Chapter 8

The value of online corpora for the analysis of variation and change in the Caribbean

John R. Rickford
Stanford University

In recent years, corpora have proven to be very powerful tools, impacting not only the field of linguistics, but other fields like Oral Literature, Folklore, History, Anthropology, Sociology, Education, Speech Recognition and Criminal Justice. However, there are still very few large-scale corpora available for research on Caribbean and other Creole settings. This paper reviews six examples that illustrate the multiple contexts in which corpora have proven to be a meaningful instrument for understanding language variation and change. The examples come from theoretical and descriptive areas and from applied areas. Particular focus will be paid to how online corpora can benefit criminal justice. Together the examples discussed here suggest many future ways in which online corpora can serve as an invaluable for furthering research on language variation in on Caribbean communities.

1 Introduction

It's an honor and a pleasure to contribute to this volume in honor of Donald Winford, because he has been at the forefront not only of developments in creole and language contact studies (see his books and articles and his long editorship of the *Journal of Pidgin and Creole Languages*), but also of the study of language variation and change in Anglophone Caribbean varieties and African American Vernacular English (AAVE). His 1972 University of York dissertation, *A sociolinguistic description of two communities in Trinidad*, was a masterpiece, demonstrating the value of quantitative analysis of social stratification and stylistic differentiation

John R. Rickford. 2022. The value of online corpora for the analysis of variation and change in the Caribbean. In Bettina Migge & Shelome Gooden (eds.), *Social and structural aspects of language contact and change*, 213–232. Berlin: Language Science Press. DOI: 10.5281/zenodo.6979333

in St. James and Mayo. And virtually everything he has written since exemplifies the value of excellent fieldwork and careful and insightful analysis. He is also a good friend, and this paper is intended as a tribute and a gift to him.

One of the best ways of honoring Winford and the other pioneers in the study of the Caribbean is not only by looking back, but also by looking forward, creating and making available online corpora of Anglophone Caribbean language varieties (and of course Francophone and other varieties too). In recent years, corpora have proven to be very powerful tools, impacting not only the field of linguistics, but other fields like Oral Literature, Folklore, History, Anthropology, Sociology, Education, Speech Recognition and Criminal Justice. As a result, it would be incredibly important to equip future researchers with this important resource. Six examples in particular (four from theoretical and descriptive areas, two from applied areas) illustrate the multiple contexts in which corpora have proven to be a meaningful instrument for understanding language variation and change, and in turn suggest future ways in which online corpora of Caribbean language could be invaluable. The sixth case I will discuss at some length, because it shows how online corpora can benefit criminal justice and it has not been presented in any other publication to date.

2 Variation in the use of Creole variants in Jamaican popular music, 1962–2011

A corpus was an essential tool in Byron Jones' (2019) examination of the use of Creole variants in Jamaican popular music for his PhD thesis. Jones compiled data for the Corpus of Popular Jamaican Music (COPJAM) and quantitatively analyzed it.[1] Through this corpus, he was able to explore a variety of factors including linguistic variable, gender, genre, decade, theme, and others, which he discusses with very sophisticated quantitative tools in his thesis, *Beyond de riddim: Language use in Jamaican popular music*, University of the West Indies, St. Augustine, Trinidad.[2] The availability of such a corpus allowed him to track Creole variant usage over time, in the language of the top 20 Jamaican songs each decade from 1962–2012, which showed a dramatic *increase* in the frequency of Creole forms across this period from 3.7% in 1962–1971 to 63.4% in 2002–2011. (See Table 1.) This research was especially important for expanding the conversation surrounding language use and attitudes into the domain of music. It dramatically

[1] As far as I know, COPJAM is not available for general or public use.

[2] I was privileged to serve as external examiner on Jones' excellent thesis.

shows how the existence of a corpus can greatly extend the possibilities for the study of linguistic variation and change.

Table 1: Relative frequency of Creole variants in Jamaican songs from 1962–2011. (From table 9, p. 155, Jones 2019)

Decade	% Creole in song lyrics	# of Creole variants out of total
1st: 1962–1971	3.7	111/2,987
2nd: 1972–1981	19.8	692/3,501
3rd: 1982–1991	47.1	2,849/6,045
4th: 1992–2011	56.9	4,181/7,353
5th: 2002–2011	63.4	4,687/7,388

3 Variation and change in the verbal coda of "as far as" noun phrases

Stanford's online "Searcher" corpus of English literature from 1800 to 1959 was similarly helpful in our study of the variable absence and loss of *is/are concerned* or *go(es)* in *as far as* phrases in English (Rickford et al. 1995). For years, as part of interest in relatively understudied syntactic variation, I had been collecting examples like:

(1) *As far as* filling out the details ∅, that isn't a problem.

(2) People think I'm constantly in motion, *as far as* making films ∅.

I was curious about what influenced whether the sentence coda (*goes* or *is concerned*) was omitted and if the omission after the NP was increasing. I asked Tom Wasow about the syntax of the NP after *as far as*, and his interest piqued, he joined the project as well. We added two students to the team – Norma Mendoza-Denton and Julie Espinoza. Our data were from the "Searcher" online corpus and we also elicited intuitions, made new recordings, collected overheard examples, and perused usage manuals, collecting over 1200 *as far as* tokens. We coded 1065 of these tokens for various factors, looking at the NP following "as far as", the mode of communication, and the speakers themselves. The factors that had a statistically significant effect on the observed variation are shown below along with their likelihood of affecting verb absence.

