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

Abstract

In this paper, 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. Introduction

Minority-languages have always existed and will continue to exist. That is an historical, sociological and political fact. But nowadays, electronic devices, internet communication and computer storage enable to keep track of their existence and, more, to study their evolution. In this perspective, our project aims at setting up a methodology and tools facilitating the study of these languages through NLP technologies. The project will take as example the American-Caribbean creoles, largely spread over the Caribbean area and considered as a cultural vehicle, but not yet an official language except for Haiti. This language has not yet attained normalization, as it lacks sufficient lexicographical resources and a real political strategy.

We will present in this paper the first steps to setup a dictionary of American-Caribbean Creoles, mirroring the effective use of this language in web corpus. It will use the up-to-date Natural language processing tools developed to study the major languages. We will first present the existing lexicographical resources for this language, then will detail the main steps of the project: compilation of existing and available dictionaries, use of web corpora to complete and tune the lexicographical resources.

2. Existing Lexicographical Resources in French-based Creole

In this section, we will present the main existing lexicographical resources in French-based Creoles. Our goal is twofold: first to study how lexicographers have dealt with the specific Creole situation, in terms of macro and microstructure, as well as in terms of number and nature of words included; second, to explicit which dictionaries can be reused for numerical purposes, in terms of availability, copyrights and ease of compilation.

This section is naturally divided into two parts: the first one will detail the main historical dictionaries, in paper format; the second one will explicit electronic resources. It will conclude by identifying the resources that we can use.

2.1 Paper-based dictionaries

The first lexicographical works date from the end of the 19th century and were mainly represented by bilingual lexicons.

Specific Creoles had focused attention such as the Haitian Creole and other varieties of Caribbean Creole languages (in Guadeloupe or Martinique). Albert Valdman, Annegret Bollée, Robert Chaudenson, Hector Poullet and Raphaël Confiant are the precursors of the lexicographical development on Creole languages. They produced several dictionaries that contributed to the lexicographical description. In 1885, Lafcadio Hearn compiled hundreds of proverbs of various types of Creole languages with various cultural information. A few decades later, research teams coming from several universities initiated ambitious projects whose achievements would become the major reference in Creole languages lexicography. First, lexicographers have focused on Haitian and Indian Ocean Creole languages, as more resources were available. Then, Caribbean Creoles, with the works of Hector Poullet, Sylviane Telchid and Raphaël Confiant have been developed.

These various attempts are characterized by the disparity of macro and microstructures. The first difference resides in orthographic conventions: some authors remain closer to French while others recommend a system allowing to break the speech continuum with its parent language, due to the process of decreolization. Other differences in microstructure are more common in lexicography, depending on the description objectives (insertion of spelling variants, available examples, translation of these examples in other languages, parts of speech, origin of entry words, etc.). We must also point out specific Creole languages descriptive elements, which have been specified by (Hazaël-Massieux, 2002): presence of false friends ("fig" = figue), erroneous and uncontrolled diversion (*chokatif - chokativité*), creation by erroneous integration of French words (*abònman* / abonnement).

Macrostructure's main problems are also noticed by the same author: treatment of diatopic variation (all the spelling variants should be specified); demarcation, sometimes complicated, between French-based Creole and its parent language; and selection of the technical and scientific vocabulary. The lack of monolingual[1] works also constitutes one of the most considerable lack in Creole languages.

The main etymology dictionaries: DECA and DECOI

The main lexicographical resources in French-based Creoles derive from two projects aiming at studying etymology, one dealing with Indian Ocean Creoles (IO => DECOI) and the other American-Caribbean Creoles (AC => DECA). The second project is managed by Annegret Bollée (University of Bamberg in Germany) and Ingrid Neumann-Holzschuch (University of Regensburg), with the collaboration of Dominique Fattier (University of Cergy-Pontoise), inspired from (Chaudenson, 1979).

