Building bilingual dictionaries for minority and endangered languages with Mediawiki

2^{nd} Workshop on Collaboration and Computing for Under-Resourced Languages - 'Towards an Alliance for Digital Language Diversity'

George Dueñas Diego Gómez

Caro y Cuervo Institute - Muysc cubun

May 23, 2016

Outline

1 Dictionaries

2 Mediawiki

3 Colombian Endagered Languages

4 Recovery Information (SMW)

5 Contributions

6 Conclusions and perspectives

Outline

1 Dictionaries

2 Mediawiki

3 Colombian Endagered Languages

4 Recovery Information (SMW)

5 Contributions

6 Conclusions and perspectives

Outline

1 Dictionaries

2 Mediawiki

3 Colombian Endagered Languages

4 Recovery Information (SMW)

5 Contributions

6 Conclusions and perspectives

Outline

1 Dictionaries

2 Mediawiki

3 Colombian Endagered Languages

4 Recovery Information (SMW)

5 Contributions

6 Conclusions and perspectives

Outline

1 Dictionaries

2 Mediawiki

3 Colombian Endagered Languages

4 Recovery Information (SMW)

5 Contributions

6 Conclusions and perspectives

Outline

1 Dictionaries

2 Mediawiki

3 Colombian Endagered Languages

4 Recovery Information (SMW)

5 Contributions

6 Conclusions and perspectives

Basic Issues about Dictionaries

Atkins et al., 2008

- A dictionary is "a description of the vocabulary used by members of a **speech community**".
- The types of dictionaries are monolingual, **bilingual** (unidirectional or bidirectional), and multilingual.
- The means of disclosure (printed, electronic, or **web-based**).

The dictionary has been used to describe the vocabulary of many languages of the world both extinct and living, as well as to describe the inventory of words from different areas of knowledge.

The Americas (Colombia)

First dictionaries

- The creation of dictionaries by hand began with the arrival of the Spaniards and the Catholic missionaries that undertook the evangelization of the Indians.
- They created some kind of grammars called “artes” and bilingual vocabularies between Spanish and native american languages.

Summer Institute of Linguistics

- They created typewritten dictionaries composed of two parts: (i) an explanation about grammar, alphabet, and sounds; (ii) a list of words in indigenous language (part of speech, meaning in Spanish, an example, and an image).

Electronic Lexicography

Lexicography

- On the practical side, it is associated with creating a dictionary (planning, resourcing, compiling, writing, and editing).
- On the theoretical side, it is associated with developing models on the structural and semantic relationships between words.
- **Electronic Lexicography** became an interesting research subject when *The Random House Dictionary of the English Language* was published (Urdang, 1966).
  - Information into computer: 1) illustrations, 2) pronunciations, inflected forms, and part of speech, 3) definitions, 4) variants, 5) etymologies, 6) words that are not defined because their meanings are self-evident, and 7) synonyms, word lists and usage notes.

Mediawiki (MW)

MW is a free and powerful wiki engine to process and display stored data in an easy and fast way.

Advantages (Granger, 2012)

- Using Wikitext.
- Integrating multimedia content as audio, images, and video.
- Managing easily content between different versions.
- Providing a collaborative framework, and not requiring prior knowledge by the users in HTML or CSS to insert information.

- Users need to be registered before they can edit (MW: Restrict anonymous editing).

