Issues from corpus analysis that have influenced the on-going development of various Haitian Creole text- and speech-based NLP systems and applications

Marilyn Mason

Mason Integrated Technologies Ltd (MIT2)
P.O. Box 181015, Boston, Massachusetts 02118 USA
MIT2USA@aol.com

Abstract

This paper describes issues that are relevant to using small- to large-sized corpora for the training and testing of various text- and speech-based natural language processing (NLP) systems for minority and vernacular languages. These R&D and commercial systems and applications include machine translation, orthography conversion, optical character recognition, speech recognition, and speech synthesis that have already been produced for the Haitian Creole (HC) language. Few corpora for minority and vernacular languages have been created specifically for language resource distribution and for NLP system training. As a result, some of the only available corpora are those that are produced within real end-user environments. It is therefore of utmost importance that written language standards be created and then observed so that research on various text- and speech-based systems can be fruitful. In doing so, this also provides vernacular and minority languages with the opportunity to have an impact within the globalization and advanced communication needs efforts of the modern day world. Such technologies can significantly influence the status of these languages, yet the lack of standardization is a severe impediment to technological development. A number of relevant issues are discussed in this paper.

1. Introduction

A certain number of extra-linguistic factors, including sociolinguistic and psycholinguistic variables, come into play for ensuring that robust, user-independent systems can be functional and usable in real end-user environments within the context of vernacular language and minority language situations. One issue concerns the standard vs. non-standard status of any given language and its significance with respect to the development of such systems for minority and vernacular languages. Another key issue is how to deal with high levels of intra-textual and inter-textual linguistic variation that seem to permeate the entire lexicon of such languages, especially during their transition phase from being only an oral language to becoming an oral and written language. We demonstrate the idea that standardizing the written lexicon for these languages is a critical step for implementing written language in publications-based environments (e.g., authoring, newspaper, translation, public relations, marketing, etc). We present tangible and concrete results of R&D efforts over the past few years that have been conducted on human language technology (HLT) applications for Haitian Creole. In addition to indicating research-oriented systems (Brown, 1998; Lenzo et al., 1998; Decrozant and Voss, 1999; Hogan, 1998) that have been developed for HC, we will also present the Mason Method of Haitian Creole Orthography Conversion (MMHCOC) (Mason, 1991; Mason, 1999) that has been developed and is being put to market by a commercial language consulting service. We show the linguistic and extra-linguistic points that have been encountered by academic and commercial development teams for HC and how these issues can impede the development of systems for this language, for other minority and vernacular languages (Baker et al; Somers, 1998), and potentially for other lesser-commonly taught languages (Ostler, 1999).

2. Necessary Steps to Text Processing

Many language technology tools have been created for the purpose of improving and speeding up the communication process, often including both authoring and translation aspects of a documentation process workflow. Since standardization is necessary at multiple linguistic levels, spanning from the general vocabulary individual lexical item, to the multiple-word technical term, to the phrase, to the clause, to the sentence, and possibly beyond, it is necessary that the standardization process be correctly, coherently and thoroughly applied across the data that is used/created. One of the most common and practical by-products of natural language processing (NLP) R&D, whether it be academic, corporate or industrial, is the standardization of the lexicon. For some cases, this might only be a monolingual lexicon, whereas for others it may be a multilingual terminology databank consisting of two or three to a few dozen languages. Lexicon clean-up and validation is often conducted at both the general vocabulary and the domain-specific vocabulary levels. When focusing on technical documentation written in one of the international languages, the compilation and standardization processes for the common core general vocabulary is a rather straightforward process for what concerns spelling. A more difficult issue is checking the different meanings of the individual verbs, nouns, adjectives, etc, and how they already function in an existing domain. Yet, unlike international languages, the process of standardizing and consistently using the same orthography/spelling systems and lexical items within the written code of a vernacular language is not a trivial matter. Given that the majority of the world's languages are in fact vernacular languages, this issue has significant impact on future efforts of preparing many of the world's minority and vernacular languages to be adequately processed via the very techniques (i.e., Controlled Language, Machine Translation, Translation Memory, Authoring Memory, Text-to-Speech, Speech Dictation, Speech Translation, etc) that are demonstrated at this conference.

