Morphologically annotated corpora of Pomak

Ritván Jusúf Karahóǧa Panagiotis Krimpas Vivian Stamou Vasileios Arampatzakis
Dimitrios Karamatskos Vasileios Sevetlidis Nikolaos Constantinides
Nikolaos Kokkas George Pavlidis Stella Markantonatou
Institute for Language and Speech Processing, Athena R.C.
{ritvan.karachotza, p.krimpas, vistamou, vasilis.arampatzakis, dkaramatskos,
vasiseve, n.konstantinidis, nikolkok, marks, gpavlid}@athenarc.gr

Abstract

The project Philotis is developing a platform to enable researchers of living languages to easily create and make available state-of-the-art spoken and textual annotated resources. As a case study we use Greek and Pomak, the latter being an endangered oral Slavic language of the Balkans (including Thrace/Greece). The linguistic documentation of Pomak is an ongoing work by an interdisciplinary team in close cooperation with the Pomak community of Greece. We describe our experience in the development of a Latin-based orthography and morphologically annotated text corpora of Pomak with state-of-the-art NLP technology. These resources will be made openly available on the Philotis site and the gold annotated corpora of Pomak will be made available on the Universal Dependencies treebank repository.

1 Introduction

In Philotis[1] we aim at supporting the researchers of living languages to develop annotated (linked) spoken and textual resources without external technical aid: ideally, speakers of the documented language and eager linguists alone would suffice. To this end, we take advantage of open-source NLP tools, semantic web technologies, annotation tools and universally adopted annotation and codification schemes. Pomak is our case study of endangered oral language. We make available existing and new textual and oral material of Pomak and develop annotated spoken and textual corpora. Here, we provide some information about Pomaks and their language and, briefly present our experience from the development of a Latin-based orthography and a morphologically annotated corpus of Pomak.

Several researchers have highlighted the interdisciplinary nature of language documentation work (to mention but a few: Woodbury 2003; McDonnell 2018; Rice 2018; Bird 2020) because different linguistic specialisations are required and linguistic activity can hardly be considered independent of its social and situational settings. Furthermore, the technical problems of resource development should not be underestimated. Back in 2003, Woodbury (2003) explained that ideally, language technology should support multimodal data and multilayered annotations that would be linked to each other so that they could be studied simultaneously. We would add that technical solutions have to be flexible, among other things because different languages may pose different documentation problems, in particular if the linguistic communities want to exploit legacy material.

State-of-the-art tools and methods greatly facilitate traditionally hard tasks such as morphological annotation of corpora (Anastasopoulos et al., 2018) and speech to text transcription (Lane et al., 2021). We have taken advantage of this technology and received excellent results but the overall experience was not devoid of problems. We proceed by introducing Pomak as an endangered oral language; next we discuss our experience with the development and morphological annotation of the corpora of Pomak.

2 About Pomak

Pomak (endonym: Pomácky, Pomácko, Pomácku or other dialectal variants) is a non-standardised East South Slavic language variety. Pomak is spoken in Bulgaria and Greece (mainly the Rhodope Mountain area), in the European part of Turkey and, in the places of Pomak diaspora (Constantinides 2007: 35). Pomak is included in the map of the European Languages Equality Network[2]. As is the case with all East South Slavic varieties, several of the linguistic features that appear in the Pomak dialectic continuum are due to mutual interaction and convergence with non - Slavic languages

[1] https://philotis.athenarc.gr/

[2] https://elen.ngo/languages-map/

Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 179 - 186

of the Balkan Sprachbund (Papadimitriou 2013: 23), mostly Latin (Solta 1980) and Greek (Krimpas 2020). In comparison to all East South Slavic languages, Pomak seems to exhibit a more profound phonological, morphological, morphosyntactic and lexical influence by Medieval and Modern Greek (Krimpas 2020: 196) and, due to the predominantly Muslim religion of its speakers, a more profound lexical and phonotactical influence by Ottoman and Modern Turkish.

There is no widely accepted orthography of Pomak. The language is not taught in any of the countries where Pomaks reside.

Table 1 describes Pomak with the six factors of language vitality and endangerment proposed in Brenzinger et al. (2003). Note that "A language that is ranked highly according to one criterion may deserve immediate and urgent attention due to other factors" (Brenzinger et al. 2003: 9).

