NLP for Arbëresh: How an Endangered Language Learns to Write in the 21st Century

Giulio Cusenza, Çağrı Çöltekin
University of Tübingen
giuliocusenza@gmail.com, ccoltekin@sfs.uni-tuebingen.de

Abstract

Societies are becoming more and more connected, and minority languages often find themselves helpless against the advent of the digital age, with their speakers having to regularly turn to other languages for written communication. This work introduces the case of Arbëresh, a southern Italian language related to Albanian. It presents the very first machine-readable Arbëresh data, collected through a web campaign, and describes a set of tools developed to enable the Arbëresh people to learn how to write their language, including a spellchecker, a conjugator, a numeral generator, and an interactive platform to learn Arbëresh spelling. A comprehensive web application was set up to make these tools available to the public, as well as to collect further data through them. This method can be replicated to help revive other minority languages in a situation similar to Arbëresh's. The main challenges of the process were the extremely low-resource setting and the variability of Arbëresh dialects.

Keywords: Extremely low-resource language, Data gathering, Spellchecking, Automatic inflection, Arbëresh

1. Introduction

With the recent shift in communication from the oral dimension to digital media, many minority languages suffer from their speakers' inability to write. This ultimately leads to vocabulary loss and overall language decline (**?**). One of these languages is Arbëresh [aɾb’ɾeʃ] (**?**), on which this study centers its focus. This work explores the development of straightforward and easily accessible tools that may enable speakers of linguistic minorities to learn how to write in their native language.

Arbëresh is spoken in southern Italy and related to Tosk, the group of southern Albanian dialects (**?**). The Arbëresh people are the descendants of Albanian refugees that settled in Italy between the 14th and the 18th centuries as the Ottoman Turks conquered the Balkans. Although Arbëresh dialects exhibit loanwords from languages such as Italian, Sicilian, Neapolitan, or other, varying by region, they are often regarded as a conservative version of nowadays Albanian, untouched by Turkish influence. Arbëresh morphology is rather complex: nouns and adjectives inflect for number, gender, case, and definiteness, while verbs inflect for person, number, mood, tense, and voice. It is hard to establish how many Arbëresh speakers are there today, **?** reported an estimation of roughly 80.000 speakers.

The presented work produced the very first machine-readable data of contemporary Arbëresh (*Corpus Arbëresh*), as well as a spellchecker, a conjugator, a numeral generator, and a web application (*Arbor*) to deliver these tools to the public along with interactive spelling lessons. The app can be used by individuals interested in writing in Arbëresh, or employed by experts in educational contexts. It will also be a source of further data coming from the use the tools. This paper traces a strategy that may be applied to other minority languages to foster revitalisation, from the data gathering process to the deployment of the tools. More specifically, the adaptations to the edit-distance based spellchecker may prove applicable to other situations in which speakers' attempts at writing are influenced by the spelling standards of a majority language. The main challenges of such process are represented by the extremely low-resource setting and the variability typical of minority languages, which hinder standardisation.

This work was possible thanks to the first author's knowledge of the language as son of an active speaker. This eased communication with the community of Piana degli Albanesi, an Arbëresh town in Sicily, whose institutions and local businesses were so kind to promote the initiative through social networks and flyers.

2. Background

For centuries, the Arbëresh people managed to preserve their traditions and language with limited influence. More recently, Arbëresh has experienced a substantial decline in vocabulary with each generation, and is nowadays used in speech alongside Italian and southern Italian languages (**??**), through different mechanisms of "linguistic fusion" (**?**). The main causes of this decline may be traced to "the introduction of Italian into all layers of society, the massive spread of secondary education, of media and all modern means of communication" (**?**), as well as demographic shift (**?**). In some towns, Arbëresh has completely disappeared, while in others it has managed to survive among today's youth (**?**). Arbëresh dialects exhibit rather high mutual intelligibility, with the main differences appearing in phonological phenomena and borrowed vocabulary. These aspects make it challenging to establish an Arbëresh *koine*. Despite this, Arbëresh shares a standard "phonemic" alphabet with Albanian, designed during the Congress of Manastir in 1908.

A common tendency among those working toward a revival of Arbëresh is to refer to old Arbëresh or Albanian, avoiding most Romance loans. The project of an ideal Arbëresh, distant in time and space from contemporary spoken language, is ambitious, but the utility thereof can be disputed. In the work presented here, this prescriptive approach was relaxed, and resources were directed toward distinguishing between morphologically integrated loanwords and code-switching cases, with no stigma attached to Romance loans.

3. Resources

Arbëresh has a long literary tradition, including one of the oldest texts in an Albanian language.[1] Literary works consist mainly of ecclesiastic and folkloric texts, vastly unintelligible to today's Arbëresh speakers, as they include vocabulary that has been lost or that the average person probably never used, such as Greek loans and *hapax legomena*.

