A Multilingual Parallel Corpus for Aromanian

Iulia Petrariu, Sergiu Nisioi

Human Language Technologies Research Center
Faculty of Mathematics and Computer Science
University of Bucharest
iuliapetrariu@gmail.com, sergiu.nisioi@unibuc.ro

Abstract

We report the creation of the first high-quality corpus of Aromanian, an endangered Romance language spoken in the Balkans, and the equivalent sentence-aligned translations into Romanian, English, and French. The corpus is publicly released at https://github.com/senisioi/aromanian and the Aromanian parts are published using several orthographic standards. Additionally, we provide a corpus-based analysis of Aromanian linguistic particularities and the overall demographic and political context that impacts the development of the language.

Keywords: Aromanian, parallel corpora, machine translation, low-resource, endangered languages

1. Introduction

Aromanian (ISO 639-3 - rup) is an Eastern Romance language that is part of the larger family of macro-Romanian languages consisting of Modern Standard Romanian and its regional varieties (ron), Megleno-Romanian (ruq), and Istro-Romanian (ruo) (Papahagi, 1974; Saramandu, 1984). Aromanian is an endangered[1] language that currently lacks corpora and electronic resources that can potentially contribute to the preservation of its cultural heritage. Several sources (Salminen, 2007; Moseley, 2005; Campbell et al., 2022) state that children are no longer learning the language as their mother tongue and the total number of speakers is estimated to be between 300,000 and 500,000 (Lewis, 2009; Atanasov, 2002; Nevaci, 2013). Official censuses from North Macedonia, Alabania, Romania, and Greece count less than 90,000 people, Kahl (2006) estimating this to be the total number of fluent speakers. Due to reasons of national identity, Romanian censuses count Aromanians as a subcategory of Romanians, and Greek censuses consider Aromanians as native Greeks that speak a Romance language. Both groups predominantly identify with their categories, and it is therefore difficult to obtain accurate statistics from the two countries where the largest Aromanian communities reside (Prentza and Kaltsa, 2020).

Working with dialectal languages comes with several challenges: (1) Aromanian has several varieties / dialects that have been influenced by the contact with Greek, Romanian, Turkish, Albanian and South Slavic languages (Caragiu-Marioțeanu, 1975). (2) The writing and vocabulary (beyond basic terms) have not been standardized in a widely accepted institutional manner and several types of spelling are currently in use (Caragiu-Marioțeanu, 1997; Cunia, 1997; Nevaci, 2008; Ballamaci, 2010), and (3) the language has historically been transmitted orally with usage within families or small communities (Gica et al., 2009; Maiden, 2016), public usage being limited to a few localities from Albania and North Macedonia.

With this work, we create the first corpus[2] of Aromanian and contribute to the efforts of preserving an endangered Romance language that is part of the Latin cultural heritage. More specifically, we release a high-quality corpus of written Aromanian texts consisting of short stories. Our data is based on the work of Candroveanu (1977) who collected an anthology of prose and traditional Aromanian tales together with the equivalent original translations into Romanian. We further enhance this data by providing annotations and parallel sentence alignments between Aromanian and translations in Romanian, English, and French.

Grammars and word inventories have been produced since the *XIX*th century (Boiagi, 1813; Capidan, 1932), while a more modern approach to phonology and structural morphology has been proposed in the seminal work of Caragiu-Marioțeanu (1968). Several recent linguistic studies cover verbs, morphology, syntax, and various aspects of the language (Bara et al., 2012; Todi and Nevaci, 2012; Maiden, 2016; Manzini and Savoia, 2018), but overall digital resources remain scarce. They consist mainly of multilingual word-aligned lists (Nisioi, 2014; Cristea and Dinu, 2020; Beniamine et al., 2020; Fourrier and Sagot, 2022) to study language evolution, language contact, historical linguistics, and the relationship to other Romance languages. The only corpus of written Aromanian is the roa-rup Wikipedia which is incidentally included in the lan-

[1] https://endangeredlanguages.com/lang/963

[2] Dataset publicly available at https://github.com/senisioi/aromanian

LREC-COLING 2024, pages 832–838
20-25 May, 2024.

guage identification dataset (Thoma, 2018). However, at the time of writing this work, Wikipedia quality is low, the orthography is inconsistent, the majority of articles contain one sentence or a stub, the total number of articles is relatively low (approximately 1300), and there is a proven small user engagement (Alshahrani et al., 2023).