1. Factor 1. "(4)" is defined as: "Most but not all children or families of a particular community speak their language as their first language, but it may be restricted to specific social domains (such as at home where children interact with their parents and grandparents)." (Brenzinger et al. 2003: 9).

2. The value of factor 2, and consequently of factor 3, is an estimation (Adamou and Fanciullo, 2018).

3. Factor 4. "(3)" is defined as: "The language is used in home domains and for many functions, but the dominant language begins to penetrate even home domains." (Brenzinger et al. 2003: 10).

4. Factor 5. "(1)" is defined as: "The language is used only in a few new domains." (Brenzinger et al. 2003: 11).

5. Factor 6. "(2)" is defined as: "Written materials exist, but they may only be useful for some members of the community; and for others, they may have a symbolic significance. Literacy education in the language is not a part of the school curriculum." (Brenzinger et al. 2003: 12).

Factors of language vitality and endangerment	**Scores for Pomak**
1. Intergenerational Language Transmission	4
2. Absolute Number of Speakers	35000
3. Proportion of Speakers within the Total Population	3,2 %
4. Trends in Existing Language Domains	3
5. Response to New Domains and Media	1
6. Materials for Language Education and Literacy	2

Table 1: Factors of language vitality and endangerment for the Pomak language as of 2021.

3 Compiling textual corpora of Pomak

An oral/endangered language may have some textual and audio legacy (Gerstenberger et al., 2017). There are sporadic transcriptions and recordings of Pomak folk songs and tales; in addition, there are very few modern texts (journalistic texts and translations from Greek and English into Pomak). The texts are in a variety of alphabets ranging from Cyrillic to Greek to an English-based Latin alphabet. We collected these dispersed resources via a network of native speakers and Greek scholars who are close to the Pomak community. Following the requirements of the Pomak community, selected parts of this material was included in the developed corpora and the original material will be made available exactly as it was received. Our research center and the copyright owners (authors, publishing houses) have agreed, according to the Greek law, to ensure free distribution of the material for research purposes. Eventually, a corpus of about 130000 words was compiled. Table 2 shows the types of text included and the size of the respective corpora in words. Where possible, the geographical origins of the texts are given as a reliable indication of the dialect represented in the text.

Mature open-source NLP technology that would take full advantage of archived textual material is not available yet (Hutchinson 2020). Undoubtedly, a detailed TEI-conformant encoding of this material is the optimum approach but, at the moment, we have given priority to (spoken) material collection. We are in the process of defining Dublin Core and TEI-conformant metadata to declare the origins of the material in the corpus and to develop links of medium granularity between the resources.

Text types	Words	Geographical origins
Folk tales	43.817	Emonio, Glafki, Dimario, Echinos, Myki, Pachni, Oreo
Language description	19.524	mixed
Journalism	25.236	Myki
Translations into Pomak	24.208	Myki - Pachni
Folk songs	18.434	mixed
Proverbs	550	mixed
Other	5.325	Myki

Table 2: Pomak corpus: type, size and geographical origins of texts.

In addition to the above, an oral/endangered language may have resources that can be used to effectively improve the quality of its morphosyntactic annotation. In our work with Pomak we had the benefit of the electronic lexicon Rodopsky[3], which contains approximately 61.500 lemmas that correspond to about 3.5 x 10^6 unique forms (i.e., combinations of a lexical token and a PoS symbol) annotated for lemma, PoS and morphological features (Figure 1). We exploited Rodopsky to obtain mature morphological annotation of the corpora so we needed morphological annotation and evaluation facilities separate from the syntactic ones.

It goes without saying that, since Pomak has been sparsely documented by individuals who employed largely incompatible orthographies and in order to take advantage of Rodopsky, which also employs its own orthography, text homogenisation work was deemed necessary. The first step to this direction was the Krimpas et al. (2021) alphabet (K&K alphabet from now on, illustrated in Table 5). First Rodopsky was transcribed automatically to the K&K alphabet and corrected manually; the procedure helped us better define the orthography applied to the corpora. Finally, the morphological annotation of the corpus required additional orthographic refinements. The various orthographies used in the corpora were automatically mapped on the K&K orthography and the output was corrected manually.