IO Creoles were first studied because more documentation was available. The project ended in the DECOI, the Etymological Dictionary of Indian Ocean French-based Creoles (1993), managed by Bollée, with the cooperation of Patrice Brasseur, Robert Chaudenson and Jean-Paul Chauveau. Composed of four volumes, it is divided into two parts: the first one devoted to French-originated words and the second to words with other and unknown origins. A large part of the etymological information comes from (Chaudenson, 1974). The last volumes of the dictionary appeared in 2007. The DECA then began. This last work, still in progress, is largely based on Haitian Creole material from Albert Valdman (Creole Institute, University of Indiana); the *Haiti Linguistic Atlas* (HLA), elaborated by (Fattier, 1998) and supervised by Robert Chaudenson, is also one of the sources of the DECA; Félix-Lambert Prudent who prepares a dictionary of Creole from Martinique contributes also with the creation of this dictionary. The purpose of this work is not only to establish etymological indications of Haitian Creole, but also to compare varieties of Creoles, (mainly from Guadeloupe and Reunion regions). One of the most important difficulties is the orthographic variation in Creoles (Allen, 1998). The DECA will also include an electronic version, in TEI format.

2.2 Electronic format dictionaries

The most exhaustive on-line dictionaries are Krengle[2] and Webster's online dictionary[3].

Krengle (approximately 18000 entries) is an Haitian Creole – English dictionary, developed by Eric Kafe, a computational linguist (University of Copenhagen). The English section is connected to Wordnet, so every entry offers definitions and semantic relations such us hyperonyms, hyponyms, synonyms and antonyms, sometimes accompanied with examples. The last update of this resource was made in July 2008. This dictionary can be partly downloaded for free: a list of English - Creole and Creole - English words, with some pairs of sentences in English - Creole.

Webster's online dictionary is an online multilingual Thesaurus offering more than 1200 languages. The Creole part is the result of a revision of the work supervised by the Haitian Creole specialist Noah Porter in 1913. The dictionary can be enriched by users via moderation. This tool has interesting characteristics as it recognizes several varieties of Creole, so that users can point out lexical resemblances and dissemblances in the same family of Creole languages. In addition, it includes synonyms and refers to other lexicographical sites.

Other lexicographical databases are available on the web, offering less information than aforementioned, but easier to retrieve, such as glossaries and lexicons of words of frequent use, which partially supplied information in the database. See (Carrión, 2011) for more details.

3. Project Architecture

This quick overview of existing lexicographical resources show that research and development are still in its infancy, and far from exhaustive, whereas our goal in this project is to build an electronic dictionary covering most of the general-purpose vocabulary, as well as identifying spelling variants and an environment to maintain it through continuous corpora analysis. As a result we have setup an architecture enabling to build up and maintain dictionaries through corpus analysis and an iterative process between dictionary and corpora, as described in figure 1.

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.

[1] There is only a monolingual dictionary in Mauritian Creole, elaborated by Arnaud Carpooran: *Diksioner Morisien*, Koleksion Text Kreol, Ile Maurice, 2009.

