Methodological issues regarding the semi-automatic UD treebank creation of under-resourced languages: the case of Pomak

Stella Markantonatou Nicolaos Th. Constantinides Vivian Stamou
Vasileios Arampatzakis Panagiotis G. Krimpas George Pavlidis
Institute for Language and Speech Processing, Athena R.C.
{marks, n.konstantinidis, vistamou, vasilis.arampatzakis, p.krimpas, gpavlid}@athenarc.gr

Abstract

Pomak is an endangered oral Slavic language of Thrace/Greece. We present a short description of its interesting morphological and syntactic features in the UD framework. Because the morphological annotation of the treebank takes advantage of existing resources, it requires a different methodological approach from the one adopted for syntactic annotation that has started from scratch. It also requires the option of obtaining morphological predictions/evaluation separately from the syntactic ones with state-of-the-art NLP tools. Active annotation is applied in various settings in order to identify the best model that would facilitate the ongoing syntactic annotation.

1 Introduction

The development of the Pomak UD (Universal Dependencies) treebank was carried out as a case study of the project PHILOTIS, which aims at providing the infrastructure for the multimodal documentation of living languages.[1] Pomak is an endangered oral Slavic language of historical Thrace (South Balkans). Morphological and syntactic annotation are carried out in two distinct settings because the first one uses existing resources and the second one starts from scratch.

Sections 2 and 3 briefly present the current situation of Pomak language, the script/orthography adopted for the development of its treebank and the available resources. In Section 4 a short linguistic description of Pomak is given in the UD framework. The annotation procedure is discussed in Section 5. Conclusions and future plans are presented in Section 6.

2 About Pomak

Pomak (endonym: Pomácky, Pomácko, Pomácku or other dialectal variants) is a non-standardised East South Slavic language variety.[2] Pomak is spoken in Bulgaria and Greece (mainly in the Rhodope Mountain area), in the European part of Turkey and in places of Pomak diaspora (Constantinides 2007: 35). The Pomak dialect continuum has been influenced by Greek and Turkish due to extensive bilingualism or trilingualism (Adamou and Fanciullo, 2018).

Pomak scores low on all six factors of language vitality and endangerment proposed by UNESCO (for more details see Brenzinger et al. 2003). In short, there is little written legacy of merely symbolic significance for the speakers of Pomak, the language is not taught at school, it is mainly used in family settings, which are increasingly penetrated by the dominant language(s) (Greek, Turkish).

Furthermore, Pomak showcases certain issues involved in the development of NLP resources for an oral, non-standardised language with some legacy such as texts and/or lexica of some type; this is the case for a number of, European at least, language varieties (Gerstenberger et al., 2017; Bernhard et al., 2021). The exploitation of the linguistic knowledge contained in such legacy may require (i) the transcription/transformation of the textual sources to the right processable format (ii) some adaptation of the processing pipelines offered by open-source state-of-the-art NLP tools. Both types of action were required for the development of the Pomak UD treebank.

3 Pomak textual sources and scripts

There are sporadic transcriptions and recordings of Pomak folk songs and tales as well as very few modern texts (mostly journalistic texts and translations from Greek and English into Pomak). The existing texts are written in a variety of scripts, ranging from Bulgarian-based Cyrillic to Modern Greek to an English-based Latin alphabet. PHILO-

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

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

Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023), pages 27 - 35

TIS collected these scattered resources through a network of native speakers and Greek scholars who, for various reasons, are close to the Pomak community. Selected parts of this material are included in a corpus of about 130,000 words. The corpus will be made available on the web to institutions and individuals for research and innovation purposes; at the moment, it is available for the same purposes after personal contact with the first and/or last author. Table 1 shows the text genres included in the corpus and the size of the respective texts in words. Where possible, the geographical origin of the texts is also given as a hint to the dialect used in the text.