In light of these technologies and applications, we are persuaded that the successful integration of any NLP system into a work environment is dependent upon sociolinguistic and psycholinguistic variables; these factors in the modern computer science field have simply often been renamed, such as Human Factors (HF) and Human-Computer Interaction (HCI). Such extra-linguistic factors in 'vernacular' languages (e.g., Haitian Creole,

which is our test case) must be considered when attempting to provide automated processing techniques for languages in which linguistic variation permeates the entire lexicon of the language.

Within NLP, it is necessary to apply some of the principles that have come out of the fields of sociolinguistics and language planning, namely the distinction that is made between the standardization of an orthography and the normalization of its use for those who wish to write in a given language. It has been noted that there are many vernacular languages that are currently undergoing stages of standardization (Tabouret-Keller et al. 1997, p. 6). There is a distinction between the stage of standardization of the language (determining what forms should be used) and normalization of the language (implementing what has been decided): standardization is the decision-making process and normalization is putting it into practice. We refer readers to the literature on this topic for a much longer and in-depth discussion of issues at hand for the standardization and normalization of the Haitian Creole (abbreviated as HC) language (Dejean, 1977; Valdman, 1978: 349-350; Valdman, 1988: 76; Allen, 1998; Allen, 1999). Although the thrust of the language 'standardization' process of HC in Haiti has taken place over a period of several decades, that thrust unfortunately has mainly focused on 'orthographic standardization'. In essence, the orthographic issues of standardization were more or less resolved in the late 1970s and early 1980s with the creation of the 'official' Institut Pédagogique National (IPN) orthography (Bernard, 1980; Dejean, 1989: 46; Valdman, 1988: 77). Yet, the reality of the matter is that over many decades "in Haiti, there have often been two or more competing orthographies in the same territory, varying mainly in their representation of nasalized vowels" (Baker, 1997: 120). Also to be taken into consideration is that "as of 1980 eleven proposed spelling systems could be identified" (Schiefflin & Doucet, 1992: 431) for HC. This does not include the dozen attempts -- at least the known ones -- at orthographic standardization that have resulted in the creation of many hybrid spelling systems for this language. Of all of the orthographies developed, the IPN orthography is the 'official' orthography and is consequently the most widely accepted for HC today. However, there is no guarantee 1) that all present-day texts follow the same orthography, or 2) that the HC written language will naturally and automatically pass through the stage of wider-use normalization whereby the lexicon standardizes itself in written form. Standardization of the lexicon, and not simply just of the orthography, is therefore a crucial issue in what concerns the use of the written form of the language in all potential areas influenced by authoring, publishing, translation, etc.

3. Lexical Variation in a Minority Language

Allen & Hogan (1998) have provided detailed frequency counts on variation found for 27 HC lexical items within texts collected from 13 independent sources. Their initial study on variation in HC spelling is limited only to the context of nasalization. A few examples, provided below, are taken from that study (Allen & Hogan, 1998: 1-2).

Frequency	*Written form*	*speech-to-text phonetic interpretation based on phoneme /grapheme rules:*

(1) The word for "enemy"

457	lènmi	{lEnmi}
2	lènnmi	{lEn:mi}
9	lenmi	{le~mi}
5	lennmi	{le~nmi}
9	ènmi	{Enmi}
6	enmi	{e~mi}
7	ennmi	{e~nmi}

(2) The word for "week"

295	semèn	{semEn}
11	semènn	{semEn:}
20	semen	{seme~}
28	semenn	{seme~n}
2	senmenn	{se~me~n}

(3) The word for "government"

10	gouvèman	{guvEma~}
8	gouvèmnan	{guvEmna~}
7	gouvènmam	{guvEnmam}
924	gouvènman	{guvEnma~}
5	gouvènnman	{guvEn:ma~}
20	gouvenman	{guve~ma~}

These authors and their colleagues have also conducted complementary analyses in which hundreds of additional examples demonstrate a high level of variation in the HC lexicon, whereas there is a significantly less amount of variation in English (Allen & Hogan, 1998; Allen 1999; Allen & Hogan, 1999; Hogan & Allen, 1999; Eskenazi, Hogan, Allen & Frederking, 1998).

