Proceedings of 1st Workshop on Language Technologies for Historical and Ancient Languages, pages 22–27
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Lemmatization and POS-tagging process by using joint learning approach. Experimental results on Classical Armenian, Old Georgian, and Syriac

Chahan Vidal-Gorène, Bastien Kindt
École Nationale des Chartes-PSL, Université catholique de Louvain
65, rue de Richelieu (F-75002 Paris), Institut orientaliste - Place Blaise Pascal, 1 (B-1348 Louvain-la-Neuve)
chahan.vidal-gorene@chartes.psl.eu, bastien.kindt@uclouvain.be

Abstract

Classical Armenian, Old Georgian and Syriac are under-resourced digital languages. Even though a lot of printed critical editions or dictionaries are available, there is currently a lack of fully tagged corpora that could be reused for automatic text analysis. In this paper, we introduce an ongoing project of lemmatization and POS-tagging for these languages, relying on a recurrent neural network (RNN), specific morphological tags and dedicated datasets. For this paper, we have combine different corpora previously processed by automatic out-of-context lemmatization and POS-tagging, and manual proofreading by the collaborators of the GREgORI Project (UCLouvain, Louvain-la-Neuve, Belgium). We intend to compare a rule based approach and a RNN approach by using PIE specialized by Calfa (Paris, France). We introduce here first results. We reach a mean accuracy of 91,63% in lemmatization and of 92,56% in POS-tagging. The datasets, which were constituted and used for this project, are not yet representative of the different variations of these languages through centuries, but they are homogenous and allow reaching tangible results, paving the way for further analysis of wider corpora.

Keywords: POS-tagging, Lemmatization, Morphological Analysis, Classical Armenian, Old Georgian, Syriac

1. Introduction

Classical Armenian, Old Georgian and Syriac are still poorly digitally resourced. Some major corpora already exist, for instance the *Digital Syriac Corpus* (DSC) for Syriac; *Digilib*, *Arak29*, *Calfa* and *Titus* for Classical Armenian; and *Titus* and the *Georgian Language Corpus* for Georgian[1]. These corpora, when they are really specialized on the ancient state of these languages, are mainly composed of plain texts or texts analyzed out of context (all possible analyses are given for each token and polylexical[2] word-forms are not fully described). Accordingly, scholars are still waiting for corpora enhanced with complete and reliable linguistic tags. Concerning the modern state of these languages, the Universal Dependencies (UD) provide annotated corpora for Armenian and Georgian, with the same limitations as described above. Furthermore, the modern and the ancient states of each language are usually quite different, so that digital resources built for either are inadequate to process the other.

Usual techniques for the lemmatization of these corpora rely on sets of rules and dictionaries. Such a method is unable to handle unknown tokens, or to readily process data in context. We have initiated experimentations to complete these operations using a neural network (RNN) and purpose-built corpora dedicated to this very task (Dereza, 2018). The task is particularly complex for these aforenamed languages due to their wealth of polylexical forms. In this paper, we present experimental results achieved through the application of state-of-the-art technologies to these languages. This research depends on the data and tools developed by both the GRE*g*ORI (henceforth GP)[3] and Calfa[4] projects. The texts all derive from the database of the GP, which consists of texts written in the main languages of the Christian East and already published in the form of critical editions.

The scope of this paper is limited to the three already quoted languages. The datasets described below have all previously undergone automatic out-of-context lemmatization, and manual proofreading (see *infra* 3. Data Structure).

2. Datasets

D-HYE: Classical Armenian is an Indo-European language. This dataset contains 66.812 tokens (16.417 of which are unique) originating from three different corpora: Gregory of Nazianzus (Coulie, 1994; Coulie and Sirinian, 1999; Sanspeur, 2007; Sirinian, 1999) (GRNA), the *Geography of the Indian World* (Boisson, 2014) (GMI), and the *Acta Pauli et Theclae* (Calzolari, 2017) (THECLA). GRNA gathers the text of the Armenian versions of Gregory of Nazianzus' *Discourses*, already published in the *Corpus Christianorum* series. Gregory of Nazianzus (†390 AD) is the author of 45 *Discourses*, more than 240 letters, as well as theological and historical works in verse.