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

We took advantage of the Pomak electronic lexicon Rodopsky, which contains about 61.500 lemmas corresponding 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.[3] Rodopsky, the existing textual sources and the fact that Pomak is a Slavic language have helped us solve several issues regarding Pomak orthography, namely identification of words and their grammatical function and the identification of inflectional paradigms. Still, a lot of work was required to adapt the linguistic information in Rodopsky and the textual legacy of Pomak to contemporary linguistics and UDs.

Text homogenisation work was necessary because Pomak texts, including Rodopsky, employ various orthographies. The Latin-based alphabet proposed by Ritvan Karahoğa and Panagiotis G. Krimpas (hereinafter: K&K alphabet), which has a language resource-oriented accented version and a non-accented all-purpose version, was used to semi-automatically transliterate Rodopsky and the corpus. The K&K alphabet has been developed along the following lines (Karahóğa et al., 2022): (i) Portability of the alphabet (use of UNICODE) (ii) Phonetic transparency (iii) Easily learned representations of sounds due to the use of similar diacritics for same articulation sounds and the absence of digraphs (iv) Consistent spelling not affected by predictable allophony; for instance, the de-voicing of b [b], d [d], g [g] in word-final position or before a voiceless consonant is not shown for the sake of consistency across the declension/conjugation paradigm, which is why *hlæb* 'bread (NomlSg)' is spelled with a b although it is actually pronounced [hlɛʊp] in order to ensure consistency across the clitic paradigm *hlæbu* 'of/to (the) bread', *hlæbove* 'breads (NomlPl)' etc. (v) The K&K alphabet is based on the dialect of Myki but it can also partially serve as a hyperdialectal script by allowing various predictable pronunciations of the same graph according to dialect; for instance, the vowel in the first syllable of *zømom* '(that) I take,' which is pronounced as [ø] in Myki, can also be acceptably pronounced as [jo] in Echinos or as [e] in Dimario, no matter that all three variants are spelled with an ø. This is because speakers from Echinos and Dimario have no [ø]-sound, which is why they spontaneously replace it with [jo] or [e], respectively, while speakers from Myki, if asked to read out the digraph jo or the graph e would not automatically pronounce them as [ø], given that they do have words with [jo] and [e] in their native variety. However, a hyperdialectal Pomak orthography is often not possible, given that some varieties differ also in the lexical and/or morphological level.

4 Pomak morphosyntax at a glance

A short morphosyntactic description of Pomak follows in the framework of Universal Dependencies, Version 2 (UD).[4]

4.1 Morphology of Pomak

In the orthography adopted, words are delimited by white space characters. Distributional and phonological criteria were applied regarding the place-

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

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

ment of white spaces with certain interrogative, indefinite and negative pronouns, conjunctions and adverbs that are spelled as a single word in most Slavic languages but as two words in the adopted Pomak orthography, e.g., *at kak* for *atkák* 'since', *ní kutrí* for *níkutrí* 'nobody' and *nó kadé* for *nókade* 'somewhere'. In all these cases, the first word can be independently identified as a preposition or particle, e.g., *at* 'from; out of', *kak* 'how; as; like', *kadé* 'where' and the second as an interrogative pronoun or adverb. The particles are assigned the PoS tag 'PART' and the feature 'PartTypeQpm' with one of the values 'Ind' (indefinite), 'Neg' (negative), 'Tot' (total). 'PartTypeQpm' is defined for Pomak.

In the general description of Pomak morphological features given below, certain interesting or very special cases are highlighted. The Pomak treebank uses 16 universal POS categories ('SYM' is not used).

4.1.1 The grammatical features gender, number, case and animacy

Pomak common and proper nouns, determiners, adjectives, pronouns, participles and some of the numerals are morphologically marked for gender, number, case and animacy (see below).

Gender, Case: Pomak overtly marks three genders (masculine, feminine, neuter) and four cases (nominative, genitive, accusative and vocative).

Animacy: The opposition 'Human vs. Non-human' is overt with masculine plural and rarely with masculine singular of adjectives, pronouns and participles.

