Taiwan Journal of Linguistics
Vol. 6.2, 27-58, 2008

SPOKEN CORPORA: RATIONALE AND APPLICATION

John Newman

ABSTRACT

Despite the abundance of electronic corpora now available to researchers, corpora of natural speech are still relatively rare and relatively costly. This paper suggests reasons why spoken corpora are needed, despite the formidable problems of construction. The multiple purposes of such corpora and the involvement of very different kinds of language communities in such projects mean that there is no one single blueprint for the design, markup, and distribution of spoken corpora. A number of different spoken corpora are reviewed to illustrate a range of possibilities for the construction of spoken corpora.

Key words: corpora, communication, oral history, speech analysis, language documentation

1. INTRODUCTION

Linguistics has undergone considerable changes in the last couple of decades with respect to the kinds of data that are considered relevant to the field. Data obtained from electronic corpora, in particular, have come to play an ever increasing role in the analysis of language, reflecting a more usage-based orientation on the part of linguists and spoken corpora have, arguably, a special role to play in any usage-based approach to linguistics.

Section 2 reviews the current climate in linguistics and discusses some of the considerations which have led linguists to be interested in spoken corpora. The interest that linguists have in constructing spoken corpora overlaps to some extent with an interest in various kinds of speech-based documentation of culture and local history, originating

outside of the field of linguistics, suggesting the possibility of greater collaboration between linguists and non-linguists in this area. Section 3 introduces four examples of spoken corpora to illustrate a range of possibilities in the construction of spoken corpora. The paper concludes with some summary observations in Section 4.

2. DOCUMENTING SPEECH

2.1 Why Linguists Need Spoken Corpora

It is fair to say that, over the last 50 years, the linguistic mainstream, at least in the United States, has been dominated by an over-reliance on one kind of linguistic evidence, namely "native speaker intuition", and one kind of goal, viz. the reconciliation of such data with formal models of language. There can be no doubt just how stimulating and rewarding this kind of linguistics can be. Whatever the pros and cons of working with purely intuition-based data may be, however, there comes a time when one simply has to acknowledge a greater role for other kinds of data in linguistics, in particular data drawn from actual usage. Such data has always been a feature of some subfields of linguistics, e.g., child language acquisition and historical linguistics (where there is a written record to be studied). When it comes to mainstream linguistic work in syntax and semantics, however, usage and the corpora which document that usage have been largely shunned. The field of linguistics is now experiencing a surge of interest in usage-based data, alongside other kinds of empirical data such as psycholinguistic experimental data, as part of a broader empirical turn in the field. As evidence of this shift, one may cite the words of the current Editor of Language who has observed, with reference to the contents of the journal: "…we seem to be witnessing…a shift in the way some linguists find and utilize data – many papers now use corpora as their primary data, and many use internet data" (Joseph 2004: 382). One reason for this empirical turn is a desire to correct the imbalance in the range of data which had hitherto been accepted as linguistic evidence. Another reason is simply the emergence of new opportunities to study large collections of data as a result of technological advancements in speech technology, computing hardware, and software development.

Hand in hand with this broadening of the kinds of evidence that linguists work with has come a greater diversity of goals within linguistics. Some schools of thought in linguistics, e.g., Systemic Linguistics and the Birmingham School, always worked towards broader goals than were common in mainstream North American linguistic circles. Cognitive Linguistics (as defined, say, in Evans and Green 2006:27-28) is a more recent example of a movement within linguistics with relative broad goals, concerned as it is with general principles that provide some explanation for all aspects of language. These principles may be drawn from disciplines other than linguistics, and many kinds of evidence and methodologies will therefore be relevant, including corpus data and its associated methodologies (cf. Tummers, Heylen, and Geeraerts 2005). Evans and Green (2005:108) draw attention, specifically, to the importance of usage-based data in Cognitive Linguistics: "…language structure cannot be studied without taking into account the nature of language use". Cognitive Linguistics, understood in this way, requires the incorporation of corpus data into linguistic analyses. Even without subscribing to all the tenets of Cognitive Linguistics, however, anyone open to a full understanding of the nature of language must be prepared, correspondingly, to admit a full range of data, including corpus-type data.

One effect of working with corpora has been an increase in awareness among linguists of the very different genres which typically exist in languages, especially the distinction between spoken and written genres. One does not need to look further than such well-known corpora as the British National Corpus (BNC), the American National Corpus (ANC), and the International Corpus of English (ICE) to appreciate how widespread the spoken vs. written distinction has become as a feature of corpus design. Spontaneous face-to-face conversation would seem to occupy a special place among all the genres in so far as it is represents a relatively basic kind of human interaction. It is, for example, the very first kind of language interaction that a human is typically exposed to. And it is the only kind of language interaction relevant to some speech communities where there is no written tradition. One does not necessarily have to agree that face-to-face conversation is paramount in terms of our communicative activities – and it may not be for some individuals who inhabit a highly literate cultural milieu – to accept that it is an important kind of human activity and deserving of study.

Documenting the spoken language is special, too, in terms of the technological challenges it presents, compared with the written language. It is obvious that the speech signal of speakers carries important cues as to the message intended through volume, pitch, duration, pauses, etc., hence the critical role of speech technology in capturing the high quality speech. High-quality speech recording is not always easily achieved, however, due to the difficulties of making speech recordings in some field situations. Annotating transcripts of spoken language also presents formidable, though not insurmountable, problems (cf. Gut and Bayerl 2004). Wichmann (2007) draws attention to the time-consuming nature of such transcription, as well as the difficulties of any kind of labeling of prosodic features by humans. She cites as one example Schriberg et al. (1998) who found that labeling by hand of "prominent syllables" in annotation achieved only 31.7% accuracy. Of course, there can be high quality prosodic annotation of transcripts of speech as corpora – one thinks, above all, of the London Corpus of Spoken English (Svartvik 1990) – but even in these cases, not all the acoustic information a researcher may need would necessarily have been anticipated at the time of annotation and hence included in the transcript. Clearly, an audio file of the original speech remains a vital part of studying spoken language, however difficult it may be to integrate audio data with transcripts.

2.2 Linguistics and the Study of Communication

Even with an expanded, and expanding, role for corpora in linguistics, the field as a whole is still mostly concerned with linguistic data abstracted away from the actual speech act situation. That is, linguists, for the most part, do not see communication as the object of study so much as language. There are exceptions, of course, but for the most part, speech rate, eye movement, hand gestures, body language etc. are relatively marginal as objects of study within linguistics. Linguistics can only gain from a more expansive view of its focus of study, incorporating the study of communication in its entirety (cf. also Wichmann 2007:82-83 in which the author calls for data from all channels of communication to be included in our corpora). It is, of course, possible in theory to construct video-based corpora which could form the basis for the close study of face-to-face communication, but such corpora are not widely used. Charles Goodwin, in a lifetime of publications such as Goodwin (1979, 1980, 1981) and many other

publications since then, has explored the psychological and social processes which occur during face-to-face communication and his research has great relevance to the study of language as used in communication. His research includes, for example, the study of eye gaze on the part of participants, when one participant's gaze meets another participant's eyes, withdrawing a gaze, the lowering of the volume of the voice, whether one can be certain that a stretch of talk was heard by other participants, etc. Consider the utterances in (1a-c), taken from Goodwin (1981:57), three separate utterances in which the speaker might be said to have made a "false start". The utterances are written out using some of the notational conventions developed by Goodwin. Here, we focus just on the 'X' on the line of the second participant, connected to the line above by '[', which in each case marks the point when the gaze of the second participant reaches the gaze of the first participant. The continuous line after the X indicates the stretch of time when the two participants are gazing at each other. The notation is immediately informative when it comes to understanding what is happening on the "text" line. In each case, the first participant restarts the utterance just when (s)he has secured the gaze (hence attention) of the second participant. The "false starts" are in reality "correct starts" which can only be properly studied and understood in light of the whole communicative context. Goodwin analyzes numerous snippets of conversation in this way, most of which are far more complex than the examples given here.