The Armenian version is anonymous and dates from 500-550 AD; its style has been qualified as pre-Hellenophile

[1] We only quote here some freely available data.

[2] The word "polylexical" is used here as a very generic term (but relevant for the three mentioned languages), referring to word-forms combining more than one lexeme in a single graphical unit (e.g. agglutinated forms).

[3] The GP develops digital tools and resources aimed at producing tagged corpora, at first in Ancient Greek, but now also in the main languages of the Christian East. Tagged textual data are processed in order to publish lemmatized concordances and different kinds of indexes. These tools are based on a stable standard of lexical examination (Kindt, 2018).

[4] The Calfa project develops a complete database for Classical Armenian, as well as tools for corpora creation and annotation (crowdsourcing interface and OCR technology for historical languages) (Vidal-Gorène and Decours-Perez, 2020).

(Lafontaine and Coulie, 1983). THECLA contains the Armenian version of a group of texts relating to the legend of Thecla and the martyrdom of Paul (5^{th}-14^{th} c. AD), while GMI is a very small text written around 1120 AD, enumerating cities and trading posts of the Indian world. GMI contains a lot of unique tokens, such as toponyms and personal names. **D-HYE** primarily covers texts of the Hellenophile tradition, which entails a large number of neologisms and idiosyncratic syntactic constructions. As such, for the time being, it is not entirely representative of the Classical Armenian language (see *infra* 5. Perspectives).

D-KAT: Old Georgian is a Kartvelian language. It contains 150.869 tokens (30.313 unique) from one unique corpus, made up of the texts of the Georgian versions of Gregory of Nazianzus' *Discourses* already published in the *Corpus Christianorum* series (Coulie and Métrévéli, 2001; Coulie and Métrévéli, 2004; Coulie and Métrévéli, 2007; Coulie and Métrévéli, 2011; Métrévéli, 1998; Métrévéli, 2000). Several translations from Greek into Georgian are known. The most important of which are those by Euthymius the Hagiorite (10^{th} c. AD) and Ephrem Mtsire (Black Mountain, near Antioch, 11^{th} c. AD) (Haelewyck, 2017b).

D-SYC: Syriac is a Semitic language. This dataset contains 46.859 tokens (10.612 unique). It is the most heterogenous dataset of this study, since the texts it contains relate to a variety of topics: biblical, hagiographic, and historical texts, homilies, hymns, moral sayings, translations of Greek philosophical works, etc. These texts have been lemmatized by the collaborators of the GP: the Syriac version of *Discourses* I and XIII by Gregory of Nazianzus, translated from Greek in the 6^{th}-7^{th} c. AD (Haelewyck, 2011; Haelewyck, 2017b; Schmidt, 2002; Sembiante, 2017); the *Story of Zosimus*, translated no later than the 4^{th} c. AD (Haelewyck, 2014; Haelewyck, 2015; Haelewyck, 2016; Haelewyck, 2017a); the *Syriac Sayings of Greek Philosophers* (6^{th}-9^{th} c. AD) (Arzhanov, 2018); the *Life of John the Merciful* (Venturini, 2019); and some other texts dating from the 4^{th} to the 9^{th} century, described on the GP's website.

Type	D-HYE	D-KAT	D-SYC
different tokens	66.812	150.869	46.859
unique tokens	16.417	30.313	10.612
unique lemmata	5.263	8.897	2.957

Table 1: Composition of the datasets

These datasets do not embrace the whole lexicon of these languages (as a reference, the Calfa dictionary contains around 65.000 entries for Classical Armenian). We discuss this shortcoming in parts 3. and 4.