[3] https://www.rodopsky.gr/

Figure 1: Rodopsky: Electronic lexicon of Pomak. Partial screenshot of the entry *čulæk* 'man' with morphological annotation encoded in Greek.

We proceed to a brief presentation of the adopted orthography of Pomak.

4 The orthography of Pomak

A key issue in developing the corpora of Pomak was the orthography. No alphabet of Pomak proposed so far, let alone orthography, has enjoyed any acceptability. A good alphabet would, at minimum, help maximise the possible impact of the developed resources on the sustainability of the documented language. In the case of Pomak, we have adopted the K&K alphabet (Table 5) that is the outcome of several years of manual work.

Cahill and Karan (2008) discuss good practices for developing orthographies for oral languages. Armostis et al. (2014) discuss the case of Cypriot Greek, which is a major dialect of Greek not adequately represented by the standard Greek alphabet. Their overall recommendation is that native speakers should participate in the definition of the orthography and have the final word in several decisions. Furthermore, they identify the, probably conflicting, good practices that are briefly introduced and discussed immediately below:

Phonetic transparency. The phonemic analysis of a language is indispensable for orthography design but detailed work with a multitude of languages suggests that the lexical level of phonology is also important. This is because while different phonological processes may result in given surface forms, native speakers may be aware of some phonological processes but not of others, so they may only be aware of the lexical form. After all, the script is meant to be used first and most by native speakers rather than by linguists. We exemplify the application of these ideas with the following Pomak words that form minimal pairs on the basis of sound to phoneme correspondence: *paláta* 'floor'

vs. *palta* ‘doused’, *cíkom* ‘squeak’ vs. *číkom* ‘cut; break’, *samár* ‘saddle’ vs. *šamár* ‘slap’, *som* ‘I am’ vs. *søm* ‘I sow’, *grom* ‘thunder’ vs. *grøm* ‘ I heat’, *pat* ‘under’ vs. *pæt* ‘read (past passive participle)’, *lóka* ‘valley’ vs. *lka* ‘light (adj., acc.masc.sing.), *sénem* ‘I shadow’ vs. *šénem* ‘I amuse myself’, *vris* ‘fountain; tap’ vs. *vriš* ‘you boil; you are full of’.

Systematic orthographies with reliable sound-symbol representation and consistent spelling enjoy enhanced acceptability, learnability, and usability by native speakers. Spelling should not be affected by pronunciation changes due to context. For instance, b [b], d [d], g [g] are devoiced in word-final position or before a voiceless consonant. We chose not to orthographically show this devoicing for the sake of consistency across declension (in the case of nominal forms) and conjugation (in the case of verbal forms). This is why we spell *hlæb* ‘bread (Nom|Sg)’ even though this form is pronounced [hlp] given the final position of the originally voiced consonant; in this way spelling is consistent with all other forms, e.g. *hlbu* ‘of/to (the) bread’, *chlba* ‘bread (Acc|Sg), *hlbove* ‘breads (Nom|Pl)’ etc.

Easily discriminable symbols: Similar symbols or crowd adjoining letters, mirror–image symbols, overuse of a letter as part of various digraphs (e.g., bh, dh, ...), superimposing more than one diacritic are not recommended. For example, graphs denoting palato-alveolar sibilants are consistently spelled by adding a háček above their non-palato-alveolar counterparts (as in most other Latin-written Slavic languages), while graphs denoting palatalised sonorants are consistently indicated by means of a cedilla (or comma depending on the keyboard) below their non-palatalised counterparts as in Latvian; this system was preferred to Croatian *lj* and *nj* or Slovak *l’* and *ň* since the former requires two graphs and the latter is not consistent. Examples: *cístem* ‘I clean’ vs. *čeréša* ‘cherry’, *slónce* ‘sun’ vs. *šténe* ‘puppy, cub’, *zólezo* ‘iron’ vs. *žalvá* ‘turtle; tortoise’, *kópele* ‘lad’ vs. *kókaļe* ‘bones (Pl)’, *pésne* ‘song’ vs. *kámeņe* ‘stones (Pl)’.

Portability of the alphabet. UNICODE is strongly recommended. The K&K alphabet of Pomak is encoded in Unicode.