More recently, dictionaries, grammars, textbooks, and dramas have also been published with more accessible language and in the standardised alphabet, including the following resources, which were essential for the accomplishment of this work: *Fjalor* (**?**), a rich and thorough Arbëresh-Italian dictionary; *Gramatikë Arbëreshe* (**?**), a grammar aiming to describe all Arbëresh dialects; *Udha e mbarë!* (**?**), a comprehensive Arbëresh textbook; *Fjalori Arbërisht-Italisht i Horës së Arbëreshëvet* (**?**), a short dictionary based on the dialect of Piana degli Albanesi; *Grammatica della parlata arbëreshe di Piana degli Albanesi* (**?**), a grammar on Piana degli Albanesi's dialect; Papàs Gjergji Schirò's unpublished Arbëresh translation of the christian Gospel, which helped mainly with the consultation of optative verb forms.

4. Corpus Arbëresh

4.1. Data Gathering

A data gathering web page was promoted among Arbëresh communities. The need for it was determined by the absence of data on contemporary, everyday Arbëresh, and more generally of digital Arbëresh data: as Arbëresh speakers do not write, those who constructed dictionaries and grammar books had to refer to more or less dated literary works, which fail to correctly represent modern language. *Corpus Arbëresh* appears thus to be the first machine-readable data of contemporary Arbëresh.

The web page (in Italian) includes a text field prompting the insertion of everyday sentences, a field to select a hometown, on-screen keys for non-ASCII characters, a submit button, an option for daily reminders (browser push notifications), an introductory video, and some instructions. Contributors were told not to worry about correct spelling and loanwords. Speakers were made aware of the web page through social media and a flyer campaign. Flyers included a QR code and prompts to incentivise natural data ("Donate the last sentence you uttered in Arbëresh"), as well as different themes and registers ("Donate an Arbëresh sentence you used as a child"). Currently, over 1300 sentences have been donated with 5.72 words per sentence and at least over 70 contributors estimated through anonymised web cookies. The vast majority of the sentences (about 1150) are from the town of Piana degli Albanesi, where promotion was most successful; further action should target Arbëresh communities in other Italian regions. These data should not be considered authentic speech: the main goal of this setup was to quickly collect as many sentences as possible to allow for the development of character-level tools.

4.2. Data Standardisation

Most contributors did not know standard Arbëresh orthography. Each developed a strategy based on a mix of Italian and Arbëresh-looking spelling rules; therefore, standardisation was a necessary step. Actually, there is no solid standard for Arbëresh writing. Current Arbëresh authors make use of the unified alphabet (Section 2), but differ in their exact choices for specific words (also due to dialectal variations). However, these appear to be marginal differences, so a general standardisation was nevertheless carried out referring to the resources mentioned in Section 3. Dialectal variations were, in some cases, rewritten to a single word form when similar enough or easily inferrable from the phonological environment (*bunj* $\rightarrow$ *bënj*), while in others they were kept separate (*hëngra* and *hëndra*). So far, no strategy to deal with code-switching was designed, and sentences presenting code-switching cases were skipped.

Currently, 475 sentences have been standardised and used for the current version of the tools. The data is available under the name of *Corpus Arbëresh* in CSV format with the following fields: id, raw sentence, revised sentence, town, and year.

[1] Luca Matranga, *E Mbësuame e Krështerë*, "Christian Doctrine" (1592).

The raw sentences are provided to reflect the standardisation decisions that were made, but they would also be useful to anybody else interested in developing spellchecking tools for the language. Currently, there are no aligned Italian or English translations available, although this task is certainly slated for future annotation.

5. Inflectors

With the goal of building a vocabulary-based spellchecker and Arbëresh being a morphologically complex language, it became apparent that the donated data was not enough to cover a sufficient portion of Arbëresh vocabulary. To facilitate the inclusion of all these forms in the spellchecker's vocabulary, two rule-based inflectors were set up: a conjugator and a numeral generator. Noun inflection was not yet undertaken.

5.1. Conjugator

The conjugator was developed according to the resources mentioned in Section 3. Intricate rules account for the variability of Arbëresh verbs. Any regular verb can be automatically conjugated, provided the following data: lemma, conjugation class (1st, 2nd, 3rd), transitivity (transitive, intransitive, reflexive), present root, imperfect root, simple perfect 1st person singular, imperative 2nd person singular, participle, reflexive root. The imperfect and reflexive roots need to be specified only if different from the present root, mainly to account for apophony. For regular verbs, it is sufficient to provide simple forms, as the compound ones are always regular.