3. Data Structure

The data have been prepared and analysed in the framework of the GP. For each corpus, the following processing steps were implemented:

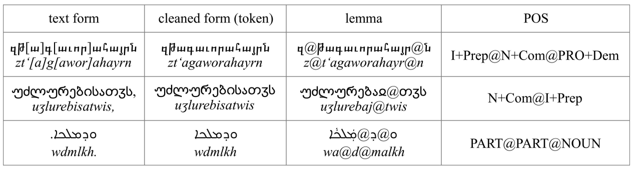
1. Cleaning up the forms of the text (removal of uppercase, critical signs used by editors, etc.). These forms constitute the column "cleaned form" of the corpus (see figure 1);

2. Morpho-lexical tagging, i.e. identifying a lemma and a POS for every cleaned-up form (token) of the text. This task is conducted through automatic comparison of the clean forms of the texts to the linguistic resources of the GP: dictionaries of simple forms and rules for the analysis of polylexical forms (see *infra*);

3. Proofreading of the results, corrections and encoding of missing analyses;

4. Enrichment of the linguistic resources for future processing of other texts.

Syriac, Classical Armenian and Old Georgian contain a large quantity of polylexical forms, combining words with different prefixes (preposition or binding particle) and/or suffixes (postposition or determiner). These forms are systematically (and automatically) segmented in order to identify explicitly each of its components. The different lexical elements are separated by an @ sign and divided into the following columns: lemma, POS and morph (see table 4; displaying a short sentence from the *Inscription of the Regent Constantine of Papeṙōn* (Ouzounian et al., 2012)). The morpho-lexical tagging follows the rules laid out for each language by the collaborators of the GP (Coulie, 1996; Coulie et al., 2013; Coulie et al., 2020; Kindt, 2004; Haelewyck et al., 2018; Van Elverdinghe, 2018). This automated analysis does not take the context into account. The resulting data are proofread manually and the proofreaders add the morphology according to the context (see table 4, columns marked GP).



text form	cleaned form (token)	lemma	POS
զթ[ա]գ[աւոր]ահայրն *zt'[a]g[awor]ahayrn*	զթագաւորահայրն *zt'agaworahayrn*	զ@թագաւորահայր@ն *z@t'agaworahayr@n*	I+Prep@N+Com@PRO+Dem
უძლურებისათჳს, *uʒlurebisatwis,*	უძლურებისათჳს *uʒlurebisatwis*	უძლურებაჲ@თჳს *uʒlurebaj@twis*	N+Com@I+Prep
ܘܕܡܠܟܗ. *wdmlkh.*	ܘܕܡܠܟܗ *wdmlkh*	ܘ@ܕ@ܡܠܟܗ *wa@d@malkh*	PART@PART@NOUN

Figure 1: Raw output from the GP system

4. Method and Experiments

Up until now, the annotation has depended on a set of rules and dictionaries, and the result has been manually corrected. The main flaw of this approach lies in the fact that this analysis only concerns the forms attested in the corpus and already included in the lexical resources (< 40% for a representative corpus of Classical Armenian like the NBHL (Vidal-Gorène et al., 2019)) on the one hand, and that it does not provide answers in case of lexical ambiguity on the other hand. We have, hence, initiated experimentations to complete the task of lemmatization and POS-tagging with a neural network.

At present, the choice has fallen on PIE (Manjavacas et al., 2019), which offers a highly modular architecture (using

Train	All token			Ambiguous token			Unknown token		
	D-ARM	D-KAT	D-SYC	D-ARM	D-KAT	D-SYC	D-ARM	D-KAT	D-SYC
accuracy	0.9307	0.9698	0.8877	0.9318	0.9354	0.8307	0.7210	0.8460	0.5914
precision	0.7067	0.8187	0.6475	0.5997	0.7104	0.5382	0.5350	0.7177	0.4131
recall	0.7076	0.8132	0.6503	0.6566	0.7367	0.5982	0.5361	0.7101	0.4094
f1-score	0.7071	0.8159	0.6489	0.6269	0.7233	0.5666	0.5355	0.7139	0.4117

