools will be designed and made available for the researchers' and wide-reaching users' community. This in turn will foster further NLP research and enhance resource developing in order to process written and oral Armenian multivariational data in particular and cross-linguistically for other under-resourced languages in general.

Armenian language preliminaries, existing resources and corpora

Despite being a language with a multisecular written tradition, Armenian lacks significantly digital resources for NLP and linguistic research. Several important projects for particular Armenian varieties exist, as well as a growing interest in NLP resources is observed.

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

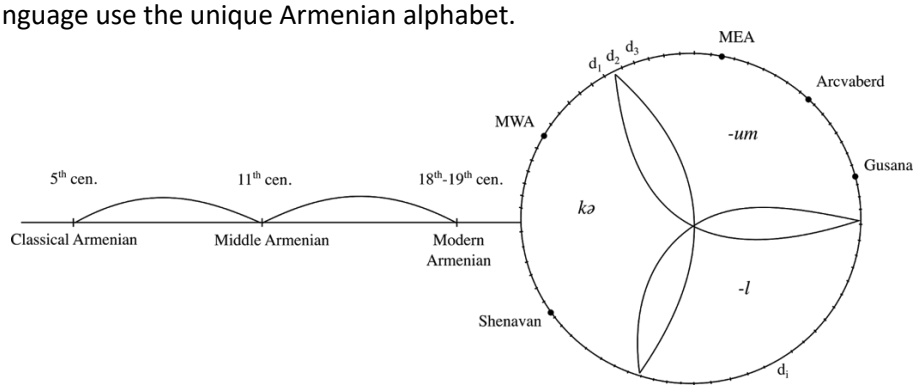

Figure 1: Armenian diachronic and synchronic varieties with d_i corresponding to a dialectal variety (Vidal et al. 2020:91).

The varieties may vary from partial to no mutual intelligibility, with in particular Classical Armenian being a flexional language with rich morphology (almost completely unintelligible for today's speakers), Middle Armenian being in between Classical and Modern varieties and containing significant variation (almost completely unintelligible for today's speakers) and finally Modern Western and Eastern Armenians and the three target dialects with more agglutinative morphology. Contrary to the two standards between which a considerable mutual intelligibility exists, the dialects are greatly incomprehensible for the speakers of the two standards (for more details on Armenian varieties and dialects see (Donabedian 2018; 2021), and (Martirosyan 2018)).

The most significant *Classical Armenian* resources include the Digital Library of Armenian Literature (Digilib) with its significant plain-text database; the Bible corpus (approximately 630.000 tokens with 60.000 unique tokens and 12.000 lexemes) and certain canonic texts with full morphological annotation and English alignment by the Arak29 foundation; a more modest linguistic corpus (66.812 tokens with 16.000 unique tokens) specialized in Hellenophile Classical Armenian texts (6th-7th cen.) by the GREgORI project (UCLouvain), the Calfa project with its comprehensive online Classical Armenian reference dictionary platform (1.3 million tokens, 190.000 unique tokens), and several other Classical Armenian searchable databases of the Bible and certain historical and hagiographical texts with limited annotation (TITUS project, the Leiden Armenian Lexical Textbase).

To the best of our knowledge no dedicated *Middle Armenian* corpora are available.

To the best of our knowledge no dedicated *Modern Western Armenian* (MWA) corpora are available. 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).

The largest resource for *Modern Eastern Armenian* (MEA) is the Eastern Armenian National Corpus (EANC), an open-access, comprehensive corpus with about 110 million tokens (mid-19th cen. to the present) provided with full morphological annotation using a rule-based approach. The texts/transcripts have full morphological, semantic and metatext annotation and they are provided with English translations. EANC is searchable for making complex lexical, morphological queries. The EANC annotation relies on a rule-based approach, which combines a wordlist (about 80.000 lexemes compiled from different dictionaries) with a morphological model (Khurshudyan et al. 2009, 2021).

The project of Universal Dependencies for MEA provides 2.502 manually annotated sentences with about 53.000 tokens [v.2.5] with morphological and syntactic annotations in the form of a complete dependency tree bank (Yavrumyan 2020). Several other resources provide MEA and/or MWA OCRed or scanned plain-text databases (Armenian Wikisource project, Fundamental Scientific Library of the National Academy of Sciences of the Republic of Armenia, etc.). Rare tools such as spellcheckers and orthography converters exist for the two modern standards.

The only *Armenian dialectal* corpus has been designed within the framework of EANC as a pilot project. It includes fully annotated 250 000 tokens in transcripts corresponding to about 40 hours of recordings.

While current ASR systems for Armenian do exist, they are either commercial or not accessible for research, such as Google's. Incidentally, with the availability of open-source ASR toolkits (namely HTK, Kaldi (Povey et al. 2011) and ESPnet (Watanabe et al. 2018)), there is an increasing number of ASR systems and pre-trained models that are accessible, such as DeepSpeech (Hannun et al. 2014), wav2vec 2.0 (Baevski et al. 2020), and even user-friendly like Elpis (Foley et al. 2019, Adams et al. 2021).

The existing resources are heterogeneous in terms of accessibility, formatting, linguistic background and they are usually specialized in only one type of a tool/resource (scanned text and/or plain-text databases, dictionaries, annotation models/tools, annotated corpora and datasets etc.). However, they do not cover any Armenian variety completely (with the exception of EANC for MEA), let alone the whole Armenian variation heritage. Furthermore, basic NLP resources are even more scarce (see *Table 1* for a summary of the details on the existing and new resources to be developed within the framework of DALiH).

Resources	Project	Plain-text data	Resources/Tools
Classical Armenian (CA)			
Existing resources	Arak29	AD: ~ 1m. tokens (Bible+liturgy)	AD: List of 12.000 lexemes/ 60.000 unique annotated wordforms e.g. mard noun.gen.dat.abl.sg.def
	Gregori	AD: 66.812 tokens	
	Digilib	AD: ~ 3,5 m.	
	Calfa	AD: ~ 1.3m. tokens	AD: Dictionaries, e.g. mard@n 1.NOUN:abl.sg@DEF, 2.N+Com: s@DEF
	Titus	AD: ~ 600 000 tokens (Bible)	
Resources to be created	**DALiH**	10m. tokens	1. CA annotated corpus (10m. tokens) 2. CA annotation models (RB, RNN, transformers, hybrid) 3. Grammatical dictionary (with about 50 000 lexemes and full grammatical and semantic tagging) 4. CA golden annotated dataset (~ 500 000 tokens) 5. Aligned annotated corpora (English, French, Russian, Greek)
Middle Armenian (MA)			
Resources to be created	**DALiH**	1m. tokens	1. MA annotated corpus (1m. tokens) 2. MA annotation models (RNN, transformers, hybrid) 3. MA annotated dataset (~ 50 000 tokens)
Modern Eastern Armenian (MEA)			
Existing resources	Eastern Armenian National Corpus (EANC)	A: 110m. tokens AD: ~ 10m. tokens	1. MEA annotated corpus (110 m.) 2. MEA full morphological modelling ex. 1. ergel (V,intr/tr), cvb, pfv 'sing', 2. ergel (V,intr/tr), inf 'sing'
	UD	AD: 53 000 tokens	53 000 tokens (2500 sentences) with full annotation ex. ergel Aspect=Perf, Polarity=Pos, VerbForm=Part, Voice=Act
	Ruscorpora	A: ~ 2,4m. tokens	1. MEA-RU annotated corpus (2,4m. tokens)
	Wikipedia	C: ~ 274m. tokens	
Resources to be created	**DALiH**	110 +10m. tokens	1. MEA annotated corpus (10m. + 274m. WIkipedia+ 110m. EANC) 2. MEAannotation models (RB, RNN, transformers, hybrid) 3. Updated grammatical dictionary (~ 70 000 lexemes with full grammatical and semantic tagging) 4. MEA annotated written dataset (~ 0,5m. tokens) 5. AD: aligned annotated corpora (English, French, Russian, Greek) 6. AD: MEA oral sound-aligned corpus (~ 0,5m. tokens) 7. AD: MEA oral sound-aligned dataset (~ 50 000 tokens) 8. AD: MEA ASR model
Modern Western Armenian (MWA)			
Existing resources	Digilib	AD: ~ 3 m.	
	Wikipedia	C: mixed with MEA	
Resources to be created	**DALiH**	15m. tokens	1. MWA annotated corpus (15m. tokens) 2. MWA annotation models (RB, RNN, transformers, hybrid) 3. MWA Grammatical dictionary (~ 50 000 lexemes with full grammatical and semantic tagging) 4. MWA annotated dataset (~ 0,5m. tokens) 5. aligned annotated corpora (English, French, Russian, Greek) 6. MWA oral sound-aligned corpus (~ 100 000 tokens) 7. MWA oral sound-aligned dataset (~ 10 000 tokens) 8. MWA ASR model
Armenian Dialects			
Existing resources	EANC	A: ~ 300 000 tokens	A
Resources to be created	**DALiH**	~ 300 000 tokens	1. Three dialectal annotated corpora (100 000 tokens x 3) 2. Dialectal annotation models (RNN, transformers, hybrid) 3. Three dialectal annotated dataset (~ 50 000 tokens) 4. Three dialectal sound-aligned corpora (~ 100 000 tokens x 3) 5. Dialectal ASR models
Multivariational			
Resources to be created	**DALiH**	~ 450m. tokens	1. Multivariational annotated corpora (~ 450m. tokens) 2. Multivariational annotation models (RB, RNN, transformer, hybrid) 3. Multivariational ASR models