The resources from Section 3 name past forms using names from Italian traditional grammar, but these names fail to correctly reflect tense and aspectual features. This work substitutes these names with more fitting ones, inspired by Spanish grammars (e.g., "remote past" → "simple perfect").

5.2. Numeral Generator

A program was designed to generate a dictionary for numbers up to 999. A separate function uses this dictionary to convert integers into words, forming higher-order numbers with the terms for "thousand", "million", "billion", etc. The process was applied to both cardinal and ordinal numerals.

6. Spellchecker

6.1. Machine Learning Experiments

Different versions of an encoder-decoder model with Bahdanau attention (**?**) based on bidirectional *Gated Recurrent Units* (**?**) were trained on 1831 raw-to-revised word pairs. Given a misspelled word, the model was tasked to generate its correction.

Because of the scarcity of the data and poor results during evaluation (Section 6.3), and because the model often generated non-words – which would harm more than help the final users –, efforts were directed toward the development of an edit-distance algorithm.

6.2. Edit-distance Algorithm

The algorithm includes three edit operations: deletion, insertion, and substitution or copy. The cost of each operation is determined by edit weights extracted from the data. To account for the highly frequent misspelling of the bigram "nj" as "gn", a preprocessing step substitutes all occurrences of "gn" in the misspelled word with "nj" ("gn" is not a possible bigram in Arbëresh, so there is no risk of spoiling the input).

It is important to mention that this is not a usual spellchecking scenario. Users are not making occasional typos: they are attempting to write under the influence of other spelling standards. Therefore, a Levenshtein distance algorithm – albeit weighted – will have the problem of being biased toward fewer edits, although in some cases a couple more operations might be needed to map between two words (e.g., "c" → "çë", [t͡ʃə]). To address this, it is possible to normalise the weighted edit distance by the number of edits, thus obtaining the average edit cost. This method also proved itself problematic, as misspelled words can get mapped to much longer or shorter correction candidates. A better formula would thus be somewhat sensible to word length, while still allowing for light-edit candidates to close the gap with few-edit candidates. This can be achieved by taking the logarithm of the number of edits. The following score function was hence designed (a lower score corresponding to a better candidate):

$$score(c, m) = \frac{WD(c, m)}{1 + \log\left(D(c, m) + 1\right)}$$

where c is the correction candidate word, m is the misspelled word, WD is the function giving the weighted distance between them, and D is classic Levenshtein distance. D was chosen over the number of edits in the weighted distance because it gives a further advantage to words that undergo fewer edits in the weighted version of the function compared to the unweighted one. The function is adjusted to avoid division by zero and logarithm of zero. In the case of candidates with the same score, the system picks the one with higher frequency in *Corpus Arbëresh* (Section 4).

6.3. Evaluation

The systems were evaluated on 304 unique misspelled-correct word pairs (none of them were out-of-vocabulary words). Each system predicted a correction candidate for each misspelled word. For the systems based on edit-distance, the closest candidate was taken as their prediction. The metric used was the percentage of correct predictions over all words. The results are reported in Table 1.

system	score
baseline (Levenshtein dist.)	57.2%
encoder-decoder model	26.0%
weighted Lev. dist.	65.1%
score function	66.1%

Table 1: Evaluation results

The final tool will present the user with various correction candidates. Therefore, to gain a more comprehensive insight into the performance of different systems, the number k of top correction candidates considered can be increased. In other words, if the expected word lies within the top k candidates, the system is deemed successful. Figure 1 illustrates how the score function's performance increases rapidly with very low k, slowing down as k becomes bigger. Conversely, the performance of the system using weighted Levenshtein distance rises rather constantly after $k = 2$. This highlights the score function's proficiency in ranking correct candidates higher, a significant advantage not readily discernible from Table 1. Moreover, while the impact of the score function might appear marginal at first glance, it crucially influences the outcome for some of the most frequent Arbëresh words, hence noticeably affecting the perceived quality of the tool.

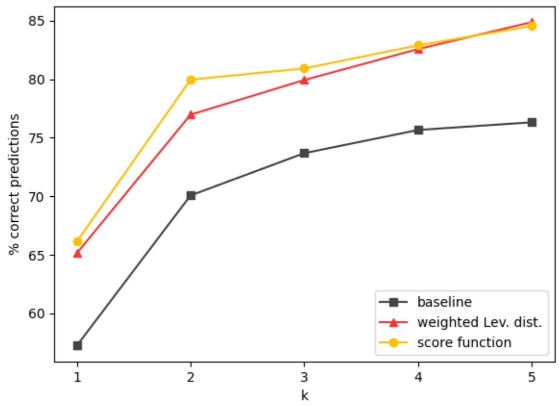


Figure 1: Performance of the different systems with increasing k (number of top-ranked correction candidates considered)

As our encoder-decoder model provides only one correction, it was excluded from this analysis. Despite its poor performance, further exploration of neural approaches is still worthwhile: out-of-vocabulary words represent a challenge for solutions based on edit-distance, while a more successful generative model should be able to generalize and deal with them accordingly.