[2] http://www.krengle.net

[3] http://www.websters-online-dictionary.org/browse/

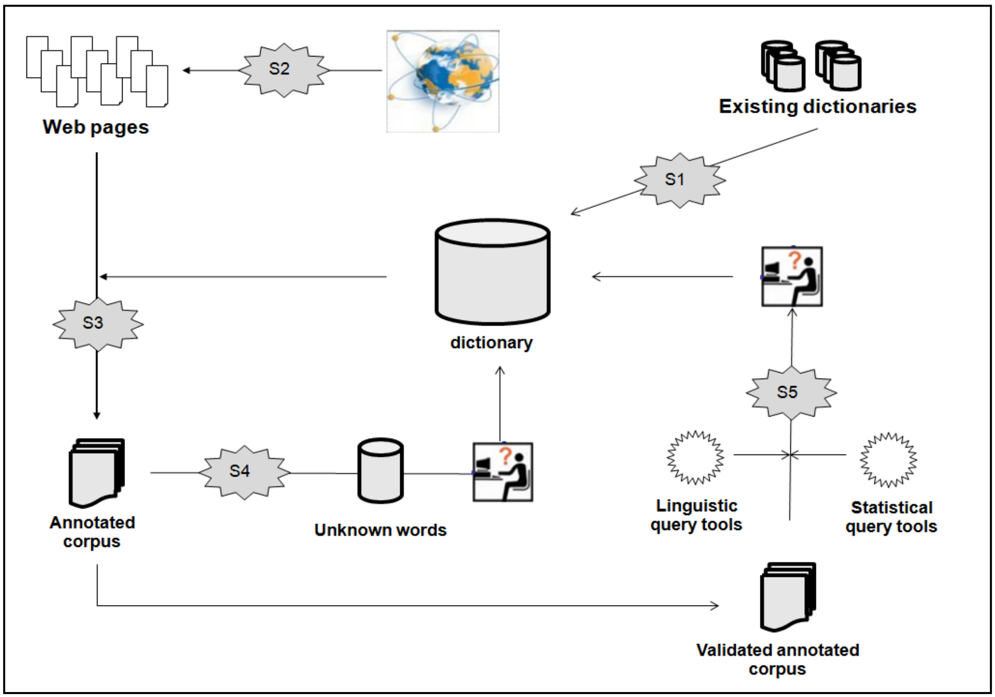

Figure 1 : project architecture

4. Dictionary building: existing resources compilation

Among existing resources, only a few ones are electronically available. Among these, two cases: electronically-based online dictionaries, digitized available dictionaries, originally designed for the paper format. We will detail the main steps used to deal with the data.

4.1. Electronically-based online dictionaries

One of the main problems of on-line lexicographical databases derives from their educational purpose, for most of them. These tools are often reduced to small lexicons which don't offer enough linguistic information (mostly only the entry and its definition or translation). We have retrieved around 2000 entries from these resources:

Lexilogos[4] (161 entries): this site is connected which other Creole languages sites and includes a glossary of Caribbean Creole. Examples and variations are available as the main information of every entry, but no morphological marks on inflected marks are included.

Ecrit créole[5] (90 entries): a Caribbean Creole lexicon which offers the translation or the definition for each entry, but no morphosyntactic or inflected information.

Pédagogie [6] (250 entries): a lexicon from Reunion compiled from Jean Albany's « P'tit glossaire » (Ed. Hi-Land O.I.), Jules Bénard's « Petit Glossaire Créole » (Ed. Alizées) and « Dictionnaire illustré de la Réunion ». Its spelling is based on French language.

Petit lexique créole antillais [7] (107 entries): the definition or the translation of every entry is the only available information.

Antan Lontan[8] (124 entries): created by Marie-Andrée Blameble, this Caribbean Creole lexicon contains only a definition or a translation for each entry. Examples are sometimes available.

Dictionnaire créole[9] (477 entries): this lexicon contains word-forms from Haiti, Guadeloupe, Martinique and Reunion Creoles. It does not offer morphological information, examples or references. However, several meanings are often indicated for one entry.

Choubouloute[10] (60 entries): small list of Creole words from Martinique. Sometimes, spelling variants of the entry are included, but no other linguistic information.

Potomitan [11] (614 entries): Raphaël Confiant's lexicon from Martinique. It only includes the translation of every entry.

Créole Réunionnais[12] (112 entries): small lexicon from Reunion which refers to an on-line dictionary / translator containing altogether 5168 words and Creole expressions.

4.2. Paper-based / digitized dictionaries

Although less numerous, these digitized resources contain much more linguistic information. They are generally bilingual resources that require more complex data processing.

[4] http://www.lexilogos.com/creole_langue_dictionnaires.htm
[5] http://ecrit.creole.free.fr/lexique.html
[6] http://pedagogie2.ac-reunion.fr/clglasaline/Disciplines/Creole/lexiquecreole.htm
[7] http://www.ieeff.org/creole.html
[8] http://antanlontan.chez-alice.fr/motscreo.htm
[9] http://www.dictionnaire-creole.com/
[10] http://www.choubouloute.fr/Lexique-Creole.html
[11] http://www.potomitan.info/dictionnaire/francais.php
[12] http://www.mi-aime-a-ou.com/le_creole_reunionnais.htm

Kwéyòl Dictionary[13] (about 4000 entries): this bilingual lexicon (creole-english / english –creole), focuses on the Saint Lucia Creole. Several meanings are indicated, as well as phrases containing the entry. For each entry is indicated: word part of speech, examples in Creole and their English translation, spelling variants, cross-references, semantic relations (synonyms, antonyms) and etymology. This dictionary microstructure has been used as initial model for the database development.

English Creole Dictionary [14] (7500 entries): this bilingual dictionary (english - creole / creole – english) is bsed on the translation of the Bible in Haitian Creole. The microstructure only contains the translation of each lexical entry; the word part of speech is sparsely specified.

Haitian Creole-English Dictionary[15] (8221 entries): this bilingual dictionary of Haitian Creole -English offers a rich microstructure: word part of speech, examples in Creole and English, spelling variants, semantic relations, etymology of the word, polysemy indications, phrases and sometimes other indications (pejorative, familiar word, euphemism, etc.)