To the best of our knowledge, there is no comparable previous work presenting a multilingual, sentence-aligned, high-quality corpus of Aromanian texts. Our paper's contributions cover roughly the following points:

1. we highlight a series of particular dialectal challenges of Aromanian and how they impact the development of language, its orthography, the availability of resources, and the preservation efforts of the language,
2. we release publicly the first high-quality corpus of Aromanian and make it accessible to users in several writing standards,
3. and last but not least, sentence-aligned translations into Romanian, English, and French are provided for future comparative studies, for training low-resource dialectal machine translation, or for building future NLP processing pipelines for Aromanian.

2. Population Statistics

Local populations identify their language by the following terms: Armâneaști, Armânească, Armâneșce, Rrămăneaști, Machiduneaști, Aromână, Vlașki, Vlach and their ethnicity by Armân, Aromân, Rămăn, Arumanian, Vlach, Arumun, Aromunian, Macedo-Romanian, Macedo-Rumanian, Țiți, Român, Grec, Macedonean. The main regions in the Balkans where populations of Aromaninans reside are Thessaly, Epirus, Central Albania, North Macedonia, Greek Macedonia, Bulgaria, Serbia, and Romania (Dobruja) (Lozovanu, 2008; Lewis, 2009). The only recent official censuses that record Aromanians are from Albania[3] and North Macedonia[4] where the language has an official status in local communities, but the total number does not exceed 18,000 members in total in both countries. The largest communities are in Greece, according to Kahl (2006), and the numbers are estimated at the order of hundreds of thousands. However, the last census of 1951, which numbered Aromanians, found only 40,000 people self-declaring as such[5]. The second largest community is likely to be in Romania, where Aromanians have been given land through colonization and Romanization of the Dobruja region, since the end of *XIX*th century to the beginning of the *XX*th century (Gica et al., 2009; Clark, 2015). The last Romanian census that included Aromanians was held in 2002 (Lozovanu, 2008) and numbered approximately 25,000 people.

Except for the notable exceptions of Albania and North Macedonia where specific localities have Aromanian designated as official language, the language is rarely spoken outside private homes and family contexts (Lewis, 2009), access to public schools is severely limited, and therefore many Aromanians lose language skills and become assimilated by the dominant linguistic group. Given this situation, here we reiterate the urgency to develop linguistic resources that can contribute to the preservation of the Aromanian cultural heritage.

3. Corpus Creation

Our primary objective is to create the first set of digital resources for Aromanian and thus to increase preservation efforts and to facilitate future comparative studies. The corpus is based on an anthology of prose gathered by Candroveanu (1977) which consists of original Aromanian tales and short prose together with equivalent translations into Romanian. The anthology exists only in printed form, and it has not been republished since 1977. The author traced its sources from texts published since the Grammar of Boiagi (1813), previous anthologies (Papahagi, 1922), and magazines up to 1948. In our data, we include 47 short stories that in total count to around 360 pages and a total size of approximately 75,000 words. Two short stories have been removed from the official release of the corpus as they contain highly racist stereotypes which could have a negative impact when training downstream models.

Caragiu-Marioțeanu (1997) classifies the Aromanian sub-dialects into two types: type F – the variants that resemble the Farsherot, predominantly found in Albania and some parts of Greece (mostly in Metsovo and Kastoria) and type A, which includes all the other variants which are more widely spread. She argues that type F sub-dialects are spoken by smaller communities, which have been influenced by type A. The dataset collected covers mainly type A sub-dialects with several texts being visibly influenced by standard Romanian lexicon and syntax.

[3] 2011 Albanian census https://www.instat.gov.al/

[4] 2021 North Macedonian census https://www.stat.gov.mk

[5] 1951 Greek census http://dlib.statistics.gr/Book/GRESYE_02_0101_00030.pdf

Corpus	DIARO	Cunia	Greek	IPA	Example	Example IPA
(1977)	(1997)	(1997)	(2005)			
î/â	î/â	ã	α/â	[ɨ]	cându / when	/kɨnˈdw/
ă	ă	ã	α/â	[ə]	tricură / passed by	/triˈkuˈrə/
l’	l’	lj	λ	[ʎ]	oclĭ / eyes	/ˈokʎ^j/
ñ	ń	nj	ν	[ɲ]	ńelu / lamb	/ˈɲnelw/
ș	ș	sh	σ	[ʃ]	aroşu / red	/aˈroʃw/
ț	ț	ts	τσ	[t͡s]	țeru / sky	/ˈt͡serw/
dz	ḑ	dz	τζ	[d͡z]	ḑinire / son-in-law	/d͡ziˈniˈre/
γ	y	gh	γ	[h]	yespi / wasps	/ˈɣesp^j/

Table 1: Several orthographic standardizations and variants currently in use. Other spellings may be encountered as a mixture of the above.

