Technological tools for dictionary and corpora building for minority languages: example of the French-based Creoles

Paola Carrión Gonzalez(1,2), Emmanuel Cartier(1)
(1)LDI, CNRS UMR 7187, Université Paris 13 PRES Paris Sorbonne Paris Cité
(2)Departamento de Traducción e Interpretación, Facultad de Filosofía Y Letras, Universidad de Alicante
E-mail : pccg1@alu.ua.es , ecartier@ldi.univ-paris13.fr

We present a project which aims at building and maintaining a lexicographical resource of contemporary French-based creoles, still considered as minority languages, especially those situated in American-Caribbean zones. These objectives are achieved through three main steps: 1) Compilation of existing lexicographical resources (lexicons and digitized dictionaries, available on the Internet); 2) Constitution of a corpus in Creole languages with literary, educational and journalistic documents, some of them retrieved automatically with web spiders; 3) Dictionary maintenance: through automatic morphosyntactic analysis of the corpus and determination of the frequency of unknown words. Those unknown words will help us to improve the database by searching relevant lexical resources that we had not included before. This final task could be done iteratively in order to complete the database and show language variations within the same Creole-speaking community. Practical results of this work will consist in 1/ A lexicographical database, explicitating variations in French-based creoles, as well as helping normalizing the written form of this language; 2/ An annotated corpora that could be used for further linguistic research and NLP applications.

1. Existing Lexicographical Resources in French-based Creole

[19th century -1885] Lafcadio Hearn (Creole proverbs) | bilingual lexicons

- 1. Haitian / Indian Ocean Creole languages
- 2. Caribbean Creoles
- Precursors: Albert Valdman, Annegret Bollée, Robert Chaudenson, Hector Poullet, Sylviane Telchid, Raphaël Confiant

Main problems of lexicographical development [Hazaël-Massieux : 2002]

- Disparity of macro and microstructures
- Orthographic conventions → continuum
- Content → spelling variants, available examples, parts of speech, origin of entry words, (…)
- Falsefriends ("fig" = figue), uncontrolled diversion (*chokatif - chokativité*), erroneous integration (*abònman* / abonnement)
- Diatopic variation

Importance of monolingual works

- Arnaud Carpooran: *Diksioner Morisien*, Koleksion Text Kreol, Ile Maurice, 2009

DECOI

- A. Bollée (University of Bamberg) [2007] / R. Chaudenson's project [1979]
- 4 Volumes (3 french origin / 1 unknown origin) | French word forms → creole variations

Example

- **paille**,n.f.
- lou.**pay,lapay,lepay**n.'paille;cosse,enveloppe,spathe(demaïs,delacanneàsucre,etc.);styl esetstigmatesdumaïs',*kaselapay*'rompreuneamitié'(AVaDLC)haï.**pay**n.'feuillesèche,spath e,chaume,son'(ALH1557),'chaff,straw[drygrass],anydrycombustibleplant;hay;husk;bran; thatch;waste,scraps;misery,suffering,trouble;low,badorlosingcard',*fèpay*'tobefrequent;t obenumerous',*fèyonmounpasepay*'toannoy,bother;tomistreat';**pay**adj.'weak,verylight'(A VaHCE),*kafe(a)pay,paykafe*'cafétrèsléger'(ALH790)
- gua.**pay**n.'paille'(LMPT)

DECA

- A. Bollée / Ingrid Neumann-Holzschuch (University of Regensburg)
- Sources: Philip Baker, Albert Valdman, Robert Chaudenson
- DECOI structure
- Paper-format/ Electronic database (XML)
- In process