Petit Lexique du Créole Haïtien[16] (408 entries): the poet Emmanuel Védrine includes as main information for each lexical entry the syllable division, the translation or definition in French, examples in Creole and French, the word part of speech, spelling variants, semantic relations and sometimes, word etymology.

As a result, we have downloaded more than 20 000 entries from these on-line or digitized resources, with the following microstructure: part of speech, examples in Creole and their English translation, spelling variants, cross-references, semantic relations (synonyms, antonyms) and etymology, source dictionary.

5. Corpora to improve the dictionary

Web Corpora to help lexicographical studies has been a thorough trend since the advent of internet (see for example Kilgariff, 2003; Baroni, 2009). Some systems have also focused on web corpora for minority languages (see for example Scannel, 2007)

As mentioned in the architecture, 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) and extracting unknown words as a first step. Clearly, iteration of this process will end up with a more exhaustive dictionary, following the Zipf's law.

In this section, we will detail the steps of this iterative process: corpora downloading, automatic analysis, and dictionary improvement. A last point will detail an on-going development that will enable to maintain the dictionary by scanning on a regular basis the web corpus.

5.1. Corpora Building and Feeding

Corpora has been mostly built from Haitian Creole resources, as other varieties are not yet sufficiently represented on the web. We have scheduled three main corpora: a first corpus, limited to 1 million words, will be setup to tune the morphosyntactical analysis and the dictionary improvement process, with in mind the general-purpose language; a second one uses the whole translated Bible, so as to complete and tune the dictionary with a specific vocabulary; the last experiment builds an evolving corpus, so as to maintain the dictionary and setup an adequate environment for lexicographers.

5.1.1. First and second corpora

The first corpus consists in 15 different sources, either manually or automatically downloaded. This corpus aims at rendering as much as possible the varieties of French-based Creole (see (Carrión, 2011 for details on this corpus). The documents all belong to the "general-purpose language". One part has been manually downloaded from websites, whereas automatic download has been setup for two newspapers (« Voa News 2 » and « Alter Presse 3 », both containing Haïtian creole). This corpus finally consists of 1 208 862 words.

The second corpus, the whole Bible translated in Haitian Creole, has been automatically downloaded from the BibleGateway [17] with httrack, with a filter on the webpages link, each containing the keyword "HCV" (*Haitian Creole Version*). It consists of about 4 million words.

Retrieval and cleaning of web pages

The automatic download has been done with Httrack, an open-source software. After the download, it was necessary to convert html files into a text format; this step comprises three different tasks: identification of the encoding and conversion into UTF-8 if applicable, identification and retrieval of the textual zone of interest into the html page, conversion from html to txt. These tasks have been solved by perl scripts, using some previous done work and state-of-the-art techniques (Cartier, 2007, 2009).

Morphosyntactic analysis of the corpus, word frequency of unknown words

The aim of this step is to improve the dictionary by analyzing the given corpus and identifying unknown words sorted by frequency. A Perl program has been used for this task (see Cartier, 2007 for details). From the annotated first and second corpora, we have generated statistics as follows:

Corpus	Recognized words	Unknown tokens	Unknown words / Unique words
Corpus 1(1 208	631984	576878	345653 (28,6%)

[13] www.linguafranka.net/saintluciancreole/dictionary/dictionaryfrontmatter.pdf

[14] www.ngohaiti.com/disaster/downloads/creoledictionary.pdf (*English-Creole Dictionary / Kreyòl-Angle Diksyonè*, publié par Eastern Digital Ressources, 2005)

[15] www.dunwoodypress.com/148/PDF/HCED_sample.pdf

[16] http://www.potomitan.info/vedrine/lexique.pdf

[17] http://www.biblegateway.com

862 tokens)	(52,3%)	(47,7%)	/ 243 892 (70,55%)
Corpus 2 (4 505 442 tokens)	3722628 (82,6%)	782814 (17,4%)	343657 (7,6%) / 336 896 (98,03%)

Table 1: corpora statistics

The following remarks apply to these figures:

1/ The first outstanding element concerns the quick lexicographic coverage of the dictionary: whereas the first corpus consists of 28,6% unknown words, with about 70% unique words, these figures fall down with the second corpus (where these unknown words are integrated in the dictionary) to 7,6%, with about 98% of unique words; this is a confirmation of Zipf’s law (see Manning et al, 1999) : about 20% of words represent about 90% of word occurrences, and about 80% of words represent about 10% of occurrences; the consequence is also that whereas a relatively small corpus enable to cover 90% of lexicographic entries, only a really huge corpus enable to tend to 100% coverage.