Table 1: Summary table of the existing Armenian corpora/resources (A = open access, D = downloadable)

AAPG2021	DALiH		PRC
Coordinated by:	Victoria Khurshudyan	42 months	
scientific evaluation committee 8.6 (38)			
Interest for TAP « AquaticPollutants »* □			

The DAliH project is organized into seven Work Packages which regroup different tasks and different partners. The current sanitary situation has been taken into account and the project is structured as shown in the Gantt chart below:

WP	Tasks/Month	Year 1											Year 2											Year 3											Year 4					
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42
WP0 DALiH project coordination	Task 0.1 DALiH general project coordination												M1						M2					M3							M4						M5				0.1	M6
WP1 Armenian multivariational plain-text, aligned/comparable and oral database compilation	Task 1.1 Classical Armenian database compilation (10mln tokens)											1.1																													
	Task 1.2 Middle Armenian database compilation (1 mln tokens)												1.2																													
	Task 1.3 MWA database compilation (15 mln tokens)																	1.3																									
	Task 1.4 MEA database compilation (10 mln + EANC 110 mln tokens)												1.4																													
	Task 1.5 Dialectal database compilation (~ 100 000 tokens x 3 dialects)																	1.5																									
	Task 1.6 Aligned and comparable database compilation																	1.6																									
WP2 NLP and linguistic research and resources on grammatical annotation	Task 2.1 Unsupervised Language model with multivariational embeddings																						2.1																				
	Task 2.2 Rule-based morphological annotation models for CA and MWA																	2.2																									
	Task 2.3 RNN-based model with active-learning and cross-lingual transfer																						2.3																				
	Task 2.4 Hybrid annotation model based on embeddings, RNN and rule-based models																															2.4											
WP3 NLP research and resources on Automatic Speech Recognition / Speech-to-text	Task 3.1 Oral data pre-processing and segmentation																				3.1																						
	Task 3.2 Speech-to-text alignment																							3.2																				
	Task 3.3 ASR model building and evaluation																													3.3													
	Task 3.4 Annotated downloadable oral datasets: MEA - 50k tokens, MWA - 10k tokens, dialects - 3x5k tokens																		3.4											3.4													
WP4 Language variation identification and distance measuring	Task 4.1 Language variation identification and distance measuring through linguistic and statistical rules and forms																																			4.1						
	Task 4.2 Language variation identification and distance measuring through speech-to-text																																			4.2						
WP5 Armenian multivariational	Task 5.1 Annotation platform												5.1											5.1													5.1				5.1	
	Task 5.2 Search engine setup for Armenain multivariational corpora												5.2											5.2													5.2					
	Task 5.3 Website development and content												5.3											5.3													5.3					5.3
WP6 DALiH Dissemination	Task 6.1 DALiH workshops												6.1												6.1										6.1			6.1			
	Task 6.2 DALiH shared tasks																								6.2											6.2					
	Task 6.3 DALiH conference																																							6.2		
	Task 5.5 DALiH public launching																																							5.5		
	Task 6.4 DALiH publications												6.4												6.4												6.4					6.4

AAPG2021	DALiH		PRC
Coordinated by:	Victoria Khurshudyan	42 months	
scientific evaluation committee 8.6 (38)			
Interest for TAP « AquaticPollutants »* □			

WP0: DALiH project coordination, 1-42 (month)

Responsible/Participant: Victoria KHURSHUDYAN, SeDyL, INALCO

Work Package Tasks	Objectives / Deliverables
Task 0.1. DALiH project coordination	○ Coordinating the overall activity of the project ○ Regular reports

Work Package 0 includes the overall coordination and management of technical, financial and human resource aspects of the project including the management of the consortium, all other collaborators and outsourcing over its whole duration. All coordination and management issues will be carried out by the project coordinator Victoria Khurshudyan (SeDyL/Inalco). The coordinator will be supported by an administrative board composed of the scientific leaders. Regular trimestral and annual meetings will be held between the consortium members to state and to discuss the progress of the project followed up by accompanying reports.

WP1: Armenian multivariational plain-text, aligned and oral corpora compilation, 1-18 (month)

Responsible/participant(s): SeDyL, Digilib, Vladimir Plungian, LIPN

Work Package Tasks	Objectives / Deliverables
Task 1.1 Classical Armenian data compilation	Classical Armenian database of ~10m. tokens with metadata
Task 1.2 Middle Armenian data compilation	Middle Armenian database of ~1m. tokens with metadata
Task 1.3 MWA data compilation	○ MWA database of ~15m. tokens with metadata including ○ MWA database of ~ 100 000 tokens of audio transcribed
Task 1.4 MEA data compilation	○ MEA database of ~10m. + EANC 110m. tokens with metadata ○ MEA database of ~ 0,5m. tokens of audio transcribed
Task 1.5 Dialectal data compilation	○ Three dialectal databases of recordings of 3x15h ○ Three dialectal databases of transcripts of 3x~100 000 tokens
Task 1.6 Aligned and comparable data compilation	○ MWA and MEA comparable Wikipedia corpus (~274 m. tokens) ○ Armenian-multi-lingual comparable Wikipedia corpus ○ Armenian multivariational aligned translated corpus (~5 m. tokens) ○ Armenian-multi-lingual aligned translated corpus (~5 m. tokens)

Work Package 1 covers tasks related to the compilation of source text and oral databases, their pre-processing and metadata tagging which is the first step in corpus development. Since there are several potential sources of Armenian texts, the compilation task will aim at reusing existing downloadable texts/databases, web crawling of selected sites, OCRing new texts and including oral discourse recording and transcripts. Each different source of text calls for its own particular methodology and also demands a different amount of effort pertaining to its inclusion in the corpus. Oral data is fundamentally important for any corpus, as it represents spoken language.

DALiH compilation tasks for all the target Armenian varieties will be particularly meticulous for the quality assessment parameters for a corpus database which include size (number of tokens it contains), representativeness (texts of different kinds, e.g. genres, authors, time periods etc.) and balancedness (representativeness in comparable proportions).

Seven target Armenian varieties will be processed to various degree of granularity in the framework of DALiH: Classical Armenian (10m. tokens), Middle Armenian (1m. tokens), Modern Western Armenian (15m. tokens), Modern Eastern Armenian (10m. tokens update), three dialects (Agulis, Getashen and Stepanakert) (100k tokens for each).

Task 1.1 Classical Armenian data compilation

Responsible/participant(s): SeDyL, Digilib

The compilation task of Classical Armenian written database will start by amassing the existing databases by Arak29 (Bible and liturgy texts, ~ 1m. tokens) and the Digilib (texts of various genres, ~ 3,5m. tokens). In order to have a comprehensive database it will further be completed by new texts through OCRing by outsourcing (~ 6m. tokens). The compiled and OCRed databases will be pre-processed and the Digilib will provide the metadata processing of each text (author, date of creation, genre etc.) for which they have a long and sound experience. The output of the task will be the processed database (~10m. tokens) provided with metadata which will serve as the basis for NLP and linguistic tasks in WP2, WP4 and WP5.

Task 1.2 Middle Armenian database compilation

Responsible/participant(s): SeDyL, Digilib

Middle Armenian having no data available in any resources, the data compilation task will consist in mainly OCRing the target data (~ 6m. tokens) through outsourcing. The compiled data will be pre-processed and Digilib will furnish metadata for each text. The output of the task will be the processed database (~1m. tokens) provided with metadata which will serve as the basis for NLP and linguistic tasks in WP2, WP4 and WP5.