3.1. Orthography and Phonology

A challenge in collecting good quality data for Aromanian appears due to the multitude of standards that have been proposed (Nevaci, 2008) which can cover variations of the Latin or Greek alphabet. In our case, the original texts were published in Romania before official Aromanian standardizations were proposed, here denoted as *DIARO* after the Aromanian-Romanian Dictionary by Caragiu-Marioțeanu (1997) and *Cunia* after the author of the Aromanian dictionary (Cunia, 2010). In the original source of our corpus (Candroveanu, 1977), yet a different spelling was used based largely on the Romanian standard at that time with several particularities:

- the palatal lateral aproximant consonant [ʎ] in the text is represented as: [l̓]; Caragiu-Marioțeanu (1997) DIARO standard advocates for the same letter by adding a diacritic to the dental variant [l], similar to the Slovak alphabet; Cunia (1997) advocates for the usage of [lj] group, similar to the Romanization of Macedonian alphabet

- the post-alveolar affricate [d͡z] in the text is represented as: [dz]; Caragiu-Marioțeanu (1997) DIARO advocates for the letter [ḑ], a letter that has been historically in the Romanian alphabet during its transition from Cyrillic script; Cunia (2010) advocates for the letter group [dz]

- the palatal nasal [ɲ] which is represented in the text as [ñ], similar to Spanish; Caragiu-Marioțeanu (1997) proposes a diacritic over the letter dental counterpart [ń] while Cunia (2010) follows the Cyrillic transliteration rules to argue for the group [nj]

- the voiced velar non-sibilant fricative [ɣ] is uniquely represented in our texts as the Greek letter [γ]; the DIARO standard argues that loanwords from Greek have [ɣ] and proposes the usage of letter [y] for the sound while Cunia argues for using the group [gh]

- the closed central vowel [ɨ] is represented in the text with letter [î], the Romanian standard at that time, with the exception of proper nouns referring to Aromanians (armân); Caragiu-Marioțeanu (1997) DIARO follows the current standard in Romania to propose an alternation between letter [î] at the beginning / end of a word and [â] mid-word; (Cunia, 2010) advocates for a simpler standard where both closed central and mid-central vowels are represented by [ã] because these tend to interchange in sub-dialectal variants of Aromanian from different regions.

The standardization of Cunia (1997) follows the Boiagi (1813) grammar and the Romanized Macedonian script. One of the main arguments for using this orthography consisted in the *technical difficulties of publishing with characters using diacritical signs* (Cunia, 2010) with Microsoft Word in the early 2000s. This type of spelling has been criticized for the lack of linguistic and etymological grounding by Caragiu-Marioțeanu (1997) and Nevaci (2008), as the writing may omit diphtongs [e̯a] (/ˈkʎe̯aˈmə/ vs. c**lja**mã, English: *they call*) and it can also omit the glide [j] as a plural marker in nouns (e.g., /ˈokʎ^j vs. oc**lj**, English: *eyes*). Despite the criticism, the publication of an exhaustive dictionary using this standard (Cunia, 2010) and the low effort to type on any English keyboard made it a more popular choice in online electronic resources (including Wikipedia).

The DIARO standardization of Caragiu-Marioțeanu (1997), while being more linguistically rigorous, did not gain a similar wide-spread popularity in online materials, possibly because the DIARO dictionary was not complete and the published version covers words only to the letter *D*. Recent Aromanian learning text books still show variations, predominantly with respect to the preference of using [dz] instead of [ḑ] (Ballamaci, 2010) or the usage of [l’] with an apostrophe[6]

[6] See `https://roa-rup.wikipedia.org`, accessed in October, 2023.

corpus	#words	#types	ttr	sent
Aromanian	76,000	9,000	0.119	35.87
Romanian	84,000	8,700	0.104	39.66
English	85,000	5,400	0.063	40.42
French	86,000	7,800	0.091	40.56

Table 2: Statistics regarding the corpus size in words (values are rounded for readability). The original Aromanian texts have shorter average sentence length and higher lexical richness, as estimated based on type-token ratio.

instead of [ľ] with a caron diacritical mark.