Decisions might be needed as to where *word delimiters* should be put, often in the cases of compounds, clitics, pronouns, and prepositions. Distributional and phonological criteria are applied. For instance, various interrogative, indefinite and negative pronouns, conjunctions and adverbs, the first element of which is originally a preposition or a particle are normally used as a single word in most Slavic languages. However, given that there are quite a few cases where components are written as separate words in given contexts e.g., *at* ‘from; out of’, *kak* ‘how; as; like’, *kadé* ‘where’), we chose to write them as two words irrespective of context. So, instead of writing *atkák* ‘since’, *níkutrí* ‘nobody’ and *nókade* ‘somewhere’ we write *at kak* ‘since’, *ní kutrí* ‘nobody’ and *nó kadé* respectively.

Dialectical issues. Most languages consist of dialect continua often exposing systematic phonological and morphosyntactic differences across dialects. In the uni-lectal approach one dialect serves as the basis for the written form and the others make a mental adjustment while reading and writing. In the multi-lectal approach the dialects are accommodated via consideration of the various varieties (Cahill and Karan, 2008). Pomak has several dialects. The K&K alphabet stands somewhere between the two approaches. For example, the vowel in the first syllable of *zmom* ‘(that) I take’ is pronounced as [ø] in Myki, as [jo] in Echinos, and as [e] in Dimario. However, we chose to spell it as *ø* irrespective of dialect, given that speakers from Echinos or Dimario automatically pronounce [ø] as [jo] or [e], respectively, while speakers from Myki, if asked to read out the spellings *jo* and *e* respectively, would not automatically pronounce them as [ø], given that they do not have the [jo] and [e] sounds; moreover, there are words that are spelled and pronounced with [jo] or [e] in all dialects, e.g. *med* ‘honey’, *jok* ‘non-’. Of course, since Pomak dialects are numerous and geographically dispersed, major vowel differences cannot sometimes be spelled by means of a ‘neutral’, i.e. hyperdialectal orthography.

5 The gold morphologically annotated corpus

We have already said that in our work with Pomak we had the benefit of the electronic lexicon Rodopsky (Fig. 1), which contains approximately 3.5 x 10^6 unique forms annotated, among other things, for lemma, PoS and morphological features. In order to take advantage of this rich source of linguistic knowledge of Pomak, some adaptation work was required: apart from transcribing it to the K&K orthography, the morphological annotation had to be mapped on the Universal Dependencies frame-

work (UD)[4] and the CONLLU format had to be adopted. UD was chosen as a morphosyntactic annotation framework because of its large inventory of annotation features and because it is recognised by several open-source, state-of-the-art NLP tools that we planned to use for the morphosyntactic annotation of the corpus.

The mapping on UD revealed problems of which the most important were:

1. The analysis in Rodopsky did not include the UD PoS DET(erminer) and X(other). In addition, re-assignment of PoS to several lemmas was required, e.g., which participles would be considered adjectival or verbal forms.

2. Additional morphological features were necessary to describe (i) Degree modification of nouns, adjectives and adverbs (Degree modification should not be confused with Comparison), (ii) Determiners and adverbs that are formed with one of the particles *náe / nó, ní, sé*; these are assigned the new feature "particle type" with values "indicative", "negative" and "total".

3. The tense and aspect system of Pomak required extra attention in order to be described with some accuracy.

The mapping of the morphological annotation of Pomak in Rodopsky on the UD framework was carried out by native speakers and linguists and the results will be uploaded on the UD language specifications area. Furthermore, it revealed interesting parallel phenomena of Greek and Pomak, in particular in the verb and the Degree modification systems that deserve a closer study.

Once Rodopsky was transcribed into a UD and CONLLU compatible form and was manually corrected, it was mapped on the corpora (both Rodopsky and the corpora had been transcribed into the K&K orthography). This initiated an about 30-days long cycle of manual corrections, this time of 6350 sentences and 86700 words selected from the Pomak corpus to form the gold tagged corpus that would be used for training and evaluating the NLP tools. This part of the annotation was performed by a native speaker and a linguist fluent in Pomak but not in UD, so the manual annotation time reported includes their training in the framework (Interannotation agreement kappa scores on 476 sentences: PoS tags 0.90, features 0.87, lemmas 0.93). The corpus will be uploaded to the UD language repository.