Number: In addition to singular and plural number, Pomak also has:
(i) plurale tantum, e.g., *pantóly* 'pants', *diláve* 'fire tongs', *nallamý* 'pincers', collective nouns ending in *-ja* are always plural (the feature has not yet been implemented in the UD treebank)
(ii) count plural, used with masculine nouns after numerals; etymologically, this is a relic of the dual form, e.g. *dva balóna* 'two ballons', *dva kámenæ* 'two stones'
(iii) collective/mass/singulare tantum; collective nouns ending in *-(j)e*, despite having always plural (collective) meaning, can be either grammatically singular (a less frequent case) or grammatically plural, depending on the speaker's perception of the set of objects as a whole or as distinct items (dialectal variation is possible), e.g., *balóṇe / baloṇá* 'multitude of ballons".

With possessive determiners both the number of the possessor and the possessed object are encoded.

4.1.2 Diminutives; the tripartite enclitic definite article

Like most Balkan/Slavic languages, Pomak has a rich inventory of diminutive and augmentative forms of nouns, adjectives, adverbs and certain passive participles; the feature has been implemented in the UD treebank.

Pomak is special in that it uses a **tripartite enclitic definite article** *-s, -t, -n* (Adamou and Fanciullo, 2018; Krimpas, 2020) that occurs with nouns, adjectives, strong types of pronouns, certain numerals, adverbs and passive participles and denotes deixis and definiteness as follows:
(i) Proximity to the speaker, annotated as 'Deixis=Prox' and 'DeixisRef=1', e.g., *čulékos* 'the man close to the speaker'
(ii) Proximity to the listener, annotated as 'Deixis=Prox' and 'DeixisRef=2', e.g., *čulékot* 'the man close to the listener'
(iii) Distance from both the speaker and the listener, annotated as 'Deixis=Remt', e.g., *čulékon* 'the man who is away from both the speaker and the listener'.

The feature 'DeixisRef' has been defined for Pomak because the attested opposition between "proximity to the speaker" and "proximity to the listener" could not be modelled with the values available in UD for the feature 'Deixis' that do not distinguish among reference points.

4.1.3 Auxiliaries

The auxiliary *som* 'to be' is used to form perfect verb tenses and the passive voice. *som* is considered a verb (and bears the dependency relation 'root') when it means 'to exist' (1), or heads an impersonal clause with a phrasal subject (2).

(1) je górmon ad pó napréš itám
is forest from more near there
'A forest is nearby.'

(2) tébe tí je jálnis
you to you is only
da rečéš krívo
to speak wickedly
'All you can do is to speak wickedly'

šom/štom and *še/ša* express possibility and, like the Greek θα, precede indicative verb forms to form the tense 'Future' (3).

(3) ja še tí dam halvá
I will you give halva
'I will give you halva (a kind of a candy)'

The question particle *li*, e.g., *dojdéš li* 'do you come?', is assigned the PoS label 'PART' and the dependency 'aux:q'.

4.1.4 Verbs

Modal verbs, personal and impersonal verbs, participles, infinitives and converbs are assigned the PoS 'VERB'.

Verbs have finite and non-finite forms. *Finite verbs* are marked for 'Mood' with values 'Ind' (indicative) or 'Imp' (imperative), one of the four values of 'Number' (see above) and one of the three values of 'Person': '1', '2' or '3'. Verbs in the 'Ind' mood are marked for one of the two values of 'Tense', namely 'Past' or 'Pres' (present).

As in all Slavic languages, *aspect* is either a lexical or a morphological feature of the verb; it is described with the values 'Imp'(imperfective) or 'Perf'(perfective) of the feature 'Aspect', e.g., *kázavom*, *kážom* 'to say/to narrate' respectively.

There are three types of *nonfinite verb forms*: converbs, participles and infinitives. Only passive participles are assigned the pair 'Voice=Pass'; all other verb forms are assigned the pair 'Voice=Act'.