(1) a. Debbie: Anyway, (0.2) Uh:, (0.2) We went t- I went ta bed
[
Chuck: X_____

b. Barbara: Brian, you're gonna ha v- You kids'll *ha*ve to go
[
Brian: X_____

c. Sue: I come in t- I no sooner sit down on the couch
[
Diedrie: X_____

Even with the detail provided for in some modern transcripts of conversational corpora, one does not usually see this level of detail and

yet, clearly, such information provides essential insight into the processes at work in face-to-face conversation. Software for facilitating the transcription and retrieval of such information is available which greatly facilitates this task, in particular multimedia annotation tools such as CLAN (http://childes.psy.cmu.edu/), ELAN (http://www.lat-mpi.eu/tools/elan) and Anvil (http://www.dfki.de/~kipp/anvil/). With tools such as these now available, the prospects are good for the inclusion of more gestural and gaze information into spoken corpora.

The sample transcriptions in (1) lead to a further observation that needs to be made about transcription, in general, and with respect to spoken corpora in particular, namely, that a transcription embodies a multitude of assumptions about the data. These assumptions, in turn, will influence the analysis and the results of research. Decisions about representing speech, for example, are closely tied to theoretical stances about the separability of the prosodic level of speech from the analysis of words in orthographic representation. The importance of recognizing the underlying theoretical bias of a transcript has been emphasized, in particular, by Elinor Ochs (Ochs 1979). She draws attention, for example, to the practice of linking adjacent turns and utterances in conversational speech, a practice reflected in corpus software which facilitates the expansion of a turn in a transcript to the immediately preceding and following turns. This may seem highly desirable and theoretically sound in many cases, but Ochs raises the question of whether it is good practice in the case of studying children's speech at the stage where they are still acquiring adult patterns. She argues, in fact, that transcripts of such speech should be "relatively neutral with respect to the contingency of children's talk" (Ochs 1979:47).

2.3 Speech Recordings by Non-linguists

Speech recordings can be motivated by many kinds of considerations, extending well beyond the realm of linguists. One valuable kind of speech recording is the category of video/audio recordings which document one or more aspects of culture. An example of such documentation is oral history. Oral histories record the past in the words of the people who have experienced it. Never has so much oral history been collected and disseminated as now, thanks to the ease with which sound recordings can be made and the availability of the internet to

disseminate such recordings.[1] Of course, the oral history movement itself has been a key factor in the development of such histories, too, providing the academic legitimization of such story-telling. The scope and content of oral history collections can vary greatly, but generally these collections would be best described as "archives" of speech, rather than "corpora" in the sense that linguists are accustomed to. Oral history collections of speech recordings typically allow users to listen to individual recordings, without the benefit of easily searching a topic or pattern across all the recordings in an archive. In this sense, they are similar to the book holdings of a library which allow the user to borrow individual books without allowing the user to search for patterns across all texts contained within the books. It seems useful here to maintain a distinction between "archives" which can be accessed on an item-by-item basis and "corpora" which allow information to be retrieved efficiently from all items in a collection. Nevertheless, oral history collections do vary in how they are stored and some collections can indeed be similar to the kind of corpora that linguists are accustomed to.

As one example of an oral history collection, consider the Oral Histories of the American South project (http://docsouth.unc.edu/sohp/). This project, based at the University of North Carolina at Chapel Hill, builds upon an existing program – the Southern Oral History Program (SOHP). SOHP began in 1973 with the aim of documenting the life of the American South in tapes, videos, and transcripts. According to the website, this project will ultimately make 500 oral history interviews available over the internet (the website reports that 400 are already available), selected from the 4,000 or so oral history interviews carried out by SOHP over thirty years. The interviews selected for SOHP cover a variety of topics in recent North Carolina history, particularly civil rights, politics, and women's issues. As of writing, the index contains a list of 496 topics. Interviews can be read as text transcript, listened to (or downloaded) with a media player, or both simultaneously. Transcripts are also available in XML format.

As if Oral Histories of the American South were not impressive enough as a record of the American South, the University of Carolina at Charlotte (another of the 16 campuses of the University of North

[1] For a list of links to oral history projects around the world, see the numerous links at http://www.bl.uk/collections/sound-archive/ohlinks.htmlarchive/ohlinks.html#oral-history-in-the-uk.

Carolina) offers yet another oral history collection of comparable scope and quality in the form of New South Voices (http://newsouthvoices.uncc.edu/about/index.php). New South Voices is a project of the Special Collections Unit, J. Murrey Atkins Library, University of North Carolina at Charlotte and provides online access to a collection of over 600 interviews, narratives and conversations documenting the Charlotte region in the 20th century. According to the website, the interviews cover a wide range of historical subjects, from African American churches and Billy Graham crusades to women's basketball and World War II. New South Voices also pays particular attention to the experiences and language of recent immigrants to the area. Interviews may be available as wav, mp3, video, HTML, or pdf files, though not all formats are usually available for any one interview.

When there is such professional documentation of oral histories, it seems natural for linguists to exploit them for linguistic research. And indeed, in the case of New South Voices, transcripts from that collection have been included in the ANC. The 198,295 words of the Charlotte Narrative and Conversation Collection of New South Voices is, in fact, the only material in the "face-to-face" category of the spoken part of the ANC. Of course, not all oral history collections are as easily available or as fully documented as the two mentioned above. The Millennium Memory Bank, a BBC initiative involving oral history interviews is a case in point (http://www.bl.uk/collections/sound-archive/millenni.html). Clearly, this project is massive in scale, with a total of 6,000 people interviewed. These recordings are of great value, potentially, to linguists, but they are not as immediately useful as are the recordings and transcripts of New South Voices (there are no transcripts available on the website for the Millennium Memory Bank, for example).[2] A similar kind of initiative undertaken by the New Zealand National Broadcasting Service has given rise to a uniquely important corpus: the Origins of New Zealand English (ONZE) Corpus (http://www.ling.canterbury.ac.nz/onze/index.html). This corpus (Gordon, Maclagan, and Hay, to appear) includes sound recordings made by a mobile unit which travelled around New Zealand between 1946-1948, collecting reminiscences from about 300 people

[2] Audio clips from the Millennium Memory Bank can be heard over the internet as part of the British Library's COLLECT BRITAIN website at http://www.collectbritain.co.uk/collections/dialects/ as well as at its more pedagogically oriented and highly interactive website http://www.bl.uk/learning/langlit/sounds/index.html. Audio clips from these recordings can also be heard at the BBC's VOICES website http://www.bbc.co.uk/voices/.

born between 1851 and 1910. This part of the collection (the Mobile Unit sub-corpus of ONZE) is particularly significant in light of the history of European settlement in New Zealand which first took place in the 1800's and most of it post-1850. It means that we have an unusually rich resource for the study of the evolution of the features of New Zealand speech.