Last but not least, various Greek spellings are also commonly used (Malabakēs, 2005; Thakis, 2020) in the Greek-speaking regions (Nevaci, 2008), which may occasionally include Latin letters for the central unrounded vowel as seen in a more detailed comparison in Table 1.

Given this situation, it is clear that not all spellings are mutually translatable without losing some degree of detail. We created several conversion codes and provide the corpus in all formats including the Greek script so that readers familiarized with a specific spelling can interact with it more easily. However, for natural language processing purposes, we recommend using the linguistically motivated orthography proposed by Caragiu-Marioțeanu (1997).

3.2. Processing the Data

The process of extracting parallel data is labor intensive and uses proficient manual annotators who split long Aromanian phrases into smaller sentences and align each sentence to the equivalent Romanian translation available in the original book (Candroveanu, 1977). We mention here that word-to-word translations between Aromanian and Romanian preserve the meaning of the original text, but sound un-natural to a Romanian reader, therefore the Candroveanu often chose to adapt the resulting texts to resemble the modern spoken Romanian. A recurring pattern is the translation of simple perfect (which is predominant in Aromanian) to the passé composé form in Romanian. A more particular example is the Aromanian: *ș-u băgă tu minte* [he put it in his mind] which can be literally translated to Romanian (își băgă în minte), but the translator chose *s-a gândit ce s-a gândit el* [he thought about it for a while].

Romanian is a medium-resourced language already covered in the current state-of-the-art proprietary DeepL machine translation system[7]. We use the proprietary DeepL MT system available through an API to translate each Romanian phrase into English and French. After this process, proficient bilingual readers verify the translations into English and French against the originals and manually introduce post-editing corrections to ensure the translation is correct and that it preserves the meaning and style of the original Aromanian phrase. Editors are advised to preserve the original sentence splitting and to focus on lexical adaptations.

We acknowledge the fact that semi-automatic translations via a pivot language such as Romanian might not give the same quality results as direct source-to-target translation, but we believe this process to be useful nonetheless as a first step. subsection 3.2 provides the basic statistics regarding the size of our corpus. Romanian and Aromanian tokenization has been done with a blank spaCy[8] model (Honnibal et al., 2020) while French and English with the corresponding large models.

Additionally, we have experimented with various methods of training neural machine translation systems between Romanian and Aromanian inspired by recent WMT Low-resource Shared Tasks (Przystupa and Abdul-Mageed, 2019). Our goal is to generate more content in the endangered language and not the other way around so we only thoroughly tested one translation direction. We report here that our attempts ended up in failure with BLEU scores ranging between 7 and 8 points maximum, but even more important, the output is rarely relevant to a native speaker. Our experiments include training from scratch a fariseq model with convolutional neural networks (Ott et al., 2019), training from scratch a Transformer MarianMT model (Junczys-Dowmunt et al., 2018), and fine-tuning gpt-3.5-turbo with prompts to translate from Romanian to Aromanian. The 8 point BLEU score was obtained by fine-tuning a French-Romanian pretrained transformer model released by HelsinkiNLP group on huggingface (Tiedemann and Thottingal, 2020) to translate French to Aromanian (initializing the Aromanian decoder with Romanian), however, the results are not substantial.

Aromanian is highly similar to the Romanian language, but there are also differences that we could observe at the corpus level which, in addition to orthography and phonology, makes transfer learning from Romanian a non-trivial task. Among the particularities we count the morphological system which has a stronger declination than the Romanian for the possessive pronoun, and the conjunctive. The conditional is synthetic, linked to the word, as compared to Romanian, where it is analytic and different morphemes are used. The plus-que-parfait indicative is analytic, whilst the one in Romanian is synthetic.

[7] https://www.deepl.com

[8] Using the latest spaCy version == 3.7.0.

4. Conclusions

We hope that our research will encourage future studies on Eastern Romance low-resource languages such as Aromanian, Istro-Romanian, or Megleno-Romanian. The corpus that we release together with the experimental framework can have a significant impact for machine translation and NLP researchers, for translation studies and historical linguistics, and most importantly, for the Aromanian community and language preservation, by raising awareness and potentially increasing young people's interest in Aromanian, and by enhancing the degree of access to electronic resources and information.