The infinitive forms the prohibitive imperative (4) when it appears after the particles *na/ne* and *namój* (sing.)/*namójte* (pl.) 'not'.

(4) namój barzá
not you rush
'do not rush'

Interestingly, Pomak has another, innovative form of infinitive, which may be called the **morphologically reduplicated infinitive** ending in *-titi*, crystallised in a small number of imperfective verbs that are repeated as bilects denoting the continuous/monotonous/rythmic repetition of a motion,'e.g. *čúktiti čúktiti* 'hit and hit'.

To summarise, Pomak uses the UD morphological apparatus extensively, including features for diminutives, and defines two new Pomak-specific features, namely 'PartTypeQpm' and 'DeixisRef'.

4.2 Syntax of Pomak

The Pomak treebank implements most UD dependency relations (hereinafter: "dependencies"). So far, not used dependencies include: 'cop' (copula), and 'dep' (unspecified dependency). As syntactic annotation of Pomak is still ongoing, modifications may occur in future editions of the treebank. The introduction of the following two dependencies is among our plans: (i) 'cop', as in the standing edition of the Pomak treebank auxiliaries depend on content words with the dependency 'aux' for reasons of uniformity and, (ii) 'compound:lvc' (light verb construction).

4.2.1 Pomak: a nominative-accusative language

Subjects (dependency 'nsubj') are typically marked with the nominative case and objects (dependency 'obj') with the accusative, although some verbs select objects in the genitive case. Indirect objects ('iobj') are marked with the genitive/dative case, which is morphologically based on the Slavic dative case. Ethic datives are tagged with the dependency 'obl', e.g., **dečómne** *drago* ... 'the children like to ...'.

When the strong and the weak type of the personal pronoun cooccur, the strong type is assigned the dependency 'obl' (oblique) and the weak type the dependency 'expl' (expletive) (Figure 1).

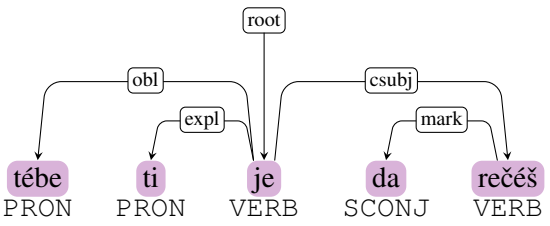

Figure 1: **tébe' ti** *je da rečéš* (literally: you to.you is to speak) 'it is up to you to speak'

The dependency 'expl:pass' is reserved for reflexive pronouns attached to transitive verbs as voice markers. Finally, the dependency 'expl:pv' is reserved for reflexive pronouns (*so, sa, se, si, su*) attached to verbs used as reflexives. In Pomak the dependency occurs with intransitive and certain transitive verbs (5).

(5) kopélkata **si** mýje rakýne
girl-the herself washes hands-the
'the girl washes her hands'

The dependency 'expl:impers' (expletive impersonal) is reserved for the reflexive pronoun (*só, sí, sé*) in impersonal constructions.

4.2.2 Compounds and fixed phrases

The dependency 'compound:redup' (reduplicated compounds) is used between pairs of identical words; in (6) reduplication serves emphasis purposes.

(6) adín sítan sítan dožd letæšo
a soft soft rain was raining
'a very soft rain was falling'

The dependency 'fixed' essentially assigns a flat structure to fixed (multiword) expressions that behave like function words or short adverbials (Figure 2, Figure 3).