Task 1.3 MWA database compilation

Responsible/participant(s): SeDyL, Digilib

Similar to Classical Armenian MWA data compilation will include compiling the existing databases, namely, Digilib MWA database (~ 3m. tokens) as well as web crawling for other possible plain-texts (e.g Wikipedia, which contains a limited number of articles in MWA with the majority of Armenian Wikipedia being, however, in MEA). Once the database is completed by new OCRed texts (~ 5m. tokens), it will be pre-processed and text metadata tagging will be added by Digilib.

Besides the written database, an oral one (~100k tokens) will be compiled to supplement it. The oral audio data exists already from previous SeDyL members' fieldworks and it will be transcribed and completed if necessary.

The output of the task will be the processed written (~ 15m. tokens) and oral (~100k tokens) databases provided with metadata which will serve as the basis for NLP and linguistic tasks in WP2, WP3, WP4 and WP5.

Task 1.4 MEA database compilation

Responsible/participant(s): SeDyL

MEA database will mainly contain EANC database (~110m. tokens), completed by new texts from different available resources (various media sites, target fiction projects, publishing houses, Wikipedia etc.) (~10m. tokens + 274m. Wikipedia). A transcribed oral database from EANC (total oral corpus is ~ 3,5m. tokens) will be maximally catalogued with the available sound. The written (~400m. tokens) and oral (~ 0,5m. tokens) databases will further get pre-processing and metadata tagging to be reused in WP2, WP3, WP4 and WP5.

Task 1.5 Dialectal database compilation

Responsible/participant(s): SeDyL

Three dialectal databases will be compiled by first recording the data (3 x 15 hours) through two fieldworks in Armenian. The recordings will be further transcribed manually and (3x~100k tokens) provided with metadata (place, time of recording, respondent's sex, age etc.). The pros and cons of different transcription methods will be weighed up to choose between a phonetic, standardized and semi-standardized transcription with the last two ones allowing to apply existing standard variety NLP resources to dialects (Arkhangelskiy & Georgieva 2018; Waldenfels et al. 2014). The three databases containing transcripts and audio recordings will be pre-processed and transmitted for other tasks in WP2, WP3, WP4.

Task 1.6 Aligned and comparable database compilation

Responsible/participant(s): SeDyL, Digilib, Vladimir Plungian, Vahram Atayan, LIPN

This task will include two subtasks 1. aligned and 2. comparable data compilation. The aligned corpus corresponds to the aligned translated versions of the same text (either in another Armenian variety or in another language), whereas the comparable database will include Wikipedia articles which do not contain exactly the same text in different languages. SeDyL and the foreign partners including Digilib, Vladimir Plungian (RAS) and Vahram Atayan (Heidelberg University) will be in charge of aligned data compilation. More particularly SeDyL will assume Armenian multivariational and multilingual aligned corpora (~ 5m. tokens), Vladimir Plungian will be in charge of Russian-Armenian and Armenian-Russian fiction and poetry aligned corpora (~ 2,4m. tokens) as a collaboration with the Russian National Corpus. Vahram Atayan will provide German-Armenian and Armenian-German aligned corpora (~ 2m. tokens) within the framework of his ongoing project "Philology – Technology – Translation" funded by the German Academic Exchange Service. Finally, LIPN will process Wikipedia Armenian multivariational (~274 m. tokens) and multilingual comparable databases. All the processed databases will be reused in WP2, WP4 and WP5.

WP2: NLP and linguistic research and resources on grammatical annotation, 1-41 (month)

Responsible/participant(s): LIPN, ERTIM, SeDyL, PostDoc, PhD, Vladimir Plungian, Petr Kocharov, Digilib, Timofey Arkhangelsky

Work Package Tasks	Objectives / Deliverables
Task 2.1 Unsupervised Language model (LM) with multivariational embeddings	Transformer-based LMs for all the target varietiesLivrable
Task 2.2 Rule-based morphological annotation models for CA and MWA (+MEA)	○ CA Grammatical dictionary with full morphological and semantic tagging ○ MWA Grammatical dictionary with full morphological and semantic tagging ○ CA modelling of the Inflection types (paradigms) ○ MWA modelling of the Inflection types (paradigms) ○ CA RB (UniParser) morphological analyzer ○ MWA RB (UniParser) morphological analyzer
Task 2.3 RNN-based model with active-learning and cross-lingual transfer	○ RNN-based annotation models for all the target varieties ○ cross-lingual transfer LMs for all the target varieties
Task 2.4 Hybrid annotation model based on embeddings, RNN and rule-based models	○ Hybrid LMs for all the target varieties ○ Annotated downloadable datasets for all target varieties: CA - 0,5m. tokens, MA – 50k tokens, MEA - 0,5m. tokens, MWA - 0,5m. tokens, dialects – 30k tokens

Work Package 2 focuses on the research issues concerning grammatical annotation models for multivariational data, blending NLP and linguistic perspectives. The objectives are multiple: first to provide state-of-the-art resources and tools for full morphological contextual annotation for the Armenian multivariational data; second, to make available essential Armenian-based end-products to a large community for various types of reutilization, and last to further research by introducing new perspectives on the methods and risks of processing languages with 'scarce' data as well as advancing alternatives to quantity-data oriented tendances (cf. Bender et al. 2021).

The research will be conducted in three directions:

1. A more linguistic *rule-based* approach (UniParser) using language modelling through a set of rules and associating it with the grammatical dictionary, and
2. Deep-learning-based approaches with various language models (LM): *a neural network-based* LM, in particular Recurrent Neural Network [RNN] built on top of fine-tuned transformer-based embedding and trained on annotated corpora and allowing predictions for new data.
3. and finally, a *hybrid* model which exploits the advantages of other models while adapting their drawbacks, and finally cross-fertilizing them. The hybrid model could be a fallback for specific Armenian varieties less prone to taking advantage of transfer learning methods due to the lack of data quantity or data quality.

Task 2.1 Unsupervised Language Models and Multivariational embeddings

Responsible/participant(s): LIPN, ERTIM, PostDoc

Recent developments in NLP deep-learning domain converged on the advantages of pre-trained language models, based on *word embeddings* (Mikolov et al., 2013) and even more recently on *contextual embeddings*, such as BERT and m-BERT (Devlin et al., 2019). These last models are trained on a large amount of un-annotated texts, taking into account the bi-directional context. The resulting representations capture various useful linguistic properties, from morphology to semantics. As for now, only a *word embeddings* model exists for Armenian, based on (mainly) Eastern Armenian Wikipedia (Grave et al., 2018). It is relatively small and does not provide sufficient coverage for all varieties. The same corpus is included in the training data of multilingual language models such as m-BERT, where a single model is learned from several monolingual corpora, and all languages are embedded in the same vector space. The multilingual approach has two advantages. It avoids training a model for each language or variety; it also benefits from structural similarities among languages to learn cross-lingual features. This kind of training can be considered as a form of transfer from resource-rich languages to poorer ones.

In DALiH the training of both *monolingual* and *multilingual* LMs will be extended to additional multivariational corpora. These additional resources either already exist but were not used in existing models, or will be collected during the project (see Table #ref). To tackle the resource scarcity issue, information from different levels of granularity will be incorporated into our models (Ma et al. 2020). This includes characters, character n-grams and word segments which do not necessarily correspond to a meaningful morphological analysis. Units of increasing sizes can be composed through the neural

architecture and used alongside word representations. This typically increases the lexical coverage of the model and reduces the effect of out-of-vocabulary tokens.

The result will be a set of transformer-based language models which vary on two axes: a) the choice of tokenization and the resulting vocabulary; b) the training corpus varieties and whether the model is part of multilingual training and thus subject to transfer from other languages.

Resulting language models will be evaluated using the standard perplexity measure. The associated embeddings will also be evaluated, intrinsically on word analogy tasks (Avetisyan et al., 2019) and extrinsically on downstream morphological and syntactic analysis tasks.

Task 2.2 Rule-based morphological annotation models for CA and MWA (+MEA)

Responsible/participant(s): SeDyL, PhD, Vladimir Plungian, Petr Kocharov, Digilib, Timofey Arkhangelsky

So far, rule-based approaches (RB) have prevailed in the annotation of the Armenian varieties (Khurshudyan et al. 2009; 2021) and they proved to show good results for one target variety. Within the framework of DALiH this approach will be applied to the "standardized" Armenian varieties, i.e. Classical Armenian and MWA and also MEA for which such a model already exists.

