Approaches to Searching for Language and Diversity in a 'Whitebread City' Digital Corpus: The Charlotte Conversation and Narrative Collection

Stephen Westman *(srwestma@email.uncc.edu)*
University of North Carolina at Charlotte
Boyd Davis *(bdavis@email.uncc.edu)*
University of North Carolina at Charlotte

Macaulay comments that

> Dialects, like languages, have both a unifying and a separatist function. We speak the way we do to be like those we wish to associate with and to distinguish ourselves from others. When that association is based on where we live. . . a distinctive form of speech is likely to survive. However, we need to look at the whole configuration of linguistic features and not a few features which may or may not be the critical ones for the speakers.
>
> (239)

Why? In addition to the "grammatical, phonetic, and lexical" features traditionally posited as characterizing a dialect, Macaulay adds

> prosodic features and possibly also voice quality and discourse characteristics. There is no reason to believe that dialects have fewer features than other forms of language, and we do not know in advance which features will be important to distinguish the dialect.
>
> (229)

In a discussion of features of southern style that warrant further investigation, Barbara Johnstone cites several which can be searched at word- and text-level: these lexicogrammatical features contribute to the reader/hearer's assessment of style, and include rhetorical genres triggered by particular discourse markers; style shifts into regional colloquialism, stylization and self-parody signaled by shifts into nonstandard verbs, for example, or a judicious sprinkling of double modals to suggest temporary intimacy. She asks for an investigation of regional styles of interacting that "makes strategic use of nostalgia for neighborhood, local community, or region." (206) Well, they said southernly, we have a gracious plenty of data that accommodates such investigation; the issue is, of course, how to access, identify, and retrieve it.

The corpus that we are using to investigate these questions is *Project MORE*'s expanded *Charlotte Conversation and Narrative Collection* (*CNCC*), which is part of the 11.5 million-words in the First Release of the *American National Corpus* (*ANC*). Considered a satellite corpus to the core of the *ANC*, which parallels the organization of the *British National Corpus* (Reppen et al.), the *CNCC* goal is far more modest, but still one of difficulty: to create a corpus of conversation and conversational narration in a New South city at the beginning of the 21st century. And that, of course, brings us smack up against issues of region (Macaulay) and representativeness (Douglas), of dialect diversity (Wolfram & Dannenberg), and distinctions between rural and metropolitan features (Tillery, Bailey & Wikle).

The *CNCC* is hybrid in some ways; similar to the *ONZE* corpus in its evolution through multiple formats and purposes (Gordon et al.). In addition to being a part of the *ANC*, it is also included in the *New South Voices* (*NSV*) digital resource housed at the University of North Carolina at Charlotte Library. *NSV* includes interviews that cover a wide range of historical subjects, from African American churches and Billy Graham crusades to women's basketball and World War II. Other interviews, narratives and conversations document the experiences and language of recent immigrants to the area. As such, it seeks to address a wide variety of audiences from local historians and historic preservationists to public school students. By using *NSV*, we are able to expand the number and range of interviews available for linguistic as well as for historical analysis.

If the corpus is to be inclusive of the range of spoken styles that conglomerate in the elastic borders of a New South city, it must begin by identifying what they are. In today's Charlotte, today's North Carolina, this is no longer simple. As Tillery, Bailey & Wykle note, metropolitanization, foreign and domestic migration, and expanding ethnic diversity have "eliminated many of the vestiges of traditional regional culture and . . . are radically reshaping the United States" (228). Their painstaking study of what they see as the impact of demographic change on American speech is keyed to 22 socio-demographic and linguistic variables: 14 phonological features, 3 lexical, and 5 that are lexicogrammatical. They see a balkanization (241; cf 244) with increased divergence of rural and urban ways of speaking; they ask will "old towns with new populations" — such as Charlotte — create new communities and new ways of speaking?

Investigation of these phenomena within a database environment requires a variety of tools and approaches if we are to extract the information contained in these transcripts. The reason for this is due to the nature of the types of information we need to

obtain from these interviews and then to the question of how we can best obtain that information.