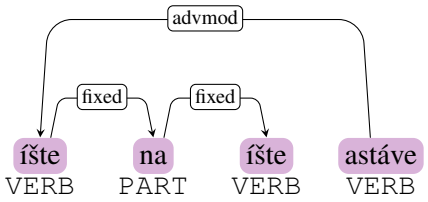

Figure 2: **íšte na íšte** *astáve gi faf kavenǿno* (literally: willing or not willing leaves them at café-the) 'willy-nilly he leaves them at the café'

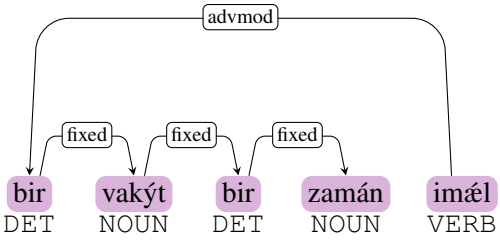

Figure 3: **bir vakýt bir zamán** *imæl je adín čülék* (literally: one time one era has been a man) 'once upon a time there was a man'

Summing up, Pomak uses the UD syntactic apparatus extensively and so far, no new dependencies have been defined.

5 Development of the Pomak treebank

The UD Pomak treebank is developed in two distinct steps: morphological annotation and syntactic annotation. Different methodologies have been adopted for each step because background knowledge about Pomak morphology was available (in Rodopsky) while the description of Pomak syntax is an ongoing process intertwined with the annotation of the Pomak corpus with UD syntactic dependencies. Naturally, morphological annotation preceded syntactic annotation so the two steps are discussed below in this order.

5.1 Morphological annotation

Rodopsky was transcribed into the K&K orthography, the CONLLU format was adopted and the original morphological annotation was mapped semi-automatically on the UD framework by one native speaker and two linguists, one of them expert in UDs and the other in Slavic languages and Pomak. The transcribed and annotated Rodopsky was mapped on 6350 sentences (86,700 words) selected from the Pomak corpus to form the gold annotated corpus. Although in the case of endangered languages often there is a shortage of annotators, we were able to employ a native speaker and a Greek linguist fluent in Pomak who edited the corpus with very good interannotation agreement kappa scores on 476 sentences (PoS tags 0.90, features 0.87, lemmas 0.93) (Karahóğa et al., 2022). The gold corpus (hereinafter: 'QPMcorpus') has been uploaded on the UD language repository and included in the UD treebanks on which the recent edition of the Stanza tool has been trained.[5]

The procedure of assigning morphological annotation to the Pomak gold corpus was designed to exploit the resource Rodopsky. Although non standardised/oral languages and dialects may not be endowed with such legacy, when it exists, it is valuable and should be exploited; in fact, several European non-standardised languages have some textual legacy (Gerstenberger et al., 2017; Bernhard et al., 2021). In the merits of the selected approach are (i) the development of the morphologically annotated gold corpus proceeded faster because the annotators only edited good quality morphological tags (ii) the use of dedicated resources mitigated the effect of imposing knowledge from other languages onto the documented one through shared training language models (Bird, 2022) (see also the discussion on syntactic annotation) (iii) it made room for the active participation of the community in the documentation process of their native language. On the processing front, the existence of an independently created relatively substantial morphologically annotated gold corpus allowed us to test various open-source NLP tools, namely

[5] https://github.com/stanfordnlp/stanza/blob/main/stanza/models/common/constant.py

spaCy v3.2.2[6] (Honnibal et al., 2020), Stanza[7] (Qi et al., 2020), UDify[8] (Kondratyuk and Straka, 2019) and UDPipe[9] (Straka et al., 2016) (for details see (Karahóǧa et al., 2022)) and select Stanza for its accuracy results in order to annotate our Pomak corpora.

A comment is due here: Like many open-source NLP tools (Nguyen et al., 2021), Stanza did not allow for the independent assignment and evaluation of morphological and syntactic annotation. Thus, an incremental corpus creation (active annotation) was not properly supported. Working with an unstudied language, like Pomak, in a project that targeted active corpus building, revealed that the morphological and syntactic annotation processes should be independent. Thus, we manipulated the Stanza code in order to separate the two annotation processes. We also reported the issue to the Stanza development team and, as a result, the updated Stanza version provides an approach for the required separate annotations.[10]

5.2 Syntactic annotation

In this section we describe the ongoing syntactic annotation of the QPMcorpus that will eventually yield the Pomak morphologically and syntactically annotated gold corpus (Pomak UD treebank).

