Challenges in endangered language lexicography

Antonia Cristinoi, François Nemo

Challenges in endangered language lexicography

Antonia Cristinoi, Laboratoire Ligérien de Linguistique, Université d'Orléans
François Nemo Laboratoire Ligérien de Linguistique, Université d'Orléans

Abstract

When it comes to endangered minority languages, lexicography is faced with specific limits and challenges. Based on our field lexicography experience and on the writing of a Palikur (Arawakan, French Guyana) dictionary, this paper aims to present some of the shortcomings of dictionary production, and what should and could be done to address the specific challenges which have to be met. First of all, we shall discuss several frequent limits (of) small language dictionaries, namely: i) the absence/scarcity of available corpora; ii) the use of an onomasiological approach for word collection, through a kind of enlarged Swadesh list methodology; iii) the consequent reduction of dictionaries to mere lexical lists; iv) the consequent risk of minoring the lexical specificities of a language by using pre-determinate lists of things or notions to be named. Secondly, we shall advocate for an approach to field lexicography centred on the collection of the most rapidly vanishing parts of the vocabulary, and insist on the strategies which can be used in order to do so, defending the necessity to overcome the lexicographer's absence of expertise in many fields of knowledge by promoting multidisciplinary field work. This issue will be illustrated by the example of the "biolexicon" (plants, insects, etc.) in Palikur. Finally, we shall tackle word translation issues, and the risk to produce either hyperonymic lexicographical descriptions or precise (but unusable for the reader) equivalents. We shall show how these constraints should lead to a semi-encyclopaedic lexicographical approach.

Keywords: bilingual dictionaries, field lexicography, lexical erosion.

1. Introduction

Regardless of their geographical, economic or political context, dictionaries are important tools for describing, promoting and defending languages. However, in the context of endangered languages they play an even more crucial role, as they help saving what is left of rapidly vanishing languages and cultures by recording valuable information that would be lost otherwise. Moreover, in many cases, the existence of a dictionary helps reviving a language and modifies the speaker's attitude towards it (as we clearly noticed during our field research in French Guyana), encouraging him/her to use it more often.

It must be noted, though, that writing a dictionary for an endangered language presents particular challenges that distinguish it from regular dictionary making (i.e. writing dictionaries for widely spoken languages) and requires rather different methodologies and lexicographical strategies. In this context, it is important to point out the specificities of endangered language dictionaries as opposed to major language dictionaries, which will lead us to a critical discussion of the methodological trends in use (with a special focus on Palikur, an Arawakan language spoken in French Guyana and Brazil) and to suggesting alternative methods both for data collection and lexical description.

2. What is so special about endangered language dictionaries?

Endangered language dictionaries differ from what we might call "classical dictionaries" from various points of view, ranging from the number of people involved in their conception to funding, lexicographical descriptions and ideological stakes.

2.1 Authors

It is a commonly accepted idea that lexical collections (glossaries, bilingual wordlists, thematic dictionaries, etc.) concerning endangered languages are produced by one (or in the best of cases a small group) non-native author, most often a linguist, an ethnographer or a missionary, while dictionary projects for major languages are usually carried out by a

professional staff of trained native speakers. The time allotted to these enterprises is consequently different. It is also worth mentioning that the author's being a non-native speaker and the multi-cultural/multi-linguistic context of vulnerable languages entail almost exclusively that an endangered language dictionary will necessarily be bilingual.

2.2 Funding

Dictionaries for widely spoken and thoroughly described languages are definitely a profitable commercial enterprise, and are systematically re-printed and enriched, but this doesn't apply to endangered language ones, where there are hardly ever any new improved versions. Consequently, writing a dictionary for an endangered language is a one time shot with no update possibilities, which entails complicated methodological choices that will be discussed later on.