On the one hand, we need to be able to perform textual analysis on the interviews and to examine subjects' speech patterns and linguistic characteristics. This in turn requires that we be able to extract information that is embedded within discursive text — looking at how they use language '*in situ*'. On the other hand, there are discrete pieces of information — metadata if you will — about the participants (place of origin, current residence, gender, ethnicity, etc.) to which we need access if we are to do anything meaningful with what we discover from our textual analysis. This dichotomy pertains to any area doing textual analysis. As noted by Ronald Bourrett, the roots of this dichotomy lie in the two types of information with which we are dealing: document-centric and data-centric.

Due to the different nature of the two types of information — linguistic analysis and descriptive metadata — we have found that a single approach does not allow us to fully explore the types of correlations we were seeking. While our XML database allows us to find useful things in searching tagged information within the interviews, it does not provide sufficient flexibility in searching data-centric information. On the other hand, with relational database technology we have exactly the opposite situation.

In addition, during the course of our investigation, we discovered that there were certain types of textual information — such as word- and phrase-frequency; retrieving and isolating particular words and phrases within their context in a document; and looking for particular words and/or phrases within certain proximity of each other — that were amenable neither to an XML database, nor a classic relational database, approach. To address this need, we decided that a third option — inverted indexes — was needed to allow us to look for such patterns. As noted in Zaïane, this technique greatly enhances the ability to search textual-based information.

Therefore, in designing our database system, we decided to adopt a mixed approach, one that allowed us to utilize the strengths of each system without running afoul of its limitations. In doing so, we use both XML and inverted indexing to do textual analysis and then a relational database to correlate that information with relevant demographic criteria. The result is a hybrid that allows us to do more than any single approach could provide.

This paper will first present how we are using readily available tools to implement a searching system that supports demographic correlation with textual features (including some features of proximity search and frequency of occurrence). These tools, all of which are part of the Open Source tools, allow us to build and configure with ease a system that not long ago would require extensive and non-trivial programming. They include:

- *eXist* (XML) and *MySQL* (relational) database managers
- php, perl and Java programming languages
- *Apache* Web server

As a way of concluding, we will then earnestly solicit assistance on how we can best make this collection of roughly 1,000 transcribed oral interviews, conversations and narratives more useful to any researcher, particularly in the area of text-based, online searching.

Bibliography

Douglas, Fiona. "The Scottish Corpus of Texts and Speech: problems of corpus design." *Literary and Linguistic Computing* 18 (2003): 23-37.

Gordon, Elizabeth, Margaret Maclagan, and Jennifer Hay. "The ONZE corpus. Manuscript." *Models and Methods in the Handling of Unconventional Digital Corpora. Volume 2: Diachronic Corpora.* Ed. J.C. Beal, K.P. Corrigan and H. Moisl. Houndsmills: Palgrave, Forthcoming.

Hudson-Ettle, Diana. "Nominal that clauses in three regional varieties of English: a study of the relevance of text type medium, and syntactic function." *Journal of English Linguistics* 30 (2002): 258-273.

Johnstone, Barbara. "Features and uses of southern style." *English in the Southern United States.* Ed. S. Nagle and S. Sanders. Cambridge: Cambridge University Press, 2003. 189-207.

Kjellmer, Goeran. "A modal shock absorber, empathizer/emphasizer and qualifier." *International Journal of Corpus Linguistics* 8 (2003): 145-168.

Macaulay, Ronald. "I'm off to Philadelphia in the morning." *American Speech* 77 (2002): 227-241.

Reppen, Randi, and Nancy Ide. "The American National Corpus: Overall goals and the first release." *Journal of English Linguistics* 32 (2004): 105-113.

Tillery, Jan, Guy Bailey, and Tom Wikle. "Demographic change and American dialectology in the twenty-first century." *American Speech* 79 (2004): 227-249.

Wolfram, Walt, and Clare Dannenberg. "Dialect identity in a tri-ethnic context: The case of Lumbee American Indian English." *English World-Wide* 20 (1999): 179-216.

Zaïane, Osmar. *Inverted Index for Information Retrieval (Slides keyed to Chapter 22 of unlisted textbook: CMPUT 391: Database Management Systems).* University of Alberta, 2001. Accessed 2005-04-11. <http://www.cs.ualberta.ca

```
/~zaiane/courses/cmput391-02/slides/Lect7
/>
```