5.2.1 Data, tools and methods

Drawing on our experience from morphological annotation, we use Stanza to support the syntactic annotation of the QPMcorpus. We have adopted the active annotation method (Settles, 2009; Anastasopoulos et al., 2018; Shi et al., 2021) because, contrary to Pomak morphology, there is no prior 'formal' approach to Pomak syntax. As a result, a formal description of the syntactic properties of Pomak is developed as the annotation of parts of the QPMcorpus advances. Active annotation, as it is shown schematically in Figure 4, unfolds in cycles where an initial model is trained on an available dataset, it is then applied on unseen data, its output is edited manually, the data on which the model is re-trained include the original material and the edited one and so on. This procedure only partially corresponds to the actual annotation procedure of a language for which prior knowledge is not available. This is because at each annotation cycle, the annotators' knowledge about the language increases and, possibly, the annotation guidelines are modified enforcing the editing of all the material used so far to train the model (and not only of the output of the previous cycle). We still hope that active annotation will minimize annotation workload but we intend to study this issue more systematically in the immediate future with more annotation cycles. Annotation is performed by a Greek linguist fluent in Pomak who is advised by native speakers, an expert in Slavic languages and Pomak and a computational linguist familiar with the UD framework. As opposed to morphology, in the case of syntactic annotation we were not able to employ more than one expert mainly because there were no background extensive studies of Pomak syntax.

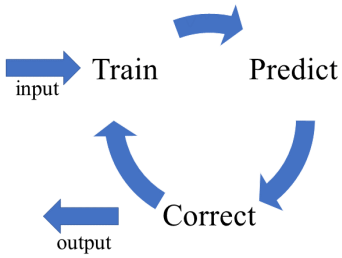

Figure 4: The active annotation procedure.

To better understand the effect of pre-existing knowledge from languages similar to Pomak on the model's performance, we created two corpora: (i) the "sl+po" corpus, comprising the QPMcorpus and annotated text retrieved from the UD treebanks of other languages in the South Slavic language group to which Pomak belongs, namely Bulgarian, Croatian, Serbian, and Slovene plus Slovac which is a West Slavic language whose alphabet is very similar to the Pomak one; treebanks in Cyrillic scripts were transliterated into Latin script with UROMAN[11] (ii) the "bg+po" corpus, consisting of the Bulgarian UD treebank only.[12] In order to have a more balanced dataset in terms of size, we copied the available syntactically annotated Pomak sentences as many times as needed to reach the same order of magnitude, namely about 4000 sentences.

In addition, we created word embeddings with the "sl+po" corpus, which is a superset of the

[6] https://spacy.io/.
[7] https://stanfordnlp.github.io/stanza/.
[8] https://github.com/Hyperparticle/udify.
[9] https://ufal.mff.cuni.cz/udpipe/1/models
[10] https://stanfordnlp.github.io/stanza/new_language.html.
[11] https://github.com/isi-nlp/uroman.
[12] https://universaldependencies.org/

"bg+po" corpus. We used UROMAN to transliterate into Latin those treebanks that employ scripts based on the Cyrillic alphabet.[13] With the fasttext2 tool (skipgram method) set to default parameters we created 16738 x 100 word embeddings. The 16738 words originate in the mixture of the "sl+po" and the original Pomak corpus (130000 words) from which the QPMcorpus was extracted.

5.2.2 Experiments

We followed two lines of experimentation:
1. *Train all processors* for both morphology and syntax, resulting in the following three models:

- base (only Pomak)
- bg + po (Bulgarian + Pomak)
- sl + po (5 Slavic languages + Pomak)

2. *Train for syntax only*; we loaded our best morphological model (indicated with the label "gm") that was trained on the QPMcorpus[14] (Karahóǧa et al., 2022). Results of this process were the following three models:

- base-gm (only Pomak)
- bg + po-gm (Bulgarian + Pomak)
- sl + po-gm (Slavic + Pomak)

Each training used a typical 80%—10%—10% data split for the training, validation and testing sets. In Table 2, the labels "**a**" and "**b**" indicate the manually annotated Pomak corpora used for the first and second active annotation cycles respectively. We report on the following metrics: Unlabeled Attachment Score (UAS), Labeled Attachment Score (LAS), Content-word Labeled Attachment Score (CLAS), Morphology-aware Labeled Attachment Score (MLAS) and Bi-LExical dependency Score (BLEX) (Zeman et al., 2018).