Collections of oral histories are relevant not just on account of their potential as spoken corpora, but also because of the value that society at large places upon cultural documentation. Linguists have now become used to discourse about the value of linguistic and cultural documentation in connection with endangered languages where the documentation of narratives, legends, and traditional stories can be a significant cultural asset. Of course, for many endangered languages there simply isn't any agreed-upon writing system which could be a natural vehicle for conveying this kind of material in writing and so one has no choice but to rely upon audio or audio-visual recordings as the appropriate medium. In any case, the collected narratives and stories from endangered language communities, whatever format they may be in, are often seen as being of great value to the language community. But the success and popularity of the oral history movement shows that, even in the context of large nations speaking world languages, recording the stories of the past in the words of the people who have experienced this history has great value too. And hearing the spoken words of the story tellers, using language which belongs to the past, is an important part of this history.

Apart from collections of speech where the collection is intended as part of cultural documentation, as with the above examples, there are many other kinds of transcripts which have come about as by-products of other activities and where the transcript is not really the main point of the exercise. Above all, one thinks of transcripts of radio and TV shows, where it is really the show itself, at the time of broadcasting, which constitutes the main activity, not the transcribed record. Mark Davies has made excellent use of such transcripts for the 72+ million-word sub-corpus of spoken language in the BYU Corpus of American English (http://www.americancorpus.org/). Davies has relied upon transcripts of unscripted TV and radio programs for the spoken language sub-corpus: All Things Considered (NPR), Newshour (PBS), Good Morning America (ABC), Today Show (NBC), 60 Minutes (CBS), Hannity and Colmes (Fox), Jerry Springer (syndicated), etc. On the website, Davies

provides (convincing, I think) justification for this choice of data. Among other things, he addresses the question of just how "unscripted" some of the material is. He concludes: "The question is whether you would rather have a 76+ million word spoken corpus with about 5% scripted material (but still leaving more than 70 million words of unscripted material), or a 'completely pure' corpus that is so small (1-2 million words) that it is unusable for many types of research. We opted for the former." Davies also offers some justification for the use of these materials in terms of fair use, noting in particular that the restrictions on accessing data through his website should satisfy any concerns of this type. Davies refers to the guidelines of the Stanford Copyright and Fair Use Center which describe the factors relevant to decisions about fair use in the USA.[3] Presumably, the corpus use of these transcripts count as "transformative", leading to new insights and understandings about the original spoken events – this is considered a key factor in defending the scholarly uses of material.

Much of the data and corpora on offer from the Linguistic Data Consortium (LDC) at http://www.ldc.upenn.edu/, too, is associated in some way with projects which have extra-linguistic, or at least ambiguously linguistic, goals. The documentation for projects such as the extensive CALLHOME (lexicon, speech, transcripts) and CALLFRIEND (speech) data, for example, both explicitly acknowledge that the data was collected in support of a project sponsored by the U.S. Department of Defense. In other cases, the documentation references a branch of the U.S. Department of Defense. The documentation for the SWITCHBOARD corpus, for example, states that it is "collected at Texas Instruments with funding by DARPA" (LDC website). DARPA stands for the Defense Advanced Research Projects Agency of the U.S. Department of Defense and states its mission in the following terms: "It manages and directs selected basic and applied research and development projects for [the Department of Defense], and pursues research and technology where risk and payoff are both very high and where success may provide dramatic advances for traditional military roles and missions" (DARPA website, http://www.darpa.mil/). This is not to suggest any insidious intent on the part of LDC organization here.

[3] See http://fairuse.stanford.edu/Copyright_and_Fair_Use_Overview/chapter9/9-b.html for these guidelines.

The point is simply that these resources, as valuable as they are to linguists, are not simply made "for and by" linguists.

3. HOW WE MAKE SPOKEN CORPORA

Spoken corpora present very particular technical and procedural challenges, as mentioned above, and determining "best practice" in terms of how to construct such corpora is not so easy, despite the availability of informative guides such as Wynne (2005), McEnery, Xiao, and Tono (2006), and the resources available from LDC. This fact complicates any attempt at a "one size fits all" approach to constructing corpora. Furthermore, not all languages are at the same state of development as far as linguistic understanding of their structure is concerned. For some languages, we do not yet have fully articulated linguistic analyses and even basic questions, such as where a word begins and ends, remain unanswered.[4]

Rather than try to describe one model of corpus construction and utilization – it is unrealistic to think that such a model exists – four spoken corpus projects will be reviewed in turn: The (North Carolina) Sociolinguistic Archive and Analysis Project (SLAAP), the spoken component of International Corpus of English (ICE), the Wenzhou Spoken Corpus (WSC), and the Dinka Narratives Corpus (DNC). These four projects are quite different to each other in so far as their rationale, intended audience, institutional support, and the state of linguistic research on the language are concerned. SLAAP and ICE are predicated upon very substantial research funding and relate to a world language (English). WSC and DNC have not had the benefit of any substantial research funding and relate to lesser known languages. SLAAP and DNC incorporate audio in an essential way into the corpus, whereas ICE and WSC are designed around the analysis of the transcript more than the audio. In light of this diversity among the projects, reviewing them here seems a suitable way to introduce and discuss a variety of issues relating to spoken corpora.

[4] For both Mandarin and Chinese dialects, we have an increasingly large pool of spoken corpora to choose from (cf. Yang 2006).

3.1 Sociolinguistic Archive and Analysis Project

The Sociolinguistic Archive and Analysis Project, or SLAAP, (http://ncslaap.lib.ncsu.edu/index.php) arises out of a primary interest in sociolinguistic aspects of language use.[5] SLAAP represents an exciting new model for the integration of audio and text (for an overview, see Kendall 2007).

SLAAP is an extension of the North Carolina Language and Life Project (NCLLP), a sociolinguistic research initiative at North Carolina State University, focusing on Southern American English, under the direction of Walt Wolfram. That project holds the audio and video tape records of some 1,500 sociolinguistic interviews conducted from the late 1960s up to the present, with about 100 interviews being added each year. The NCLLP website specifies four main goals of the project (http://www.ncsu.edu/linguistics/ncllp/index.php):

- to gather basic research information about language varieties in order to understand the nature of language variation and change;
- to document language varieties in North Carolina and beyond as they reflect varied cultural traditions;
- to provide information about language differences for public and educational interests;
- to use research material for the improvement of educational programs about language and culture.

SLAAP is digitizing and making accessible through the internet much of this material and is a joint initiative of NCLLP and the North Carolina State University Libraries. At the time of writing, 640 interviews are reported as being included into SLAAP, corresponding to 260,000 words of orthographically transcribed speech.[6]

Clearly, SLAAP has been constructed for the benefit of linguists and, above all, sociolinguists. The strong sociolinguistic orientation of the

[5] I have been advised that the acronym to be used in future is SLAAP, rather than the acronym which currently appears on the website (NCSLAAP), hence I will use SLAAP here.

[6] I am indebted to the project coordinator Tyler Kendall for providing this and other information about SLAAP.

project makes it unusual, and hence particularly interesting, as a spoken corpus. Even so, it is worth noting the broader public and educational interests which the project aims to serve and the Community Outreach page (http://www.ncsu.edu/linguistics/ncllp/outreach.php) of NCLLP contains many interesting examples of such community engagement.

The sociolinguistic interviews which are the basis for SLAAP are constructed around a variety of tasks. Part of the interview is similar to an informal oral history where the interviewee is asked to describe what it was like to have grown up in the area. Other leading questions, given as examples on the NCLLP website, are "What games did you and your friends play when you were kids?" and "Do you have any friends that aren't from around here? What are they like?". In addition to encouraging casual, relaxed conversation, the interviewer also elicits responses to questions which target items of specific sociolinguistic, dialectal interest such as "What do you call people who aren't from town? That is, do you have a word (or words) for visitors?". Also, the interviewer may ask for pronunciations of specific words.