Test	All token			Ambiguous token			Unknown token		
	D-ARM	D-KAT	D-SYC	D-ARM	D-KAT	D-SYC	D-ARM	D-KAT	D-SYC
accuracy	0.9044	0.9628	0.8817	0.8620	0.8235	0.8460	0.6864	0.8220	0.6274
precision	0.6630	0.784	0.6211	0.4411	0.4261	0.6106	0.5074	0.6775	0.4112
recall	0.6711	0.7761	0.6215	0.5211	0.4928	0.6591	0.5118	0.6702	0.4072
f1-score	0.6670	0.7800	0.6213	0.4778	0.4570	0.6339	0.5096	0.6738	0.4092

Table 2: 1. Best scores for the training step of the lemmatizer on **D-HYE**, **D-KAT** and **D-SYC**; 2. Evaluation of the lemmatizer on the **D-HYE**, **D-KAT** and **D-SYC** Test datasets

bidirectional RNN). PIE enables, in particular, to process ambiguous or unknown forms by integrating contextual information, and to increase accuracy of the lemmatizer and the POS-tagger (Egen et al., 2016). Even though PIE allows simultaneous annotation of lemmata and POS, we have decided here to conduct the tasks independently. We use the default hyper parameters proposed by Manjavacas and applied on twenty different corpora from UD, without tailoring them in any way to the dataset under consideration[5]. For the lemmatization task, we have followed the default structure provided by PIE. We are working at the char level, and we include the sentence context. We use an attention encoder-decoder.

For the POS-tagging task, we have compared the Conditional Random Field (CRF) provided by LEMMING (Müller et al., 2015) and the linear decoder implemented in PIE.

We have divided **D-HYE**, **D-KAT** and **D-SYC** into three sets: Train (80% of data), Validation (10%) and Test (10%). The distribution was implemented automatically on a sentence basis.

Results on lemmatization

The results achieved are consistent with the representativeness and the size of the corpora studied, and the results provided by Manjavacas on similar datasets (see *infra* 5. Perspectives). **D-HYE** is the most homogenous dataset, despite the numerous unique toponyms. Thus, there is little variation regarding vocabulary and expressions, which is why we achieve a very good accuracy during training, almost as good as with **D-KAT**, but for a corpus twice as small. By contrast, **D-SYC** is more representative of all the language state of Syriac.

The results on ambiguous and unknown tokens are quite low, however they make it possible to already process automatically a larger number of cases.

The train set for Armenian contains 17% of unknown tokens, due to the high proportion of proper nouns from GMI, whereas the proportion of unknown tokens is 14% in Georgian and 20% in Syriac, the latter being penalized twice, by its size and this proportion of unknown tokens. The confusion matrix reveals that mistakes are concentrated on homographic lemmata (e.g. *mayr* (mother) and *mayr* (cedrus)).

Besides, these languages exhibit numerous polylexical forms: these are similar in form but they differ in their analysis. We had identified the homographs beforehand, in order to disambiguate them (e.g. իւր (իւրոց) and իւր (իւրեանց)), but the lack of data results in a more complex task for the network. Besides, 50% of mistakes are localized on polylexical forms, such as demonstrative pronouns or prepositions. This is made clear in table 4, where no pronoun has been predicted. The same applies for the task of POS-tagging.

Results on POS-tagging (crf / linear)

The Linear Decoder achieves better results for the task of POS-tagging, except for the task of tagging ambiguous and unknown tokens during training. Nevertheless, the linear decoder remains better than the CRF decoder (LEMMING) on the test datasets, except for unknow tokens in Old Georgian and Syriac. The issue of the ambiguous tokens is the same as for the task of lemmatization. The confusion matrix for **D-HYE** shows that mistakes are essentially concentrated on common nouns (21%, generally predicted as verbs) and verbs (12%, generally predicted as common nouns). Vocalic alternation in Classical Armenian appears to create ambiguities between declined and conjugated tokens.

As regards **D-KAT**, mistakes are essentially concentrated on common nouns (30%) and V+Mas (12%)[6], which are generally confused with each other.