2/ Dictionary coverage : at the end of the two processes, the dictionary is composed of 123245 unique words-forms; this is congruent with the dictionary coverage of main language, considering that Creole language is not morphologically rich; nevertheless, this coverage has to be tuned with the fact that our processes do not recognize phrases, whereas they represent at least 50% of the vocabulary (see Sag et al, 2002, for example); it has also to be tuned with the fact that our processes integrate unknown words without linguistic information, as the processes has been automated. Finally, spelling variants have still to be gathered.

Analysis of unknown words

Analysis of unknown words is just a first step of our process. In fact, to really connecting dictionary to corpora, it would be necessary to have automatic procedures to track the meaning evolutions, rather than word forms existence. This step is presented in the next section.

Unknown words retrieved from corpus belong to various categories: words from other language (specifically from English, in our case), misspelled words, proper names, specific notations, real unknown words. Clearly, we have to remove all but the last category. This first filtering results in the figures of the third column, table 1. 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.

5.1.2. A live-corpus and an environment for lexicographers

The experiments done have enabled to complete the dictionary substantially. But our goal is not only to complete the dictionary but also to maintain it, that is check continuously the life of words: emerging, stabilization, meaning changes, disappearance. Towards this goal, we have glued existing environments dealing with corpora handling. Among existing systems, we have retained two systems: SketchEngine (Kilgariff, 2004) and the IMS Workbench (Christ, 1994; Evert, 2011); the first one is certainly the most complete, as it proposes a web crawler and several statistical and linguistic tools to search the corpora; but its main drawback is that it is not freely available. The second environment is also really interesting because it is free and it uses one of the most powerful Linguistic search tools: CQP. But it does not include any tool to retrieve the web nor tools to convert it to the environment internal format.
As a result, we have decided to combine various tools available: a customized web crawler to retrieve corpora; Textbox for conversion to XML and morphosyntactical analysis, mwetoolkit to generate statistics and CWB to search for the corpus, as well as CQPWeb to have a web-based graphical environment for lexicographers. As this project is on-going, we will not detail it in this paper.

6. Conclusion

This paper has presented an on-going project whose goals are: 1/ to explicit a methodology to improve NLP and linguistics development and research for minority-languages, focusing on the American-Caribbean creoles; 2/ to setup procedures and finally an environment for lexicographers to store and maintain lexicographical data from existing resources and web corpora.

It is also important to specify that this project would provide not only a normalizing educational tool, but a translation tool that may be of great help for “mixed” literatures translation and understanding.

According to the general architecture of the project, we have first compiled existing lexicographical resources, either paper or electronically-based; this step permitted to gather about 20 000 lexicographical entries, but exhibited complex-to-solve sparsity problems, as quality, quantity diverge from one resource to another, and macro and micro-structures are far from unified. We have then build a Part-of-Speech tagger from this resource and, using web corpora, have begun to complete the dictionary essentially from unknown words. This step has revealed to be a good procedure, and has to be continued with a live corpus, so as to attain the Zipf's law limit. Finally, we have begun to setup a web-based environment to maintain the dictionary and study lexicographical phenomena through several iterative processes: morphosyntactical analysis and retrieving of unknown words; continuous downloading of web pages; statistical measures over the corpora. 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.

In the near future, we have in mind two main tasks: inclusion of the main existing dictionary in the field, the DECA; a theoretical study of the microstructure for the dictionary. We essentially hope that this contribution will help minority-languages to be more considered and studied, through NLP procedures already in use for the major languages, and crucial tools for lexicographers and language practitioners.

7. BIBLIOGRAPHY

Allen, J. (1998) *Lexical variation in Haitian Creole and orthographic issues for Machine Translation (MT) and Optical Character Recognition (OCR) applications.* First Workshop on Embedded Machine Translation systems of the Association for Machine Translation in the Americas (AMTA), Philadelphia, 28 octobre 1998.

Baker P. And Hookoomsing V. Y. (1987) *Diksyoner kreol morisyen: morisyen-English-français*, l'Harmattan

Baroni M. Et Bernardini S. (2004), “BootCaT: Bootstrapping Corpora and Terms from the Web”,in *Proceedings of LREC 2004,*Lisbon, Portugal.

