Building a Community-Grounded Spell Checker for the Kabyle Language Using NLP Powered by AI

Massil Aoudj[1], Tahar Boukhenoufa[1], Samia Bouzefrane[1], Yulliwas Ameur[1], and Kamal Naït-Zerrad[2]

[1]*CEDRIC Lab, Cnam, Paris, France,* {massil.aoudj, tahar.boukhenoufa, samia.bouzefrane, yulliwas.ameur}@lecnam.net

[2]*LACNAD (EA 4092), INALCO, Paris, France,* kamal.nait-zerrad@inalco.fr

Keywords: Kabyle (Taqbaylit), Amazigh/Berber, spell checking, low-resource languages, community-based NLP, language revitalisation

Context

Using innovative technologies today is an effective solution for the revitalisation of endangered and/or under-resourced languages. Some of these languages, although they have a significant number of speakers, suffer from a limited presence in the digital sphere. This is the case with Kabyle, one of the Berber languages spoken in North Africa and within the diaspora, for which digital resources and tools remain limited and inadequate.

Problem Statement and Objectives

To improve the use of the Kabyle language in digital tools and AI, large data corpora are required. However, such data are often not standardised and require linguistic correction. To address this crucial issue, we have undertaken a project to develop an automatic corrector using natural language processing techniques that follow the linguistic norms taught at INALCO, capable of processing large volumes of textual data. Beyond corpus cleaning, the corrector is designed to be integrated into everyday writing tools such as keyboards and word processors, so that it directly benefits Kabyle writers, teachers, and learners.

State of the Art

Automatic spelling correction is usually organised into three types of methods [1]: rule-based models, distance-based models, and context-based models that learn error patterns from data. Key elements include the edit distance [2,3], phonetic keys like Soundex, and n-gram context models. For Amazigh, Chaabi and Ataa Allah [4] combined the Damerau-Levenshtein distance with n-grams, but their system focuses on Moroccan Standard Amazigh in Tifinagh and does not model context-dependent morphosyntactic constraints such as the state of annexation. Community Hunspell dictionaries for Kabyle also exist as Firefox and LibreOffice extensions, but they perform context-free lexical lookup and encode no grammatical knowledge. Kabyle resources themselves remain scarce and are mostly community-driven, as with Mozilla Common Voice [5]. Our work builds on these methods while adding an explicit encoding of the state of annexation and a community-in-the-loop design.

Method and Experiments

Our approach is based on several modules. First, we use deterministic methods to obtain a strong base, combining the Levenshtein distance with a Soundex system adapted to Kabyle phonology, which maps a misspelled word onto the closest forms in our dictionary. In the second step, we train n-gram models on valid corpora provided by the community, in order to capture local context. Finally, we apply linguistic engineering that encodes the Kabyle state of annexation [6, 7], so that the proposed form is grammatically correct for the position of the word in the sentence. To validate

and improve the system, we keep the user in control by exposing the five closest correct forms. This approach lets us reuse the users' choices as a statistical signal for improvement and keeps the community involved in the project. All resulting resources will be shared in line with the FAIR [8] and CARE [9] principles, keeping the Kabyle community in control of its data.

Results

We evaluated the tool with standard spell-checking metrics, complemented by a noise-injection protocol that corrupts words already present in our dictionaries and then measures their recovery. The first results show:

- Sentence exact match: 75%
- Word error rate: 15.9% to 10.1% (relative reduction: 36%)
- Detection: precision 61.5%, recall 88.9%, F1 72.7%
- Correction: precision 53.8%, recall 77.8%, F1 63.6%, F0.5 57.4%
- Correct word among the top-5 candidates: 90% (isolated), 95% (in context)
- Top-5 when the automatic choice is wrong: 87% (isolated), 93% (in context)

These figures are based on synthetic noise; we are now building a test set of real errors taken from OCR and ASR outputs, in order to evaluate the system in the conditions that motivate it.

Future Work

Our next goal is to capture wider context in order to move from a static correction to a semantic one. We plan to build a semantic dictionary based on WordNet [10], specifically its free derivative WOLF [11], using French–Kabyle and English–Kabyle translation dictionaries, together with a root dictionary that groups all words sharing the same root. The ultimate step is to enable an LLM to understand the meaning and generate high-accuracy text corrections. To reach this goal, we intend to involve students of the Master's programme in Berber languages at INALCO.

AI-Use Statement

Claude (Anthropic, Opus 4.8) was used to draft the State of the Art section and isolated sentences elsewhere, for English language editing, and for LaTeX typesetting, including bibliography formatting. The research itself (the system, its design, and all experiments and results) was conceived and carried out by the authors.

References

[1] Daniel Hládek, Ján Staš, and Matúš Pleva. Survey of automatic spelling correction. *Electronics*, 9(10):1670, 2020.

[2] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[3] Fred J. Damerau. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176, 1964.

[4] Youness Chaabi and Fadoua Ataa Allah. Amazigh spell checker using Damerau-Levenshtein algorithm and N-gram. *Journal of King Saud University - Computer and Information Sciences*, 34(8, Part B):6116–6124, 2022.

[5] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common Voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 4218–4222, Marseille, France, 2020.

[6] Salem Chaker. *Manuel de linguistique berbère*. ENAG, Algiers, 1988.

[7] Kamal Naït-Zerrad. *Grammaire moderne du kabyle*. Karthala, Paris, 2001.

[8] Mark D. Wilkinson, Michel Dumontier, et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018, 2016.

[9] Stephanie Russo Carroll, Ibrahim Garba, et al. The CARE principles for indigenous data governance. *Data Science Journal*, 19(1):43, 2020.

[10] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[11] Benoît Sagot and Darja Fišer. Building a free French wordnet from multilingual resources. In *Proceedings of the OntoLex 2008 Workshop*, Marrakech, Morocco, 2008.