Ken Decker (1994: section 3.2) states that in "B[elize] C[reole] texts, I have often found the same word spelled different ways in the same text, or even the same sentence." Pierre-Louis Mangeard (personal e-mail communication, 15 October 1998), speaking of Reunion Creole, indicates that "la variation graphique atteint ici 100 % des unites lexicales" (our translation: every lexical item of [the language] has instances of graphemic variation). It has been clearly shown (Allen & Hogan, 1998; Allen, 1998; Allen, 1999) that variation in HC spelling for the same lexical items has not only been

found to be 'inter-textual' (i.e., between the many different editorial teams writing in Creole), which is something probably to be expected, but also that variation is very frequent at the 'intra-textual' level (i.e., within the same texts produced by the same editorial team). This tendency toward a high level of lexical variation in publishing and authoring environments has led us to consider how to develop automated authoring tools for vernacular and minority languages as has been raised in several other articles (Ostler, 1999; Baker et al. 1998; Somers, 1998).

Yet, variation in a traditionally oral language is not something new but is rather a well-known fact. Current speech-data-collection projects for German (Burger & Schiel, 1998) and Spanish (Moreno et al., 1998), which happen to be normalized written and spoken languages, indicate that local dialects of these languages provide much information on phonetic variation. However, when written language is heavily influenced by the spoken dialect, the issue at hand from a computational perspective is how to deal with a significant amount of lexical variation found in written data of a non-normalized vernacular language.

4. Types of Systematic Variation in Haitian Creole

Despite the high level of variation in HC, it is quite systematic and computationally usable when compiled into machine-readable form (Allen, 1998; Hogan, 1998; Lenzo et al, 1998; Allen & Hogan, 1998; Allen & Hogan 1999; Hogan & Allen, 1999; Mason, 1991a; Mason, 1991b; Mason, 1994; Mason, 1998; Mason, 1999a; Mason, 1999b; Mason, 1999c). The examples of variation given below are fully backed with supportive evidence of numerous linguistic studies as described in Allen (1998) and Allen (1999).

A) alternation in vowel height
   (The alternation in vowel height primarily concerns:
   1) variation between the front mid-high e [e] and front mid-low è [E] vowels; and
   2) variation between the back mid-high o [o] and back mid-low ò [O] vowels)

B) alternation between the velar fricative **r** and the labial glide **w**

C) nasalization (oral vs. nasalized vowels)

Algorithms can be appropriately configured for HC as has been done by both the DIPLOMAT project of Carnegie Mellon University (http://www.lti.cs.cmu.edu/Research/Diplomat) and Mason Integrated Technologies, Ltd (http://hometown.aol.com/mit2usa/_Index2.html). Yet, as recognized in CL efforts mentioned during panel discussions and question/answers sessions at CLAW98, it is necessary to obtain approval and ownership of CL from the users in order for the implementation task to be truly successful. User ownership of the process, and not just forced acceptance of it resulting from tools developed by outsider groups, is a significant issue to consider for appropriate implementation. This is extremely important for minority languages, as stated by Mason (1999d), since decisions on language choice should be made by appropriate powers (i.e., native-speaking HC representatives from the community of linguists, educators and government officials) with regard to the standardized forms to be used.