	Corpus	Train	Dev	Test
Sentences	**a**	184	16	16
Tokens		2033	178	208
Sentences	**b**	342	42	42
Tokens		3956	489	546

Table 2: Manually annotated Pomak corpora used in the two active annotation cycles.

We set as a baseline the UPOS, UAS and LAS values obtained with the first cycle of training on corpus **a** only, as reported in the first line of Table 3. In the second cycle we did not use the "sl+po" corpus, because in the first cycle it resulted in lower metric scores than those attained by the "bg" corpus (see Table 3). The results of the "sl" model suggest that it may be better to rely on models of few, or even one, very similar languages than models obtained from a branch of languages (including the branch to which the studied language belongs). However, our results do not suggest that Bulgarian is the language most similar to Pomak among the East South Slavic languages because we have not experimented with each one of the remaining languages in the "sl" model. Another reason for avoiding training on the "sl+po" corpus was the considerably long processing time required, due to its large size.

Model	UPOS (%)	UAS (%)	LAS (%)
base	84.62	73.56	58.65
base-gm	97.12	77.88	63.46
sl+po	87.50	75.48	64.42
sl+po-gm	97.12	79.81	68.27
bg+po	83.65	76.44	60.10
bg+po-gm	97.12	**82.69**	**69.23**

Table 3: UPOS, UAS, and LAS obtained with models trained and tested on corpus **a**.

The results of the two annotation cycles are summarised in Table 4 (in %), where boldface numbers denote the best model per task and per cycle. These results were obtained with a test set of 42 annotated sentences (extracted from corpus **b**).

Two annotation cycles with the same guidelines do not provide enough evidence for reliable conclusions. However, some interesting observations can be made:
1. Impact of gold morphology (model "gm") on metrics: syntactic predictions were improved considerably in all settings. This result supports our choice to exploit the resource Rodopsky and propose the modification of the process pipelines offered by the NLP tools.
2. Impact of increasing amounts of manually annotated data at the second annotation cycle (indicated with the "b" subscript):
2.1. As expected, the metrics of the models obtained from manually annotated Pomak data only ($base_a$, $base_b$) are improved as the annotated data increase in size (Anastasopoulos et al., 2018). However, one may notice that the best results in cycle **b**

[13] https://github.com/isi-nlp/uroman.
[14] In this approach, we attained a UD Part of Speech tags (UPOS) accuracy of 98.73% and a UD morphological features (UFEATS) accuracy of 95.23%.

Model	UPOS	UFeats	AllTags	Lemmas	UAS	LAS	CLAS	MLAS	BLEX
$base_a$	86.81	73.08	69.96	81.50	75.64	62.45	54.83	32.07	37.93
base-gm_a	98.72	98.35	97.25	99.27	83.70	71.61	65.07	59.59	64.73
sl+po_a	89.74	75.09	68.32	70.51	79.12	69.96	64.24	38.19	43.06
sl+po-gm_a	98.72	98.35	97.25	99.27	84.43	74.18	67.79	62.08	67.45
bg+po_a	85.53	71.98	67.77	80.59	73.99	61.72	53.04	30.41	39.86
bg+po-gm_a	**98.72**	**98.35**	**97.25**	**99.27**	**88.28**	**79.67**	**73.29**	**69.18**	**72.95**
$base_b$	91.03	81.50	78.57	86.63	81.50	70.88	62.80	45.39	51.19
base-gm_b	98.72	98.35	97.25	99.27	84.80	75.27	67.59	63.10	67.24
bg+po_b	91.58	84.25	80.40	86.26	82.05	71.61	66.43	47.20	55.59
bg+po-gm_b	**98.72**	**98.35**	**97.25**	**99.27**	**86.63**	**76.92**	**72.57**	**66.67**	**72.22**