Table 2: VARBRUL (variable rule) weights for significant factors in *as far as* verb absence. Sample size per factor is indicated in parentheses. Originally Table 3 in Rickford et al. (1995), reprinted with the permission of the Linguistic Society of America.

SYNTACTIC COMPLEXITY OF THE NP		
Noun, with or without modifiers	0.31	(679)
Conjoined NPs and NPs with PPs	0.46	(163)
Sentential NPs	0.86	(314)
MODE		
Speech	0.62	(732)
Electronic mail, or written exams	0.33	(322)
Writing (newspaper, articles, books)	0.21	(95)
AGE OF SPEAKER/WRITER		
≤ 19	0.69	(17)
20–39	0.56	(306)
40–59	0.44	(180)
≥ 60	0.24	(31)
PROSODIC STRUCTURE OF THE NP		
Branching	0.57	(682)
Nonbranching	0.40	(483)
SEX OF SPEAKER/WRITER		
Male	0.47	(670)
Female	0.56	(295)
POSITION OF *as far as* PHRASE IN SENTENCE		
Initial	0.54	(550)
Noninitial	0.46	(605)

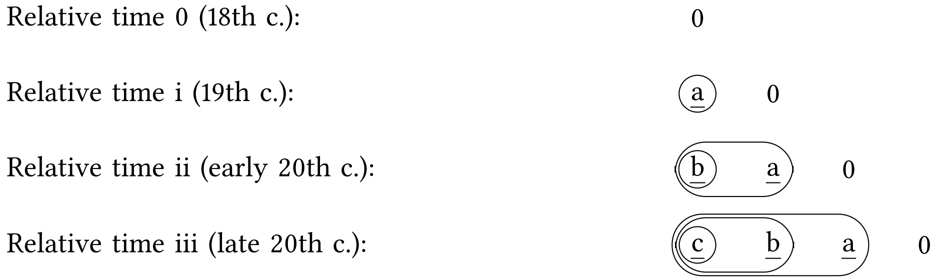
Relative time 0 (18th c.):
0
Relative time i (19th c.):
a
0
Relative time ii (early 20th c.):
b
a
0
Relative time iii (late 20th c.):
c
b
a
0

Figure 1: Spread of the rule deleting the verbal coda in topic-restricting *as far as* constructions, depicted in terms of the Baileyan wave model. (Adapted from Fig. 2 [Bailey 1973: 68], in which, as Bailey's caption notes, 'The letters represent successively later, or lighter-weighted, environments in which the rule operates.'). In our case, a, the earliest environment affected by the rule, refers to sentential N̄Ps, as in ex. (3) above; b, the next environment affected by the rule, is prepositional or conjoined NPs, as in ex. (4) above, and c, the most recent environment affected by the rule, includes simple NPs, as in ex. (5) above. Source: Fig. 5 in (Rickford et al. 1995), reprinted with the permission of the Linguistic Society of America.

The "Searcher" corpus also gave us access to data that allowed us to examine the historical development of the *as far as* NP *be concerned* construction and allowed us to see the historical evidence for wavelike spread of verb absence in *as far as* phrases. According to our corpus data, the verb absence first appeared in *as/so far a*s phrases with a sentential NP, as shown in (3) below. This later spread to *as/so far a*s phrases with prepositional or conjoined NPs as in (4). And it finally spread to *as far as* phrases with simple NPs, as in (5).

(3) And I will own to you, (I am sure it will be safe), that *so far as* our living with Mr. Churchill at Enscombe ∅, it is settled. [1816, Jane Austen, *Emma*, p. 460]

(4) The cabin … was in perfect condition *so far as* frame and covering ∅ until 1868. [1939, Henry Seidel Canby, *Thoreau*]

(5) *As far as* the white servants ∅, it isn't clear. [Renee Blake, 22, 1987, (p.c.)]

This spread to new environments with increasing frequency is depicted in terms of the Baileyan wave model in Figure 1.

The "Searcher" corpus was an invaluable tool in understanding these factors in the variability and change of verb absence in *as far as* phrases, greatly extending the number of tokens and the span of time we were able to analyze.

4 The rise and fall of quotative *all*

Multiple corpora were critical to our study of the rise and fall of quotative *all*, i.e. using *all* instead of *like, go,* or *say* to introduce a quotation (Rickford et al. 2007), as in (6) and (7) below:

(6) He's *all*, "Let me see your license; is that your car?" (Latino Male)

(7) She's *all*, "What do you mean, gum?" (White female)

For this study, we used data from multiple sources and corpora:

- 1990/1994 recordings of native California adolescents & young adults collected by Ann Wimmer (Stanford undergrad senior thesis) & Carmen Fought (Pitzer College, LA area)

- New 2005 recordings of high school & college students from Palo Alto, Stanford, & San Francisco

- A multi-source corpus: examples from conversation, but also from publications (Waksler, *American Speech* 2001), web pages, TV series (*Buffy the Vampire Slayer*) & movies (*Clueless*). Lots of *all* tokens (253 quotatives), but not accountable like recorded corpora (cf. Labov 1972:72), since we did not have corresponding examples of where other variants besides *all* were used.