A discussion of the pros and cons of phonemic orthographies with respect to beginning and advanced literacy campaigns is described in Allen (1998) and Allen (1999) for the case of written HC. The data provided in the present paper and in many of the articles mentioned above simply infer that some level of psycholinguistic influence is involved when literate (in HC) Haitians produce written texts in HC; this is based on statistical analysis of a 1.2 million word HC database written by 13 independent authoring teams. We believe that there are certainly psycholinguistic issues at work, most likely influenced largely by issues such as 1) very low funding, 2) written language acquisition at a late age, and 3) lack of adequate educational support. These issues, and their direct correlation with resulting lexical variation in written HC, require a significant amount of additional (ethno-) linguistic research that has never yet been undertaken. If funding is obtained for research on the psycholinguistic and sociolinguistic issues, we would hope to participate in such research, but we clearly state that our current research does not investigate these issues and does not necessarily determine all of the reasons for such variation, because this sociolinguistic issue is widespread in languages across the globe. We are thus currently concerned with the fact that there is a high level of variation of lexical forms in written texts. This variation clearly indicates that there is a need to develop an automated process to assist native speakers of vernacular languages in processing the text (including the multiple orthographic forms) that they must work with on a daily basis in various computer applications (e.g., OCR software, MT systems, spell checkers, text-to-speech synthesis) which are designed to decipher and analyze written text.

5. Computational Work on Vernacular Languages

The numerous linguistic and extra-linguistic issues described above present a myriad of factors that have been considered in the DIPLOMAT project at Carnegie Mellon University (http://www.lti.cs.cmu.edu/Research/Diplomat) on which there has been testing of Example-based Machine Translation, Optical Character Recognition (OCR), speech synthesis, and speech recognition systems for HC and English. During a significant database-compilation stage (from November 1996 through September 1998) to collect speech and written data for HC, the issue of lexical variation became very obvious (Allen, 1998; Hogan, 1998; Allen & Hogan, 1998; Allen, 1999; Hogan & Allen, 1999; Allen & Hogan, 1999; Hogan, 1999). Projects like DIPLOMAT that aim at developing rapid-implementation and rapid-deployment speech MT systems are unable to wait around for a language to standardize and normalize. For these cases, such a "system must know at least the inventory of the [Source Language] SL's parts of speech, their orthographical (e.g., variant spellings), morphological (e.g., word formation), and syntactic (e.g., subcategorization) properties" (Nirenburg, 1998: 740). It is clear that variant orthographic issues are common to all projects for vernacular languages. Other projects also confirm (Aduriz *et al.*, 1998: 821; Camara *et al.*, 1995:

section 3.2) that lexical variation at both oral and written levels is a significant issue for minority and vernacular languages.

6. A New Solution for Lexical Standardization via Orthography Conversion

Although research has been conducted by several institutes on how to process minority and vernacular language written text, no academic research project has thus far produced a usable, functional, tool for end-user native speakers of these types of languages. On the other hand, outside of academia a software program has been in the making for nearly two decades (Mason 1991; Mason 1994; Mason 1998) and is now just making its public appearance (Mason 1999a-d; Mason 2000a-c). The prototype of a flexible, semi-automated process for converting texts written in earlier HC orthographies to conform to the Institut Pédagogique National (IPN) orthography (i.e., the legal standard established by the Orthography Law of 1979) was initially completed in 1991 (Mason, June 18, 1991). The algorithm is based on the 1979 Law that established a core of fixed phonemic-to-graphemic rules along with a set of other rules for the use of apostrophes, hyphens, contractions, punctuation, capitalization, proper names, and nasalization.

Like all NLP systems, a benchmark test was conducted (back in 1991) in order to validate this system. The benchmark for this orthography conversion process was based on the Bible in HC which is one of the largest corpora in existence for HC (Allen & Hogan, 1998; Allen & Hogan, 1999; Mason, 2000c). Other reasons for the choice of this text are explained in Mason (2000, forthcoming). Using copies of the digitized HC Bible texts, the initial experimentation process was based on a "character matching" approach of speed editing away from older orthographies toward IPN -- using the standard editing tools of "over-the-counter" word processing software. The prototype allowed one to convert one orthography to another (ie, Pressoir-Faublas text to IPN text, IPN text to McConnell text, etc). Examples of the conversion of Pressoir-Faublas text to IPN text can be found in Figures 1 and 2.