6.4. The Vocabulary

The quality of such a system is ultimately strictly tied to its look-up vocabulary. The current vocabulary consists of 2892 word types coming from four sources. Table 2 shows how many word types each source provides.

source	n types
Corpus Arbëresh	638
Conjugator	1710
Numeral generator	347
(**?**)	437

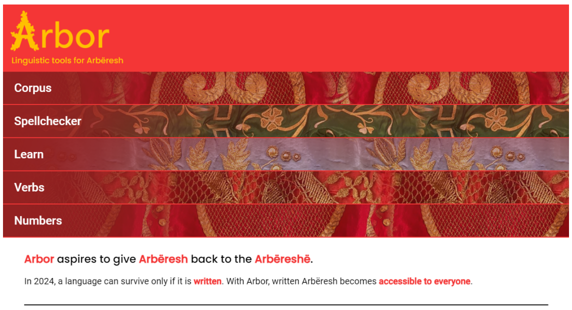
Table 2: Sources of vocabulary word types

Corpus Arbëresh is the most valuable resource, being the best reflection of everyday speech. The inflectors are able to generate hundreds of word forms, but most of them are seldom used. Finally, **?** includes some texts from which it was possible to extract words, but this resource might be dropped in future versions as it also contains a few “artificial” Albanian loans, normally not used in Arbëresh. A future version would ideally be paired with a loanword detection system to avoid the mapping of loans onto Arbëresh words.

7. Arbor

A web application by the name of *Arbor* was set up to deliver the tools to Arbëresh communities. The name was inspired by the Latin word for “tree” (*arbor*), because of its phonetic resemblance to the word “Arbëresh” and because of its symbolic meaning of community, tradition, as well as language structure. *Arbor* includes:

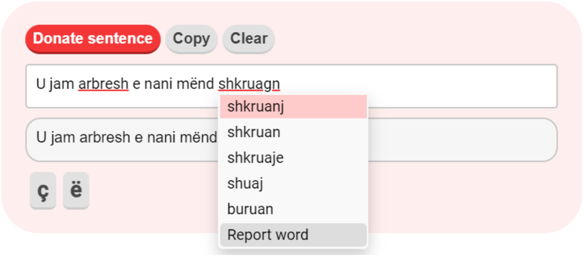
- A home page (Figure 2) with navigation buttons, a motto, an introductory video, a share button, and a news section.
- A page dedicated to *Corpus Arbëresh*, where it is possible to donate further sentences and read how they can be used for the development of the tools.
- An interface for the spellchecker (Figure 3), where each out-of-vocabulary word is underlined in red. The top five correction candidates are suggested for each misspelled word; alternatively, users can report the word as missing from the vocabulary. Users can also decide to donate the sentences to Corpus Arbëresh.



Arbor aspires to give Arbëresh back to the Arbëreshë.

In 2024, a language can survive only if it is written. With Arbor, written Arbëresh becomes accessible to everyone.

Figure 2: Top part of *Arbor*'s home page



U jam arbresh e nani mënd shkruagn

U jam arbresh e nani mënd

Figure 3: Interface of the spellchecker

- A page with interactive spelling lessons, inspired by the *Duolingo* language learning application.[2]
- A portal to consult verb conjugations (generated by the conjugator).
- An interface for the numeral generator.

It also provides a feedback module, contact options, and instructions for those who would be interested in collaborating.

8. Discussion

One week after its launch, *Arbor* had been visited by over 260 different users, with the home page viewed 697 times. Promotion so far has been conducted mainly on social media (Facebook) and through a few blogs that wrote articles about it, but it was effective only in Sicilian communities. Further promotion is currently being planned for communities in other Italian regions.

Ideas for future development of the platform include the improvement of the tools through newly collected data, collaboration with schools and local administrations, as well as the creation of a forum for Arbëresh speakers from different regions to ask questions and get in contact.

If *Arbor* will be used extensively by different Arbëresh communities, it will significantly facilitate the efforts to standardise the language and identify an Arbëresh *koine*, allowing for digital bridges between otherwise isolated communities and ease revitalisation. Such a scenario, albeit hard to achieve, was the main inspiration of this work, with the hope that positive results will further inspire other projects aiming at language revitalisation.

9. Material

Arbor available at: aarbor.web.app. *Corpus Arbëresh* data available at: aarbor.web.app/corpus/CorpusArbëresh.csv.

10. Bibliographical References

[2] www.duolingo.com