- **The Google News groups Corpus, 1981–2005**: Billions of words including at least 354 examples of quotative *all.*

This search of corpora yielded some interesting observations. For example, the first time this usage appeared was in 1982. In fact, the first Switchboard Corpus, collected in 1988–1992, and the Santa Barbara Corpus of Spoken American English, part I, collected in 1988, each has only 1 example of quotative *all.* Clearly, the use of quotative *all* was new. In addition, our data from the Google News groups corpus, 1981–2005, suggested that quotative *all* peaked in 1999 and then declined steeply, as shown in Figure 2 (originally figure 4 in Rickford et al. 2007).

This peak and decline was supported by evidence from our other corpora. For example, in the 1990/94 corpus, *all* was the primary quotative introducer (*all* used 46% of the time, *like* 17%, unframed 16%, *say* 11%, Other 8%, and *go* 2%). In contrast, in our new 2005 corpus, *all* was much less frequent as the quotative introducer (4%), overtaken by *like* (69%), and the quotative *all like* had emerged.

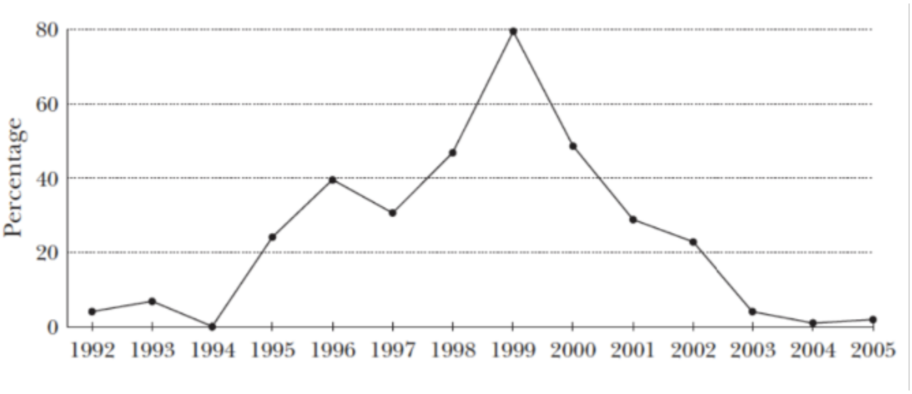

Figure 2: Frequency of quotative *all* over time, normalized for number of posts per year over a composite of very frequent words (*word, other, make, look, write, see, number, way, people, first, the,* and *is*)

Through our corpora, we were also able to look at what factors favored the use of quotative *all* at different points in time. VARBRUL (variable rule) analysis showed that in the 1990/94 corpus, the primary favoring factor is present tense, then Quoted Speech (vs. Thought), then Perseverance (quotative *all* in 5 preceding lines). In the 2005 corpus, tense was not significant, and while Quoted Speech (vs. Thought), still favored quotative *all*, Perseverance *dis*favored it. Once again, the availability of corpora gave us the opportunity to see the feature across time and in a variety of environments, allowing for a more robust understanding of this usage's variation and *change*.

5 FAVE alignment, DARLA, and the *Voices of California* project

These possibilities for linguistic study have expanded even further with the advent of new automatic processing and analysis technologies which make rapid analysis of large corpora achievable. Tools such as FAVE align/extract[3] and DARLA[4] are aligners that allow for audio samples and transcripts to be paired and aligned. With the help of lexicons like the Carnegie Mellon University (CMU) Pronouncing Dictionary,[5] these aligners facilitate the rapid measurement of

[3] https://github.com/JoFrhwld/FAVE/wiki/FAVE-align

[4] https://linguistics.dartmouth.edu/research/darla-dartmouth-linguistic-automation

[5] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

vowels for acoustic characteristics. Measuring acoustic characteristics of vowels which historically was more time consuming and individualized can now be completed on a large-scale, quickly. Put simply, we currently have technology that allows for swift, large-scale linguistic analysis and broader generalizations.

For example, the *Voices of California* (VOC) project,[6] directed by my faculty colleagues Penny Eckert and Rob Podesva at Stanford, has drawn on recordings with more than 1000 speakers across California to look at broad-scale language variation and change. Mengesha (2020) is one recent VOC project that showcases the speed and detail that FAVE aligning and extracting offer for large scale analysis. Looking at FEEL-FILL mergers in Bakersfield and Sacramento, the author analyzed a total of 48 vowel tokens from African Americans, and 330 tokens from whites, force-aligning word list data into word and sound segments using FAVE, and creating PRAAT scripts to take 11 measurements across the rhyme (EEL or ILL). Among other things, Mengesha found that FEEL is lowering over time and becoming more monophthongal among African Americans, particularly among those with college and graduate degrees. FILL is also lowering and becoming more monophthongal among African Americans, also with education and gender effects; African American women maintain a monophthongal FILL while African American men maintain a diphthongal FILL.