Alternatively, we could have proceeded with the morphological annotation of gradually bigger corpora (Anastasopoulos et al. 2018). However, the selected procedure had clear merits:

1. We proceeded faster since the annotators worked on texts that were assigned morphological annotation of good quality.

2. Dedicated resources mitigate the effect of imposing knowledge from other languages onto the documented one through shared training language models.

3. It made room for the active participation of the community in the documentation process of their native language.

On the downside of the procedure are:

1. The overall procedure of transcribing Rodopsky into CONLLU cannot be generalised and made useful to other languages.

2. We faced extra problems with the NLP tools because some of them do not offer the option of separate morphological and syntactic annotations (see below).

6 Morphological annotation of the corpus of Pomak

The gold morphologically annotated corpus was used to train and evaluate NLP tools that would, in turn, be used to assign morphological annotations to the entire Pomak corpus and to future material from the spoken corpora. We conducted a series of experiments with four tools in an effort to identify the one that would yield the best morphological annotation results for Pomak.

The situation with state-of-the-art open-source NLP tools reminded of the description by (Arkhipov and Thieberger 2018:141): "... although basic principles are quite straightforward to master, the details of use of particular tools and interaction between tools in different setups are highly specific and can often be a source of frustration. Thus, not only an effort is required from the LD practitioners to invest in learning, but considerable effort is also required from the developers to invest in harmonisation of tools and making workflows more straightforward and robust."

[4] https://universaldependencies.org/

Our experience confirms that even people with a training in programming must spend considerable time on state-of-the-art NLP tools. We ran four open-source tools, all implemented in Python. All tools provided a command line interface, but:

A. Instructions often were problematic: (a) Outdated compilation instructions (b) Instructions for training a model of a new language from scratch: (i) some tools provided insufficient documentation of the addition of languages new to the UD framework, and (ii) the alignment of the processes included in the pipeline was a hard task with some tools with incomplete instructions (c) Outdated README instructions required missing files; we had to correct the code.

B. Both the separation of morphological from syntactic annotation and the independent evaluation of the two annotation levels were hard.

We used Rodopsky for the morphological annotation of the Pomak corpus and we wanted to evaluate morphological annotation only, however some tools did not allow for this. Also, all tools assigned both morphological and syntactic annotation which may not be always desirable because when a new language is documented, the various levels of analysis (morphology, syntax, semantics etc) have not reached the same stage of maturity. Morphology is the basic annotation level and it is reasonable to address it first. We think that the unified annotation should be an option and not the rule. We had to rewrite the code of some NLP tools and comment out the parts handling dependency relations in order to obtain evaluation results for the morphological annotation.

This said, we would like to note that, probably, the assignment of false dependency relations might eventually be of no or little harm. We plan to compare the unified and the two-stage annotation strategy with future experiments on the corpora of Pomak.

There is a keen interest in incorporating contextual word embeddings as a functionality (Nguyen et al., 2021) but at the moment, pretrained transformer models are available with few tools only. Amongst the ones we tested, spaCy v3.2.2 allows for transformer based autoregressive models, while Udify supports only Bert like models.

One might note that pretrained multi-language models can be used by just one openly available NLP library. However, languages with no annotated corpora, such as Pomak, must have access to pretrained multi-language models in order to be assigned a reasonable (first) morphosyntactic annotation (Anastasopoulos et al., 2018).

We investigated the performance of the tools spaCy v3.2.2[5] (Honnibal et al., 2020), Stanza[6] (Qi et al., 2020), UDify[7] (Kondratyuk and Straka, 2019) and UDPipe[8] (Straka et al., 2016) on the gold morphologically annotated corpus of Pomak that was further split into training, development and test set (80:10:10). (Table 3).

Corpus	Train	Dev	Test
Sentences	5000	671	679
Tokens	67345	9736	9701

Table 3: Statistics on the training, development and test sets.

We experimented with the tasks of lemmatisation, PoS tagging and morphological annotation. The performance of each tool on the Pomak corpus is illustrated in Table 4.