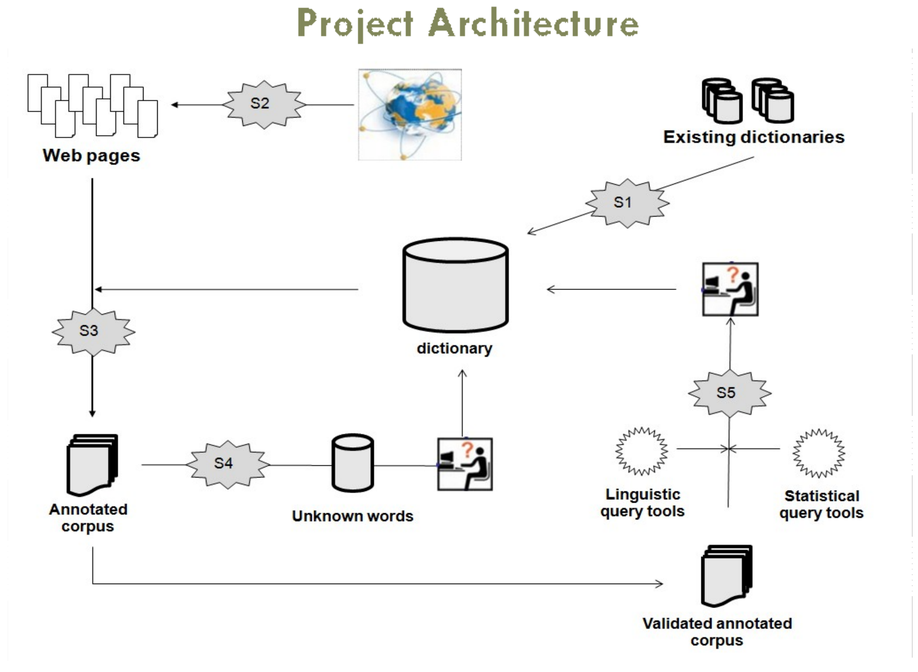
KRENGLE (18000 entries) Haitian Creole – English dictionary, by Eric Kafe (University of Copenhagen). Connected to Wordnet (semantic relations). Partly downloaded for free [http://www.krengle.net]

English: manufacture
Creole: fabrike - izin -
Pawol mo - Root word: manufacture
Synonyms / Sinonim
Definition / Definisyon
Example / Egzanp

WEBSTER'S ONLINE DICTIONARY (multilingual Thesaurus, 1200 languages, Noah Porter, 1913). It includes Haitian Creole, varieties of Creole, enriched by users via moderation. It recognizes several resemblances and dissemblances in the same family of Creole languages.
[http://www.websters-online-dictionary.org/browse/]

izin (manufacture), **frabrik** (manufacture), **fabrike** (do, make, manufacture). Additional references: Haitian Creole, Dominican Republic, Haiti, manufacture, (volunteer & more translations)

Project Architecture



This architecture comprises five main steps:
1. Step 1 (S1): this preliminary step aims at building a first electronic compilation from existing lexicographical resources. This implies gathering the existing resources, either digitized resources from paper-based dictionaries, or fully electronic resources. This also implies identifying available resources. This initial step is detailed in section 4;
2. Step 2 (S2) : this second step aims at setting up an infrastructure enabling corpora building and feeding; this means identifying web available resources as well as setting up automatic procedures to retrieve on a regular basis these documents; it is detailed in section5.1.
3. Steps 3 and 4: (S3-S4) morphosyntactic analysis of the corpora to maintain the existing dictionary; automatic analysis of corpora will explicit unknown words, and some of them will have to be included in the initial dictionary; iteration of this procedure will permit to complete the existing dictionary, as well as enabling morphosyntactic annotation of the corpora; this step is detailed in section 5.2.
4. Step 5 (S5) : this step, out of the scope of this paper, will be implemented as soon as the dictionary is sufficiently completed; annotated corpus could then be validated and then queried using linguistic and statistical tools, so as to improve information in the dictionary. It will be evoked in the section 5.3.

2. Dictionary building: existing resources compilation

Objectives :
- retrieve available (free) data from the web
- two main sources : paper-based and electronically-based dict.

Steps followed :
- compilation of available ressources
- scripts to retrieve data and combine them into a unique dictionary keeping track of source and

Electronically-based dict.: about 2000 entries
- purpose of the dictionaries : mainly educational or popularisation
- small lexicons
- weaknesses of linguistic information (mostly reduced to entry and definition / translation)

Paper-based dict.: about 20 000 entries
- less numerous
- contain much more linguistic information (pos, examples, translation, spelling variants, cross-references, semantic relations, etymology)
- generally bilingual resources
- require more complex data processing.