2.3 Users and uses

While classic dictionaries (either monolingual or bilingual) are distributed world-wide, endangered language dictionaries will only count a small number of users, most of them *scholars*. However, this situation is currently changing (and modern language documentation trends fully show it) as dictionary makers are more and more aware that the role of a dictionary in a fragile language context is no longer exclusively academic. Field lexicographers (be they linguists, missionaries or anthropologists) are more and more aware that their work must have a positive impact for the people they work with and not only for scientists. Therefore, dictionaries play a crucial role in preserving and transmitting linguistic and cultural knowledge for *indigenous people* who constitute a second category of users. The third category consists of *non-native speakers* interacting with indigenous people for various reasons.
Endangered language dictionaries will thus be used:

- for research purposes;
- to document a specific language and more importantly a specific culture;
- to preserve a linguistic and cultural heritage which would disappear without written material;
- to help indigenous people communicate in a foreign (often dominant) language by finding the proper equivalents for indigenous words;
- to help non-native speakers understand native-speakers and their cultural background;
- to provide a stable orthography for the whole lexicon.

These elements will also have a crucial influence on methodological choices, in terms of both, lexicon and lexical descriptions.

Paradoxically, while endangered language dictionaries broaden their scope, classic dictionaries co-exist with more and more specialized ones, usually directed towards very specific audiences (learner's dictionaries, specialized dictionaries, historical ones etc.). The challenges of building an all-in-one dictionary as opposed to something like "a railroad dictionary" one will be very different once again.

2.4 High ideological stakes

Besides obvious economic motivations, dictionaries concerning major languages are written and published for a variety of reasons:

- keeping track of language evolutions;
- documenting a specific field (i.e. astronomy);
- enforcing terminology and language use;
- defending national languages and obviously
- encouraging and facilitating cross-cultural communication.

The situation is more complex as far as endangered languages are concerned, since dictionaries often involve a transition from oral to written language practices which will entail

political consequences like standardizing language, making orthographical choices and, more importantly, deciding which language variety will constitute THE standard.

2.5 Data collection

The differences between endangered language dictionaries and "regular" ones can also be seen in data collection methods. In the first case, the scarcity of available corpora for indigenous languages results in the use of elicitation methods derived from ethnographical work, while in the second one the problem would rather be the selection of a representative corpus sample out of an infinity of possible choices.

Data collection is also linked to the single/multiple author issue and to the compiler's linguistic and encyclopaedic competence (whether or not the field lexicographer is familiar with botany, for instance) which will drastically influence both field practices and subject choice (one will more likely work on familiar themes).

2.6 Information formatting

All the elements mentioned before have a crucial influence on the amount and the type of information selected for lexical descriptions. Financial issues will be determinant for the length of a dictionary, the number of entries and even for the layout. The number of authors and their linguistic and non linguistic competence (and for endangered language lexicographers the time they spend with indigenous people) will also play a role in data selection, as well as political issues (words that one can or cannot include the dictionary) and user profile (which may be the most delicate aspect).

3. Endangered language dictionaries: state of the art

All the issues discussed above have a noticeable impact on dictionary making and consequently on dictionaries. Even if the language documentation perspective is increasingly present in lexicographical practices on endangered languages, the resulting dictionaries still have a variety of flaws, deriving from:

- **data collection:**
  - partial coverage of the lexicon, usually general vocabulary (often reduced to mere lexical lists);
  - lexical artefacts, i.e. words that are not used by the speakers, whose referents are absent from the cultural background;
  - few examples of language in use if any;
- **data description:**
  - complex structure, which makes dictionaries hard to use by untrained users;
  - complicated writing systems which make it impossible for non-specialists to find information;
  - equivalents or equivalent explicative structures (as we are dealing bilingual dictionaries) that are either to precise or too vague and thus exclude some user categories;
  - little cultural/encyclopaedic information on important subjects.

4. Collecting data: upside down methodology

4.1 A plea for lexicographically relevant corpora

The main reasons for faulty on incomplete lexical inventories are the absence/scarcity of available corpora and the use of inadequate elicitation methods.

While corpus absence is a self-evident matter, the presence of some forms of corpora raises several questions like lexical variety and corpus relevance for a given language. Since most available corpora are the result of grammar descriptions, they will certainly fail to show the lexical complexity of a language. The matter of corpus relevance can be anti-illustrated by the example of our work on Palikur, where the only available corpus (at least this was the case when we began our work on the Palikur-French dictionary) was the bible translated into Palikur

by American missionaries (who also did a considerable linguistic work on the language). Using such a corpus as dictionary basis is more than problematic, as we are dealing with translation from a source text with a completely distinct historical, cultural and geographical background, produced by a non-native speaker.