Table 4: Metrics (%) of the two cycles (corpora **a** and **b**) obtained with the same test set, namely 42 sentences extracted from corpus b.

are a bit lower than those of cycle **a**, which seems counter-intuitive. It is our understanding that this is due to the instability of the learning process at the initial cycles, which deal with a limited number of samples (sentences) available for training. Nevertheless, the results are indicative and are expected to stabilize and improve in the next annotation cycles.

2.2. The difference between the scores obtained with models base-gm_b and bg+po-gm_b (for instance, for the UAS metric: 86.63%-84.80%=1.83%) is less than the respective difference between base-gm_a and bg+po-gm_a (for the UAS metric: 88.27%-83.70%=4.57%). This may be an encouraging development because it suggests that a point will be reached where a supportive language (here, Bulgarian) will not be necessary in few additional annotation cycles.

6 Discussion and Future work

We have presented the procedure we adopted to develop a UD treebank of Pomak, an endangered oral language of the East South Slavic group. The task is a case study of the project PHILOTIS and was supported by a group of computational linguists, linguists fluent in Slavic languages and Pomak and engineers as well as by the Pomak community.

Pomak exploited the UD inventory of labels and exposed unique linguistic phenomena regarding the system of Deixis and the verb system; modelling of Deixis led to the definition of a new UD morphological feature.

In this work we had the opportunity to apply two different annotation methods, one exploiting background knowledge (morphology) and one developing knowledge from scratch. The exploitation of background knowledge led to excellent accuracy scores with minimal annotation effort, however, few languages are endowed with the required resources. Therefore, an evaluation of the active annotation method that assumes no previous (morphological and/or syntactic) knowledge may be of more general interest. As the syntactic annotation of Pomak is still going on, a better understanding of the method, e.g., its impact on annotation time and costs, is among our immediate plans.

Acknowledgements

We acknowledge support of this work by the project "PHILOTIS: State-of-the-art technologies for the recording, analysis and documentation of living languages" (MIS 5047429), which is implemented under the "Action for the Support of Regional Excellence", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

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.

Delphine Bernhard, Anne-Laure Ligozat, Myriam Bras, Fanny Martin, Marianne Vergez-Couret, Pascale Erhart, Jean Sibille, Amalia Todirascu, Philippe Boula de Mareüil, and Dominique Huck. 2021. Collecting and annotating corpora for three under-resourced languages of France: Methodological issues. *Language Documentation & Conservation*, 15:316–357.

Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829, Dublin, Ireland. Association for Computational Linguistics.

Matthias Brenzinger, Akira Yamamoto, Noriko Aikawa, Dmitri Koundiouba, Anahit Minasyan, Arienne Dwyer, Colette Grinevald, Michael Krauss, Osahito Miyaoka, Osamu Sakiyama, et al. 2003. Language vitality and endangerment. *Paris: UNESCO Intangible Cultural Unit, Safeguarding Endangered Languages.*

Nicolaos Th. 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.

Ritván Jusúf Karahóǧa, Panagiotis G Krimpas G., Vivian Stamou, Vasileios Arampatzakis, Dimitrios Karamatskos, Vasileios Sevetlidis, Nicolaos Constantinides Th., Nikolaos Kokkas, George Pavlidis, and Stella Markantonatou. 2022. Morphologically annotated corpora of Pomak. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 179–186.

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 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.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A lightweight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90, Online. Association for Computational Linguistics.

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.

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Tianze Shi, Adrian Benton, Igor Malioutov, and Ozan Irsoy. 2021. Diversity-aware batch active learning for dependency parsing. *arXiv preprint arXiv:2104.13936*.

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).

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.