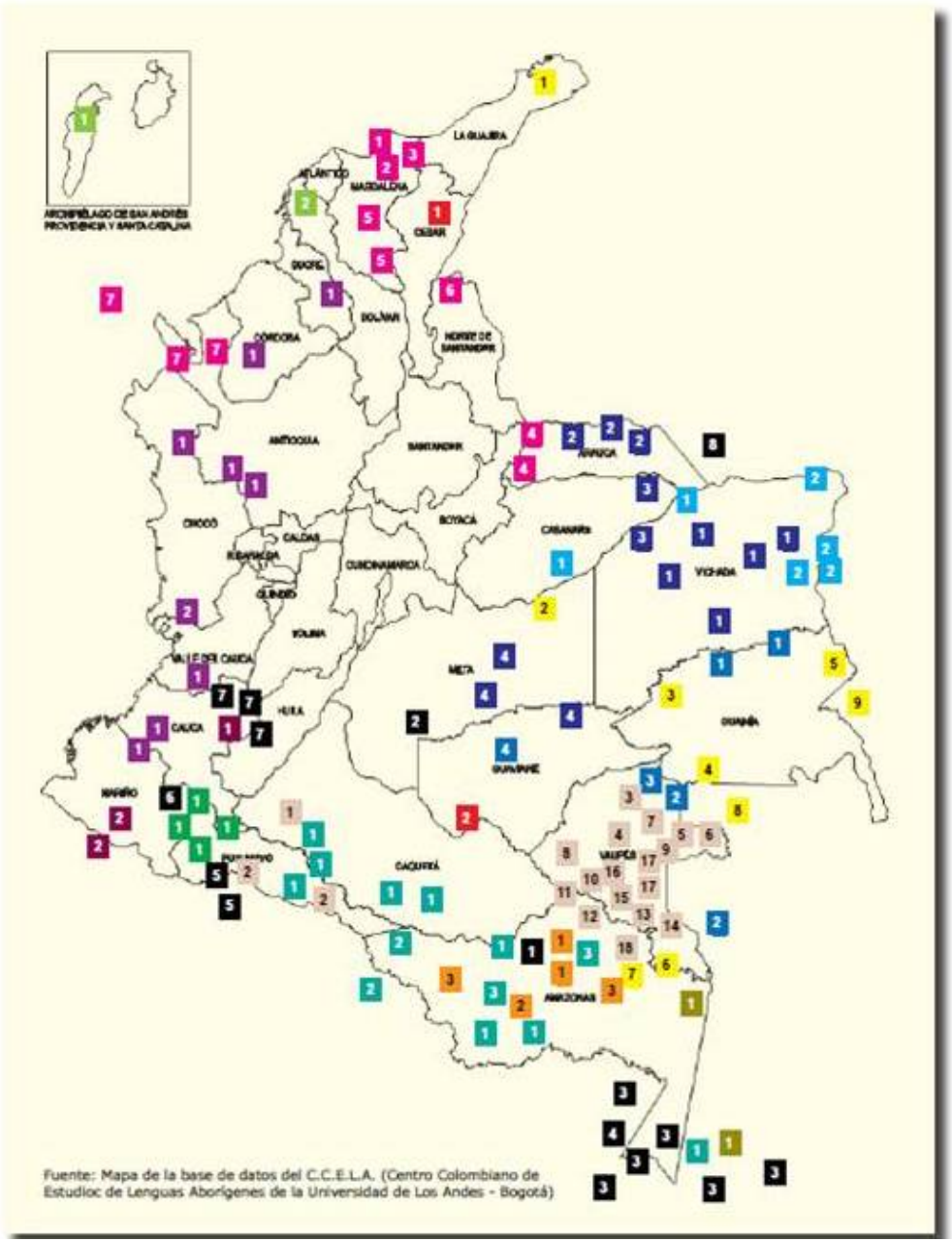
Semantic Mediawiki (SMW)

An extension that generates automatically lists and perform more efficiently searches in a wiki-dictionary.

Krotzsch, 2007

- This extension allow us to query the dictionary as long as users tag the wiki's contents with explicit, machine-readable information, i.e. the main prerequisite of exploiting semantic technologies is the availability of suitably structured (semantic) data.

Credit: *Addicted04 et al.*



FAMILIAS LINGÜÍSTICAS

ARAWAK
1. Wayuu
2. Achagua
3. Piapoco
4. Curripaco
5. Baniwa
6. Kawiyari
7. Yukuna
8. Tariano
9. Baniba

BORA
1. Muinane
2. Bora
3. Miraña

CARIBE
1. Yuko
2. Karijona

CHIBCHA
1. Kogui
2. Ika
3. Damana (Wiwa)
4. Uwa-Tunebo
5. Chimila
6. Bari
7. Cuna

CHOCÓ
1. Embera
2. Waunan

GUAHIBO
1. Sikuani
2. Hitnu
3. Kuiba
4. Guayabero

MAKU
1. Puinave
2. Yuhup
3. Cacua
4. Nukak

SALIBA
1. Saliba
2. Piaroa

WITOTO
1. Uitoto
2. Okaina
3. Nonuya

QUECHUA
1. Inga

CRIOLLOS
1. San Andrés
2. Palenque

TUPI
1. Cocama

TUKANO
1. Coreguaje
2. Siona
3. Kubeo
4. Pisamira
5. Piratapuyo
6. Wanano
7. Dsano
8. Carapana
9. Tucano
10. Tatuyo
11. Taiwano
12. Barasana
13. Bará
14. Macuna
15. Tuyuka
16. Yuruti
17. Siriano
18. Tanimuka

BARBACOA
1. Guambiano
2. Awa-Kwaiker

AISLADAS
1. Andoke
2. Tinigua
3. Tikuna
4. Yagua
5. Cofán
6. Kamsá
7. Paez-nasa
8. Yaruro

Colombian Endagered Languages

There are 64 or more indigenous languages (13 linguistic families), two creoles, romani, Colombian sign language, and Spanish.