For poorly documented languages, corpus doesn't seem to be the best way to achieve an exhaustive lexical inventory. However, corpus (as a result of some form of active eliciting) can be a really interesting tool for lexical description as it can provide three types of data: examples illustrating language in use, information on word polysemy (to some extent) and valuable encyclopaedic and cultural information which is vital for a "documenting" dictionary.

The method we used for the creation of such a corpus (mainly for bird, plant and insect names) involves long term fieldwork and interdisciplinary specific missions, like for instance, one week trip in the forest with two native informants and a botanist. The informants would be asked to give as much information as possible on one object (plant, animal, etc.): physical description (for a plant leaf and stem textures, sap colour and consistence, taste), identification criteria, different uses, symbolic role and so on. The final corpus will be in fact a collection of individual corpora that would provide a constellation of new words and valuable information to facilitate description. The presence of a specialist who can provide additional scientific information makes this kind of corpus even richer. This method can also be combined with the use of a specific type of word list.

4.2 Word lists for lexicographers?

Generally speaking, the use of pre-determinate lists of things or notions to be named and to a lesser extent even some aspects of direct/active eliciting can have a negative impact on data collection. Using the onomasiological approach for lexical inventories, through a kind of enlarged Swadesh list methodology often leads to the reduction of dictionaries to mere bilingual lists of words. The word list methodology has three other major inconveniences: the consequent risk of missing the lexical specificities of a language by using an ethnocentric method (indeed adapted for short-term fieldwork), the creation of lexical artefacts and the omission (and eventually loss) of a great part of the lexicon, concerning highly specialized concepts. The second and third points need further discussion.

Here is a very relevant example of lexical artefact, extracted from Green's Palikur-Portugese dictionary:

ihpaki ku ka aynsima uhokriviyenevwi ay ***hinduismo*** [s.m.]
(Green 2000 :160)

literally, faith in more the one god, which is clearly a word list construct as the Palikur entry is not even a word but a whole sentence.

However, the main flaw of the word list method is that it focuses on the most common and used parts of the vocabulary, which are actually the least endangered. The vanishing words, those corresponding to very specific cultural realities and only known by the most experienced speakers, are the most threatened ones. These words name biological entities (animals, plants), ritual practices, mythical entities, traditional medication and objects which are no longer used in everyday life. The same process of lexical erosion applies not only to words but also to word meanings and word uses. So, linguists are clearly doing things in the wrong order, for obvious reasons: short time field work and thus lack of confidence from native speakers, no multidisciplinary competence, lack of funding, etc..

One way of avoiding at least the ethnocentric character of word lists is to elaborate context specific lists with specialists of other fields.

4.3 Saving disappearing words: inter-active eliciting

The best way to grasp little used, specialized words, is of course observation on a daily basis, which involves learning the language, building a long-term relationship with the community and of course working with the right informants. This can only be done through long-term field work which allows the linguist him/herself to acquire a deep knowledge of the

context. However, as far as natural subjects are concerned, we have developed an alternative methodology which does not require extreme field work (forest expeditions and so forth), the *multi-stimulus approach*. While working on bird names in French Guyana we realized that the use of images alone is not a reliable method as Palikur people do not identify all birds according to their appearance but according to habitats, behaviour and calls/songs. Consequently, we used images and recorded bird songs combined (and sometimes information on habitat and behaviour that can be found in good bird guides) for a thorough identification, which lead to more effective results (completed by random *in situ* elicitation).

The same method can be easily applied to frog names and to a lesser extent to plants (by using images and descriptions which include sap colour and consistence and medicinal uses, whenever they are recorded, of course). However, for insects, the best approach is direct fieldwork, notably because of the lack and un-readability of drawings, preferably with a trained entomologist and at least two native informants for *in situ* cross-checking. The multi-stimulus method (involving this time touching or feeling things) is also interesting when working on adjectives concerning physical attributes (soft, stingy, etc.), tastes or textures.

5. Finding the right equivalent(s) in bilingual dictionaries: not to do list

Another delicate aspect of building an endangered language dictionary is the description the lexical inventory. Providing thorough and accessible descriptions is a real challenge and the subject has been vastly documented, so the only aspect we shall focus on here is lexical equivalence. Finding the most accurate equivalent structure means taking into account the variety of users mentioned in 2.3, i.e. providing information for scientists (via scientific names), a direct equivalent in the other language for native speakers and some sort of description where the equivalent is not self-evident for non native speakers. It is definitely a lot to ask from one dictionary and some dictionaries sometimes fail to meet all these requirements as shown in the three examples below.