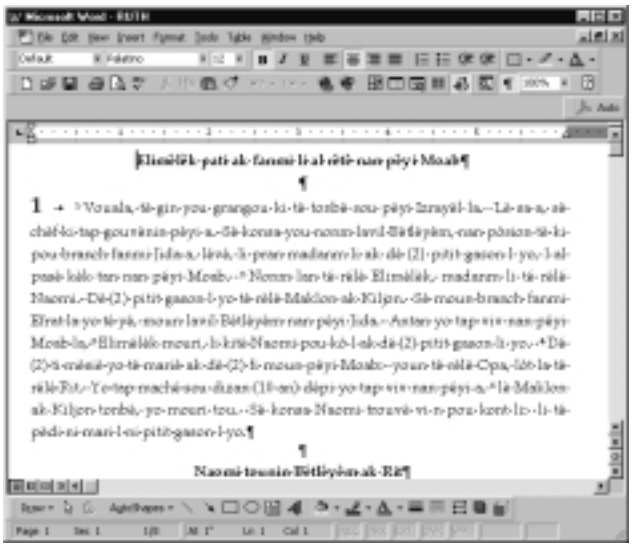

Figure 1. Text in Pressoir-Faublas orthography BEFORE orthography conversion

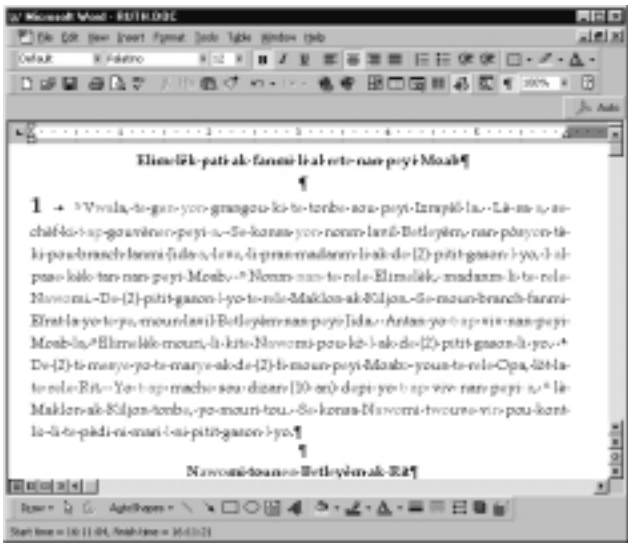

Figure 2. Text in IPN orthography AFTER orthography conversion

Over a number of years of development, the process matured from a semi-automated process taking 2 hours to convert a 250-page book, to a fully-automated process requiring less than 2 1/2 minutes to convert that same 250-page book.

The second step was testing the program by expanding the research to include new HC texts that had not been used to train the system. As a result of scanning, with the use of customized optical character recognition (OCR) software (Mason, June 28, 1991), printed texts of varying age and print quality produced by other writers and authoring teams were used to test the software program. As a result, CreoleConvert™ (Mason, 1999d) has been developed and used to automatically and successfully convert the outdated orthographies of samples from periodicals such as *Boukan, Jé Nou Louvri* and *Chanmòt la* to the IPN orthography (samples available at http://hometown.aol.com/mit2haiti/Index4.html).

7. Going Beyond Vernacular Languages to International Languages

Even beyond vernacular languages, we know that some international languages have recently undergone orthographic modifications. Note the orthography reforms listed below:

German:
http://www.sfs.nphil.uni-tuebingen.de/linguist/issues/7/7-455.html
http://www.mailbase.ac.uk/lists/german-studies/1999-02/0005.html
http://www.intl.com/english/faq/v5a10.htm

Dutch:
http://www.kun.nl/celex/subsecs/section_spell.html

Norwegian:
http://simba-s.online.no/ud/nornytt/uda-315.html
http://www.dokpro.uio.no/sprakrad/engord.html

Swedish:
http://www.deeptrans.com/deeptrans/swedish.html

Greenlandic:
http://buchholdt.com/EbenHopson/czm/1977cz/Feb1977/index.html
http://www.buchholdt.com/EbenHopson/icc/ICCBooklet.html).

