MORPHOLOGICALLY ANNOTATED CORPORA OF POMAK

ATHENA Research & Innovation Information Technologies

Ritván Jusúf Karahóğa Panagiotis Krimpas Vivian Stamou Vasileios Arampatzakis
Dimitrios Karamatskos Vasileios Sevetlidis Nikolaos Constantinides Nikolaos Kokkas George Pavlidis Stella Markantonatou

In a nutshell..

Philotis: Platform for the easy development and distribution of state-of-the-art spoken and textual annotated resources of living languages. Case studies: Greek, Pomak– an endangered, East South Slavic language.
Procedure: A group of native speakers, linguists, computational linguists and engineers took full advantage of Pomak legacy and state-of-the-art NLP technology.
Outputs: Morphologically annotated Pomak corpus openly available on Philotis; gold annotated corpus on the UD treebank repository.

About Pomak

Pomak is spoken in Bulgaria, Thrace/Greece, European part of Turkey and, in the places of Pomak diaspora. Factors of language vitality and endangerment for the Pomak language as of 2021 [2]:

- Most but not all children or families speak Pomak as their first language in specific social domains.
- Absolute number of speakers: 35,000 (estimated [1]).
- The dominant language begins to penetrate even home domains.
- The language is used only in a few new domains.
- Written materials are useful for some members of the community. The language is not a part of the school curriculum.

Developed corpora of Pomak: 130,000 words

Texts:

- in a variety of alphabets (Cyrillic, Greek, English-based Latin alphabet)
- sporadic transcriptions and recordings of Pomak folk songs and tales
- few modern journalistic texts and translations from Greek and English
- collected via a network of native speakers and Greek scholars
- copyright issues resolved according to the Greek law to ensure free distribution for research purposes
- semi-automatically homogenised with the K&K orthography

Text types	Words	Places in Thrace
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.

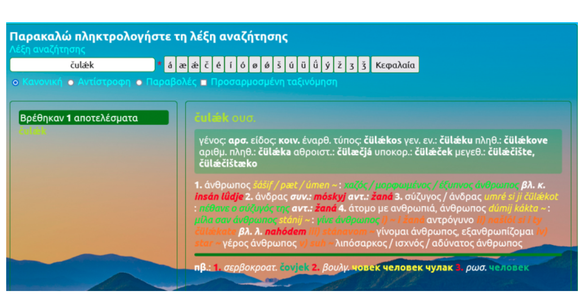

Figure 1: Rodopsky `https://www.rodopsky.gr/`: *čulæk* 'man' with morphological annotation in Greek. 61,500 lemmas, 3,5 x 10^6 unique forms (i.e., combinations of a lexical token and a PoS symbol) annotated for PoS and morphological features.

The orthography of Pomak

No widely acceptable alphabet or orthography of Pomak exists.
K&K orthography has been developed by native speakers and linguists and follows good practices [3]:

- *Phonetic transparency*. Sound to phoneme correspondence.
- *Systematic orthographies*. Spelling independent of pronunciation changes due to context: E.g., in *hlæb* 'bread (Nom|Sg)' [b] is devoiced [hlæp] in the final position but spelling is consistent with all other forms, e.g., *hlæbu* 'of/to (the) bread', *hlæba* 'bread (Acc|Sg) etc.
- *Easily discriminable symbols.*
- *Portability of the alphabet*. Encoded in Unicode.
- *Placement of word delimiters*: E.g., formations with a preposition/particle are spelled as two words for distributional reasons: *at kak* 'since' instead of *atkák* etc.
- *Dialectical issues*. As many dialects as possible are covered.

Gold morphologically annotated corpus

Rodopsky: Transfer to CONLLU format, map PoS and morphological features on Universal Dependencies (UD) framework `https://universaldependencies.org/`, map Rodopsky on corpus. Issues:

1. UD PoS not in Rhodopsky: DET(erminer) and X(other).
2. PoS re-assignment, e.g., participles: ADJ or VERB?
3. New morphological features for UD (i) Diminutives, augmentatives for nouns, adjectives and adverbs, (ii) Part(icle) Type for determiners and adverbs formed with *næ / nó, ní, sæ*; values "indicative", "negative" and "total".

6,350 sentences (86,700 words) of the Pomak corpus were manually edited by a native speaker and a linguist. IA kappa scores on 476 sentences: PoS tags 0.90, features 0.87, lemmas 0.93.
Merits: Good quality annotation, Less imposed knowledge from other languages through shared training language models, Active participation of the community.
Downside: Non generalisable procedure.

Morphological annotation of the entire corpus

Table 2: Accuracy scores for lemmatisation (LEMM), PoS tagging (UPOS) and morphological feature assignment (FEATS).

Parser	Model	LEMM	UPOS	FEATS
spaCy v3.2.2	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 v1.2	UDPipe	92.04	95.94	90.39

Table 3: Statistics on the training, development and test sets (80:10:10).

Corpus	Train	Dev	Test
Sentences	5,000	671	679
Tokens	67,345	9,736	9,701

Issues with the NLP tools:

- Outdated compilation and README instructions required missing files; we had to correct the code.
- Insufficiently documented alignment of (i) the processes, (ii) the addition of new languages.
- Few tools offer pretrained transformer models.
- Pretrained multi-language models are used by one tool only.
- Some NLP tools do not separate morphological from syntactic annotation assignment and evaluation; to obtain evaluations of the morphological annotation, we rewrote the code. Separation of morphological from syntactic annotation should be an option when:
  - good quality morphological annotation is otherwise possible
  - not equally mature morphological and syntactic analyses are available

References

[1] E. Adamou and D. Fanciullo. 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, 2018.

[2] M. Brenzinger, A. M. Dwyer, T. de Graaf, C. Grinevald, M. Krauss, O. Miyaoka, N. Ostler, O. Sakiyama, M. E. Villalón, A. Y. Yamamoto, and O. Z. U. A. H. E. G. on Endangered Languages). Language vitality and endangerment. Paris, 10-12, 03 2003.

[3] M. Cahill and E. V. Karan. Factors in designing effective orthographies for unwritten languages. In *SIL Electronic Working papers 2008-001*, 2008.

Acknowledgement

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).

European Union European Regional Development Fund
HELLENIC REPUBLIC MINISTRY OF DEVELOPMENT AND INVESTMENTS SPECIAL SECRETARIAT FOR ERDF & CF PROGRAMMES MANAGING AUTHORITY OF EPAnEK
EPAnEK 2014–2020 OPERATIONAL PROGRAMME COMPETITIVENESS ENTREPRENEURSHIP INNOVATION
ΕΣΠΑ 2014-2020 Partnership Agreement 2014 - 2020
Co-financed by Greece and the European Union