The first one provides very little information for native speakers and for non-specialists: we know that the word refers to a creeper with a thin stem and nothing else, so the entry is clearly addressed to a scholar.

.ɨpɔkasilisili **N**

♦ Liane (sp.), *Mesechites trifida* (Jacq.) Mül. Arg.
et *Condylocarpon guianense* Desf. (Apocynaceae)
◊ Etym.‖ liane *Odontadenia grandiflora/* fine ‖
□ La tige de cette espèce est très fine. Grenand (1989 :198)

In the example below we have no scientific name, and no equivalent (it is true though that sometimes equivalents simply do not exist), just a defining gloss (a kind of caterpillar that eats cassava crumbs), which makes it impossible for any user to look for any further information on the referent.

khalise, *n.* ◊ ZOO., espèce de chenille qui mange les restes de cassave (JPB)
Patte (2011 :143)

The third case illustrates a lacking description, as the only things we find in the entry are scientific and vernacular names, which do not provide any idea about the referent. However, these names give the user some hints for further research.

SABUATETE, *n.* ◊ ZOO., *(Melanerpes cruentatus)* pic à chevron d'or.
Patte (2011 :191)

One other potential danger in finding equivalents is hyperonymy, i.e.the use of a generic term as an equivalent for a very specific. The example discussed here comes from our own field practice and unfortunately neither long-term fieldwork nor complicated expeditions could have prevented it. The problematic issue is the Palikur word *kasis* which we translated by *fourmi*, the French equivalent of *ant* and thought to be used as a generic term for all ant-like species. However, during field sessions on ethnoentomology with a trained specialist we realized that what the Palikur called *kasis* was in fact a very specific ant, that the term does not

have a generic use and that the equivalent was more difficult to find (in this case, two out of three types of information were possible: scientific name and defining gloss but no vernacular name). This situation was obviously due to our own ignorance of aunts in and also of the way they are perceived by Palikur people.

Last, but not least, in a multi-lingual, multi-ethnic context like French Guyana (where the official language is French but people also speak French Creole and Brazilian Portuguese) bilingual dictionaries easily become multi-lingual ones, which raises further questions about lexical descriptions and the amount of information one can decently insert in a dictionary.

6. Conclusion

To conclude with, dictionaries for endangered languages are extremely complex projects and all the challenges they have to meet are a huge task for one researcher. Consequently, they can hardly be flawless.

It is however possible improve them considerably by: i) bearing in mind when working on endangered languages that it is crucial to focus on the most vulnerable part of the lexicon which is the most likely to vanish rapidly; ii) targeting more than one category of users and thus working not in a strictly bilingual but an encyclopaedic perspective; iii) working if possible in multidisciplinary teams; iv) creating areal (e.g. Amazonian) multi-stimulus tools that could be made available for the individual lexicographer in order to facilitate his/her work on specialized portions of the lexicon. All these elements combined may succeed in transforming the dictionary into a genuine language documentation tool.

References

Austin, P. & Sallabank, J. (eds.). *The handbook of endangered languages*. Cambridge: CUP

Green, D. & Green, H. 2010. *Yuwit kawihka dicionário Palikúr - Português.* SIL.

Grenand, F. 1989. *Dictionnaire wayãpi-français, lexique français-wayãpi.* Paris : Peeters/Selaf.

Grenand, F. (Chief-Editor) .2009. *Encyclopédies palikur, wayana, wayãpi : langue, milieu et histoire,* fascicule Encyclopédie des Amérindiens de Guyane. Paris : PUO- CTHS.

Hartmann, R. R. K. 2003. *Lexicography, critical concepts.* New York: Routledge.

Launey, M. 2003. *Awna parikwaki. Introduction à la langue palikur de Guyane et de l'Amapa*. Paris : IRD Éditions.

Mosel, U. 2011b. Lexicography in endangered language communities. In Austin, Peter & Sallabank, J. (eds.). *The handbook of endangered languages*. Cambridge: CUP, pp. 337-353

Patte, M. F. 2012. *La langue arawak de Guyane. Présentation historique et dictionnaires arawak-français et français-arawak.* Paris : IRD.