PHONEMIC AND ORTHOGRAPHIC REALIZATIONS OF 'R' AND 'W' IN HAITIAN CREOLE

Christopher Hogan* and Jeffrey Allen†
** Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA*
† European Language Resources Association, Paris, France

ABSTRACT

This paper presents a synchronic perspective on the phonemic status of the orthographic forms 'r' and 'w' that appear in Haitian Creole (HC) texts. A survey of existing linguistic literature is presented on the status of 'r' and 'w' in HC. Other HC language researchers have postulated two phonemes (*i.e.*, /r/ and /w/) with roundness/labialization conditioning the various allophones of /r/. Such evidence, however, is consistently contradicted by the orthographic representations of these phonemes in written corpora produced by numerous independent sources. Our analyses take into account the variation in HC found in extensive written corpora and in spoken data that we have compiled for this language. From this work, we aim to determine the status and distribution of the related phonemes and phonetic realizations in HC. Our findings have considerable bearing on speech recognition and speech synthesis systems that are currently under development for Haitian Creole.

1. BACKGROUND

This study is based upon two significant aspects of the Haitian Creole DIPLOMAT project[1] undertaken at Carnegie Mellon University from November 1996 to November 1998. The DIPLOMAT project is an experiment in rapid-deployment, wearable, bi-directional speech translation systems. An implementation of the complete recognition and synthesis systems should be available at an acceptable level of quality within a few weeks after initial recording corpus design, with continual, graceful improvement to a good level of quality over a period of months. Our present analyses are limited to a study of the written and spoken forms of 'r' and 'w' in Haitian Creole within the framework of this project. Written forms are taken from 15 independent sources of texts including newspapers, books, the Bible, training manuals, Internet pages, published speeches, *etc.*; an exhaustive list of source texts and reference materials used in our work can be found in [3] and [2]. Speech data used in our current analyses has been randomly selected from digitally recorded read speech of 150 Haitian Creole speakers from recording sessions conducted in Port-au-Prince (Haiti), Paris (France), New York City, and Pittsburgh (Pennsylvania). An exact account of the procedures followed for corpus design, recording conditions, quality control, pre-/post-processing, and literacy level assumptions in recording new populations, as well as resulting statistics, have been thoroughly discussed in [7] and [9]. Our study compares the variation of alternation between orthographic forms 'r' and 'w' in the written texts with phonetic realizations found in our read speech data.

2. PREVIOUS LINGUISTIC WORK

2.1. Phonetic and phonemic analyses

With respect to the phonetic units [w], [r], [ɣ], [r ^{w}], and [ɣ^{w}] that are found in the speech of HC speakers, Tinelli [12] indicates that r, "instead of the IPA symbol ɣ for typographical reasons" is a velar or post-palatal continuant. Valdman [13] and Hazaël-Massieux [8] both state that the occurrence of [w] and [r] in Haitian Creole is the result of progressive and regressive conditioning by labial and non-labial vowels. Valdman, provides several examples (*e.g.*, /ri/ vs. /wi/; /rete/ vs. /wete/; /sekre/ vs. /sekwe/) that represent the typical case of lexico-semantic differences through the use of minimal pairs. Orjala [10] and Dejean [6], however, both state that the same lexical items reveal the replacement of /r/ by /w/.

2.2. Socio-psycholinguistic factors

It has been briefly noted in [3] and then more substantially in [2] that a significant number of psycholinguistic factors are involved when literate (in HC) Haitians produce written texts in HC, thus resulting in extensive variation of lexical items in texts produced by different Haitians. A complete set of literacy statistics are provided in [2]. The variation in written texts may have to do with the "psycholinguistics of literacy campaigns done in a rush and with extra-low funding, and with the difficulties involved learning an orthography at a late age and/or with little educational background/support" [5].

We do not wish to discuss this point in detail in this paper but rather we clearly acknowledge that lexical variation in written HC may be due to psycho-socio-linguistic processes at work among Haitians who write in HC.