In **D-SYC**, mistakes are more diversified: adjectives (11%), tokens composed by a particle followed by a name

[5] The hyperparameters we used are: batch size: 25; epochs: 100; dropout: 0.25; optimizer: Adam; patience: 3; learning rate: 0.001; learning rate factor: 0.75; learning rate patience: 2.

[6] The tag “V+Mas” (“Masdar Verb”) is used for Georgian Infinitives corresponding to the conjugated verbs.

Train	All token			Ambiguous token			Unknown token		
	D-ARM	D-KAT	D-SYC	D-ARM	D-KAT	D-SYC	D-ARM	D-KAT	D-SYC
accuracy	0.9403 0.9485	0.9773 0.9769	0.9203 0.9126	0.9418 0.9435	0.9452 0.9424	0.9330 0.9088	0.7794 0.6594	0.8923 0.8854	0.6970 0.6594
precision	0.7704 0.7725	0.7057 0.6993	0.6424 0.6612	0.7473 0.7528	0.7771 0.7390	0.8011 0.7151	0.4207 0.4159	0.4417 0.3935	0.4369 0.4159
recall	0.7242 0.7408	0.6536 0.6733	0.6133 0.6456	0.7417 0.7215	0.7284 0.6938	0.8026 0.7445	0.4100 0.4029	0.4504 0.3764	0.4047 0.4029
f1-score	0.7466 0.7563	0.6787 0.6861	0.6275 0.6533	0.7445 0.7368	0.7520 0.7157	0.8018 0.7295	0.4153 0.4093	0.4460 0.3848	0.4202 0.4093

Test	All token			Ambiguous token			Unknown token		
	D-ARM	D-KAT	D-SYC	D-ARM	D-KAT	D-SYC	D-ARM	D-KAT	D-SYC
accuracy	0.9238	0.9718	0.8813	0.9145	0.8694	0.8775	0.7441	0.8632 0.8647*	0.6067 0.6463*
precision	0.6513	0.7604	0.5832	0.6306	0.5790	0.6516	0.2920	0.4215 0.4550*	0.3128 0.3433*
recall	0.6264	0.6979	0.5725	0.6501	0.5847	0.6884	0.3124	0.3991 0.4146*	0.3431 0.3495*
f1-score	0.6386	0.7278	0.5778	0.6402	0.5818	0.6695	0.3019	0.4100 0.4339*	0.3273 0.3464*

Table 3: 1. Best scores for the training step of the POS-tagger on **D-HYE**, **D-KAT** and **D-SYC** with a CRF decoder (a) and a Linear Decoder (b); 2. Evaluation of the POS-tagger (linear decoder) on the **D-HYE**, **D-KAT** and **D-SYC** Test datasets. For the "unknown token" on **D-KAT** and **D-SYC**, the CRF decoder (LEMMING) gives better results (displayed in the table*)

token	lemma GP	lemma pred.	POS GP	POS pred.	Morph. GP
շինեցաւ *šinec'aw*	շինեմ *šinem*	շինեմ *šinem*	V	V	BÎJ3s
տաճարս *tačars*	տաճար@ս *tačar@s*	տաճար *tačar*	N+Com@PRO+Dem	N+Com	Ns@ø
սուրբ *surb*	սուրբ *surb*	սուրբ *surb*	A	A	Ns
փրկչին *p'rkč'in*	փրկիչ@ն *p'rkič'@n*	փրկիչ *p'rkič'*	N+Com@PRO+Dem	N+Com	Gs@ø
և *ew*	և *ew*	և *ew*	I+Conj	I+Conj	ø
անապատս *anapats*	անապատ@ս *anapat@s*	անապատ *anapat*	A@PRO+Dem	A	Ns@ø
հրամանաւ *hramanaw*	հրաման *hraman*	հրաման *hraman*	N+Com	N+Com	Hs
և *ew*	և *ew*	և *ew*	I+Conj	I+Conj	ø
ծախիւք *caxiwk'*	ծախ *cax*	ծախ *cax*	N+Com	N+Com	Hp
թագաւորահաւրն *t'agaworahawrn*	թագաւորահայր@ն *t'agaworahayr@n*	թագաւորահայր *t'agaworahayr*	N+Com@PRO+Dem	N+Com	Gs@ø
կոստանդեայ *kostandeay*	կոստանդին *kostandin*	կոստանդեայ *kostandeay*	N+Ant	N+Ant	Gs