Spanish:
http://www.xlation.com/mailing-lists/xr/Nov1999/doc00002.doc

Even a reform proposal was established for the French language in the early 1990s, but these efforts did not end up being successful. For all languages that are currently undergoing such written orthography standardization processes, it is now possible to integrate a cross-platform system that will conduct the conversion process. This software is fully compatible with Microsoft Word and is Y2K compliant. The interface is completely user-friendly and is already available for integration and implementation within contexts that could use such software. Orthographic and lexicon standardization are the base elements for correctly implementing texts for various purposes in an authoring/translation workflow environment. Mason Integrated Technologies Ltd offers the technology that can provide international as well as minority languages with consistent and coherent lexical and orthographic standardization strategies in view of being optimally coherent for authoring and translation processing tasks.

8. Conclusion

Lexical standardization is one of the base issues for working with language resource data in all types of NLP environments. For some of the international languages, such standardization has been achieved over time and with the recent help of integrated spelling checkers in Microsoft Word and other applications. Yet, we note that even the international languages are less coherent and stable than would first appear. The majority of the world's languages, being minority and vernacular languages, have not been able to benefit from such advantages of the modern technological world. Through the efforts of Mason Integrated Technologies Ltd, it is now possible for many of the world's languages to achieve lexical standardization at the written level with the use of existing and upcoming corpora. By applying these technologies to the standardization of corpora, the next step would be to produce additional multilingual documentation technologies and corpus products for minority languages. However, if techniques are not developed and implemented to provide for something as simple as lexical standardization and spell-checking, these minority languages of today and tomorrow will suffer greatly and will be unable to meet the needs of the authoring and translation sectors that are so critical in a modern world of globalization.

9. References

ADURIZ, Itziar, IZASKUN ALDEZABAL, Olatz Ansa, XABIER ARTOLA, Arantza Días de Ilarraza, and Jon Mikel INSAUSTI. 1998. EDBL: a Multi-Purposed Lexical Support for the Treatment of Basque. In *Proceedings of the First International Conference on Language Resources and Evaluation*, 28-30 May 1998, Granada, Spain. Vol. 2, pp. 821-826.

ALLEN, Jeffrey. 1998. Lexical variation in Haitian Creole and orthographic issues for Machine Translation (MT) and Optical Character Recognition (OCR) applications. Paper presented at the workshop on Embedded MT systems of the Association for Machine Translation in the Americas (AMTA) conference, Philadelphia, 28 October 1998.

ALLEN, Jeffrey. 1999. La standardisation du créole haïtien par l'intermédiaire de la linguistique computationnelle. Paper presented at the "Orthography: Between Myth and Reality" Workshop at the 9è Colloque du Comité International des Études Créoles. Held at the Université de Provence, Aix-en-Provence, France, 24 - 29 June 1999.

ALLEN, Jeffrey and Christopher HOGAN. 1998. Evaluating Haitian Creole orthographies from a non-literacy-based perspective. Paper presented at the annual meeting of the Society for Pidgin and Creole Linguistics, New York City, 9-10 January 1998.

ALLEN, Jeffrey and Christopher HOGAN. 1999. Le 'r' et le 'w' en créole haïtien: 1, 2 ou 3 phonemes? Paper presented at the 9è Colloque du Comité International des Études Créoles. Held at the Université de Provence, Aix-en-Provence, France, 24 - 29 June 1999.

BAKER, Paul, MCENERY, Tony, SEBBA, Mark, and Lou BURNARD. 1998. Minority Language Engineering. In ELRA Newsletter, Vol. 3 N4, Nov 1998, p. 10.

BAKER, Philip. 1997. Developing Ways of Writing Vernaculars: Problems and Solutions in a Historical Perspective. In Tabouret-Keller *et al. Vernacular Literacy: A Re-evaluation.* pp. 93-141.

BERNARD, Joseph. 1980. Ki Jan Mou Ekri Kreyòl Ayisyen [Reprint of communiqué on HC's official orthography]. Études Créoles 3.1: 101-105.