5. Acknowledgements

This work has been supported by POCIDIF project in Action 1.2 ”Romanian Hub for Artificial Intelligence”. We would like to thank Cristian Stanescu for his valuable work in building this corpus and the anonymous reviewers for their constructive comments.

6. Limitations and Ethical Considerations

Among the limitations of our paper we enumerate the overall small size of the corpus and the fiction genre, both bringing issues in cases where NLP applications and tools have to be developed, as they will not perform as well on news or texts that can have a broader reach to the community. Regarding the computational experiments, these are limited to incipient experiments that are constrained by the size of our data. Last but not least, the topics covered in the data might not be representative for how the language is used verbally in communication.

7. Bibliographical References

Ti Alkire and Carol Rosen. 2010. *Romance languages: A historical introduction*. Cambridge University Press.

Saied Alshahrani, Norah Alshahrani, and Jeanna Matthews. 2023. Depth+: An enhanced depth metric for wikipedia corpora quality. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 175–189.

Petar Atanasov. 2002. *Meglenorumaenisch. Lexikon der Sprachen des europaeischen Ostens*. Wieser.

Evangelos Bakalis and Alexandra Galani. 2012. Modeling language evolution: Aromanian, an endangered language in greece. *Physica A: Statistical Mechanics and its Applications*, 391(20):4963–4969.

Iancu Ballamaci. 2010. *Metoda aromână/vlahă*. Predania.

M. Bara, T. Kahl, and A.N. Sobolev, editors. 2012. *Die südaromunische Mundart von Turia (Pindos)*. Peter Lang.

Sacha Beniamine, Martin Maiden, and Erich Round. 2020. Opening the romance verbal inflection dataset 2.0: A cldf lexicon. In *12th Conference on Language Resources and Evaluation [postponed due to Corona]*, pages 3027–3035. European Language Resources Association (ELRA).

Victoria Bobicev, Cătălina Mărănduc, and Cenel Augusto Perez. 2017. Tools for building a corpus to study the historical and geographical variation of the Romanian language. In *Proceedings of the First Workshop on Language technology for Digital Humanities in Central and (South-)Eastern Europe*, pages 10–19, Varna. INCOMA Inc.

Mihail G Boiagi. 1813. *Grammatikī romanikī ītoi macedonovlachikī [Romance and Macedono-Vlach Grammar]*. Tipografia Nationale a lui Stefanu Rasidescu [Stefanu Rasidescu National Tipography].

Lyle Campbell, Nala Huiying Lee, Eve Okura, Sean Simpson, and Kaori Ueki. 2022. The catalogue of endangered languages (elcat). Database available at http://endangeredlanguages.com/userquery/download/, accessed 2022-08-28.

Hristu Candroveanu, editor. 1977. *Antologie de Proza Aromână [Aromanian Prose Anthology]*. Editura Univers.

Theodor Capidan. 1932. *Aromânii. Dialectul aromân. Studiu lingvistic [Aromanians. The Aromanian Dialect. Linguistic Study]*. Monitorul Oficial și Imprimeria Statului [Official Journal and National Printing Company].

Matilda Caragiu-Marioțeanu. 1968. *Fonomorfologia aromână [Aromanian Phonomorphology]*. Editura Academiei Republicii Socialiste România [The Press of the Academi of the Socialist Republic of Romania].

Matilda Caragiu-Marioțeanu. 1975. *Compendiu de dialectologie română [Compendium of Romanian Dialectology]*. Editura Științifică și Enciclopedică [Scientific and Encyclopedic Press].

Matilda Caragiu-Marioțeanu. 1997. *Dicționar aromân (macedo-vlah) DIARO. A-D. Comparativ (român literar - aromân) [Aromanian (Macedo-Vlach) Dictionary. Letters A-D. Comparative dictionary literary romanian - aromanian]*. Editura Enciclopedică [Encyclopedic Press].

Matilda Caragiu Marioțeanu and Nicolae Saramandu. 205. *Manual de aromână. Carti trâ învițari armâneaști [Aromanian Manual. Book for Learning Aromanian]*. Editura Academiei Române [Romanian AcademyPress].

Ioana Chițoran. 2002. *The phonology of Romanian: A constraint-based approach*, volume 56. Walter de Gruyter.

Roland Clark. 2015. Claiming ethnic privilege: Aromanian immigrants and romanian fascist politics. *Contemporary European History*, 24(1):37–58.