The workflow of CA and MWA RB model processing will, thus, include, the compilation of an extensive grammatical dictionary of 1. CA with a wordlist of over 54 000 headwords (based on NBHL dictionary (Awetik'ean et al. 1836-37)), and MWA (started by the SeDyL team) containing a wordlist of some 100 000 headwords (based on (Čērēčean et al. 1992) and (Malxaseanc' 1944-1945) dictionaries). CA and MWA morphological modelling and an exhaustive system of description of all the inflectional types will be carried out to provide each headword with morphological (POS; inflectional type; full morphological annotation (e.g. case, number, tense, mood, etc.) and semantic tagging (e.g. inanimate-animate (±human), toponym-first/family name (±F/M) etc.) as well as with English and French translations. Further on, language modelling will be carried out in the UniParser[1] format designed by a DALiH collaborator Timofey Arkhangelsky (University of Hamburg), which has been already tested on the MEA data (on the basis of a grammatical dictionary processed by EANC team and more particularly by Victoria Khurshudyan). The recall of the analyzer on MEA literary texts is about 93%, e.g. 93% tokens receive at least one analysis. The morphological analyzer will, thus, include a list of inflectional paradigms and a grammatical dictionary of the two target varieties. The dictionary will be composed of a lexeme list, with each lexeme detailing the information about its stem, part-of-speech tag and some other semantic/lexical information, its inflectional type (paradigm), and their English translation. The analyzer will also provide stem glossing which is the most conventional format for presenting language data by the linguists.

In the framework of DALiH MEA grammatical dictionary will be updated and reapplied to the target data. Besides, MEA as well as Middle Armenian and dialectal data will be annotated in a hybrid RNN-transformer-rule-based approach recycling the grammatical dictionaries and annotated corpora of the three other target varieties of the project (CA, MEA, MWA).

The linguistic background of the task will be realized by the SeDyL team, Petr Kocharov, Vladimir Plungian and Digilib for CA, whereas MWA subtask will be conducted by the SeDyL team and Vladimir Plungian. The morphological analyzer will be processed by Timofey Arkhangelsky with the SeDyL team.

Task 2.3. RNN-based model with active-learning and cross-lingual transfer

Responsible/participant(s): LIPN, PostDoc

Lemmatization and part-of-speech (POS) tagging of CA and MEA has been addressed in previous work by members of the project (Vidal, Kindt, 2020; Vidal et al. 2020). The main architecture considered is the bi-directional RNN architecture proposed by PIE (Manjavacas et al., 2019). PIE proposes a contextual analysis using an encoder-decoder adapted to the case of poorly endowed languages. PIE has already been evaluated on a specialized dataset of classical Armenian for lemmatization and POS-tagging tasks, with scores of 90.44% (accuracy) in lemmatization and 92.38% (accuracy) in POS-tagging, with a dataset of 66 000 tokens. This model was then fine-tuned with the data by EANC (see Table #ref) to evaluate the compatibility of a model of one linguistic variation with another (transfer from Classical Armenian to

[1] https://bitbucket.org/timarkh/uniparser-grammar-eastern-armenian/src/master/

Eastern Armenian, and from Eastern Armenian to dialects and Western Armenian), with a good average reproducibility of 85%, despite obvious incompatibilities between datasets. This work will be extended in three ways.

a. **Active learning for assisted annotation.** First, the feasibility of transferring the RNN model trained on the available subset of varieties to all varieties will be studied and its performance on a manually annotated dataset for each variety will be evaluated. Specialized (target variety trained) and unspecialized (non-target variety pre-trained) training and testing datasets for each Armenian variety will be established to evaluate the intervariational transfer for the annotation in the different neural architectures selected and the possible strategies to make the models polyvalent. This amounts to model transfer between varieties, which can also be seen as a zero-shot learning scenario. This transfer will allow us to get a first round of annotations for the newly created corpora for varieties unseen during training. Since the objective is to assist manual annotation of text using the model, uncertainty sampling in an active learning scenario is used. In this approach to active learning, the model selects a portion of the automatically annotated data which is to be examined and corrected manually by linguists. These data are then used to enhance the model. All created data will be labeled by iterating model training and manual annotation through this approach. An annotation tool integrating active learning will be used (see WP5).

b. **Multitask learning.** In the second extension the annotation model will be enhanced itself. The model is extended to *jointly* perform lemmatization, morphological analysis and POS tagging on un-annotated corpora. Explicitly modeling the interactions between these tasks is beneficial for all of them especially when annotated resources are scarce. Such joint modeling relies on multitask learning with deep neural networks. Neural networks are particularly suited for multitask scenarios and have been consistently showing positive results. The basic and most-studied mechanism is based on modeling each task separately while sharing parameters of general-purpose hidden layers between models. Different tasks are then trained separately in a supervised way and shared parameters are adapted to all tasks. The tasks are organized in a predefined architecture based on linguistic hierarchies where increasingly complex tasks happen at successively deeper layers.

c. **Cross-lingual model transfer.** Besides model transfer through pre-trained multilingual representations from LMs such as mBERT, models trained for morphology and syntax can be transferred from a resource-rich language to Armenian. In fact, mBERT has been shown to be effective in zero-shot cross-lingual transfer. Task-specific annotations in a *foreign* language (morphological annotation for English for instance) are used to fine-tune mBERT which is then evaluated in Armenian without any annotated data. Transfer has been successful for some morphological and syntactic tasks, but not for all languages (Pires et al., 2019). This transfer is more likely to succeed if the foreign language is from the same family or if there is significant lexical overlap. Not only the transfer from multiple languages to Armenian will be tested but also the multilingual performance of mBERT will be improved by aligning the embeddings from different languages with Armenian embeddings prior to the transfer. Two approaches will be considered: (a) the first is based on learning a linear transformation of embeddings from a foreign language space to Armenian space, while (b) the second directly aligns language subspaces within mBERT. Both approaches rely on existing bilingual dictionaries and parallel corpora. In this context available bilingual dictionaries and translation-aligned corpora between Armenian, Russian and other languages will be used. Evaluation of all models developed in this task will be conducted on manually annotated datasets for each variety and annotation task.

Task 2.4 Hybrid annotation model based on embeddings, RNN and rule-based models

Responsible/participant(s): LIPN, ERTIM, SeDyL, PostDoc, PhD

Besides evaluating the relevance of the three approaches on the basis of the same datasets, a new hybrid approach (with a system of rules allowing RNN-based predictions and improving unknown token processing) will be designed and evaluated. The variational corpora will be iteratively annotated (automatic annotation <> manual correction) to gradually improve the annotation models and the entries of compiled dictionaries. The experiments to create new RNN models trained on certain dialect and Modern Eastern Armenian data show considerable relevance in model reuse for Armenian diachronic and variational data (about 74% on average (Vidal et al. 2020; 2021)). The new hybrid approach combining expertise in digital

humanities, machine learning and linguistics will, thus, allow to strike a balance between time-consumption and high-quality annotation.

Since rule-based (RB) models are considerably time-consuming and not entirely reusable for other varieties of the target language, deep-learning and/or hybrid models will be used as a fallback to cushion possible RB malfunctioning alongside non-target RB models testing for Middle Armenian and dialectal data which have more fluctuations in regards to its the variational forms. Furthermore, the RNN approach has proved to be a valid alternative to the rule-based one for under-resourced multivariational languages (Vidal et al. 2020; 2021; Vidal & Kindt 2020) considering its flexibility and rapidity of application to linguistically and structurally various datasets, as well as its ability to make predictions for unknown tokens and to carry out contextual disambiguation. Transformer-based models appear to be even more efficient and flexible cross-linguistically showing better results in annotation predictability than RNNs using unsupervised or supervised data, however, their efficiency is proportionally dependent on the data quantity which is problematic for most of the Armenian varieties and RNNs have been proven to be more efficient than transformers with "less" data. The four different approaches with their data quantity- and quality-biased advantages and disadvantages will be thus applied to the Armenian varieties with extremely unequal resource distribution to check their viability and to ensure fallback solutions.

Once the annotation models are designed and the data annotation completed, datasets for all the target varieties will be processed by iterative cleaning-up and downloadable datasets will be made available in open-access and open-source. More particularly the annotated datasets will include: 0,5m. tokens for Classical Armenian, 50k tokens for Middle Armenian, 0,5m. tokens for MEA, 0,5m. tokens for MWA and 30k tokens for the three dialects.