3. Corpora to improve the dictionary

Web Corpora can help building/maintaining lexicographical resources
Our project will build its dictionary not only from existing electronic resources, but also by morphosyntactically analyzing corpora (with the help of the previously setup dictionary) (Kilgariff, 2003; Baroni, 2009 for example; and for minority languages Scannel, 2007)

Steps :
- corpora downloading (on a regular basis),
- automatic morphosyntactic analysis,
- manual dictionary improvement.

Iterative process :
1/ extracting unknown words as a first step
2/ checking meanings associated to recognized words
=> will end up with a more exhaustive dictionary, following the Zipf's law.

Three main corpora:
- first : 1 million words, setup to tune the morphosyntactical analysis and the dictionary improvement process, with in mind the general-purpose language;
- second : the Bible, to complete and tune the dictionary with a specific vocabulary;
- third : evolving corpus, to maintain the dictionary and setup an adequate environment for lexicographers.

Analysis :
1/Quick lexicographic coverage of the dictionary: from 28,6% unknown words with about 70% unique words (first corpus) to 7,6% and about 98% of unique words (second corpus)
=> whereas a relatively small corpus enable to cover 90% of lexicographic entries, only a really huge corpus enable to tend to 100% coverage. (Zipf's law)
2/ Dictionary coverage : 123 245 unique words-forms; coverage to be tuned with :
- **phrases** (about 50% of the vocabulary, see Sag et al, 2002, for example);
- **unknown words without linguistic information**
- **known words linguistic information**
- **spelling variants.**

3/ unknown words
it would be necessary to have automatic procedures to track the meaning evolutions, rather than only word forms existence.
-words from other language (specifically from English, in our case),
-misspelled words,
-proper names,
-specific notations,
-real unknown words.
=> The resulting list has been included in the dictionary without any linguistic information, except for a small part of it with information taken from web-based dictionary (that could not be retrieved globally, but can be used for individual word search) or paper-based dictionaries (see Carrión, 2011 for the list of these dictionaries). For each of these words, we have decided, in a first step, to include only part of speech, translation in French and English, and varieties if applicable.

First corpus : 15 different sources (Haïtian Creoles), either manually or automatically downloaded (« Voa News », « Alter Presse »). 1 208 862 words.
Second corpus : Bible in Haitian Creole (BibleGateway) downloaded with httrack. *about 4 million words.*
Third corpus : RSS feeds and newspaper sites. Continuous downloading. *In progress*

Retrieval and cleaning of web pages (Cartier, 2007, 2009)
- convert html files into a text format and to UTF-8;
- identification and retrieval of the textual zone of interest
- morphosyntactic tagging
- postprocessing : unknown words frequency, CQP search engine…

Word frequency of unknown words
identify unknown words sorted by frequency (see Cartier, 2007 for details).

Corpus	Recognized words	Unknown tokens	Unknown words / Unique words
Corpus 1(1 208 862 tokens)	631 984 (52,3%)	576 878 (47,7%)	345653 (28,6%) / 243 892 (70,55%)
Corpus 2 (4 505 442 tokens)	3 722 628 (82,6%)	782 814 (17,4%)	343657 (7,6%) / 336 896 (98,03%)

4. Conclusion

On-going project whose goals are to:
1/ **Explicit a methodology** to improve NLP and linguistics research for minority-languages, focusing on the American-Caribbean French-Based Creoles;
2/ **Setup procedures and an environment for lexicographers** to store and maintain lexicographical data from existing resources and web corpora.

Results / Perspectives
1/ **Compilation of existing lexicographical resources** : about 20 000 lexicographical entries, but exhibited complex-to-solve problems : quality, quantity diverge from one resource to another, macro and micro-structures are far from unified. => Discussion to work with the DECA team;
2/ **Dictionary completion and maintenance through web corpora analysis** :
=> In the course of setting up a web-based environment to maintain the dictionary and study lexicographical phenomena through several iterative processes:
- Live corpus download (BootCat, RSS Corpus Builder);
- Cleaning and normalization of web pages;
- POS tagging with existing lexicographical resources;
- Web-based environment for lexicographers to study corpus-based lexicological phenomena; (IMS Open Corpus Workbench);

This project has generated two crucial elements for the American-Caribbean creoles: a POS annotated corpus and a POS tagger. These data and tools will be soon released as open-source.