Saliba language

- Colombia (Vichada, Meta, and Casanare) and Venezuela.
- Between 1488 and 3035 in Colombia (González, 2011). Only 2 % speak in everyday life (Ramírez, 2010).

Carijona language

- Colombia (Guaviare, Vaupes, and Caqueta).
- Between 30 and 425 speakers in Colombia (González, 2011), who do not use the language in everyday life.

Severely endangered

Critically endangered

Dictionary based on MW

First Template - *Lang1-Lang2*		
Bilingual bidirectional dictionary		Bilingual unidirectional dictionary
Saliba-Spanish	Spanish-Saliba	Carijona-Spanish
{{SAL-ESP}}	{{ESP-SAL}}	{{CAR-ESP}}

{{SAL-ESP}}

{{ESP-SAL}}

{{CAR-ESP}}

Dictionary based on MW II

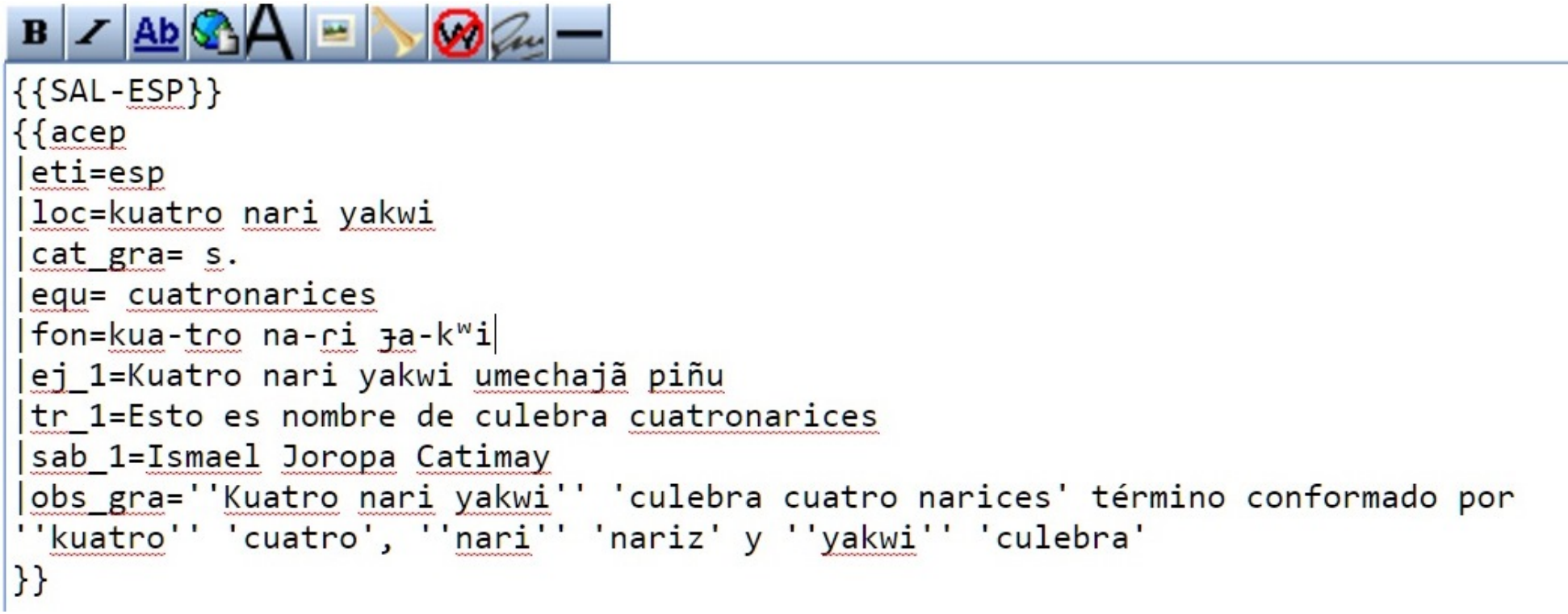
Second Template - *acep*		
SAL-ESP	ESP-SAL	CAR-ESP
{{acep |eti= |loc= |cat_gra= |equ= |fon= |ej_1= |tr_1= |sab_1= |obs_gra= }}	{{acep_es |loc= |cat_gra_es= |equ= |fon= |ej_1= |tr_1= |sab_1= |obs_gra= ... }}	{{acep |loc= |cat_gra= |equ= |fon= ... ... }}