WP3: NLP research and resources on Automatic Speech Recognition / Speech-to-text, 1-41 (month)

Responsible/participant(s): ERTIM, SeDyL, PhD, research engineer

Work Package Tasks	Objectives / Deliverables
Task 3.1 Oral data pre-processing and segmentation	Oral data subset pre-processing and segmentation for WP2: MEA -~0,5 m., MWA - ~10k, dialects - ~10k tokens x3
Task 3.2 Speech-to-text alignment	Speech-to-text alignment for all target oral varieties: MEA, MWA and 3 dialects
Task 3.3 ASR model building and evaluation	ASR model building and evaluation all target oral varieties: MEA, MWA and 3 dialects
Task 3.4 Annotated downloadable oral datasets	Annotated oral datasets: MEA – 50k tokens, MWA - ~10k tokens

Work Package 3 focuses on the oral data processing of the Armenian varieties. Oral data compiled in WP1 including EANC oral corpus for MEA (~ 3 m. tokens of spontaneous and public speech, TV shows, task-oriented discourse etc.) will be used to train an Automatic Speech Recognition (ASR) system in order to pre-transcribe the data for annotators and experts to review.

Task 3.1 Oral data pre-processing and segmentation

Responsible/participant(s): SeDyL, ERTIM, PhD

The first task will be to deliver a clean spoken corpus of Armenian (~ 540 00 tokens) for WP2 and it is a two-stage goal: MEA data will be processed within the first year of the project, whereas MWA and dialectal data will be delivered by the third year due to the fact that MEA oral data is already available via EANC. MEA data will be pre-processed and segmented into utterances. This last step is necessary because speech turns may contain several utterances, which hinders POS taggers' performance as they are trained on written sentences. For the other varieties, the model trained on MEA data will be first applied, using an iterative method and providing our experts with data with a decreasing Word Error Rate (WER) at each new iteration. Pre-processing and segmentation rules will be applied on the whole corpus (0,5 million tokens for MEA, 0,1 million for MWA and 100k for each dialect).

Task 3.2 Speech-to-text alignment

Responsible/participant(s): SeDyL, ERTIM, PhD

In order to train an ASR system and study Armenian acoustically, the segmented transcripts from Task 3.1 will be taken a step further and aligned with the speech. This process will be done semi-automatically, using a forced alignment tool to pre-process the data that DALiH experts will realign properly. Forced alignment consists in matching a given transcript to the sound, commonly on the word level, and sometimes with the

help of automatic phoneme identification. In DALiH aeneas will be used since it relies on a text-to-speech engine that was implemented (naively) for Armenian (MEA and MWA) and can be fine-tuned. The resulting sound-aligned corpus is what linguists need to study speech as it allows automatic phoneme identification as well as the computation of diverse acoustic measures (pitch levels, f0 etc.). It will also serve as the training corpus for Task 3.3.

Task 3.3 ASR model building and evaluation

Responsible/participant(s): research engineer, Damien Nouvel, Ilaine Wang

The main challenge of WP3 is the training and evaluation of one or several ASR models for the Armenian varieties. Most state-of-the-art ASR tools require hundreds or thousands of transcribed data as the training dataset, but the recent rise of interest for low- and medium-resource languages such as Armenian pushed some of them to address the challenge to offer models that require a restricted transcribed dataset. Among those tools, wav2vec 2.0 catches our attention not only as it seems to work with very few transcribed data (1 hour or even less (Baevski et al. 2020)) but also because it should be easy to implement considering its availability on the very well documented platform HuggingFace. By using wav2vec 2.0 iteratively together with a decoder (a language model that matches the acoustic embeddings with words) such as wav2letters++ (Pratap et al. 2019) more newly corrected data can be adda increasingly at each iteration and help speed up manual correction. Once the training set reaches a substantial size, other approaches will be possible to be tested, including *transfer learning* from a high-resource language, as studies showed that they give good results if fine-tuned with at least 20 (Mohamud et al. 2021) or 35 hours (Hjortnæs et al. 2020) of transcribed data of the target language. Interestingly, Mohamud et al. (2021) showed that applying a self-supervising model trained on a given language as the backbone produces "indistinguishable results on languages originating from the same family". Thus, a model trained on MEA can be expected to work efficiently on other Armenian varieties too, including the dialects.

An evaluation will be done on the impact of the approach used, the size of the training data (for example, 1h (~5k), 2h (~10k), 5h (~25k) etc.) and the varieties included in the training data. The standard evaluation metrics for this task will be used: Word Error Rate (WER) and Character Error Rate (CER).

Task 3.4 Annotated downloadable oral datasets

Responsible/participant(s): SeDyL, ERTIM, PhD

Providing the linguistic community with a clean sound-aligned corpus for the acoustic study of spoken Armenian is the underlying task of this work package. All preceding tasks are designed to deliver these datasets: 50k tokens for MEA, and 10k tokens for MWA and 3x5k tokens for dialects. The availability of these datasets will open a wide range of applications for Armenian beyond automatic transcription (voice user interface, voice recognition etc.). On a broader perspective, the research and the resources on ASR worked out for the project aim at fostering oral Armenian data processing in general and giving a recording perennity and visibility to Armenian varieties (i.e. dialects) with no written tradition. The results of the project could be reused for processing oral data of other low-resourced languages with variation.

WP4 Language variation identification and distance measuring, 1-36 (month)

Responsible/participant(s):, SeDyL, LIPN, ERTIM, PostDoc, PhD

Work Package Tasks	Objectives / Deliverables
Task 4.1 Language variation identification and distance measuring through linguistic and statistical rules and forms	Linguistic and NLP methods and tools on language variation identification and distance measuring through statistical rules and forms, experimental survey
Task 4.2 Language variation identification and distance measuring through speech-to-text	Linguistic and NLP methods and tools on language variation identification and distance measuring through speech-to-text, experimental survey

Task 4.1 Language variation identification and distance measuring through linguistic and statistical rules and forms

Responsible/participant(s):, SeDyL, LIPN, PostDoc, PhD, Vladimir Plungian, Petr Kocharov

Variety identification and distance measuring have been approached by both Linguistics and NLP. Linguistic typology studies have produced several databases of structural and semantic features enabling to classify languages. Noteworthy is the World Atlas of Language Structure (WALS) (Dryer and Haspemath, 2013) describing 2,676 languages with 192 universal phonological, morphosyntactic and semantic features.

Western and Eastern Armenian are included with 43 and 129 features respectively. DALiH will focus on discriminative and operationalizable features by injecting them in the detection system. Complementary promising approaches have emerged due to the partial coverage of existing resources, either by inducing missing features from (possibly small) annotated corpora, by inducing features from already documented ones in the given language or in typologically closed ones or even in a totally unsupervised manner, for example from contextual embeddings. These approaches present the main advantage of better reflecting the gradient nature of languages and their intra-language variation (Ponti et al. 2019: 568-580).

In Armenian linguistics, distance between varieties is usually estimated according to various linguistic features (isoglosses). The issue of linguistic distance between varieties is especially relevant in genealogical classification of languages and in dialectology. Armenian dialectology school classifies varieties according to linguistic and extra-linguistic criteria (geographical, morphological, phonological), and through a multiparameter statistical approach (Martirosyan 2018). Applying morphological annotation of a particular variety upon typologically close varieties (diachronically or synchronically) and measuring the variational distance between two varieties according to formal parameters would allow revisiting existing classifications and putting forward new hypotheses. Our interdisciplinary research will start by comparing the value of different formal, measurable parameters (lexemes, unique wordforms) on the one hand, and their frequency mapping over different varieties on the other hand. The hypothesis is that the more the mapping overlaps, the closer the distance between the varieties is. Each parameter having a different value, linguistic research will aim to set a hierarchy between them, and to assess it through NLP resources.

Variety identification is essential for downstream tasks such as machine translation or information extraction. The first question of interest in this work package is therefore how to build a system that automatically identifies the variety of a segment of speech or text of any size (word, sentence, paragraph, etc.). Modeling this task on different granularity levels is important since Armenian varieties use the same script and share a significant part of the vocabulary, but differ nevertheless in phonology, morphology and syntax. This granular modeling allows us also to handle *code-switching* phenomena in corpora. This task will be addressed using two complementary approaches: 1) a rule-based one, mostly relying on lexical information such as the presence of certain words or affixes, and 2) a statistical one, based on labeled texts in each variety collected before and during the project.