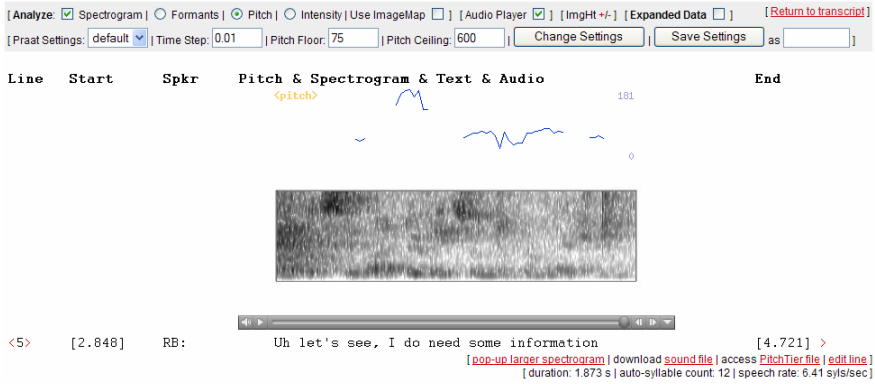
SLAAP provides many features associated with acoustic analysis, reflecting the crucial role of the speech analysis software Praat in the construction of the corpus. Even the transcripts in the case of this corpus are exported and time-stamped Text Tiers of Praat files, as discussed below. The possibilities for acoustic analysis of the audio files are therefore a particular strength of this corpus and the interface, whereas the analysis of the text itself is relatively limited (though under development) compared with, say, ICE. Four aspects of the user interface of SLAAP deserve special mention and are discussed further below: (a) audio integration, (b) transcript format, (c) user annotation of the transcript, and (d) other corpus tools.

(a) Audio integration

As well as allowing the user to listen to or download an audio file of an interview, SLAAP has integrated audio and transcript so that the user can undertake an acoustic analysis of any line of the transcript, where a "line" of transcript corresponds to an unbroken stretch of speech. The acoustic analysis is made possible by Praat scripts which run "on the fly" and no additional software is required on the user's computer. An example of such an analysis, for the utterance "Uh let's see, I do need some information", is shown in Figure 1. The top line shows various

parameters which can be selected. In the present instance, pitch, spectrogram, text, and audio have all been selected. Users have the option of downloading the wav file of this segment as well as the Pitch Tier file, as generated by Praat.

Figure 1. Screenshot of audio analysis of an utterance in SLAAP



.

(b) Transcript format

Transcripts in SLAAP are generated from the Text Tiers within Praat and can be viewed in various formats. Three of these formats are shown in Figure 2a-c. Figure 2a makes use of a vertical format with each successive speaker beginning a new row (blank lines against a speaker's name indicating an absence of speech). Figure 2b illustrates the column format in which all the utterances by a speaker are shown in a single column, with different columns for different speakers. Figure 2c shows a Henderson Graph format in which the duration of each utterance corresponds to the length of a horizontal line, together with the beginning and end times of the utterance. Square brackets indicate overlapping speech and in Figure 2a an additional option of indenting overlaps is used whereby the overlapping segments, e.g., [front and it] and [I will], are aligned vertically. A more exact record of overlapping speech can be read off from the beginning and end times of each utterance which are shown on the left and right of the displays in Figures 2a and 2b.

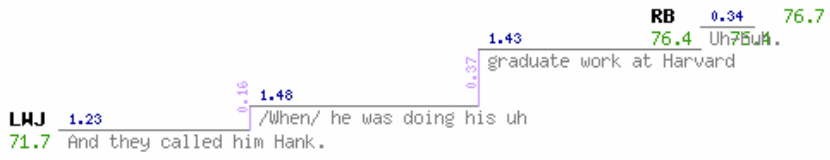
Figure 2a. Screenshot of a transcript in SLAAP in vertical format

```
[66.73] RB:   [/unintelligible/]                                                [67.34]
[67.06] LWJ:  [You] look right in the [front and it] tells where they're stationed-   [70.44]
        Henry L.
[67.34] RB:                                                                     [67.84]
[67.84]                              [I will]                                   [68.38]
[68.38]                                                                         [76.39]
[70.44] LWJ:                                                                    [71.72]
[71.72]       And they called him Hank.                                         [72.95]
[72.95]                                                                         [73.11]
[73.11]       /When/ he was doing his uh                                        [74.59]
[74.59]                                                                         [74.96]
[74.96]       graduate work at Harvard                                          [76.39]
[76.39]                                                                         [77.75]
[76.39] RB:   Uh-huh.                                                           [76.73]
```

Figure 2b. Screenshot of a transcript in SLAAP in column format

```
[66.73]                                      [/unintelligible/]                 [67.34]
[67.06] [You] look right in the [front and it]
        tells where they're stationed- Henry L.                                 [70.44]
[67.34]                                                                         [67.84]
[67.84]                                      [I will]                           [68.38]
[68.38]                                                                         [76.39]
[70.44]                                                                         [71.72]
[71.72] And they called him Hank.                                               [72.95]
[72.95]                                                                         [73.11]
[73.11] /When/ he was doing his uh                                              [74.59]
[74.59]                                                                         [74.96]
[74.96] graduate work at Harvard                                                [76.39]
[76.39]                                                                         [77.75]
[76.39]                                     Uh-huh.                             [76.73]
```

Figure 2c. Screenshot of a transcript in SLAAP in Henderson Graph format



(c) User annotation of the transcript

Sociolinguistics has been much concerned with sociolinguistic variables, tracking certain features of language in the speech of speakers by age, gender, special status etc. SLAAP makes provision for users to mark these features for later reference, in effect "annotating" the transcripts even though the actual transcripts are not changed. A certain amount of markup of such features has already been carried out. A partial summary of the distribution of one variable, the 3Sg *–s* suffix of the English present tense, is shown in Figure 3.

Figure 3. Screenshot of first 7 records of the 3Sg *–s* variable in one file of SLAAP

Tab#	Time	Speaker	Context	Abs./Pres.	Prec. Env.	Subj.	Foll. Env.	Clause	Confidence
1	52.39	Keisha	she say I'm	abs	vowel	pronoun	vowel	matrix	conf
2	57.34	Keisha	she thinks everytime	pres	consonant	pronoun	vowel	matrix	conf
3	59.70	Keisha	a boy come around	abs	consonant	NP	vowel	embedded	conf
4	65.46	Keisha	she think -	abs	consonant	pronoun	pause	matrix	conf
5	66.31	Keisha	a boy come around	abs	consonant	NP	vowel	embedded	conf
6	68.94	Keisha	she think I like	abs	consonant	pronoun	vowel	matrix	conf
7	76.30	Keisha	she say [snap]	abs	vowel	pronoun	pause	matrix	conf

(d) Other corpus tools

It is possible to summarize data relating to a speaker across many/all instances of a feature, a highly desirable feature for any corpus interface. Figure 4 shows an analysis of pause duration for one speaker in 100 instances. The analysis contains summary statistics, a graph showing pitch variation line by line, and a plot of pause duration against the total duration of the preceding and following utterances. Pitch analysis, part of speech use, and speech rate analysis are also possible with this tool. Figure 5 shows an example of part of speech use for the same speaker (tagging for POS is still under development). SLAAP also allows for searches of a string of characters in transcripts with a concordance-like display of results.