Alina Maria Cristea and Liviu P. Dinu. 2016. A computational perspective on the Romanian dialects. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3281–3285, Portorož, Slovenia. European Language Resources Association (ELRA).

Alina Maria Cristea and Liviu P. Dinu. 2020. Automatic Identification and Production of Related Words for Historical Linguistics. *Computational Linguistics*, 45(4):667–704.

Tiberiu Cunia. 1997. On the standardization of the aromanian system of writing.

Tiberiu Cunia. 2010. *Dictsiunar a limbãljei armãneascã [Dictionary of Aromanian Language]*. Editura Cartea Aromãnã [Aromanian Book Press].

Raj Dabre, Atsushi Fujita, and Chenhui Chu. 2019. Exploiting multilingualism through multistage fine-tuning for low-resource neural machine translation. 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 1410–1416, Hong Kong, China. Association for Computational Linguistics.

Clémentine Fourrier and Benoît Sagot. 2022. Probing multilingual cognate prediction models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3786–3801.

Alexandru Gica et al. 2009. The recent history of the aromanians from romania. *New Europe College Yearbook*, (09):173–200.

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

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Thede Kahl. 2006. *Istoria aromânilor [The History of Aromanians]*. Tritonic.

M.P. Lewis, editor. 2009. *Ethnologue: Languages of the World, Sixteenth edition.* SIL International.

D Lozovanu. 2008. *Populația românească din Peninsula Balcanică. Studiu uman geografic*. Universitatea “Alexandru Ioan Cuza”, Facultatea de Geografie si Geologie.

Martin Maiden. 2016. Romanian, istro-romanian, megleno-romanian, and aromanian. In *The Oxford guide to the Romance languages*, pages 91–125. Oxford University Press.

Nikos Malabakēs. 2005. *Lexiko hellēno-blachiko: etymologiko lexiko. Λεξικό ελληνο-βλάχικο: ετυμολογικό λεξικό*. Pelekanos.

M Rita Manzini and Leonardo Savoia. 2018. *The morphosyntax of Albanian and Aromanian varieties: case, agreement, complementation*, volume 133. Walter de Gruyter GmbH & Co KG.

C. Moseley, editor. 2005. *Encyclopedia of the world's endangered languages*. Routledge.

M Nevaci. 2013. *Identitate romaneasca in context balcanic [Romanian Identity in the Balcanic Context]*. Editura Muzeului National al Literaturii Romane [The Press of National Museum of Romanian Literature].

Manuela Nevaci. 2008. Sisteme de scriere utilizate în limba publicațiilor aromânești actuale.

Sergiu Nisioi. 2014. On the syllabic structures of aromanian. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, LaTeCH@EACL 2014, April 26, 2014, Gothenburg, Sweden*, pages 110–118. The Association for Computer Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Tache Papahagi. 1974. *Dictionarul dialectului aromân. General și etimologic. [The Dictionary of Aromanian Dialect. General and Etymological]*. Editura Academiei Române [Romanian Academy Press].

Take Papahagi. 1922. *Antologie aromânească [Aromanian Anthology]*. Tipografia Romania Noua.

Alexandra Prentza and Maria Kaltsa. 2020. Linguistic profiling of heritage speakers of an endangered language: The case of vlach aromanian-greek bilinguals. *OPEN LINGUISTICS*, 6(1):626–641.

Michael Przystupa and Muhammad Abdul-Mageed. 2019. Neural machine translation of low-resource and similar languages with backtranslation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 224–235, Florence, Italy. Association for Computational Linguistics.

Tapani Salminen. 2007. Language diversity endangered. *Trends in Linguistics: Studies and Monographs 181*, pages 205–232.

Nicolae Saramandu. 1984. *Tratat de dialectologie românească [Romanian Dialectology Treaty]*. Editura scrisul românesc [Romanian Writing Publisher].

Leonardo M Savoia, Benedetta Baldi, M Rita Manzini, et al. 2020. Prepositions in aromanian. *Studii si Cercetari Lingvistice*, 71:149–160.

Thomas Thakis. 2020. *Μαθα νουμε τη Βλ χικη γλ σσα [We are learning the Vlach language]*. vlahika.gr.

Martin Thoma. 2018. The wili benchmark dataset for written language identification. *arXiv preprint arXiv:1801.07779*.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Aida Todi and Manuela Nevaci. 2012. Conjugation changes in the evolution of romanian (dacoromanian and aromanian) in verbs of latin origin.