BROWN, R. 1998. Improving Embedded Machine Translation with User Interaction. Paper presented at the workshop on Embedded MT systems of the Association for Machine Translation in the Americas (AMTA) conference, Philadelphia, 28 October 1998.

BURGER, Susanne and Florian SCHIEL. 1998. RVG 1 – A Database for Regional Variants of Contemporary German. In *Proceedings of the First International Conference on Language Resources and Evaluation*, 28-30 May 1998, Granada, Spain. Vol. 2, pp. 1083-1087.

CAMARA, Émile, CÉLESTIN NSTADI, Véronique Rey, and Jean VÉRONIS. December 1995. Traitement Informatique des Langues Africaines: Problèmes et Perspectives. Action de Recherche Partagée. ALAF (AUPELF-UREF) Document ALAF ALA1. Version 1.0. http://www.lpl.univ-aix.fr/projects/alaf/ALA1.html.

DECKER, Ken. 1996. "Orthography Development for Belize Creole." In 1994 Mid-America Linguistics Conference Papers, Volume II, edited by Frances Ingemann. Lawrence, Kansas: The University of Kansas. pp. 351-362.

DECROZANT, L. and C. VOSS. 1999. Building a 'Tri-Text': Steps in the Conversion of a Hard Copy Document to an On-line Resource. In ELRA Newsletter. Vol 4 issue 1; January 1999, Paris: European Language Resources Association. pp. 10-11.

DEJEAN, Yves. 1977. *Comment Ecrire le Créole d'Haïti.* Ph.D. Thesis. Indiana University.

ESKENAZI, Maxine, HOGAN, Christopher, ALLEN, Jeffrey, and Robert FREDERKING. 1998. Issues in database design: recording and processing speech from new populations (poster session). In Proceedings of the First International Conference on Language Resources and Evaluation, 28-30 May 1998, Granada, Spain. Vol. 2, pp. 1289-1293.

HOGAN, Christopher. 1998. Embedded Spelling Correction for OCR with an Application to Minority Languages. Paper presented at the workshop on Embedded MT systems of the Association for Machine Translation in the Americas (AMTA) conference, Philadelphia, 28 October 1998.

HOGAN, Christopher. 1999. "OCR for Minority Languages". In Proceedings of the 1999 Symposium on Document Image Understanding Technology, Annapolis, Maryland, April 1999, pp. 235--244.

HOGAN, Christopher and Jeffrey ALLEN. 1999. Phonemic and Orthographic realizations of 'r' and 'w' in Haitian Creole. Paper presented at the International Conference of the Phonetic Sciences (ICPhS) 99, San Francisco, 1-7 August, 1999.

KEPHART, Ronald. 1985. "It Have More Soft Words": A study of Creole English and Reading in Carriacou, Grenada. University Microfilms.

KEPHART, Ronald. 1992. Reading Creole English Does Not Destroy Your Brain Cells! In *Pidgins, Creoles, and Non-standard Dialects in Education*, edited by Jeff Siegel. Applied Linguistics Association of Australia.

LENZO, Kevin, HOGAN, Christopher, and Jeffrey ALLEN. 1998. Rapid-Deployment Text-to-Speech in the DIPLOMAT System. Poster presented at the International Conference on Spoken Language Processing. 30 November - 4 December 1998, Sydney, Australia.

MANGEARD, Pierre-Louis. 15 October 1998. Personal e-mail communication with Jeff Allen.

MASON, Marilyn. 1991a. "Novel Method for Orthography Conversion in Haitian Creole" (June 18, 1991). Unpublished internal Technical Report.

MASON, Marilyn. 1991b. "Optical Character Recognition (OCR) Technology Widens Impact of Mason Method of Haitian Creole Orthography Conversion (MMHCOC)" (June 28, 1991). Unpublished internal Technical Report.

MASON, Marilyn. 1994. "Story behind Color Coded Mason Method of Haitian Creole Orthography Conversion (CCMMHCOC)" (May 3, 1994). Unpublished internal Technical Report.