Another task in this work package employs the tools developed for variety identification in modeling the typology of Armenian varieties. Our objective is to use this model to inform the development of more effective multivariational NLP tools and to provide new resources for the linguists.

Variety identification task will be also modelled as semi-supervised text classification. The input to this approach is a collection of texts labeled with variety information collected during and before the project, as well as a larger collection of unlabeled texts collected from Wikipedia and other sources. We are particularly interested in an approach based on variational autoencoders (VAE) (Xu et al., 2016), a generative probabilistic graphical model with deep neural networks. VAE allows to compute dense representations of the sentences with disentangled factors which can be used to distinguish varieties. This approach will be compared to character and word-based language models.

Task 4.2 Modeling and measuring distance between varieties using text and speech.

Responsible/participant(s): SeDyL, LIPN, ERTIM, PostDoc, PhD

Speech gives information about not only *what* is said, but also *how*. In WP3, acoustic embeddings for Armenian will be built with the help of tools like wav2vec 2.0. Using Armenian multivariational audio data provided with acoustic embeddings will open a new and unique perspective on variety identification and distance measuring.

The linguistic background will mainly consist in comparing the exhaustive phonetic realizations as well as the phonetic representations of the lexemes and wordforms and their frequency distribution among the target varieties. Similar to the previous task the general hypothesis would be to have more overlapping through different phonetic parameters between the varieties if they are close. Acoustic embeddings also capture a wide range of phonetic information including suprasegmental pronunciation (Shah et al., 2021), while being abstract enough to be robust against domain mismatch between the training and the test sets (Ma et al., 2021), just like we do as humans. It will be interesting to see whether this balance between fine-

grained characteristics and abstraction does discriminate between the different varieties in terms of distance of acoustic word vectors.

WP5: Unified platform for Armenian multivariational corpus, 1-42 (month)

Responsible/participant(s): SeDyL, LIPN, ERTIM, Digilib, RAS

Work Package Tasks	Objectives / Deliverables
Task 5.1 Annotation platform	DALiH annotation platform
Task 5.2 Search engine setup for Armenian multivariational corpora	DALiH search engine and multivariational corpora
Task 5.3 Website development and content	DALiH website

Work Package 5 describes software needed to perform several tasks during the project. In particular, the project will need an annotation platform (task 1.1.) a storage and search/exploration engine for annotated corpora (task 1.2.). It also briefly describes the project website (task 1.3.).

Task 5.1. - Annotation Software

The active learning framework to iteratively improve the quality of linguistic annotations through collaboration between automatic processes and linguistic expertise (1-4 WPs) implies the use of an annotation platform for linguists to validate/correct the automatically tagged corpora. The software needs to tackle both speech and written textual data, enable collaborative labelling, statistical evaluation (inter-annotator agreement) and active learning, adapt the annotation scheme to Armenian language specificities and transparent export and import functions from and to NLP models. A few studies have already benchmarked several tools (among others Neves and Seva, 2021), pointing *WebAnno* as the outstanding available tool. *Inception* (Klie et al., 2018), a web-based tool, is a recent fork of this project, adding several functionalities. It will be used for written textual content. A specific instance of the software will be installed, tuned and maintained during the project.

Similarly, the semi-automatic alignment and transcription of speech data in WP2 would benefit from a dedicated platform. While a comprehensive yet simple tool like *Inception* does not exist yet for speech data, the speech community is slowly bridging the gap. Seshat (Titeux et al., 2020) and Audino (Grover et al., (under review)) are two recently built tools that are easy to use and enable full management of annotation campaigns on oral data (task assignment to pools of annotators, inter-annotator agreement measuring). Seshat has the advantage of implementing Praat, a well-known tool among the linguistic community, but Audino proposes a modern and ergonomic interface. As manual alignment and transcription are tedious tasks, an appropriate tool will be chosen relative to the annotators' profile.

Task 5.2. Software for Corpora Storage and Enriched Search and Exploration

The DALiH project will generate various reference corpora and enable further research on the Armenian varieties. Additionally, to delivering versions of the corpora in normalized formats, the DALiH project will use a state-of-the-art web platform to efficiently store the data and enable enriched search and exploration. A significant list of corpus tools is available in this respect, and we will perform a detailed comparison of available tools that can fulfill our needs: free, open-sourced and mature, ease of use and installation, web-based, extensible storage features (especially concerning the adapted annotation scheme), search and exploration functionalities (concordance search, statistics, collocation retrieval, enriched query language, etc.). A first investigation enabled us to identify three efficient tools: *NoSketchEngine* (Kilgariff et al., 2014), *CWB* (Evert and Hardie, 2011) and *Tsakorpus* (Arkhangelskiy, 2020). The first and second ones are largely used, supported and maintained by the corpus linguistics community, with a range of features: support for word-based linguistic annotations, sentence-, paragraph- and document-level metadata, enriched and versatile language query (Corpus Query language, ref) and statistical parsing (word list, collocations). The other one was notably developed for the EANC project and provides similar features, despite its narrower diffusion. It has recently been used for similar projects of managing varieties of poorly endowed languages, from different sources (written, spoken and dialectal corpora). It includes in particular sound management, and because of its more recent development, it uses state-of-the-art and up-to-date dependencies for the management of massive corpora (python 3.x, elasticsearch, etc.). It has been successfully used for about 30 rare languages.

These tools will be further investigated and tested from the start of the project to decide on the better choice and quickly trigger installation on the project web server (see *infra*).

Task 5.3. Website

DALiH with its multivariational platform regrouping all the target variety corpora will be available at a dedicated user-friendly website hosted by Huma-Num. The corpora processed within the framework of the project will be freely searchable for all the users via an online complex query interface. A set of annotated written and oral datasets as well as all other DALiH resources (wordlists, grammatical dictionaries, exhaustive language modelling, annotation LMs, ASR models etc.) will be freely available and downloadable. The website will be provided with 'content' and 'help' texts to present the project and its functionalities in details.

The DALiH project objectives and the methodologies applied cover linguistics on the one hand, targeting in particular NLP resources and corpus linguistics (morphological formalization, rule-based grammatical annotation), variational linguistics (dialectology, historical linguistics, typology of closely related languages, sociolinguistics), and Artificial Intelligence, in particular Deep Learning applied to NLP, on the other hand. The project being at the crossroads of digital humanities and linguistics, the choice of the scientific evaluation committee 8.6/38 is conditioned by its interdisciplinary character.

II. Organisation and implementation of the project

Implication of the scientific coordinator and partner's scientific leader in on-going project(s)

Name of the researcher	Person.month	Call, funding agency, grant allocated	Project's title	Name of the scientific coordinator	Start - end
Victoria KHURSHUDYAN	6,3 (15%)	ANR	Labex EFL	Barbara Hemforth	2010-2024
Nadi TOMEH	4,2 (10%)	ANR	Labex EFL	Barbara Hemforth	2010-2024
Nadi TOMEH	10,5 (25%)	ANR	Pro-TEXT	Georgeta Cislaru	2019-2023
Damien NOUVEL	12,6 (30%)	DGA RAPID	VITAL	Vincent Nibart	2021-2023
Damien NOUVEL	14,7 (35%)	ANR	TALAD	Julien Longhi	2018-2022

The project coordinator, Victoria Khurshudyan (INALCO-SeDyL), is a linguist specialized in Armenian linguistics and in NLP development resources for Armenian data processing. Starting from her PhD thesis, the coordinator has compiled and processed linguistic corpora by regular participation in various language documentation fieldwork projects providing target language corpora, and her research has been carried out mostly from a corpus-based perspective. Victoria Khurshudyan has coordinated and organized a number of significant projects (EANC, EANC dialectal corpus), conferences, workshops in corpus linguistics (Workshop on Armenian Corpus Linguistics 2007; *Digital Armenian* International conference 2019) in collaboration with most of the members of the consortium of the current project. Within the framework of DALiH, the coordinator will be in charge of general administrative and technical coordination of the project, linguistic/corpus expertise in language modelling, grammatical dictionary compilation, dialect documentation etc. (WP0-WP6).

The consortium will be composed of the core team and two partner NLP research centers, as well as a network of partners from other organizations and countries with strong expertise in the project's domains. The consortium will, thus, include researchers in 1) general, corpus and Armenian linguistics (SeDyL), 2) NLP and digital humanities (LIPN, ERTIM) and 3) international researchers and experts in philology, linguistics and Armenian studies.

