Iles d Imesli, 9, pp. 199-202

Kamusi and Amazigh: Solutions for Dialects within a Global Linguistic Data Infrastructure

Founder and Director of the Kamusi Project, an international non-profit dedicated to producing dictionary and data resources for languages Worldwide.

Martin BENJAMIN
École Polytechnique Fédérale de Lausanne, Switzerland

Agzul

Tazerawt-gi , terza amek ara d nesuffeɣ amawal n tmaziɣt ara yesdukkelen akk tantaliwin-is mebla ma tufrar-d ta ɣef tiyiḍ . Aya- nezmer ad t-nefru s tarrayt n uskenawal ara d -yeddemen akk tantaliwin-agi i yesdukkelen tameslyat akken ma llant yerna ad sishelent asuɣel ɣer tutlayin niḍen.

Abstract

The major problem concerning the making of a common Berber dictionary is that language comes in many forms, none of which can be considered representative. We suggest that these problems can be overcome through a lexicographic approach that considers language variants as a matter of data organization. Instead of searching for a recognized definitive form for a language or region, it is possible to list all the forms encountered. Finally, there might be a monolingual Amazigh dictionary that reports on all variants of the language in its unit and provides bridges of translation to other languages around the world.

Keywords: common Berber dictionary, lexicographical approach, variants, monolingual Amazigh dictionary, translation

What is the convergence between an online multilingual dictionary intended to document all the world's languages, and a monolingual Berber dictionary that seeks relevance across borders and dialects? The March 2016 International Colloquium on the Creation of Monolingual Amazigh Dictionaries, in Béjaïa, Algeria, provided the opportunity to investigate this question. As a result of this meeting, it is now possible to envision an Amazigh dictionary that accounts for all variants of the language individually and together, and that provides translation bridges to languages throughout the world.

The essential problematic of creating a unifying Berber dictionary is that the language exists in many forms, none of which can be said to be an overarching standard. A dictionary that tries to propose one language variety

as standard will be roundly rejected by all other groups. However, a dictionary that tries to include multiple varieties runs the risk of quickly becoming unusable. The problems are both political and linguistic: for example, what weight to place on Kabyle in Algeria versus Tamazight in Morocco, and whether Chaoui and Tuareg are different enough to be considered separate languages rather than dialects.

It is proposed that these problems can be overcome through a lexicographic approach that sees variants as a question of data organization. Rather than search for a form that can be asserted as definitive for a language or region, it is possible to document all forms that are encountered, with location information that can supplement or replace the nebulous category of "dialect". Through the data design of the Kamusi Global Online Living Dictionary (GOLD), we can overcome the political and linguistic issues pertaining to the boundaries between dialects, allowing us to focus our energies on the considerable challenges of actually assembling the data.

The central organizing principle of Kamusi GOLD is to disaggregate each concept/spelling pair within a language, and then to link concepts across languages. For example, the word "star" has separate entries for the celestial body, the lead actor in a movie, and the quality rating of a hotel. Each of these individual entries is linked to the appropriate terms for those concepts in other languages, such as "etoile" in French and "nyota" in Swahili for the celestial star. All languages are linked through their shared concepts, with differences noted when the concepts are not exactly parallel.

Within this structure, it is possible to mark several features for each term that pertain to variation. One could list the dialects for a language, and list all the terms within each dialect, but this is an imprecise and unsatisfactory approach. At a much greater level of precision, Kamusi GOLD makes it possible to mark a term for exact locations where it has been sighted. Similarly, pronunciations can be recorded and geo-tagged to the speaker's location. Through geo-tagging of shape and sound, a map can be developed that shows where variants overlap and where they part. However, such specificity requires a tremendous amount of data that is not immediately available, whereas the broad-brush labelling of "dialect" provides a faster if less accurate indication of difference. The Béjaïa colloquium provided the occasion to expand the data model for a middle ground that produces useful generalities about dialect on top of the potential for local specificity offered by geotagging.

For this discussion, we will use artificial names for dialects, in order to abstract the discussion from local considerations. In fact, the design of this system will work equally well for any language with multiple variants, such as the diversity of Arabic from Morocco to Iraq, or the many variants of Pulaar from Senegal to Cameroon.