There are many similar projects around the world, for instance the Linguistic Data Consortium at the University of Pennsylvania[7] and the Origins of New Zealand Corpus Project (ONZE).[8]

Unfortunately, the study of Caribbean language is way behind this curve, in part because the CMU pronouncing dictionary, which FAVE and DARLA use, is based on Standard Mainstream American English (MUSE). We will need aligners that are more specifically geared to Caribbean English words and pronunciations, whether at basilectal, mesolectal or acrolectal levels. The good news is that University of Pittsburgh Professor Shelome Gooden (2019) and others are trying to solve these problems. Some progress in the area of digital corpora of Caribbean English has been made, for example Dagmar Deuber (Münster) has "made a forced aligner that orients toward either 'neutral' or Trinidadian English" (Lars Hinrichs email 6.16.19). See also the papers by Phillip Meer (2019, 2020), who works with Deuber, on some of the specific challenges of using state of the art aligners with Trinidadian English. However, more work must be done in order to harness this useful technology for the study of Caribbean languages.

[6] http://web.stanford.edu/dept/linguistics/VoCal/index.html

[7] https://www.ldc.upenn.edu

[8] https://www.canterbury.ac.nz/nzilbb/research/onze/

With these new automated tools, analysis of large-scale corpora, provided they exist, will be within reach.

6 Automatic speech recognition by race in US high tech companies

Corpora have also been important in the application of linguistics research to the field of technology. One such example is the impact of research focusing on automated speech recognition by race in U.S. high tech companies. Using digital corpora from online corpora, a team of sound engineers and linguists at Stanford and Georgetown universities (Koenecke et al. 2020) assessed the relative accuracy of automated speech recognizers and transcribers used by Google, Amazon, Apple, and other companies for Black and White speech samples. In order to do this research, we used CORAAL (Corpus of Regional African American Language)[9] for Black speakers and VOC (Voices of California) for White speakers. With these corpora, we were able to analyze 2,141 Black snippets and 2,141 White snippets, with an average length of 17 seconds. Our results indicated that the speech recognition error rate was statistically higher for Black speakers than for white speakers in the speech recognizers for all 5 major tech companies examined, as shown in Figures 3 and 4 (corresponding to figures 1 and 2 in the original). Figure 4 in particular supports the point that we make in the paper: "if one considers a WER of 0.5 to be the bar for a useful transcription, more than 10 times as many snippets of black speakers fail to meet that standard. In this sense, the racial disparities we find are even larger than indicated by the average differences in WER alone."

Additionally, we found that the greater inaccuracy for Black speakers was related to their use of AAVE features. Using a combined phonological and grammatical Dialect Density measure, we found that Black speakers who used more AAVE features were significantly more likely to be mis-transcribed than Black speaker who used fewer such features. This use of corpora illuminates an important next step for tech companies: to train systems more on AAVE and other ethnic dialects so that they are indeed accessible to all.

Similar results have been found by Alicia Beckford Wassink (Wassink 2020) and her colleagues and students at the University of Washington. She reported that an automated transcription service known as CLOx, developed at her university, showed very different error rates for four ethnic dialects of English represented in recordings of sixteen speakers. CLOx is actually an ASR service built on

[9] https://oraal.uoregon.edu/coraal

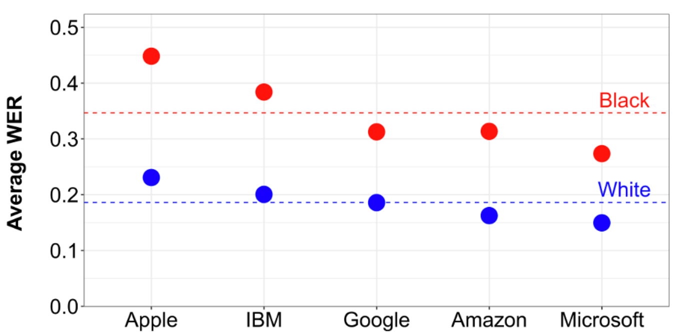

Figure 3: The average Word Error Rate (WER) across Automated Speech Recognition (ASR) services is 0.35 for audio snippets of Black speakers, as opposed to 0.19 for snippets of White speakers. The maximum Standard Error (SE) among the 10 WER values displayed (across Black and White speakers and across ASR services) is 0.005. For each ASR service, the average WER is calculated across a matched sample of 2,141 Black and 2,141 White audio snippets, totalling 19.1 hours of interviewee audio. Nearest-neighbor matching between speaker race was performed based on the speaker's age, gender, and audio snippet duration.

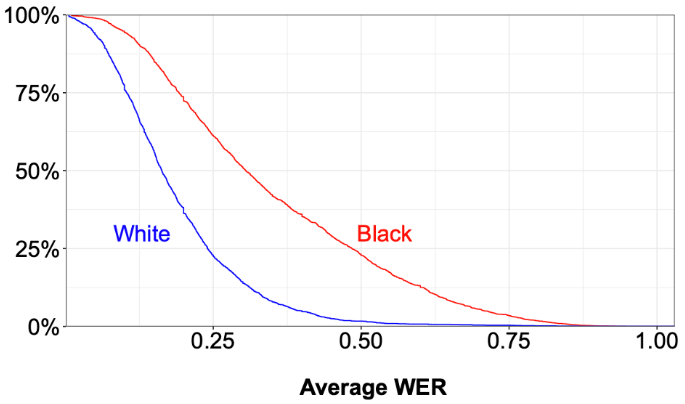

Figure 4: The Complementary Cumulative Distribution Function (CCDF) denotes the share of audio snippets having a WER greater than the value specified along the horizontal axis. The two CCDFs shown for audio snippets by White speakers versus for those by Black speakers use the average WER across the five ASR services tested. If we assume that a WER >0.5 implies a transcript is unusable, then 23% of audio snippets of Black speakers result in unusable transcripts, whereas only 1.6% of audio snippets of White speakers result in unusable transcripts.

the Microsoft Speech development toolkit, so it is using Microsoft's automated speech recognition system. Note that CLOx had 2.7 times more errors when attempting automated transcriptions of speech by African American as it did with speech by Caucasian American speakers, and 5.8 times more errors with ChicanX and Yakama English speakers than with Caucasian American speakers. These results are similar to those shown in Figures 3 and 4, but they show that it's applicable too to other ethnic dialects.