3. DATA

Our datasets of written data, provided below, substantiate the claims that orthographic 'r' and 'w' can be written interchangeably and indicate that the generalization of 'w' appearing before rounded vowels (*i.e.*, /o/, /u/, /ɔ/) and 'r' appearing before non-rounded vowels (*i.e.*, /a/, /ɛ/, /e/, /i/) is not an adequate description. Based on frequency counts on lexical items in a 1+ million word HC on-line corpus, our preliminary analyses of the written texts are then correlated with spoken data that have been collected in Haiti.

3.1. Written data

Frequency counts on written forms of individual items are given below. Note that they come from the entire database and are not

specific to an author or team of authors. The number of occurrences precedes the written form.

Written digram sequences: ‘ri’ vs. ‘wi’

The word for "firm/enterprise"
14 antrepriz
29 antrepwiz

Written digram sequences: ‘re’ vs. ‘we’

The word for "pressure"
30 presyon
58 pwesyon

Written digram sequences: ‘rè’ vs. 'wè’

The word for “congress”
148 kongrè
27 kongwè

Written digram sequences: ‘ro’ vs. ‘wo’
The word for “next”

11 prochen
144 pwochen

Written digram sequences: ‘rò’ vs. ‘wò’

The word for “too much”

10 trò
3 tròp
170 twò
94 twòp

The word for “rock/boulder”

423 wòch
36 ròch

Written digram sequences: ‘ra’ vs. ‘wa’

The words for “democracy, democratic, and democrat”

144 demokrasi
108 demokratik
21 demokrat

14 demokwasi
5 demokwatik
4 demokwat

The examples provided in the datasets above clearly show, at least in consonant clusters where the phoneme in question is in second position (*i.e.*, ##C_V), that the roundness of the following vowel is not an accurate measure for determining labialness in written forms of words. This written corpus has led us to look more closely at the phonetic realizations of recorded utterances in our database.

3.2. Spoken data

The data for this analysis was drawn from the DIPLOMAT corpus of Haitian Creole recorded speech. From the original corpus, we selected only the data that was recorded in Haiti, at the *Université Caraïbe*, in order to eliminate possible English influence. This corpus has a total of 3566 utterances with 6329 instances of orthographic ‘r’ and 4142 instances of orthographic ‘w’. From this, a sample of 197 utterances was selected, representing 99 distinct utterances and 10 speakers. Phone-level alignments with the speech signal were produced automatically [4], and each instance was carefully transcribed by hand.

Because orthographic ‘w’ unambiguously represented [w] in our corpus, we focussed our efforts on instances of orthographic ‘r’. Of these, there were 424 instances in the 197 utterances. In this analysis, we are primarily interested in whether the speaker produced the bilabial liquid [w] or the velar fricative [ɣ] for each instance of orthographic ‘r’. High-level statistics for each of the speakers are summarized in the table. The column labeled ‘other’ includes [l], [ʔ] and ∅ (deletion).

Speaker	Sex	[ɣ]	[w]	Other	Total
BH	F	42	6		48
BK	M	22	4		26
BN	F	32	15	2	49
BU	M	32	10	3	45
BV	M	30	14	1	45
CE	M	28	14		42
DJ	F	32	9	4	45
EC	M	17	8	1	26
ED	F	40	10		50
EE	M	38	8	2	48
Total		313	98	1	424

Before moving on, let us consider one of the existing analyses, and examine the conclusions we can draw about it based on our data. Consider the analysis that claims that /r/ will be realized as [w] if and only if it is followed by a rounded vowel /o/, /u/, or /ɔ/. While it is true that every instance of /r/ + [round] is realized as [w][2], the reverse implication is not true: that all instances of [w] are a result of a rounded vowel. Among the 424 instances in our corpus, we find 98 instances (23.1%) which are not followed by a round vowel, and yet are realized as [w]. This is substantially the same as the situation in the written data.

3.2.1. Trailing Context.

Nevertheless, let us follow this lead by looking at phonetic contexts that follow /r/. The table below shows that low segments are more likely to be preceded by the velar, while higher segments seem to admit more ambiguity.