Parser	Model	LEMM	UPOS	FEATS
SpaCy	XLM-Roberta-large	93.85	98.38	**95.54**
Stanza	Stanza	**97.82**	**98.73**	95.23
UDify	UDify-base	90.27	97.59	91.03
UDPipe	UDPipe v1.2	92.04	95.94	90.39

Table 4: Accuracy scores for the tasks of lemmatisation (LEMM), PoS tagging (UPOS) and morphological feature (FEATS) assignment. The highest scores in each column are in bold.

Table 4 shows that Stanza achieves the best accuracy scores in PoS tagging and lemmatisation and spaCy in feature assignment. We note that in the case of spaCy we ran (the large pretrained multilingual model) RoBERTa (XLM-RoBERTa). All tools returned reasonable PoS tagging results.

The entire annotated corpus of Pomak will be made available on Philotis. We are currently in the process of assigning syntactic annotation to the Pomak corpus according to the UD paradigm.

[5] https://spacy.io/

[6] https://stanfordnlp.github.io/stanza/

[7] https://github.com/Hyperparticle/udify

[8] https://ufal.mff.cuni.cz/udpipe/1/models

7 Conclusion

We have described the procedure of developing state-of-the-art textual resources for Pomak, an endangered, oral European language of the Slavic family. A group of linguists, computational linguists and engineers took full advantage of the Pomak legacy and cooperated closely with the native speaker community. In this way and in a short period of time (about 8 months), we produced reasonably sized morphologically annotated corpora of good quality and identified the open source NLP tools for the morphological annotation of Pomak.

We have also reported on our experience with using open NLP tools. We have observed that skilled programmers may still be needed in order to use these tools. Furthermore, powerful tools have not been fully exploited yet. In the overall, however, the huge progress in openly available state-of-the-art NLP technology has boosted the development of resources for endangered oral languages.

References

Evangelia Adamou and Davide Fanciullo. 2018. Why Pomak will not be the next Slavic literary language. In D. Stern, M. Nomachi, and B. Belić, editors, *Linguistic regionalism in Eastern Europe and beyond: minority, regional and literary microlanguages*, pages 40–65. Peter Lang.

Antonios Anastasopoulos, Marika Lekakou, Josep Quer, Eleni Zimianiti, Justin DeBenedetto, and David Chiang. 2018. Part-of-speech tagging on an endangered language: a parallel Griko-Italian resource. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2529–2539, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Alexandre Arkhipov and Nick Thieberger. 2018. Reflections on software and technology for language documentation. In Andrea L. Berez-Kroeker Bradley McDonnell and Gary Holton, editors, *Reflections on Language Documentation. 20 Years after Himmelmann 1998 Language Documentation & Conservation Special Publication*, volume 15, pages 140–149.

Spyros Armostis, Christodoulou Kyriaci, Katsoyannou Marianna, and Charalambos Themistocleous. 2014. *Addressing writing system issues in dialectal lexicography: the case of Cypriot Greek*, page 23–38.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 3504–3519. International Committee on Computational Linguistics.

Matthias Brenzinger, Arienne M. Dwyer, Tjeerd de Graaf, Colette Grinevald, Michael Krauss, Osahito Miyaoka, Nicholas Ostler, Osamu Sakiyama, María E. Villalón, Akira Y. Yamamoto, and Ofelia Zepeda (UNESCO Ad Hoc Expert Group on Endangered Languages). 2003. Language vitality and endangerment. Paris, 10-12.

Michael Cahill and E. V. Karan. 2008. Factors in designing effective orthographies for unwritten languages. In *SIL Electronic Working papers 2008-001*.

Nikolaos Constantinides. 2007. *Units of the Pomak Civilization in Greek Thrace. Brief historical review,language and identities*. Democritus University of Thrace:MA Thesis.

Ciprian-Virgil Gerstenberger, Niko Partanen, and Michael Rießler. 2017. Instant annotations in elan corpora of spoken and written Komi, an endangered language of the Barents Sea region. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages, Honolulu, Hawai*, pages 57–66.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Tim Hutchinson. 2020. Natural language processing and machine learning as practical toolsets for archival processing. *Records Management Journal*, 30:155–174.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Panagiotis Krimpas, Nikolaos Constantinides, Ritvan Karahoğa, Stella Markantonatou, and George Pavlidis. 2021. «Pomak: An idiosyncratic South East Slavic language?». In *VII Scientific Conference on "The Traditional Culture of Greece"»*, Lomonosov State University in Moscow.