Table 3: Errors in Automated Speech Recognition by CLOx as applied to samples of four ethnic Washington English dialects. nf = number of errors in corpus/total word count in the corpus x 100 words, the base of normalization. (Source: Wassink 2020)

Group	*N*	nf
Caucasian American	6,654	1.5
African American	16,276	4.1
Chicanx	3,986	8.8
Yakama	14,581	8.9

7 Corpora as relevant to criminal justice: *I'm good* 'No (thank you)'

Corpora have also proved to be useful in applied research in another way – the determination of criminal justice. For example, in a criminal case where there was contention over the meaning of *I'm good*, I was asked to provide a deposition as a Linguistics "expert" about its meaning. In this case, a Drug Enforcement Agency (DEA) agent asked two sisters if they would consent to a pat-down. While one of the sisters, Harriet, consented to a body pat-down, Tamika did not, responding *I'm good.* The DEA agent chose to interpret this as a 'yes', and did the body search despite the countervailing evidence – Tamika remained seated, while Harriet had stood up to facilitate the body search – and the prosecution at the trial was actively supporting the DEA agent. To explore the usage of *I'm good* to mean 'No', I assembled several kinds of evidence:

- Crosswords, movies
- Dictionaries (Urban Dictionary, Oxford English Dictionary)

- The Corpus of Contemporary English Usage (COCA), BYU
- Twitter
- Crowd-sourced experiments.

In this paper I'll focus on the online corpora (COCA and a small set of examples from Twitter), since these were essential to establishing compelling evidence that *I'm good* is used unequivocally to mean 'No', and never to mean 'yes' – that is, that Tamika's body search *was done without her consent.* In order to understand how each critically contributed to aiding the defendant, we will consider each individually.

The online Corpus of Contemporary American English [COCA], in 2017 "a more than 450-million-word corpus of American English," provided a window into "more than 560 million words from more than 160,000 texts, including 20 million words from each of the years 1990 through 2017" (*Wikipedia*). The corpus, created and maintained at Brigham Young University in Utah, is an invaluable resource on which scholars studying variation and change in language rely heavily, "used by approximately tens of thousands of people each month, which may make it the most widely used "structured" corpus currently available. For each year, the corpus is evenly divided between the following five genres: spoken, fiction, popular magazines, newspapers, and academic journals."

A search of COCA yielded 330 tokens of *I'm good.* A critical element that a dictionary definition does not specify, but a corpus search was able to clarify, is that the *kind* of question to which *I'm good* is a reply, affects the meaning of *I'm good.* When the question is *How are you?* or *(Are) you good/okay/fine/allright etc.?* or something similar, as in (8–9), an *I'm good* answer assumes its most literal meaning: 'I am in good shape/am feeling fine/doing ok/am content', and so on. Most of the 330 tokens in the IG (*I'm Good*) search were of this type.

(8) # MIRTHA # Hey, George. You okay? # GEORGE # Yeah. I'm fine. *I'm good.* [from *Blow*, a 2001 work of FICTION]

(9) AL ROKER (08:22:40): How are you? DANA-EISEN- (08# 22:41): *I'm good,* Al. How are you? [from, *The Today Show*, 2017, 7:00 am, EST, in the SPOKEN genre]

However, when the question is a "Yes/No" question representing an offer or request, as in egs. (10–11), *I'm good* almost always means 'No', achieving its effect by representing the respondent as satisfied with the way things are, declining the food, drink or service offered or the suggestion made, and so on. There were 80 tokens of this type (*I'm good* 'No, thanks') in the IG COCA search; note that we count repeated instances of *I'm good* in the same extract as different examples, since they need not have been repeated, indeed often are not.

(10) The waitress turned to Charlotte again. "Are you sure I can't get you anything? Maybe an appetizer or a salad?" # "*No, I'm good.* Really." (emphasis added). [from *Love, Honor and Betray*, 2011, in the FICTION genre]

(11) Guilfoyle: You don't want to comment on this, Eric? Bolling: *I'm good.* (CROSSTALK)
Guilfoyle: Right. Who can say it better? Dana? [from Marco Rubio, *The Five*, 5:00 PM EST, 2015, in the SPOKEN genre]

On the basis of these 80 tokens of *I'm good* 'No', several observations pertinent to the legal case were made. With COCA's scope of time and genre, as well as its availability of contextual information, these tokens were able to provide many unique insights with regard to the usage of *I'm good* over time, the contexts in which it appears, and its cooccurrence with the word *No*.