The core team will include Victoria Khurshudyan, MCF INALCO-SeDyL (coordinator), Anaïd Donabédian, PU INALCO-SeDyL, a PhD student to be selected through an international call and a Master student intern (WP1-3). Anaïd Donabédian will be in charge of linguistic tasks and expertise on MWA and dialectal corpora, more particularly morphology modelling and grammatical dictionary compilation, dialect transcription and annotation modelling (WP1-WP6). Together with ERTIM Anaïd Donabédian will supervise the PhD thesis which will target Armenian linguistic issues in NLP perspective according to the candidates' profile (WP1-WP4).

The two other partners of the consortium are LIPN (University Sorbonne Paris Nord) and ERTIM (INALCO) who are also involved in SeDyL/INALCO's institutional network. LIPN is represented by three

members, its scientific leader Nadi Tomeh, MCF, LIPN, as well as Emmanuel Cartier, MCF LIPN and Thierry Charnois, PU LIPN (WP1-2, WP4-6). ERTIM's team members are the scientific leader Damien Nouvel, MCF ERTIM and Ilaine Wang, research engineer ERTIM (WP3-6). ERTIM team will be completed by a research engineer to be recruited within the framework of the project and to be in charge of speech-to-text alignment. LIPN team will include a postdoctoral researcher to be hired within the framework of the project to work on various deep-learning annotation models (WP2-WP4). ERTIM's scientific leader Damien Nouvel will co-supervise the PhD student. Both partners are specialized in the NLP domain and they will focus on NLP tasks related mainly to grammatical annotation for LIPN (WP2) and automatic speech recognition for ERTIM (WP3) as well as collaborative coordination on processing word embedding models for Armenian (WP2). Together with the SeDyL team the two partners will also take part in research tasks on language identification and variety distance measuring (WP4).

Significant contribution will be provided by partners from the Russian Academy of Sciences Vladimir Plungian (Russian Academy of Sciences) (WP1-WP6) and Petr Kocharov (Russian Academy of Sciences) (WP1-2, WP4-6), as well as Chahan Vidal-Gorène (postdoctoral researcher candidate) (WP2-WP4), currently a PhD student in Paleography (École Nationale des Chartes-PSL), Timofey Arkhangelsky (Universität Hamburg) (WP2), the *Digital Library of Armenian Literature* project (American University of Armenia) (WP1, 5, 6) and the start-up Calfa specialized in ancient language digitalization (WP1).

Vladimir Plungian, the co-founder of Russian and Eastern Armenian National Corpora, will be in charge of the expertise on language modelling for RB approach and aligned corpora. Petr Kocharov will take part in Classical Armenian modelling and grammatical dictionary processing. Chahan Vidal-Gorène will focus on language model elaboration in different deep-learning approaches, whereas Timofey Arkhangelsky will provide expertise concerning the rule-based approach. Digilib and Calfa team will be urged to take part in plain-text compilation and metatext annotation tasks.

III. Impact and benefits of the project

WP6: DALiH Dissemination, 1-42 (month)

Responsible/participant(s): SeDyL, LIPN, ERTIM, Digilib, RAS + all potential collaborators

A special work package (WP6) will be dedicated to the activities and events in order to largely disseminate the research results, resources and tools of the project among the scientific, educational as well as various socio-economic spheres. The promoting and disseminating initiatives will be mainly carried out through specialized and presentational workshops, shared tasks, conferences attended and organized by the DALiH team, publications, project presentations within the framework of various university courses and seminars.

Task 6.1 DALiH workshops

Five workshops are planned to be organized over the project's timespan: with three annual professional NLP/linguistics workshops, one public workshop at the Library of BULAC, Paris, and one specialized workshop for the teenagers at the TUMO center for creative technologies in Yerevan, Armenia.

The first three annual workshops will cover NLP and linguistic issues on Armenian and all other low-resourced languages. INALCO being one of the rare Universities covering such significant number of rare languages (about a hundred of languages), the NLP and linguistic solutions for bridging the scarcity of the data and resources are of the utmost importance for preserving certain language varieties as well as for fostering research and practical developments for the target variety. Two workshops are planned to be held in Paris, and one in Yerevan to galvanize the local Armenian NLP community which is very active.

The public workshop at the Library of BULAC will be held by the end of the project and it will introduce the project to a large audience with the aim to present the results but specially to raise awareness of the importance of the accessibility of language resources, to propose solutions and methods to process language data be it oral or written.

The workshop at the TUMO center for creative technologies in Yerevan, Armenia will be intended for the teenagers aiming at particularly reusing the resources produced within the framework of DALiH to further new resources for the Armenian language. Currently, a MOOC project on Armenian alphabet

learning is in progress by the Armenian section of INALCO with Victoria Khurshudyan, Anaïd Donabédian and Chahan Vidal-Gorène and the TUMO center.

Task 6.2 DALiH shared tasks

Two international shared tasks will be organized at the end of the second and third year of the project in parallel with DALiH workshops. The first will be dedicated to compare annotation state-of-the-art models for under-resourced languages and will focus on some Armenian varieties. The second one will evaluate the ability of transfer learning for varieties of an under-resourced language. Dedicated datasets will be released to the NLP community involved in the processing of under-resourced languages, and results and awards will be released during the related workshops, and will be promoted through dedicated papers.

Task 6.3 DALiH conference

A final international conference is planned to be held at INALCO, Paris on NLP and linguistic research issues and resources for Armenian and other low-resourced languages at the end of the project. 10 invited speakers and 50 participants will be expected to present their research and results, the conference will be open to large public and the proceedings will be published.

Task 6.4 DALiH website public launching

The final result of the project will be a website which will host all the Armenian multivariational corpora, with metadata, full grammatical annotation, as well as multivariational sound-aligned and text aligned corpora with a searching platform for various types of complex search queries. All the resources processed and worked out within the framework of the project will be available in open-access. The website will be accompanied by content texts describing the existing functionalities and resources. It will be hosted at Huma-Num (TGIR), a research infrastructure for digital humanities. The launch of the website will be held at INALCO, by the end of the project.

Task 6.5 DALiH publications

DALiH team researchers will permanently present and publish the results of the project through participation at different conferences (ACL, COLING, EACL, EMNLP, ICLR, INTERSPEECH, LREC etc.) or/and by publishing in various scientific journals and books. Certain papers will be submitted in non-scientific press to popularize the target issues and results of the project.

DALiH is a pioneer in regrouping the entire Armenian language heritage, reusing all the possible existing resources, creating new ones and making them available for addressing research issues in various scientific domains (linguistics, AI, NLP, anthropology, sociology, history, philology, etc.). In addition to its scientific significance, the project delivers larger socio-economic outputs such as tools and capacity building for training Armenian language teachers, translators, editors etc. The outputs can be reusable for language maintenance and educational objectives (reference NLP tools such as dictionaries, grammatical analyzers, spell-checkers, processed texts for language teaching). A series of presentations inspired by DALiH project will be proposed in different academic courses, seminars and workshops (e.g. Digital tools and Armenian studies (Master1, INALCO), Corpus Linguistics: approaches and applications (Master2, INALCO), International School in Linguistic Fieldwork (FieldLing), CNRS etc.)

The documentation and processing of Modern Western Armenian, an endangered heritage language, and of the target dialects, the speakers of which were forced to migrate from their original lands (Artsakh and Nakhijevan), and making them available for the scientific community and for the general public, is of foremost importance not only for NLP resources but also and especially from linguistic, anthropological and social perspectives.

The renewal of NLP resources by means of various artificial intelligence resources and the processing of linguistic variation presents a significant challenge both in linguistics and in NLP, especially for under-resourced languages. Besides providing various architectures, the research experiments and models processed within the framework of DALiH will serve as a benchmark for the annotation of other under-resourced languages, especially those with non-Latin characters.

All the final products of the project will be open-source and available on an open-access platform hosted by CNRS's Digital Humanities infrastructure Huma-Num. In addition, the project results will be

accessible to scientific and socio-economic communities through publications, workshops and conferences, as well as training courses and research competitions.

Languages have usually fragile vitality when lacking natural regenerating native speakers' or standard status. Therefore, the documentation of the Armenian varieties as well as the processing of the documented data is of the utmost importance not only pertaining to scientific NLP, linguistic, anthropological and social perspectives but also for preserving and revitalizing them as a part of the world heritage.