Figure 4. Screenshot of an analysis of pause duration for a speaker in SLAAP

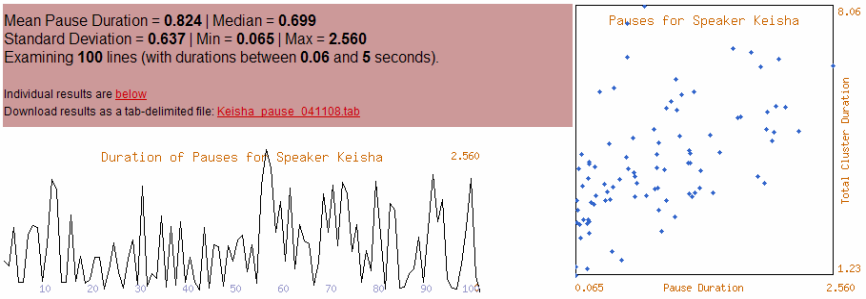


Figure 5. Screenshot of part of speech use of a speaker in SLAAP

Top 20 POS bigrams	Utterance Initial	Utterance Final
# PRP = 29	# PRP = 29	NN # = 25
NN # = 25	# CC = 21	IN # = 10
# CC = 21	# RB = 11	x? # = 9
PRP VBE1 = 17	# IN = 10	RB # = 8
CC PRP = 15	# x? = 7	PRP # = 6
PRP$ NN = 13	# DT = 6	JJ # = 5
PRP VB = 12	# OK = 5	VBG # = 5
VBE1 VBG = 11	# PRP$ = 3	NNS # = 5
# RB = 11	# WRB = 2	VB # = 4
VB IN = 11	# VB = 2	CC # = 4
IN PRP = 11	# JJ = 1	OK # = 3
PRP VBP = 10	# CD = 1	VBP # = 3
# IN = 10	# WP = 1	RP # = 2
IN # = 10	# NNS = 1	CD # = 2
DT NN = 10		VBD # = 1
PRP VBD = 9		JJR # = 1
PRP MD = 9		EX # = 1
x? # = 9		PRP$ # = 1
RB # = 8		NNP # = 1
RB PRP = 7		VBE1 # = 1
# x? = 7		WDT # = 1
		DT # = 1
		VBZ # = 1

Another useful feature of SLAAP is the summary statistics of number of lines, turns, words etc. of a transcript, as illustrated in Figure 6.

Figure 6. Screenshot of summary statistics for one transcript in SLAAP

Transcript Summary Statistics

Transcript has **2** speakers: *CM, ORH, JF, anon*

Transcript total temporal length: **2,755.75** seconds (45.93 minutes)

Transcript total line length: **3,720** lines (including blank lines, e.g., pauses)
Total non-blank lines: **1,858**

Speaker	Talk Lines[1]	Turn Lines[1]	Words	Words of Tran	Talk-Time (sec)	Talk-Time of Total Talk[2]	Turn-Time (sec)	Turn-Time of Entire Tran[3]
CM	755	1,024	2,952	39.01 %	620.61	38.64 %	825.09	29.94 %
ORH	1,007	1,505	4,207	55.59 %	882.53	54.94 %	1,272.62	46.18 %
JF	87	103	359	4.74 %	82.11	5.11 %	89.05	3.23 %
anon	9	9	50	0.66 %	21.05	1.31 %	21.05	0.76 %
Totals:	1,858	2,641	7,568	100 %	1,606.30	100 %	2,207.82	80.12 %

3.2 International Corpus of English

ICE (http://www.ucl.ac.uk/english-usage/ice/) is a global project whereby English language materials from many national varieties of English are being collected and marked up according to common guidelines. As of January 2008, there were 19 varieties of English represented, according to the website. These varieties include better known ones such as Great Britain and USA, as well as lesser known ones such as Malta, Philippines, and Sri Lanka. A full description of the project is given in Greenbaum (1996). The discussion that follows focuses on the spoken component of ICE.

According to the homepage of ICE, the primary aim of ICE is "collecting material for comparative studies of English worldwide". Clearly, the reference to "comparative studies" targets language researchers with an interest in one or more varieties of English. As such, ICE is one of the most broadly appealing collections of English language material that exists, or will exist when completed. Its focus is a major world language, as spoken around the world, guaranteeing substantial academic interest. Comparative studies of English, however, can extend beyond studies of a purely descriptive or theoretical linguistic orientation and can be undertaken for more general, cultural reasons. So, for example, relative overuse and underuse of lexical items in corpora of

different varieties of English might be used as a basis for inferring cultural differences between the societies associated with those varieties. Relevant studies using English corpora in this way are Hofland and Johansson (1982), Leech and Fallon (1992), Oakes, M. P. (2003), and Oakes and Farrow (2007). Leech and Fallon (1992), for example, compared relative overuse and underuse of lexical items in the (American) Brown Corpus and the (British) LOB Corpus to explore differences between American and British cultures. They found, among other findings, that business vocabulary (bond(s), budget, corporation, costs, stock, stockholders etc.) is more frequent in the BROWN Corpus than in the LOB corpus. Speaking for the Canadian component of ICE, ICE-CANADA, of which I am currently the project leader, there is no question that some of the funding available to ICE-CANADA has come about on account of the broad appeal of the project. The idea of a "national" corpus, reflecting Canadian usage of English, seems to strike a chord with Canadians who would otherwise have no particular interest in the concerns of linguists.

Each ICE corpus consists of one million words of spoken and written English produced after 1989, of which 600,000 words constitute the spoken component. The collection of the spoken material, carried out in the 1990s, was a large undertaking on account of the requirements of ICE. Data from various spoken genres were required, as shown in Table 1. Audio files are available for ICE-GB as a separate purchase and can be brought up and listened to for concordance lines. For every "text" (understood as a 2,000 word sample of connected speech or writing), written permission has to be obtained. Usually, this involves written permission from individuals, but sometimes, as with radio broadcasts, blanket permission from the broadcasting organization suffices. Also, metadata (age, gender, place of birth, etc.) for each speech participant is documented as far as possible and entered into a metadata file. Each file is marked up in SGML, which includes, for the spoken texts, features such as overlapping speech, unclear words, self-corrections, incomplete words, pauses, etc. Some variation in terms of just how many features are noted by SGML markers is allowed for. So, for example, all ICE corpora are expected to mark unclear words, whereas marking pauses is "recommended, but not essential". Eventually, the aim is to have all the ICE corpora tagged for part of speech as well as for grammatical relations, but not all corpora are that advanced at this point. ICE-GB is also available with all the original audio files. ICE-GB has featured as

the preferred corpus of English in much of the recent research known as "collostructional analysis" (Stefanowitsch and Gries 2003), largely due to full grammatical annotation of the corpus.

Table 1. Categories for the spoken component of an ICE corpus
Numbers in parenthesis refer to the number of texts, each text containing approximately 2,000 words.

Dialogues (180)	**Private** (100)	Conversations	(90)
		Phonecalls	(10)
	Public (80)	Class Lessons	(20)
		Broadcast Discussions	(20)
		Broadcast Interviews	(10)
		Parliamentary Debates	(10)
		Cross-examinations	(10)
		Business Transactions	(10)
Monologues (120)	**Unscripted** (70)	Commentaries	(20)
		Unscripted Speeches	(30)
		Demonstrations	(10)
		Legal Presentations	(10)
	Scripted (50)	Broadcast News	(20)
		Broadcast Talks	(20)
		Non-broadcast Talks	(10)

ICE-GB is distributed on a CD and is searchable using the ICECUP software which is packaged with it. ICECUP is a sophisticated tool for querying the corpus. Apart from displaying the usual KWIC-style concordances, ICECUP displays a full grammatical analysis of an utterance. Figure 7 is the parsed tree which the software displays for the utterance containing as its core "I mean uh what's incomplete". As can be seen, the software displays the parts of speech (e.g., PRON, V, ADJ) and grammatical relations (e.g., SU = subject) for the whole of the utterance, including parts of the utterance which have been repeated (shown with a horizontal line through them and colored red. Pauses (shown as <,>), and unclear words are also displayed if desired. The repeated parts of the utterance are shown in light shading in the parsed tree. Searches can fully exploit the grammatical markup and the

demographic and genre variables. Figure 8, for example, illustrates part of the concordance resulting from a search for adjectival phrases consisting of ADV+ADJ sequences. The search was restricted to the genre of direct conversation and further restricted to the speech of 18-25 year olds. A window at the bottom of the concordance lines reveals the logic of the query. Note how the search ignores intervening unclear words between ADV and ADJ (e.g., too <unclear word> true). The most recent release of ICECUP (3.1) includes "statistical tables" which summarize frequency information and measures of statistical significance. The sophistication and completeness of ICE-GB, a model for other ICE projects, comes at a price. Substantial funding is required to sponsor this kind of research, as described above and there is a cost for the end-user.