We begin with a simple premise: a term can be considered the same from one location to the next if it is (a) spelled and pronounced the same way, (b) spelled a bit differently but pronounced the same, or (c) spelled the same but pronounced a bit differently. If a term for a concept is spelled differently and pronounced differently, then it is a different term that should be treated as a separate dictionary entry.

For a language with three major variants, any given concept, such as a celestial body shining in the sky, has several possibilities.

1. Variant A, B, and C are all identical.
2. Variant A and B are identical, C is different.
3. Variant A is different, B and C are identical.
4. Variant A and C are identical, B is different.
5. Variant A, B, and C are all different.

It is quite possible to envision that all five conditions can occur at one time or another.

In the Kamusi structure, each entry can be marked for the variants to which it applies. In the second scenario, for example, a term can easily be tagged as belonging to Variants A and B, while the different term that is used for the same concept in Variant C is tagged as such, and the concepts linked in the same way as translations to other languages. Moreover, if Variant A and Variant B have different spellings for the same term, those spellings can be recorded and marked for their particular variants. Similarly, different pronunciations for the same term can be marked for their particular variant. In this way, the subtle differences within a term between Variants A and B can be chronicled without unnecessarily producing separate entries for essentially the same item, whereas the broader difference between how the concept is expressed in Variant C is shown through the production of a different entry.

The same logic can be applied even further, to the even muddier concept of sub-dialects. If Variant A has sub-dialects North, Central, and West, then the same 5 possibilities elaborated for dialects will pertain to the sub-dialects. In the Kamusi GOLD model, variations among sub-dialects can be tagged in exactly the same way as variations among dialects.

On the user end, the Kamusi GOLD design thus delivers a dictionary that can show all of the terms that are used within any given variant, with their local spelling and pronunciation. If a user is interested in Variant B, it does not especially matter whether a term is unique to that variant, is shared with A, is shared with C, or is common across all three; the user will see a Dictionary of Variant B. When the Variant B user wants to know how to express a term in Variant C, they will either see that the term is the same, or see the term in the other variant as a translation. Fine data, rather than course

labels, can result in dictionaries that are both broad lexicons for a grand language, and nuanced local references.

To conclude, the Béjaïa colloquium has opened the door to a comprehensive electronic dictionary that includes all Berber variations under one roof, without eliding the important differences among them. The solution is a data design that is being implemented within the Kamusi GOLD system, in conjunction with data elicitation that has yet to occur. Kamusi has developed systems to collect and align data from scholars, existing sources, and the public, which could readily be put to the service of the Berber languages (see references). Such an effort will require planning and partnerships, and is thus the subject for continued discussion. The fortunate immediate result of the Béjaïa colloquium, however, is that, from a data perspective, a successful design solution for such an effort is poised to begin.

References

(Available to download at https://people.epfl.ch/martin.benjamin):
Benjamin, Martin, 2015: Crowdsourcing Microdata for Cost-Effective and Reliable Lexicography. AsiaLex 2015, Hong Kong.
Benjamin, Martin, 2014: Elephant Beer and Shinto Gates: Managing Similar Concepts in a Multilingual Database. 7th International Global WordNet Conference, Tartu, Estonia.
Benjamin, Martin, 2014: Participatory Language Technologies as Core Systems for Sustainable Development Activities. 2014 Tech4Dev International Conference, EPFL, Lausanne, Switzerland, 2014.
Benjamin, Martin, and P. Radetzky, 2014: Small Languages, Big Data: Multilingual Computational Tools and Techniques for the Lexicography of Endangered Languages. 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland, USA.
Benjamin, Martin, 2014: Collaboration in the Production of a Massively Multilingual Lexicon. 9th edition of the Language Resources and Evaluation Conference, Reykjavik, Iceland.
Benjamin, Martin and P. Radetzky, 2014: Multilingual Lexicography with a Focus on Less-Resourced Languages: Data Mining, Expert Input, Crowdsourcing, and Gamification. 9th edition of the Language Resources and Evaluation Conference, Reykjavik, Icelan.
Benjamin, Martin, 2011:Toward a Standard for Community Participation in Terminology Development. First International Conference on Terminology, Languages, and Content Resources, Seoul, Korea.