	[ɣ]	[w]
[a]	94%	6%
[ã]	93%	7%
[ɛ]	88%	13%
[ɛ̃]	100%	0%
[e]	79%	21%
[i]	44%	56%
[j]	33%	67%
Total	76%	24%

While lowness is not a perfect predictor of the velar, most of the variation can be accounted for by a preceding /p/, as we shall see.

3.2.2. Preceding Context.

Preceding context also exhibits an interaction with /r/[3]. The following table summarizes the distribution of [ɣ] and [w] with respect to preceding context.

	[ɣ]	[w]		[ɣ]	[w]
[a]	74%	26%	[l]	83%	17%
[ã]	78%	22%	[n]	86%	14%
[ɔ]	83%	17%	[õ]	100%	0%
[b]	20%	80%	[o]	65%	35%
[d]	75%	25%	[p]	57%	43%
[ɛ]	100%	0%	#	100%	0%
[ɛ̃]	100%	0%	[t]	96%	4%
[e]	89%	11%	[u]	28%	72%
[f]	71%	29%	[v]	57%	43%
[g]	100%	0%	[j]	100%	0%
[i]	80%	20%	[z]	100%	0%
[k]	86%	14%	Total	76%	24%

The contexts that deviate most significantly from the average are [b], [o], [p], [u], [v], and possibly [f]. These labials account for a significant number of the occurrences of [w].

3.2.3. Interactions.

In the last two sections, we have shown that there are reasonable rules based on preceding and following contexts that account for some of the data. Much of the variation in these rules can be accounted for by considering the interaction between them.

Consider again the case of trailing context discussed in section 3.2.1. In that case, trailing low vowels predicted [ɣ] fairly reliably, but high vowels, while favoring [w], did so less strongly. Consider again the trailing context: [i][4], which mildly favors [w] in the preceding analysis. The following table lists the preceding contexts in this situation.

	[ɣ]	[w]		[ɣ]	[w]
[a]	50%	50%	[l]	75%	25%
[ã]	67%	33%	[n]	50%	50%
[b]	20%	80%	[o]	50%	50%
[e]	67%	33%	[p]	25%	75%
[f]	13%	88%	[t]	50%	50%
[g]	100%	0%	[u]	21%	79%
[i]	33%	67%	[v]	25%	75%
[k]	79%	21%	Total	44%	56%

There are two things to notice here. First of all, there are now very few contexts which unambiguously favor [ɣ]: only [g] remains sacrosanct. Secondly, the labial contexts noted in section 3.2.2 now favor [w] even more strongly. (*e.g.* [p] has changed from 43% to 75% in favor of [w]). The trailing [i] has influenced the choice of allophone in favor of [w].

3.2.4 Individual Variation

The statistics we have presented in the last section were aggregate over all speakers, and could be caused by variations among speakers (possibly due to dialect). This is not the case. There is as much variation in the utterances of one speaker as in those of the aggregate population. The following table is the same as the previous one, but selected from only one speaker, BH, a female of unknown dialect.

Right context	[ɣ]	[w]
[a]	100%	0%
[ã]	100%	0%
[ɛ]	100%	0%
[e]	88%	12%
[i]	56%	44%

Again it is the low contexts that strongly predict the velar, with substantial variation preceding higher segments.

3.2.5 Sex and Dialect Variation.

As in any language, it is conceivable that some of the variation in the distribution is due to sex or dialect differences. The following table summarizes these differences.

	[ɣ]	[w]
Female	78%	22%
Male	74%	26%
North	71%	29%
West/Central	78%	22%
Unknown	78%	22%
Average	76%	24%

Although the difference between the sexes may be insignificant, the difference can be accounted for by the fact that

the velar is the more prestigious variant, as it is the variant that is closer to the French uvular /R/. As such, it makes sense that females who are more conservative in their speech would make more frequent use of the velar.

The differences between dialects seem to be more significant, with the Northern Dialect exhibiting more instances of [w] than the average. However, this difference has not been noted by dialectologists [10], and may be related instead to rural vs. urban differences.