First, using COCA, we found that *the usage of I'm good to mean 'No' has been rapidly increasing in frequency over the past three decades.* From 1990–2005 (a span of 16 years) there are 12 attestations; from 2006–2011 (6 years), there are 29 attestations, and from 2012 to 2017 (6 years) there are 39 attestations. See figure Figure 4. On the evidence of COCA, it was possible to show that the encounter between the Drug Enforcement Agent and defendant Tamika took place in a six-year span at which this usage of *I'm good* was at its peak.

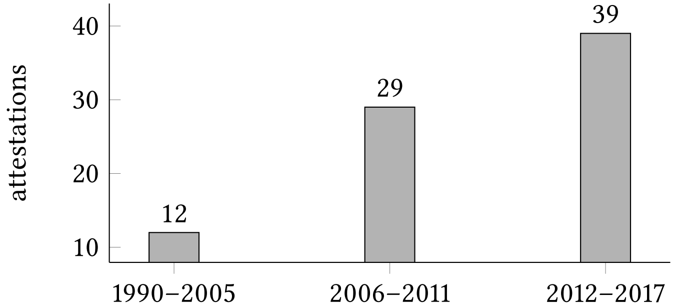

Figure 5: COCA evidence that use of *I'm good* 'No thanks' has increased steadily between 1990 and 2017

Secondly, investigation of the COCA corpus showed that while this usage may have originated in response to offers of food or drink, this is not the only appropriate context for its use. In fact, only 25% (25/80) of the examples of *I'm good* 'No' in the IG COCA search are a response to offers of food or drink. This is important because the prosecution had suggested that the agent's use of *I'm good* was not an appropriate context for signaling 'No' since it was not in a social setting in which food or drink was being offered. However, in light of the COCA evidence, where a great majority of the *I'm good* 'No' examples do *not* involve offers of food or drink, the suggestion that this use of *I'm good* is restricted only to situations involving offers of food or drink is not persuasive.

Thirdly, the COCA corpus helped call into question the Prosecution's claim that Tamika does not choose the familiar and customary phrase of negation, *No, I'm good thanks*, opting significantly to omit the operative word *No* when she says *I'm good.* This statement by the prosecution went against the evidence of the IG COCA corpus that while 54% (43/80) of the attestations of *I'm good* 'No' include a preceding *No* or other explicit negative, as in (10), another 45% (37/80) of them include no explicit negative, as in (11).

Fourthly, while *I'm good* is the primary way of signaling 'No' in the *I'm* ___-_ frame, the COCA corpus showed that *I'm okay*, *I'm fine* and *I'm cool* are other alternatives, sometimes used in conjunction with or in place of *I'm good*:

(12) "Do you want something to eat?" she calls. "*Nah, I'm cool.* I grabbed a burger on the drive." (Jeff P. Jones, Antioch Review, 2015, vol. 73, no. 3, p. 495–512, FICTION genre)

Outside of the COCA corpus, there is an interesting recent TV example in which MSNBC reporter Rachel Maddow reads this segment of the News (exhibit #4):

(13) The individual who asked for the secret filing declined to identify himself or his client and replied "I'm OK" when offered a reporter's business card to remain in touch. (MSNBC 10/25/18, 9:40 pm)

Maddow goes on to perform a mock re-creation/elaboration of the scene, using *I'm OK* 'No':

(14) # "Well, I know you don't want to talk now, but can I give you my card if you should like to talk at some point in the future?" "*Yeah, I'm OK*. I don't need your business card. If I go home with your business card in my pocket, I'll disappear. Ha ha. In fact, I was never even here!" (MSNBC 10/25/18, 9:40 pm)

Note that the presence of *Yeah* in the *I'm good* or *I'm OK*, etc. frame does not, as example (14) shows, necessarily signal assent to the request or offer to which it is a response, especially when an offer or request is repeated, as is frequently the case, and when that request or offer is unpleasant, or potentially threatening the responder's "face," sense of appropriateness, health or freedom. This is certainly the case with the "pat down" that the DEA agent sought permission to do.

Bearing these points in mind, I was able to call into question the observation of the Prosecution in their Amended Prosecution Response that in this exchange, Tamika's response of *I'm cool* to the agent's *Okay?* "gave ultimate completeness to the consent that she gave to his request to search her":

(15) DEA Agent: You give me permission just pat you down?
Tamika: I'm good.
DEA Agent: Okay?
Tamika: I'm cool, yeah.

This is certainly a possible interpretation, but in light of the following example (16) from the COCA corpus, it is also possible to see Tamika reaffirming the *I'm good* 'No' of her first response. The presence of *Yeah* here is most probably a repudiation of her initial negative *I'm good* as it is in example (16) below. And indeed, in context (a second response to a less than pleasant request) this is the more likely interpretation.

(16) JOSH-ELLIOTT-1-AB# (Off-camera) I do want to say, Melissa and I shared a car afterwards, we're going back to the hotel and she says, do you want some Advil? And she'd done it before, and I said, no, you know, I'm tired, I have an empty stomach, *I'm good.* I'll get some later. She said, you're sure? I said, yeah, *I'm good*, and I forgot to take some. [from Josh and Melissa on Wipeout, ABC, SPOKEN genre, 2011]

In addition to the COCA corpus, we also selected 75 Twitter examples from selected dates in 2015, 2016 and 2018 in which *I'm good* was used in the sense of 'No (thanks)', as in this example:

(17) You think I work my ass off to take you out on a date? Smoke you out? Buy you shit? Nah b, I enjoy my own company. *I'm good.* (BamBriaan, Nov 6, 2015)

Twitter is a very "oral" medium, which is huge and constantly growing. In this medium, the innovative usage is extremely frequent. It is also interesting that the percentage of examples in which *I'm good* is preceded by *No*, *Naw* or another negative is about the same as in the COCA corpus, 52% (39/75). About half of the time (48%, 36/75) *I'm good* retains its negative force without an accompanying overt negative form, suggesting, once again, that Defendant Tamika did not need to use an overt *No* for her *I'm good* to mean 'No (thanks)'.