Table 4: Results of lemmatization and POS-tagging on a sentence from the *Inscription of the Regent Constantine of Paperōn* and comparison with expected values manually proofread by GP

(9%), verbs (6%) and proper nouns (6%). At the moment, tokens consisting of polylexical forms are the main cause for such results (e.g. table 4).

5. Perspectives

The problems affecting our results are due to two challenges posed by the structure and the source of our data. Firstly, the amount of data remains too small to ensure representativeness of the described languages. Secondly, the large number of polylexical tokens makes processing more challenging. We intend to integrate the OCR developed by Calfa for Syriac, Old Georgian and Classical Armenian with our process, in order to increase drastically our datasets. These data will be manually proofread and pre-tagged by the previous models for training.

As regards Classical Armenian, we intend to combine the data of the NBHL on Calfa — composed in particular of more than 1.3 million tokens (190.000 of which are unique) and representative of the Armenian literary production (compilation of several hundreds of classical and medieval sources) — and lemmatized forms from the Gospels. The NBHL has already been lemmatized and the proofreading is being finalized (Vidal-Gorène et al., 2019; Vidal-Gorène and Decours-Perez, 2020). Calfa also offers a database of more than 65.000 headwords for Classical Armenian and has generated a very large number of verbal and noun forms that will be integrated into the training. Furthermore, the GP is now producing a digital corpus of all the Armenian, Georgian and Syriac texts published in the *Corpus Scriptorum Christianorum Orientalium* series.

The results presented here are a first step in the development of a lemmatizer and a POS-tagger for these languages. In particular, we only provide the results of one single neural network, but we intend to conduct a comparison with state-of-the-art technologies and rule-based approches, and to include contextual tagging at the morphological level.

We already reach a mean accuracy of 91,63% in lemmatization (84,28% for ambiguous tokens and 71,93% for unknown tokens), and of 92,56% in POS-tagging (88,71% for ambiguous tokens and 75,17% for unknown tokens). Nevertheless, these results are not robust on a wide variety of texts: resolving issue constitutes the chief objective of our upcoming experiments.

6. Bibliographical References

Arzhanov, Y. (2018). *Syriac Sayings of Greek Philosophers: A Study in Syriac Gnomologia with Edition and Translation*. Corpus Scriptorum Christianorum Orientalium, 669. Subsidia, 138. Peeters, Leuven.

Boisson, P. (2014). Précis de géographie du monde indien à l'usage des commerçants: édition et traduction annotée. In A. Mardirossian, et al., editors, *Mélanges Jean-Pierre Mahé*, Travaux et Mémoires, 18, pages 105–126. Association des Amis du Centre d'Histoire et Civilisation de Byzance, Paris.

Calzolari, V. (2017). *Acta Pauli et Theclae, Prodigia Theclae, Martyrium Pauli*. Corpus Christianorum. Series Apocryphorum, 20. Apocrypha Armeniaca, 1. Brepols, Turnhout.

Coulie, B. and Métrévéli, H. (2001). *Sancti Gregorii Nazianzeni Opera. Versio Iberica. III. Oratio XXXVIII*. Corpus Christianorum. Series Graeca, 45. Corpus Nazianzenum, 12. Brepols, Turnhout.

Coulie, B. and Métrévéli, H. (2004). *Sancti Gregorii Nazianzeni Opera. Versio Iberica. IV. Oratio XLIII*. Corpus Christianorum. Series Graeca, 52. Corpus Nazianzenum, 17. Brepols, Turnhout.

Coulie, B. and Métrévéli, H. (2007). *Sancti Gregorii Nazianzeni Opera. Versio Iberica. V. Orationes XXXIX et XL*. Corpus Christianorum. Series Graeca, 58. Corpus Nazianzenum, 20. Brepols, Turnhout.