Figure 7. Screenshot of a parsed tree display using ICECUP

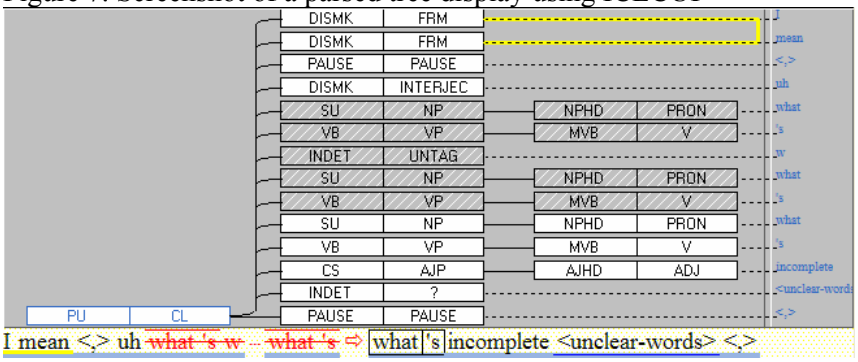
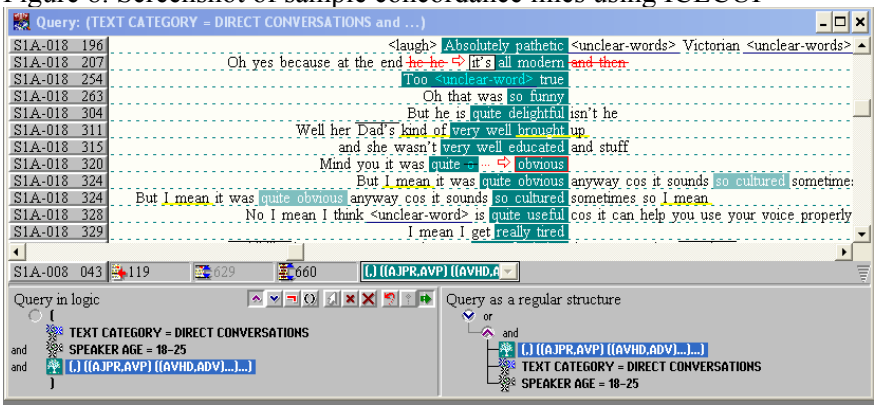


Figure 8. Screenshot of sample concordance lines using ICECUP



3.3 Wenzhou Spoken Corpus

WSC is an online corpus of transcribed spoken Wenzhou, a dialect of Chinese, developed under my supervision at the University of Alberta 2004-06 (http://corpora.tapor.ualberta.ca/wenzhou/). A full description of the corpus is available in Newman, Lin, Butler, and Zhang (2007). It is freely accessible and is not password-protected.

The impetus to create WSC came initially from Jingxia Lin, at the time an MSc student in the Department of Linguistics at the University of Alberta. Lin had, on her own initiative, collected some audio examples of Wenzhou, spoken by a population of about 7.5 million in the southeast part of Zhejiang Province. Wenzhou is Lin's own dialect and she had collected the samples of Wenzhou from her family and friends in a very informal and sporadic way. It was agreed that her MSc Project would investigate aspects of Wenzhou grammar based on the corpus she had collected.

Only very limited funding was available to this project in the form of funding for a Research Assistantship (awarded to Lin) to pay for some of the time required to transcribe and markup texts. However, and crucially, Lin was able to draw upon the resources of the digital humanities infrastructure at the University of Alberta, known as TAPoR (Text Analysis Portal for Research). TAPoR is actually a pan Canadian, multi-institutional assembly of computing infrastructure and expertise, supported by a Canada Foundation for Innovation grant (see http://tapor.ualberta.ca/). Through extended consultation with TAPoR,

agreement was reached about the procedure for Lin to follow in carrying out the transcription and markup (in XML).

As summarized in Table 2, the data collected by Lin amounts to more than 158,000 words (256544 Unicode characters), when transcribed, and falls into six genres. For genres (1)-(4) in this Table, the method of obtaining the data was to rely upon the social networks of Lin which included family members, friends, friends of friends, etc. The Internet Chat data is based on spoken exchanges, rather than typed messages. For the Story genre, the speaker was given a short story written in Mandarin (characters) to read silently and then asked to tell the story in Wenzhou without looking at the text. The informal context of the Story sessions allowed for a certain amount of free conversation, though the narration of the story dominates in each case. News Commentary was collected from the program Baixiao Jiang Xinwen 'News talk by know-it-all', a netcast provided by the Economic and Science Channel, Wenzhou TV, and used with their permission (which was readily given). The program includes relatively informal news commentaries as well as opinions offered by anonymous interviewees. Most of the corpus comes from the recorded News Commentary, reflecting the relatively easy access to this category of data. The Song category is composed of the texts of traditional Wenzhou children's songs. The data collection method was clearly "opportunistic" with a resulting over-representation of some demographic categories, as shown in Table 2. Just over 25% of the total words derive from the more spontaneous, conversational contexts (Face to Face Conversation 8%, Phone Call 13%, Internet Chat 4%) and in these categories there is a clear predominance of speakers under 34 years of age and at least high school level of education. Overall, male speakers outnumber female speakers about 3:1.

Table 2. Categories for WSC

		Word Count	Unicode Character Count
1	Face to Face Conversation	13009 (8.22%)	23582
2	Phone Call	20885 (13.20%)	36257
3	Internet Chat	7005 (4.42%)	13132
4	Story	1046 (0.66%)	2470
5	News Commentary	115293 (72.90%)	179708
6	Song	894 (0.56%)	1395
	Total	**158132**	**256544**

The corpus is encoded into Unicode characters. There is a certain amount of variation and uncertainty when it comes to writing Wenzhou in Chinese characters. We opted to represent the Wenzhou forms currently lacking Unicode Chinese glyphs in IPA transcription. The writing of sentence final particles, in particular, is problematic and these were transcribed using the IPA. The Wenzhou Fangyan Dictionary (You and Yang 1998) was used as a reference both for the Chinese characters and for the normalized phonetic transcription. Personal names and other confidential information were edited to ensure anonymity of participants. Spoken Wenzhou has a number of phonetically contracted forms which required some procedural decisions concerning transcription. The negative morpheme [fu] 不, for example, combines with certain following morphemes to produce contracted forms: [fu ha] 不勾 'not give' can be reduced to [fa], [fu hə] 不好 'not good' can be reduced to [fə]. Here, too, we followed the practice of You and Yang (1998), using a single character where the dictionary lists one (𡍮 for [fə] 'not good') but using two characters otherwise (as in the case of [fu ha] 不勾 'not give').