One perhaps relevant observation is that most of the Twitter examples come from African Americans. (Twitter examples often include a photo of the sender.) It is not my contention that *I'm good* 'No' is a distinctly African American usage; the examples in COCA demonstrate clearly that while it a relatively new usage, it is used by a wide cross section of the American public. However, Twitter is extremely popular among African Americans, and the frequency with which the *I'm good* 'No' tokens come from African Americans suggests that this usage would not have been unfamiliar to Tamika. Very likely it is something she encountered and perhaps used frequently in talking with and reading Twitter posts from other African Americans.

Three other observations about the Twitter examples that might be of more interest to linguists than the court are these. Firstly, and this may be related to the 140-character restriction of Twitter, many of the examples do not contain the Yes/No questions or assertions to which they are a response. Sometimes these are evident in the tweets of one or more other persons to which they represent a response. Secondly, a number of idiomatic spinoffs of *I'm good*, have emerged in this medium. One is *I'm good, luv, enjoy* (adding a sarcastic twist) as in:

(18) You done pissed me off for the LAST time. *I'm good luv, enjoy*! * (ayootw33t, Nov 3, 2018)

Another is the use of *I'm good* with a following prepositional phrase, as in:

(19) Too much fake love at FAMU* **I'm good off all that* * (JoseTheTre Oct 29, 2018)

Thirdly, the use of *I'm good* to reject offers that are less than pleasant and potentially self-incriminating (as was true of the body search invitation Tamika declined) is evident in examples like this one, where the tweeter suggests that the *free tattoo* might result in hepatitis:

(20) Some guy just messaged me a video on FB of him sitting at a coffee table with a tatoo machne, made it buzz and said, "Come and get a free tatt" …"I'm good on hepatitis, my dude" (_frvitbat, Oct 29, 2018)

So how did this evidence pan out for Tamika's case? In consultation with her lawyer, she opted to plea bargain rather than go to trial, as happens in more than 96% of U.S. cases, for complex reasons, including the evidence of her sister's case. However, the lawyer submitted my evidence for the usage of *I'm good* to mean 'No' in a final, extended "Expert Notice" to the court, drawing on the COCA evidence and the evidence from the Twitter examples. This apparently persuaded the prosecution to agree to more favorable terms (maximum 3 years, including the one already served) for the plea bargain of 10 years that would otherwise have been the case. As a result, her final sentence was 2 years, rather than 10, including time served, demonstrating another way in which linguistic analysis based on corpora can be very impactful in applied contexts.

8 Summary and conclusion: Towards online corpora of Caribbean Languages

As the above six examples have shown, online corpora can provide an invaluable, sometimes *essential* resource for the study of linguistic variation and change, in theoretical, descriptive, and applied contexts. They are particularly valuable when they are online and publicly available, allowing our analyses to be replicated and validated.

In the English-speaking Caribbean, we already have multiple sources for these kinds of corpora, although they are at different levels of public availability and accessibility:

- Creole recordings already in the Jamaican Language Unit (*Jumieka Langwij Yuunit*), University of the West Indies, Mona, for a long time under the direction of Hubert Devonish, now directed by Joseph Farquharson.
- ICE (International Corpus of English) Jamaica – by design, mainly acrolectal/standard.
- CCJ (Corpus of Cyber Jamaican) – all levels – Christian Mair, U Freiburg, Mair adds (in an email from 5/16/19) that the best and most comprehensive showcase of these data is a PhD thesis produced by Andrea Moll (see Moll 2015).
- Several hundred Creole recordings and transcripts I have from Guyana, Jamaica and Barbados (many made by me, a native Guyanese, others by local linguists from Guyana and Jamaica). Most of these will be available in digital form via Stanford University Library.

- Don Winford's recordings/transcripts from Sranan Tongo, Belize Creole, Guyanese Creole, and Trinidadian Creole, which he has promised to contribute to online corpora being assembled by Bettina Migge, at the University of Dublin.

- Peter Patrick's Jamaican recordings/transcripts. Of these he notes (email of 5/20/19) that "The Veeton recordings from 1989- 90 include more than 100 hours of interviews with about 75 different individuals, many in repeat recordings, plus a series of family and youth-club recordings". I also made a small series of recordings in rural locations in 1991 - principally St. Thomas and Hanover, when I was trying to revisit as many remote sites where Fred Cassidy and David DeCamp made recordings.

- Veronique Lacoste (email of 5/28/19) notes that she has about ~10GB of Jamaican Creole child language data, mostly recorded in schools.

- Other sources: Pauline Christie, Walter Edwards, Shelome Gooden (see Gooden 2003), Hazel Simmons-McDonald, Velma Pollard, Ian Robertson, and the many others who have done field work on anglophone Caribbean varieties may be willing to contribute recordings from their collection, and should be actively approached about doing so.