Coulie, B. and Métrévéli, H. (2011). *Sancti Gregorii Nazianzeni Opera. Versio Iberica. VI. Orationes XI, XXI, XLII*. Corpus Christianorum. Series Graeca, 78. Corpus Nazianzenum, 26. Brepols, Turnhout.

Coulie, B. and Sirinian, A. (1999). *Sancti Gregorii Nazianzeni Opera. Versio armeniaca. III. Orationes XXI, VIII. Oratio VII*. Corpus Christianorum. Series Graeca, 38. Corpus Nazianzenum, 7. Brepols, Turnhout.

Coulie, B., Kindt, B., and Pataridze, T. (2013). Lemmatisation automatique des sources en géorgien ancien. *Le Muséon*, 126:161–201.

Coulie, B., Kindt, B., and Kepeklian, G. (2020). Un jeu d'étiquettes morphosyntaxiques pour le traitement automatique de l'arménien ancien. *Études Arméniennes Contemporaines*. in press.

Coulie, B. (1994). *Sancti Gregorii Nazianzeni Opera. Versio armeniaca. I. Orationes II, XII, IX*. Corpus Christianorum. Series Graeca, 28. Corpus Nazianzenum, 3. Brepols, Turnhout.

Coulie, B. (1996). La lemmatisation des textes grecs et byzantins : une approche particulière de la langue et des auteurs. *Byzantion*, 66:35–54.

Dereza, O., (2018). *Lemmatization for Ancient Languages: Rules or Neural Networks?: 7th International Conference, AINL 2018, St. Petersburg, Russia, October 17–19, 2018, Proceedings*, pages 35–47. Springer, Jan.

Egen, S., Gleim, R., and Mehler, A. (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*.

Furlani, G. (1933). *Le Categorie e gli Ermeneutici di Aristotele nella versione siriaca di Giorgio delle Nazioni*. Serie VI. Vol. V. Fasc. 1. Reale Accademia Nazionale dei Lincei, Rome.

Haelewyck, J.-C., Kindt, B., Schmidt, A., and Atas, N. (2018). La concordance bilingue grecque - syriaque des Discours de Grégoire de Nazianze. *Babelao*, 7:51–80.

Haelewyck, J.-C. (2011). *Sancti Gregorii Nazianzeni Opera. Versio Syriaca V. Orationes I, II, III*. Corpus Christianorum. Series Graeca, 77. Corpus Nazianzenum, 25. Brepols, Turnhout.

Haelewyck, J.-C. (2014). Historia Zosimi De Vita Beatorum Rechabitarum ; Édition de la version syriaque brève. *Le Muséon*, 127:95–147.

Haelewyck, J.-C. (2015). La version syriaque longue de l'Historia Zosimi De Vita Beatorum Rechabitarum ; Édition et traduction. *Le Muséon*, 128:295–379.

Haelewyck, J.-C. (2016). *Histoire de Zosime sur la vie des Bienheureux Réchabites. Les versions orientales et leurs* *manuscrits*. Corpus Scriptorum Christianorum Orientalium, 664. Subsidia, 135. Peeters, Leuven.

Haelewyck, J.-C. (2017a). Histoire de Zosime sur la Vie des Bienheureux Réchabites. Les trois recensions syriaques. Édition de la version résumée. *Parole de l'Orient*, 43:175–194.

Haelewyck, J.-C. (2017b). Les versions syriaques des Discours de Grégoire de Nazianze: un processus continu de révision. *Babelao*, 6:75–124.

Kindt, B. (2004). La lemmatisation des sources patristiques et byzantines au service d'une description lexicale du grec ancien. *Byzantion*, 74:213–272.

Kindt, B. (2018). Processing Tools for Greek and Other Languages of the Christian Middle East. *Journal of Data Mining & Digital Humanities*, Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages.