MASON, Marilyn. 1998. Automated Approach to Haitian Creole Orthography Conversion. Paper presented at the Fourth International Creole Language Workshop: "Standardizing the Orthography, Vocabulary and Structure", Florida International University, Miami, FL, March 19-21, 1998.

MASON, Marilyn. 1999a. Automated Approach to Haitian Creole Orthography Conversion: Can This Methodology Be Adapted to Other Creoles? Paper presented at the "Orthography: Between Myth and Reality" Workshop at the 9è Colloque du Comité International des Études Créoles. Held at the Université de Provence, Aix-en-Provence, France, 24 - 29 June 1999.

MASON, Marilyn. 1999b. Orthography Standardisation Tools: Preparing Creole Languages for the New Millennium. Paper and demo presented at the Seychelles '99 Creole Symposium, Mahé, Seychelles, October 26-28, 1999.

MASON, Marilyn. 1999c. Kreol + Computers + Internet = A Bright Future for Kreol! Paper and demo presented at the 14th Annual Creole Festival, Mahé, Seychelles, October 23-31, 1999.

MASON, Marilyn. 1999d. "Orthographic Conversion and Lexical Standardization for Vernacular Languages". In *ELRA Newsletter*, Volume 4, Number 4, October-December 1999, pp. 5-7. Paris: European Language Resources Association (ELRA).

MASON, Marilyn. 2000a. "Authoring and Documentation Workflow Tools for Haitian Creole - a Minority Language", *Technical Communicators' (TC) Forum Magazine*, volume 1-2000 (January-March 2000), in press.

MASON, Marilyn. 2000b. "Spelling issues for Haitian Creole Authoring and Translation Workflow", *International Journal for Language and Documentation*, Volume 4, March 2000, in press.

MASON, Marilyn. 2000c. Automated creole orthography conversion. In *Journal for Pidgin and Creole Languages*, 15:1, April 2000, forthcoming.

MASON, Marilyn. 2000d. "One small step for language technologies, one big step for implementing Creole language technologies into real-user environments." Paper to be presented at the 5th International Creole Language Workshop "How the past can improve the future: Creole languages in the new millennium" to be held at Florida International University, Miami, FL, March 30-April 1, 2000.

MORENO, Asunción, HARALD HÖGE, Joachim Koechler, and José MARIÑO. 1998. SpeechDat Across Latin America: Project SALA. In *Proceedings of the First International Conference on Language Resources and Evaluation*, 28-30 May 1998, Granada, Spain. Vol. 1, pp. 367-370.

NIRENBURG, Sergei. 1998. Project Boas: "A Linguist in the Box" as a Multi-Purpose Language Resource. In *Proceedings of the First International Conference on Language Resources and Evaluation*, 28-30 May 1998, Granada, Spain. Vol. 2, pp. 739-746.

OSTLER, Nicholas. 1999. Does Size Matter? Language Technology and the Smaller Language. ELRA Newsletter, Vol 4 N1. Jan-Mar 1999. Paris: European Language Resources Association.

SCHIEFFLIN, Bambi and Rachelle Charlier DOUCET. 1992. The 'Real' Haitian Creole: Metalinguistics and Orthographic Choice. In *Pragmatics* 2:3, pp. 427-443.

SOMERS, Harold. Language Resources and Minority Languages. In *Language Today*, Number 5, 1998. Nottingham, UK: Language Publications Ltd., pp. 20-24.

TABOURET-KELLER, Andrée, LE PAGE, Robert, GARDNER-CHLOROS, Penelope, and Garbrielle VARRO (eds). 1997. *Vernacular Literacy: A Re-evaluation*. Oxford: Clarendon Press.

VALDMAN, Albert. 1978. *Le Créole: Structure, Statut et Origine*. Paris. Editions Klincksieck.

VALDMAN, Albert. 1988. Diglossia and Language Conflict in Haiti. In *International Journal for the Sociology of Language*, 71, pp. 67-80.