There should be similar efforts for the French-speaking Caribbean, the Spanish-speaking Caribbean, and so on. Indeed, there undoubtedly already are. For instance, Nicte Fuller Medina has corpora of Belizean Spanish as well as English/ Creole on her Language, Culture and History project website.[10]

Building on and using these resources will not be without challenges. We would need to determine how to collect, catalog, digitize, and safely store recordings and transcripts of Caribbean language varieties, deciding on real names or pseudonyms depending on available permission forms and local IRB (Institutional Review Board) rules. We'd also need to decide about placement: that is what would go where, whether there would be sharing between sites, and so on. Furthermore, we'd need to develop tools for the automatic analysis of digital recordings and transcripts that can handle Caribbean lexicon, phonology, and morphosyntax, along the lines that Shelome Gooden and others have begun to pioneer. And while there are numerous necessary steps to this process, we do have the people who can make it happen.

In order to accomplish this, we could perhaps start with small meetings (facilitated by funding from agencies like the National Science Foundation) of key

[10] https://nfullerm.wixsite.com/website/research

players to discuss the issues. We could also get the advice or involvement of experts like Tyler Kendall (University of Oregon, and the developer of two web-based language archives, including CORAAL) and Christian Mair.[11] But we need to begin soon, because one of the best ways we can honor pioneer linguists of the Caribbean, before they pass on, is by equipping their successors with online corpora of Caribbean languages. They are invaluable, indeed essential tools for 21st century Linguistics.

Acknowledgments

I am grateful to the following Stanford students for assistance in the preparation of this paper: Mea Anderson, Jane M. Boettcher, Susan Chang, Julia Gong, and Zion Ariana Mengesha. And as always, for the encouragement and support of Angela Eunice Rickford, my intellectual companion for more than 50 years. I also thank the two anonymous referees for their helpful comments on an earlier draft.

Abbreviations

AAVE	African American Vernacular English
ASR	Automated Speech Recognition
VOC	Voices of California
WER	Word Error Rate
COCA	Corpus of Contemporary American English
CORAAL	Corpus of Regional African American Language
ICE	International Corpus of English
IRB	Institutional Review Board

References

Bailey, Charles-James N. 1973. *Variation and linguistic theory.* Washington DC: Center for Applied Linguistics.

Gooden, Shelome. 2003. *The phonology and phonetics of Jamaican Creole reduplication.* The Ohio State University. (Doctoral dissertation).

Gooden, Shelome. 2019. Afro-American intonation and prosody: An evolving ecology. Plenary presentation at the Mervyn C. Alleyne commemorative conference, University of the West Indies, Mona, Jamaica, June.

[11] https://blogs.uoregon.edu/lvclab/people/tyler-kendall/ and https://www.researchgate.net/profile/Christian_Mair

Jones, Byron. 2019. *Beyond di riddim: Language use in Jamaican popular music, 1962-2012*. University of the West Indies, St. Augustine. (Doctoral dissertation).

Koenecke, Allison, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky & Sharad Goel. 2020. Racial disparities in automated speech recognition. In Judith T. Irvine (ed.), *Proceedings of the National Academy of Sciences*, vol. 117, 7684–7689. DOI: 10.1073/pnas.1915768117.

Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Meer, Phillipp. 2019. Sociolinguistic variation in (Standard) Trinidadian English vowels: A semi-automatic sociophonetic study of selected monophthongs and diphthongs. In *Proceedings of the 50th Congress of the Brazilian Linguistics Association, May 2–9*. Maceió, Alagoas: Brazilian Linguistics Association.

Meer, Phillipp. 2020. Automatic alignment for New Englishes: Applying state-of-the-art aligners to Trinidadian English. *Journal of the Acoustical Society of America* 147.4. 2283–2294.

Mengesha, Zion. 2020. The social meaning of vowel trajectories: FEEL-FILL merger among African Americans in California. Ms., Department of Linguistics, Stanford University.

Moll, Andrea. 2015. *Jamaican Creole goes web: Sociolinguistic styling and authenticity in a digital yaad*. Amsterdam: Benjamins.

Rickford, John R., Isabelle Buchstaller, Thomas Wasow & Arnold Zwicky. 2007. Intensive and quotative ALL: Something old, something new. *American Speech* 82.1. 2–31.

Rickford, John R., Thomas Wasow, Norma Mendoza-Denton & Juli Espinoza. 1995. Syntactic variation and change in progress: Loss of the verbal coda in topic-restricting *as far as* constructions. *Language* 71.1. 102–31.

Wassink, Alicia Beckford. 2020. Automatic speech recognition and ethnicity-related dialects. Uneven success. Paper presented at the meeting of the American Association for the Advancement of Science, February 14, 2020. https://depts.washington.edu/sociolab/publications/documents/AAAS-2020-Wassink.pdf.

Wikipedia. 2019. *Corpus of contemporary American English*. Wikimedia Foundation, 8 Apr. 2019. %7Bhttps://en.wikipedia.org/wiki/Corpus_of_Contemporary_American_English%7D.

Winford, Donald. 1972. *A sociolinguistic description of two communities in Trinidad*. University of York. (Doctoral dissertation).