Kondratyuk, D., Gavenčiak, T., Straka, M., and Hajič, J. (2018). LemmaTag: Jointly tagging and lemmatizing for morphologically rich languages with BRNNs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4921–4928, Brussels, Belgium, October-November. ACL.

Lafontaine, G. and Coulie, B. (1983). *La version arménienne des discours de Grégoire de Nazianze*. Corpus Scriptorum Christianorum Orientalium, 446. Subsidia, 67. Peeters, Leuven.

Manjavacas, E., Kádár, Á., and Kestemont, M. (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. Association for Computational Linguistics.

Métrévéli, H. (1998). *Sancti Gregorii Nazianzeni Opera. Versio Iberica. I. Orationes I, XLV, XLIV, XLI*. Corpus Christianorum. Series Graeca, 36. Corpus Nazianzenum, 5. Brepols, Turnhout.

Métrévéli, H. (2000). *Sancti Gregorii Nazianzeni Opera. Versio Iberica. II. Orationes XV, XXIV, XIX*. Corpus Christianorum. Series Graeca, 42. Corpus Nazianzenum, 9. Brepols, Turnhout.

Müller, T., Cotterell, R., Fraser, A., and Schütze, H. (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.

Ouzounian, A., Goepp, M., and Mutafian, C. (2012). L'inscription du régent Constantin de Papeṙōn (1241). redécouverte, relecture, remise en contexte historique. *Revue des Études Arméniennes*, 34:243–287.

Peshiṭta Institute. (1977). *The Old Testament in Syriac according to the Peshiṭta Version*. Peshitta. The Old Testament in Syriac. Brill, Leiden.

Sanspeur, C. (2007). *Sancti Gregorii Nazianzeni Opera. Versio Armeniaca. IV. Oratio VI*. Corpus Christianorum. Series Graeca, 61. Corpus Nazianzenum, 21. Brepols, Turnhout.

Schmidt, A. (2002). *Sancti Gregorii Nazianzeni Opera. Versio Syriaca. II. Orationes XIII, XLI*. Corpus Christianorum. Series Graeca, 47. Corpus Naizanzenum, 15. Brepols, Turnhout.

Sembiante, A. (2017). Appunti sulla tradizione siriaca delle opere di Gregorio Nazianzeno. *Koinonia*, 10:607–634.

Sirinian, A. (1999). *Sancti Gregorii Nazianzeni Opera. Versio armeniaca. II. Orationes IV et V*. Corpus Christianorum. Series Graeca, 37. Corpus Nazianzenum, 6. Brepols, Turnhout.

Van Elverdinghe, E. (2018). Recurrent Pattern Modelling in a Corpus of Armenian Manuscript Colophons. *Journal of Data Mining & Digital Humanities*, Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages.

Venturini, G. (2019). *La versione siriaca della vita di Giovanni il Misericordioso di Leonzio di Neapolis*. Corpus Scriptorum Christianorum, 679. Scriptores Syri, 263. Peeters, Leuven.

Vidal-Gorène, C. and Decours-Perez, A. (2020). Languages resources for poorly endowed languages : The case study of Classical Armenian. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*. in press.

Vidal-Gorène, C., Decours-Perez, A., Queuche, B., Ouzounian, A., and Riccioli, T. (2019). Digitalization and Enrichment of the Nor Baṙgirk' Haykazean Lezui: Work in Progress for Armenian Lexicography. *Journal of the Society of Armenian Studies*, 27. in press.

7. Language Resource References

American University of Armenia. (1999). *Digital Library of Armenian Literature*.

Arak29. (2002). *Arak29*.

Calfa. (2014). *Calfa - Enriched Dictionaries of Classical and Modern Armenian*.

J. Gippert. (2003). *TITUS Project*. Johann Wolfgang Goethe University.

Ilia State University. (2009). *Georgian Language Corpus*.

Université catholique de Louvain. (1990). *GREgORI Project - Softwares, linguistic data and tagged corpus for ancient GREek and ORIental languages*.

J. E. Walters. (2004). *Digital Syriac Corpus*.