Baroni M. Et Kilgarriff A. (2006), “Large linguistically-processed Web corpora for multiple languages”, in *Proceedings of the 11th EACL Conference*,Trento, Italy.

Baroni M., Kilgarriff, Pomikálek, Rychlý (2006), WebBootCaT: a web tool for instant corpora, *Proc. Euralex*. Torino, Italy.

Baroni, M. , Bernardini S., Ferraresi A. And Zanchetta E.. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*43(3): 209-226

Baroni, M. And Bernardini, S. (Eds.) (2006). Wacky! *Working papers on the Web as Corpus*. Bologna:

Bollee A., Brasseur P., Chaudenson R. Et Chauveau J-P. (1993) *Dictionnaire étymologique des créoles français de l'Océan Indien*, Hamburg: H. Buske

Bollee, A. (1993) *Dictionnaire étymologique des créoles français de l'Océan Indien*, Hamburg: H. Buske

Bollee, A. (2005). Lexicographie créole: problèmes et perspectives, *Revue française de linguistique appliquée* (Vol. X), p. 53-63.

Carpooran, A. (2009). *Diksioner Morisien* (version intégrale), 1017.

Carrion, P. (2011*) Le traitement automatique des langues créoles à base lexicale française*, Master Dissertation (TILDE – Traitement Automatique et Linguistique des Documents Ecrits), Villetaneuse: Université de Paris 13, 2011.

Cartier E. (2007) "TextBox, a Written Corpus Tool for Linguistic Analysis". In FAIRON Cédrick, NAETS Hubert, KILGARRIFF Adam, DE SCHRYVER Gilles-Maurice, (eds), *Building and Exploring Web Corpora (WAC3 - 2007), Cahiers du CENTAL* 4, pp. 33-42. Presses universitaires de Louvain. Louvain-la-Neuve.

Cartier E. (2009) "Corpus for linguistic resources building and maintenance (CLRBM): system architecture and first experiments”, in *5th Corpus Linguistics 2009*, 20-23 juillet 2009, Liverpool

Chaudenson, R (1974) *Le lexique du parler créole de la Réunion*, 2 tomes. Paris: Champion, 1974.

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. In *Papers in Computational Lexicography (COMPLEX '94)*, pages 22–32, Budapest, Hungary.

Colson, J.-P. (2010a). The Contribution of Web-based Corpus Linguistics to a Global Theory of Phraseology. In: Ptashnyk, S., Hallsteindóttir, E. & N. Bubenhofer (eds.), *Corpora, Web and Databases. Computer-Based Methods in Modern Phraselogy and Lexicography*. Hohengehren, Schneider Verlag, p. 23-35.

Colson, J.-P. (2010b). Automatic extraction of collocations: a new Web-based method. In: S. Bolasco, S., Chiari, I. & L. Giuliano, *Proceedings of JADT 2010,Statistical Analysis of Textual Data*, Sapienza University of Rome, 9-11 June 2010. Milan: LED Edizioni, p. 397-408.

Confiant, R. (2007) *Dictionnaire créole martiniquais-français,* Editions Ibis rouge

Crosbie P., Frank D., Leon E. Et Samuel P. (2001). *Kwéyòl Dictionary*, Castries (Sainte-Lucie), St. Lucia Ministry of Education - SIL International, 2001

Evert, S. And Hardie, A. (2011). *Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium*. Presentation at Corpus Linguistics 2011, University of Birmingham, UK.

Faaß ET AL. (2010). Gertrud Faaß, Ulrich Heid, and Helmut Schmid. Design and application of a Gold Standard for morphological analysis: SMOR in validation. In *Proceedings of the seventh LREC conference* , pages 803 – 810, Valetta, Malta, May 19 – 21 2010. European Language Resources Association (ELRA).

Fattier D. (1998). *Contribution à l'étude de la genèse d'un créole : L'Atlas Linguistique d'Haïti, cartes et commentaires.* Lille, ANRT, collection « thèse à la carte », 6 volumes. 3300p.

Fattier, D. (2007) *Le Projet de Dictionnaire Etymologique des Créoles Français d'Amérique (DECA)*, Université de Cergy-Pontoise, Séminaire du 12 octobre 2007

Ferraresi A. (2007) *Building a very large corpus of English obtained by Web crawling: ukWaC*. Master Thesis, University of Bologna

Ferraresi A., Bernardini S., Picci G. And Baroni M. (2010) “Web Corpora for Bilingual Lexicography: A Pilot Study of English/French Collocation Extraction and Translation”. In Xiao, R. (ed.*) Using Corpora in Contrastive and Translation Studies*. Newcastle: Cambridge Scholars Publishing.