IV. References related to the project

1. Adams, O., Galliot, B., Wisniewski, G., Lambourne, N., Foley, B., Sanders-Dwyer, R., Wiles, J., Michaud, A., Guillaume, S., Besacier, L., Cox, C., Aplonova, K., Jacques, G., & Hill, N. (2021). User-friendly automatic transcription of low-resource languages: Plugging ESPnet into Elpis. *arXiv:2101.03027 [cs, eess]*.
2. Anguera, X., Luque, J., & Gracia, C. (2014). Audio-to-Text Alignment for Speech Recognition with Very Limited Resources. *Fifteenth Annual Conference of the International Speech Communication Association*.
3. Arkhangelskiy, T. (2020). Web Corpora of Volga-Kama Uralic Languages. *Finno-Ugric Languages and Linguistics*, *9*(1-2), Article 1-2.
4. Arkhangelskiy, T., & Georgieva, E. (2018). Sound-aligned corpus of Udmurt dialectal texts. *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, 26-38.
5. Avetisyan, K., & Ghukasyan, T. (2019). Word Embeddings for the Armenian Language: Intrinsic and Extrinsic Evaluation. *arXiv:1906.03134 [cs]*.
6. Awetik'ean, G., Siwrmēlean, X., & Awgerean, M. (1836-1837). *Նոր Բառգիրք Հայկազեան Լեզուի (New Dictionary of the Armenian Language)*. Tparan i Srboyn Łazaru.
7. Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv:2006.11477 [cs, eess]*.
8. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623.
9. Čērēčean, G. ark'episkopos, Tōnikean, P. K., & Tēr Xač'aturean, A. (1992). *Հայոց լեզուի նոր բառարան (= New Dictionary of the Armenian Language)* (K. Tōnikean ew ortik' hratarakč'atun).
10. Collobert, R., Puhrsch, C., & Synnaeve, G. (2016). Wav2Letter: An End-to-End ConvNet-based Speech Recognition System. *arXiv:1609.03193 [cs]*.
11. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*.
12. Donabédian-Demopoulos, A. (2018). Middle East and Beyond - Western Armenian at the crossroads: A sociolinguistic and typological sketch. In C. Bulut (Éd.), *A sociolinguistic and typological sketch, in Bulut, Christiane, Linguistic minorities in Turkey and Turkic-speaking minorities of the periphery, 111/2018, Harrazowitz Verlag* (Vol. 111, p. 89-148). Harrazowitz Verlag.
13. Donabédian, A., & Sitaridou, I. (2021). Anatolia. In *The Routledge Handbook of Language Contact* (Adamou, E. (Ed.), Matras, Y. (Ed.), p. 404-433). Routledge.
14. Dryer, M.-S. and Haspelmath M., editors. (2013). WALS Online. Max Planck Institute for Evolutionary Anthropology, Leipzig.
15. Evert, S., & Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference*. Corpus Linguistics 2011, GBR. University of Birmingham.
16. Foley, B., Arnold, J., Coto-Solano, R., Durantin, G., Mark, E., van Esch, D., Heath, S., Kratochvíl, F., Maxwell-Smith, Z., Nash, D., Olsson, O., Richards, M., San, N., Stoakes, H., Thieberger, N., & Wiles, J. (2018). Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS). *The 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages*.
17. Foley, Ben, Rakhi, A., Lambourne, N., Buckeridge, N., & Wiles, J. (2019). *Elpis, an Accessible Speech-to-Text Tool*.
18. Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning Word Vectors for 157 Languages. *arXiv:1802.06893 [cs]*.
19. Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., & Ng, A. Y. (2014). Deep Speech: Scaling up end-to-end speech recognition. *arXiv:1412.5567 [cs]*.
20. Hjortnaes, N., Partanen, N., Rießler, M., & M. Tyers, F. (2020). Towards a Speech Recognizer for Komi, an Endangered and Low-Resource Uralic Language. *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*, 31-37.
21. Khurshudyan, V., Arkhangelskiy, T., Daniel, M., Levonian, D., Plungian, V., Polyakov, A., Rubakov, S. (2021). Introduction to Eastern Armenian National Corpus: www.eanc.net. *Études arméniennes contemporaines*, in press.
22. Khurshudyan, V., Daniel, M., Levonian, D., Plungian, V., Polyakov, A., & Rubakov, S. (2009). Восточноармянский национальный корпус (= Eastern Armenian National Corpus). *Proceedings of international conference "Dialog'2009"*, 509-518.
23. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*(1), 7-36.
24. Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R., & Gurevych, I. (2018). The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 5-9.
25. Köhn, A., Stegen, F., & Baumann, T. (2016). Mining the Spoken Wikipedia for Speech Data and Beyond. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 4644-4647.

26. Liptchinsky, V., Synnaeve, G., & Collobert, R. (2017). *Letter-based speech recognition with gated convnets*.
27. Ma, D., Ryant, N., & Liberman, M. (2021). Probing Acoustic Representations for Phonetic Properties.
28. Ma, W., Cui, Y., Si, C., Liu, T., Wang, S., & Hu, G. (2020). CharBERT: Character-aware Pre-trained Language Model. *arXiv:2011.01513 [cs]*.
29. Malxaseanc', S. (1944). *Հայերէն բացատրական բառարան (= Explanatory Dictionary of Armenian)*. Haykakan SSŔ Petakan Hratarakč'ut'iwn.
30. Manjavacas, E., Kádár, Á., & Kestemont, M. (2019). Improving Lemmatization of Non-Standard Languages with Joint Learning. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *1*, 1493-1503.
31. Martirosyan, H. (2018). *The Armenian dialects. In: The languages and linguistics of Western Asia: An areal perspective* (Hans Henrich Hock, Vol. 6, p. 46-105). De Gruyter Mouton.
32. Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746-751.
33. Mohamud, J. H., Thompson, L. A., Ndoye, A., & Besacier, L. (2021). Fast Development of ASR in African Languages using Self Supervised Speech Representation Learning. *arXiv:2103.08993 [cs, eess]*, *AfricaNLP2021 workshop at EACL 2021*.
34. Neves, M., & Ševa, J. (2021). An extensive review of tools for manual annotation of documents. *Briefings in Bioinformatics*, *22*(1), 146-163.
35. Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is Multilingual BERT? *arXiv:1906.01502 [cs]*.
36. Ponti, E., O'Horan, H., Berzak, Y., Vulić, I., Reichart, R., Poibeau, T., Shutova, E., & Korhonen, A. (2019). Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing. Computational Linguistics, 45-3, 1-43.
37. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. *Infoscience*, Article CONF. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.
38. Pratap, V., Hannun, A., Xu, Q., Cai, J., Kahn, J., Synnaeve, G., Liptchinsky, V., & Collobert, R. (2019). Wav2Letter++: A Fast Open-source Speech Recognition System. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6460-6464.
39. Rivière, M., Joulin, A., Mazaré, P.-E., & Dupoux, E. (2020). Unsupervised Pretraining Transfers Well Across Languages. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7414-7418.
40. Shah, J., Singla, Y. K., Chen, C., & Shah, R. R. (2021). What all do audio transformer models hear? Probing Acoustic Representations for Language Delivery and its Structure.
41. Vidal-Gorène, C., & Kindt, B. (2020). Lemmatization and POS-tagging process by using joint learning approach. Experimental results on Classical Armenian, Old Georgian, and Syriac. *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, 22-27.
42. Vidal-Gorène, C., Khurshudyan, V., & Donabédian-Demopoulos, A. (2020). Recycling and Comparing Morphological Annotation Models for Armenian Diachronic-Variational Corpus Processing. *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, 90-101.
43. Vidal-Gorène, C., Khurshudyan, V., & Donabédian, A. (2020). Modèles d'annotations morphologiques pour le traitement de données multivariées de l'arménien. In T. Poibeau, Y. Parmentier, & E. Schang (Éds.), *2èmes journées scientifiques du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT)* (p. 72-82). CNRS.
44. Waldenfels von R., Daniel M., & Dobrushina N. 2014. Why standard orthography? Building the Ustya River Basin corpus, an online corpus of a Russian dialect. In Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue". 13, pages 720–728.
45. Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N. E. Y., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A., & Ochiai, T. (2018). ESPnet: End-to-End Speech Processing Toolkit. *arXiv:1804.00015 [cs]*.
46. Xu, W., Sun, H., Deng, C., & Tan, Y. (2017). Variational Autoencoder for Semi-Supervised Text Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1), Article 1.
47. Yavrumyan, M., Danielyan A. (2020). Համընդհանուր կախվածություններ և հայերենի ծառադարանը (= Universal Dependencies and Armenian Tree-Bank), Lraber hasarakakan gitut'yunneri, 231-244.