Panagiotis G. Krimpas. 2020. Language and origin of Pomaks in the light of the Balkan Sprachbund. In A. Bartsiokas & N. Macha-Bizoumi M. Varvounis, editor, *The Pomaks of Thrace: Multidisciplinary and interdisciplinary approaches*, pages 167–204. Thessaloniki: K&M Stamoulis.

William Lane, Mat Bettinson, and Steven Bird. 2021. A computational model for interactive transcription. In *Proceedings of the Second Workshop on Data Science with Human in the Loop: Language Advances*, pages 105–111, Online. Association for Computational Linguistics.

Bradley McDonnell. 2018. Reflections on linguistic analysis in documentary linguistics. In Andrea L. Berez-Kroeker & Gary Holton Bradley McDonnell, editor, *Reflections on Language Documentation. 20 Years after Himmelmann 1998. Language Documentation Conservation Special Publication*, volume 20, pages 191–200.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *EACL (System Demonstrations)*, pages 80–90.

Panayotis G. Papadimitriou. 2013. *Dialects of the Pomaks of the Greek Rhodope. Regional Analytical Slavic and Muslim speakers in Southeastern Europe.* Thessaloniki: Balkan Peninsula Research Institute. [In Greek].

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Sally Rice. 2018. Reflections on documentary corpora. In Andrea L. Berez-Kroeker & Gary Holton Bradley McDonnell, editor, *Reflections on Language Documentation. 20 Years after Himmelmann 1998. Language Documentation Conservation Special Publication*, 15, pages 157–172. University of Hawai'i Press.

Georg Renatus Solta. 1980. *Einführung in die Balkanlinguistik mit besonderer Berücksichtigung des Substrats und des Balkanlateinischen*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

Anthony C Woodbury. 2003. Defining documentary linguistics. In Peter K. Austin, editor, *Language Documentation and Description*, volume 1, pages 35–51. London: SOAS.

A Appendices

Pronunciation	**Character**	**Example word**
[a], [ɐ]	A, a	**a**stávem 'leave'
[ɛɐ], [æ]	Æ, æ	l**æ**k 'drug'
[b]	B, b	**b**ába 'grandmother'
[tŝ]	C, c	**c**ístem 'clean'
[ʧ]	Č,č	**č**eréša 'cherry'
[d]	D, d	**d**órvo 'wood'
[e]	E, e	pr**e**dávom 'sell'
[f]	F, f	**f**átom 'catch'
[g] [gʲ]	G, g	**g**órlo 'throat'
[dẑ]	Ǵ, ǵ	**ǵ**væzda 'star'
[dʒ̂]	Ğ, ğ	**ğ**umajá 'mosque'
[x]	H, h	**h**ránem 'feed'
[i]	I, i	v**i**sok 'tall'
[j]	J, j	**j**áto 'food'
[k], [kʲ]	K, k	**k**ukóška 'hen'
[ɬ], [lʲ]	L, l	**l**ažýca 'spoon'
[ʎ]	Ļ, ļ	kókaļe 'bones'
[m]	M, m	**m**agáre 'donkey'
[n], [nʲ]	N,n	**n**us 'nose'
[ɲ]	Ņ, ņ	spaņé 'sleep'
[o], [u], [a], [ɐ]	O, o	p**o**t 'road'
[ø]	Ø, ø	sp**ø**m 'to sleep'
[p]	P, p	**p**étal 'horse shoe'
[r]	R, r	**r**ábata 'work'
[s]	S, s	**s**órce 'heart'
[ʃ]	Š, š	**š**ápka 'cap'
[t]	T, t	**t**umafíl 'car'
[u]	U, u	**u**šá 'ear'
[y], [ʲu]	Ü, ü	t**ü**rén 'train'
[v]	V,v	**v**orh 'top'
[ɨ]	Y, y	k**y**smét 'fortune'
[z]	Z, z	**z**imá 'winter'
[ʒ]	Ž, ž	**ž**alvá 'turtle'

Table 5: The A&A (2021) alphabet: phonemes, character set, usage examples.