4. DISCUSSION

Statistical analyses on only the right contexts show that previous claims regarding phonemic status of /r/ and /w/ are not necessarily accurate; this is especially important with regard to the contexts in which /r/ is realized as a velar or as a labial.

These analyses allow us to consider the possibility that orthographic 'r' and 'w' are in actually in free variation in the context that has been studied. It would be necessary to conduct more in-depth analyses that consider preceding and following contexts and word initial contexts.

5. CONCLUSION

In the context of developing rapid-deployment, speech translation systems, especially for speech-only systems, we can temporarily conclude that /r/ and /w/ can be conflated and be considered as a single phoneme. Previous analyses on the issue of /r/ and /w/ have not provided substantial evidence to adequately and quantitatively prove that these are separate phonemes. The proof is certainly inadequate for the implementation of such linguistic analyses within a natural language processing system. Further analyses should study the full set of contexts and should also consider the possibility of /r/, realized as [r] or [h], as being a 'marginal phoneme' as has been suggested in [1] for another French Creole language.

ACKNOWLEDGMENTS

We would like to extend our gratitude to the 105 students of the *Université Caraïbe* in Port-au-Prince who patiently participated in our HC recording project in late March 1997. We would also like to thank all of the external organizations (listed in [2] and [3]) that provided copies of electronic texts in HC for our research. Last, but not least, we would like to thank our team of native HC speakers at Carnegie Mellon University who participated in the compilation, analysis and production of written and spoken data on-site in our laboratory.

NOTES

1. http://www.lti.cs.cmu.edu/Research/Diplomat/
2. Instances of orthographic 'r' followed by a rounded vowel are extremely rare in our corpus, occurring only 17 times out of 6329 instances (0.3%). On the other hand, instances of orthographic 'w' followed by a rounded vowel are fairly common, occurring 1191 times out of 4142 instances (28.8%).
3. The relative strengths of the associations with preceding and following context is unclear. One measure of association, Cramer's V [11], indicates stronger association with preceding context, while another, based on entropy [11], indicates no difference between the two.
4. The other segment that predicts [w] is [j], which only occurs in the sequence /a r j/.

REFERENCES

[1] Allen, J. 1994. *Sainte-Lucie: Relexification, Décreolisation, Recréolisation ou Adlexification?* Diplôme d'Etudes Approfondies Thesis. Université Lyon 2.

[2] Allen, J. 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 in conjunction with the Association for Machine Translation in the Americas (AMTA) conference, Philadelphia, 28 October 1998.

[3] Allen, J. and Hogan, C. 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, 10 January 1998.

[4] Black, A. W. and Lenzo, K. A. 1999. *Building Voices in the Festival Speech Synthesis System: Processes and issues in building speech synthesis voices.* http://www.cs.cmu.edu/~lenzo/papers/festvox/

[5] Degraff, M. 1998. Personal communication.

[6] Dejean, Y. 1977. *Comment écrire le créole d'Haïti?* Ph.D Thesis. Indiana University.

[7] Eskenazi, M., Hogan, C., Allen, J., and Frederking, R. 1998. Issues in database design: recording and processing speech from new populations In *Proceedings of the First International Conference on Language Resources and Evaluation*, 28-30 May 1998, Granada, Spain. Vol. 2, pp. 1289-1293.

[8] Hazaël-Massieux, M-C. 1997. Talk on comparative French creole orthographies presented at the 'Des Créoles de la Caraïbe' International Workshop, Duke University, 10-12 September 1997.

[9] Lenzo, K., Hogan, C., and Allen, J. 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.

[10] Orjala, P. 1970. *A Dialect Survey of Haitian Creole*. Ph.D. Thesis. The Hartford Seminary Foundation.

[11] Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vettterling, W. T. 1986. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, Cambridge, England.

[12] Tinelli, H. 1981. Creole phonology. The Hague: Mouton Publishers.

[13] Valdman, A. 1978. *Le Créole: Structure, Statut et Origine*. Paris. Editions Klincksieck.