Dictionary based on MW III

Third Template - *cat_gra*(_es)	
SAL-ESP	<includeonly> {{#switch: {{{1}}} | num. an. = Numeral animado (animate numeral) | num. in. = Numeral inanimado (inanimate numeral) | adj. an. = Adjetivo animado (animate adjective) | adj. an. = Adjetivo inanimado (inanimate adjective) | ... }} <includeonly>
ESP-(SAL, CAR)	<includeonly> {{#switch: {{{1}}} | n. = Sustantivo (noun) | v. = Verbo (verb) | adj. = Adjetivo (adjective) | adv. = Adverbio (adverb) | ... }} <includeonly>

Example of an Entry Dictionary

Entry: http://www.yourdictionary[1]/**kuatro nari yakwi**



```
{{SAL-ESP}}
{{acep
|eti=esp
|loc=kuatro nari yakwi
|cat_gra= s.
|equ= cuatronarices
|fon=kua-tro na-ri ɟa-kʷi
|ej_1=Kuatro nari yakwi umechajã piñu
|tr_1=Esto es nombre de culebra cuatronarices
|sab_1=Ismael Joropa Catimay
|obs_gra=''Kuatro nari yakwi'' 'culebra cuatro narices' término conformado por
''kuatro'' 'cuatro', ''nari'' 'nariz' y ''yakwi'' 'culebra'
}}
```

[1] Hortensia Estrada and Camilo Robayo have gathered the data and researched linguistically these languages, respectively.

Recovery Information (SMW)

We use the parser function **#ask**, followed for a category (e.g., locution), and the latter is followed by properties. Wildcard **::+** returns all pages that have any value for the property.

The function of the vertical bar or the pipe symbol is to separate property conditions to display.

The function of the question mark followed by the property name is to display all the values assigned to a certain property.

```
{{#ask: [[locution::+]]
    | ? phonetic transcription (fonologia)
    | ? grammatical category (categoria_gramatical)
    | ? equivalence (equivalencia)
    | ? ...
}}
```

Contributions

- Bilingual dictionaries can be built by native speakers with a little help from non-native.
  - This information is stored in real-time for present and future generations (from native speakers to researchers).
- A country can use these technologies to document, to preserve and to spread widely the native languages would give a breakthrough in the struggle against inequality of linguistic rights and digital opportunities for all languages and for their speakers.
  - The advantages and disadvantages of these tools have to be shown to the communities.
- These initiatives can create synergy between native speakers and government to build and implement public policies.

Conclusions

- Documentation and revitalization can be done jointly between native speakers and researchers using MW and SMW as a friendly software to work collaboratively.
- We hope this platform promotes more involvements of the native speakers and other researchers with the language resources and technologies.
- In some cases, web-based dictionaries increase the prestige of the language among the society.

Perspectives

- It can integrate other extensions for different tasks; for example, a Semantic Maps extension for georeferencing.
- In the future, the dictionary information can be downloaded in a file in order to print.
- In the future, government websites can be written in native languages by machine translation supported by web-based dictionaries.

References

Atkins, S and Rundell, M. 2008
The Oxford guide to practical lexicography. Oxford University Press.

Urdang, L. 1996
"The systems designs and devices used to process the random house dictionary of the english language". In: *Computers and the Humanities*, 1(2), 31–33.

Granger, S. 2012
"Introduction: Electronic lexicography-from challenge to opportunity". In: *Electronic Lexicography*. Oxford University Press.

González, M. S. 2011
Manual de divulgación de las lenguas indígenas de Colombia. Instituto Caro y Cuervo.

Ramírez, H. E. 2010
"La modalidad epistémica en la lengua sáliba". In: *UniverSOS Revista de Lenguas indígenas y universos culturales*, 7, 107–118.

Thank you