There are two main search tools which are available as part of the interface to the WSC: Concordance and Collocates. Both tools provide the same set of options to restrict the search according to selections by gender, aged band, and level of education. A Concordance tool has options to allow the display of expanded context (to include preceding and following turns), as well as the complete demographic information for each speaker (gender, age, educational level, etc.). In this, we were influenced by BNCWeb, the web-based interface to the British National Corpus (Hoffmann and Evert 2006). Figure 9 is a sample of concordance lines displayed by the Concordance tool. A Collocates tool offers two options for displaying results. An "aggregated" display lists all the word

types and the number of tokens in the selected span. A "collocates by position" display shows the collocates as they occur in each word position to the left and to the right of the search word, with the collocates in each position listed in descending order of frequency. This kind of arrangement of data follows the suggestion by Stubbs (2001:87-96) who recommends such a display as a basis for "lexical profiling" of a word. A sample of collocate results by position is shown in Table 3.

Figure 9. Screen shot of a sample of concordance lines for the keyword 温州 'Wenzhou' in WSC

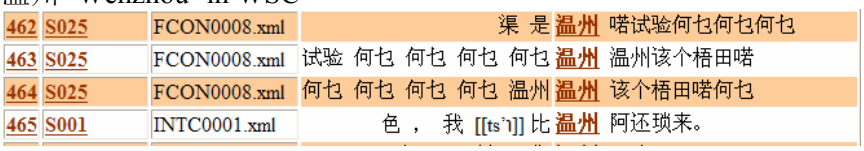


Table 3. Part of the results for the keyword 温州 'Wenzhou' using the 'collocates by position' option (* = keyword) in WSC.

-5	-4	-3	-2	-1		1	2	3	4	5
个	个	呢	是	卬你	*	市	个	个	个	个
(22)	(31)	(17)	(22)	(82)		(48)	(44)	(33)	(33)	(34)
呢	呢	个	个	宿	*	话	呢	里	呢	呢
(12)	(15)	(15)	(15)	(28)		(43)	(19)	(19)	(14)	(15)
有	温州	人	呢	个	*	人	有	一	有	人
(11)	(12)	(13)	(15)	(26)		(35)	(12)	(12)	(12)	(10)
侬	人	阿	宿	是	*	个	是	有	温州	是
(9)	(11)	(11)	(12)	(23)		(34)	(11)	(12)	(12)	(9)
该	侬	温州	侬	走	*	呢	站	阿	里	有
(8)	(8)	(8)	(11)	(19)		(19)	(10)	(8)	(8)	(9)
温州	该	讲	该日	讲	*	有	里	局	多	一
(7)	(7)	(7)	(10)	(13)		(12)	(8)	(8)	(7)	(7)
是	一	渠	一	拉	*	火车	哪	呢	侬	温州
(7)	(7)	(7)	(10)	(10)		(9)	(7)	(8)	(6)	(7)
里	我	侬	就	能界	*	大学	文明	温州	人	会
(6)	(5)	(7)	(9)	(9)		(7)	(7)	(8)	(6)	(6)

3.4 Dinka Narratives Corpus

DNC is an online collection of narratives told by members of Edmonton's Dinka-speaking community (http://ra.tapor.ualberta.ca/Dinka/). Dinka belongs to the Western Nilotic branch of the Nilo-Saharan

language family, spoken mainly in the southern Sudan. DNC is accessible online, with no password protection. No research funding was available for this project. As with the WSC, however, the project was able to draw upon the invaluable human and computing resources of TAPoR at the University of Alberta. TAPoR provided, free of charge to the project, consultation with a programmer/technician and free disk space on a server.

DNC evolved out of a Community Service Learning (CSL) initiative at the University of Alberta. CSL is a program which integrates volunteer work with classroom studies. The basic idea is that university instructors and community partners work together to design volunteer projects and experiences to meet the needs of community organizations and fulfill the objectives of academic courses. Students, in turn, are expected to reflect critically on how their experiences help them to develop both as scholars and as citizens. Campus Contact, the organization representing CSL in America, claims a membership of more than 1,100 colleges and universities. Information on CSL at the University of Alberta can be found at http://www.uofaweb.ualberta.ca/arts/CSLhome.cfm. In Fall 2004, the author offered undergraduate students in an upper-division elective course LING 324 Endangered Languages an opportunity to participate in a CSL program devised for that course. Students had opportunities to volunteer in a number of non-profit organizations in and around Edmonton where language shift and language loss were critical issues. One of the community partners in that year was the Edmonton Mennonite Centre for Newcomers where students came into contact with members of the Dinka community who had come to Canada from the Sudan. Following on from this original contact with the Dinka community through CSL, one of the students involved in that course, Kristina Geeraert, continued working with the Dinka community for her Honours Thesis in 2005-06 and as part of an Undergraduate Research Award in the summer of 2006. Her thesis (Geeraert 2006) and the research she carried out for her Research Award are unusual in the extent to which the Dinka community was consulted as to what kind of language research and development they would like to see for their language in Edmonton. Through focus group meetings with the Dinka community, an agreement was reached that the documentation of traditional Dinka stories would be a highly desirable resource for the community and would be supported by the University of Alberta. The

focus group meetings had emphasized that traditional story telling served both recreational and moral-building purposes and continues to do so, to some extent, in the Edmonton context. It was agreed that an online collection of stories would be desirable as a way of facilitating greater exposure, especially with respect to younger members of the community, to the language and the traditional values expressed through the stories. While the DNC came about in response to community wishes, it also served linguistic ends since it constituted important data underlying the analysis of aspects of Dinka, especially the tones (cf. Geeraert 2006).

DNC consists of four stories and a total of about 2,000 words. In addition, there is an untranscribed song, especially performed and recorded for the website. The inclusion of the audio for the song reflects the wishes of the participants from the community organization. Since there was no funding supporting this project, the collection of data proceeded on a purely voluntary basis, when and where it was convenient for all parties concerned. Another reason for the small size of the corpus (about the size of one single text of the 500 texts that make up an ICE corpus) is that each word is recorded, in isolation, in addition to the recording of the whole story as connected speech, making the collection a relatively time-consuming activity.

The transcription of the texts was carried out in very simple XML. The texts make use of a couple of different orthographies, a result of alternative traditions within the Dinka-speaking community and also to some extent a reflection of some dialectal differences. Some thought was given to normalizing the transcription, but, one again, the wishes of the community were respected and the different orthographic conventions desired by the community were maintained. In addition to providing a free translation of each sentence of a story, each word in a story is glossed with an approximate English translation.

DNC has a fairly basic look to it, as shown in Figure 10. Figure 10 shows the format adopted throughout the website which consists of the transcribed Dinka text together with an English translation, sentence by sentence. Clicking the speaker icons will play the audio file of the title or a whole paragraph. Clicking on individual words will play the audio file for that particular word, as spoken in isolation. Word definitions in English are available as pop-ups by "mouse-over" of words, as shown by the popup "person" when the cursor points at *raan* in Figure 10. The community was invited to contribute colored drawings appropriate for

the story as a whole and for individual paragraphs, since these could easily be added to the website, but the drawings were not forthcoming.

Figure 10. Screenshot from DNC, showing popup of the meaning of *raan* 'person'

In addition to the texts themselves, a blog was created which would allow community members to react to the stories, as well as discuss online any topic of interest. Part of the rationale for this was that the moral-building part of traditional storytelling involved discussion and informal analysis of a story after it has been told, often in a group setting. A blog could serve this purpose to some extent, allowing online follow-up to the narratives.