Ferraresi A., Zanchetta E., Baroni M. And Bernardini S. (2008) Introducing and evaluating ukWaC, a very large web-derived corpus of English. In S. Evert, A. Kilgarriff and S. Sharoff (eds.) *Proceedings of the 4th Web as Corpus Workshop (WAC-4) – Can we beat Google?*, Marrakech, 1 June 2008.

Frank D., Crosbie P., Leon E. Et Samuel P. (2001) *Kwéyòl Dictionary*, Castries (Sainte-Lucie)

Hazaël-Massieux, M.-C. (2002). *Prolégomènes à une néologie créole,* en RFLA, 2002

Hearn, L.(1885). *Little dictionary of Creole proverbs, selected from 6 Creole dialects*, translated into French and into English, with notes, complete index to subjects and some brief remarks upon the Creole idioms of Louisiana

Kilgarriff A. (2001), "Web as corpus", in *Proceedings of the Corpus Linguistics 2001* Conference, Lancaster University : 342–344.

Kilgarriff A. Et Grefenstette G. (2003), "Introduction to the Special Issue on the Web as Corpus", in *Computational Linguistics*, no 3, vol. 29.

Kilgarriff A., Rychly P., Smrz P., Tugwell D. (2004) The Sketch Engine. *Proc EURALEX 2004*, Lorient, France; Pp 105-116, (http://www.sketchengine.co.uk)

Manning C. D., Schütze H. (1999) *Foundations of Statistical Natural Language Processing*, MIT Press (1999), p. 24

Mondesir, J. E. (1992) *Dictionary of St. Lucian Creole*, Mouton de Gruyter, Berlin, Allemagne

Poullet H. Et Telchid S. (1990) *Le créole sans peine (guadeloupéen),* Assimil

Poullet H. Et Telchid S. (1999) *Le créole guadeloupéen de poche*, Assimil

Poullet H., Telchid S. Et Montbrand D. (1984) *Dictionnaire des expressions du créole guadeloupéen*, Hatier-Antilles

Rychly P. (2008). A Lexicographer-Friendly Association Score. Proc. 2nd Workshop on *Recent Advances in Slavonic Natural Languages Processing*, RASLAN 2008. Eds Sojka P., Horák A. prvni. Brno : Masaryk University.

Scannel, K. (2007), The Crúbadán Project: Corpus building for under-resourced languages, Cahiers du Cental 4 (2007), pp5-15, C. Fairon, H. Naets, A. Kilgarriff, G-M de Schryver, eds., "*Building and Exploring Web Corpora", Proceedings of the 3rd Web as Corpus Workshop in Louvain-la-Neuve*, Belgium, September 2007.

Sharoff S. (2006a), "Creating General-Purpose Corpora Using Automated Search Engine Queries", in Baroni M. & Bernardini S. (Eds), *WaCky! Working Papers on the Web as Corpus*, GEDIT, Bologna.

Sharoff S. (2006b), "Open-source corpora: Using the net to fish for linguistic data", in *International Journal of Corpus Linguistics,* no 4, vol. 11.

Targete J., Urciolo R. G. (1993) *Haitian Creole – English Dictionary with Basic English – Haitian Creole Append*, dp Dunwoody Press, Kensington, Maryland, U.S.A.

Telchid S., Poullet R.Et Anciaux F. (2009*). Le Déterville: dictionnaire français-créole,* PLB éditions

Tourneux H., Barbotin M. Et Tancons M.-H. (2009) *Dictionnaire pratique du créole de Guadeloupe: Marie-Galante: suivi d'un index français-créole*, éd. Karthala

Valdman A., Pooser C. Et Rosevelt J.-B. (1996) *A Learner's Dictionary of Haitian Creole*, Creole Institute, Indiana University, Bloomington, USA

Valdman A., Yoder S.., Roberts C.. Et Yoseph Y. (1981) *English-French dictionary,* Indiana University, Creole institute

Valdman, A. (2007) *Haitian Creole-English Bilingual Dictionary*, Indiana University, Creole Institute

Vedrine, E. W. (2005) *Petit Lexique du Créole Haïtien*, Orèsjozèf Publications, Boston, Massachusetts (USA), 2nd. ed