4. FINAL REMARKS

Corpora are a fact of life now in linguistics and represent an important and understudied source of linguistic evidence. Spoken corpora have a very special role to play within corpus-based research, especially those corpora documenting what is the most basic and the

most direct form of communication: face-to-face conversation. For languages without a written tradition, spoken corpora assume an even greater value since they document the only mode of communication. The interest that linguists have in spoken corpora overlaps, in some ways, with an interest in various kinds of cultural documentation in society at large, particularly oral history, and there can be mutual benefit to linguists and non-linguists in making speech recordings which serve larger cultural goals. DNC, modest as it is, illustrates this kind of collaborative, university-community alliance.

The four corpora considered here as examples of spoken corpora vary along many dimensions: sophistication, rationale, intended audience, etc., as described above. In terms of funding, SLAAP and ICE represent one extreme, attracting major funding, justified in terms of the potential contribution that such corpora can make to our knowledge about a major world language and its varieties. At the other extreme of funding lies the DNC, a community-centered initiative which forms part of an attempt to retain and revitalize a lesser known language in an immigrant community. In between these extremes lies WSC which was largely the result of an initiative on the part of an interested individual, working on her own, but with some research funding to elevate the project to a more professional level. Needless to say, the funding available (or not available, as the case may be) to each of these projects has been a key factor in determining the scope of the project and the level of sophistication that has been possible for each project. Obviously, a corpus project has to respond to a need, be it a need arising out of university-based research or out of a community-oriented project. And when it comes to judging the "value" of a corpus, it is important to assess the corpus with respect to the purposes for which it has been developed. From this point of view, the four corpora reviewed here could all be viewed as "successful", but in very different ways and to varying degrees, reflecting the very different contexts in which they arose.

Little has been said above about technical detail of the corpora: the markup schemes, the XML tags, etc. (see Newman et al. 2007 for technical details pertaining to WSC). The corpus projects developed at the University of Alberta have used XML (hence Unicode) as the preferred format in which transcription and markup are done. XML offers the greatest flexibility in terms of annotation, as well as having very well developed resources in support of this format. All the web pages for WSC, for example, are created directly from XML files using

the PHP programming language. Beyond advocating an XML-based approach to corpus construction, there is little else that one could insist upon in terms of design, markup, annotation, or interface for spoken corpora. One's openness to different practices in corpus construction is, of course, closely tied to a corresponding openness as to the purposes of corpus projects that linguists can, and should be, engaged in.

REFERENCES

Evans, Vyvyan and Melanie Green. 2006. *Cognitive Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.

Geeraert, Kristina. 2006. *Ten Sons and their Father: A Linguistic and Community-Oriented Study of a Dinka Story*. BA Honours Thesis, University of Alberta.

Goodwin, Charles. 1979. The interactive construction of a sentence in natural conversation. *Everyday Language: Studies in Ethnomethodology*, ed. by George Psathas, 97-121. New York: Irvington.

Goodwin, Charles. 1980. Restarts, pauses, and the achievement of mutual gaze at turn-beginning. *Sociological Inquiry* 50.3/4:272-302. (Special Double Issue on Language and Social Interaction, edited by Don Zimmerman and Candace West).

Goodwin, Charles. 1981. *Conversational Organization: Interaction Between Speakers and Hearers*. New York: Academic Press.

Gordon, Elizabeth., Margaret Maclagan and Jennifer Hay. (to appear). The ONZE Corpus. *Models and Methods in the Handling of Unconventional Digital Corpora*, *Volume 2: Diachronic Corpora*, ed. by J.C. Beal, K.P. Corrigan, and H. Moisl. Houndmills: Macmillan Palgrave.

Greenbaum, Sidney. (ed.) 1996. *Comparing English Worldwide: The International Corpus of English.* Oxford: Clarendon Press.

Gut, Ulrike and Petra Saskia Bayerl. 2004. Measuring the reliability of manual annotations of speech corpora. *Speech Prosody 2004*, ed. by Bernard Bel and Isabelle Marlien, 565-568. Available at the International Speech Communication Archive (ISCA), http://www.isca-speech.org/archive/sp2004.

Hoffmann, Sebastian and Stefan Evert. 2006. BNCweb (CQP-Edition) - The marriage of two corpus tools'. *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, ed. by Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, 177-195. Frankfurt am Main: Peter Lang. Available from http://es-sebhoff.unizh.ch/Hoffmann-Evert.pdf.

Hofland, Knut and Stig Johansson. 1982. *Word Frequencies in British and American English*. Bergen: Norwegian Computing Centre for the Humanities/London: Longman.

Joseph, Brian. 2004. The Editor's Department: On change in *Language* and change in *language*. *Language* 80.3:381-383.

Kendall, Tyler. 2007. Enhancing sociolinguistic data collections: The North Carolina Sociolinguistic Archive and Analysis Project. *University of Pennsylvania Working Papers in Linguistics* 13.2:15-26.

Leech, Geoffrey and Roger Fallon. 1992. Computer corpora – what do they tell us about culture? *ICAME Journal* 16:29–50.

McEnery, Tony, Richard Xiao, and Yukio Tono. 2006. *Corpus-based Language Studies: An Advanced Resource Book.* Milton Park: Routledge.

Newman, John, Jingxia Lin, Terry Butler, and Eric Zhang. 2007. The Wenzhou spoken corpus. *Corpora* 2.1:97-109.

Oakes, Michael P. 2003. Contrasts between US and British English of the 1990s. *Research and Scholarship in Integration Processes*, ed. by Elzbieta H. Oleksy and Barbara Lewandowska-Tomaszczyk, 213–22. Lodz: University of Lodz Press.

Oakes, Michael P. and Malcolm Farrow. 2007. Use of the Chi-Squared test to examine vocabulary differences in English language corpora representing seven different countries. *Literary and Linguistic Computing* 22.1:85-99.

Ochs, Elinor. 1979. Transcription as theory. *Developmental Pragmatics*, ed. by Elinor Ochs and Bambi B. Schieffelin, 43-72. New York: Academic Press.

Schriberg, Elizabeth, Rebecca Bates, Andreas Stolcke, Paul Taylor, Daniel Jurafsky, Klaus Ries, Noah Coccaro, Rachel Martin, Marie Meteer, and Carol Van Ess-Dykema. 1998. Can prosody aid the automatic identification of dialog acts in conversational speech? *Language and Speech* 41.3/4:443-492.

Stefanowitsch, Anatol and Stefan Th. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8.2:209-243.

Stubbs, Michael. 2001. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Svartvik, Jan. (ed.) 1990. *The London Corpus of Spoken English: Description and Research*. Lund Studies in English 82. Lund: Lund University Press.

Tummers, José, Kris Heylen, and Dirk Geeraerts. 2005. Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory* 1.2:225-261.

Wichmann, Anne. 2007. Corpora and spoken discourse. *Corpus Linguistics 25 Years on*, ed. by Roberta Facchinetti, 73-86. Amsterdam and New York: Rodopi.

Wynne, Martin. (ed.) 2005. *Developing Linguistic Corpora*. Oxford: Oxbow Books. Available for free online at http://ahds.ac.uk/linguistic-corpora/.

Yang, Xiao-Jun. 2006. Survey and prospect of China's corpus-based research. *Corpus Linguistics Around the World*, ed. by Andrew Wilson, Dawn Archer, and Paul Rayson, 219-233. Amsterdam and New York: Rodopi.

You, R. and Q. Yang. 1998. *Wenzhou Fangyan Ci Dian* [Wenzhou Fangyan Dictionary]. Nanjing: Jiangsu Jiaoyu Chubanshe [Jiangsu Education Press].

John Newman
Department of Linguistics
4-32 Assiniboia Hall
University of Alberta
Edmonton, AB
T6G 2E7 Canada
john.newman@ualberta.ca