Mining Years and Years of Speech

Final report of the Digging into Data project “Mining a Year of Speech”[1]

John Coleman	Mark Liberman
Greg Kochanski	Jiahong Yuan
Sergio Grau	Chris Cieri
Ladan Baghai-Ravary	
Lou Burnard	
University of Oxford	University of Pennsylvania
Phonetics Laboratory	Linguistic Data Consortium

1. Introduction

In the “Mining a Year of Speech” project we assessed the challenges of working with very large digital audio collections of spoken language: an aggregation of large corpora of US and UK English held in the Linguistic Data Consortium (LDC), University of Pennsylvania, and the British Library Sound Archive. The UK English material is entirely from the audio recordings of the Spoken British National Corpus (BNC), which Oxford and the British Library have recently digitized. This is a 7½ million-word[2] broad sample of mostly unscripted vernacular speech from many contexts, including meetings, radio phone-ins, oral history recordings, and a substantial portion of naturally-occurring everyday conversations captured by volunteers (Crowdy 1993). The US English material is drawn together from a variety of corpora held by the Linguistic Data Consortium, including conversational and telephone speech, audio books, news broadcasts, US Supreme Court oral arguments, political speeches/statements, and sociolinguistic interviews. Further details are given in section 2, below.

Any one of these corpora by itself presents two interconnected challenges:

(a) How does a researcher find audio segments containing material of interest?
(b) How do providers or curators of spoken audio collections mark them up in order to facilitate searching and browsing?

The scale of the corpora we are bringing together presents a third substantial problem: it is not cost-effective or practical to follow the standard model of access to data (i.e. give people copies of the corpus and let them find the bits they need). Even with the declining costs of networking and hard disks, moving terabytes of data costs money and takes substantial amounts of time. Researchers cannot reasonably distribute such corpora gratis, and a casual inquirer, wishing only to look up a few examples, would not normally be expected to spend the time or money. Consequently, we examined the

[1] We wish to acknowledge the support of JISC (in the UK) and NSF (in the USA) for their joint support of Mining a Year of Speech, under the Digging into Data Challenge. Digitization of the Spoken BNC was funded by the John Fell OUP Research Fund, the British Library and the University of Oxford Phonetics Laboratory. We thank all of these funders. We also acknowledge with thanks the invaluable support of support staff and research assistants: at Oxford, John Pybus, whose idea it was to set up a Condor cluster for the heavy-duty processing; our British Library collaborators Jonnie Robinson, Joanne Sweeney and their Sound Archive colleagues Christine Adams, Adam Tovell, Will Prentice and Tony Harris; and at UPenn, Stephanie Sessa, Cris Feo, and Dave Graff. Also, particular thanks are owed to our “Digging into Data” conference respondent, Dan Jurafsky, for his commentary on our work.
[2] The previously published XML transcriptions of the spoken part of the BNC amount to 10 million words, but as not all of the original recordings have come down to us, we estimate the audio as amounting to c. 7.4 million words.

question:

(c) How can we make very large scale audio collections accessible to others?

The large size of spoken audio corpora is a consequence of its being real-time, broad-bandwidth digital medium. Each minute of monophonic audio encoded in 16-bit uncompressed .wav data at a sampling rate of 16,000 samples/s requires 1.92 MB; that's 115.2 MB per hour, 2.77 GB per day, 1 TB per year. Stored at CD quality, and/or in stereo, requires far more storage. The very high fidelity .wav files employed in libraries and sound archives (24 bit, 96 kHz sampling rate) uses 288 KB/second, 1 GB/hr, ... 9 TB/year. Most critically, it takes a while merely to read that amount of data: disk data transfer rates are just 100 MB/s under the best conditions, so just reading or copying 9 TB takes at least 90,000 seconds: a whole day. If a user wanted to search through the data for something, a search algorithm would be expected to take even longer than that. Transferring data over a network connection is typically slower still: download times would be at least days, and perhaps weeks.

Two things are needed to make large corpora practical: a compact index, so that users can find the parts they need without paging through all the data, and a way of accessing the slices of the corpus that they find from the index. This is true both for casual users and even for people who will become intensive users: it would be foolhardy to commit to using a large corpus without a way to sample it and see what it contains.

In the case of large spoken-language corpora, potential users range from members of the public interested specific bits of content (including e.g. school students), to scientists with broader-scale interests that may be served by text search alone (e.g. legal scholars, political scientists), to phoneticians, speech engineers and others (e.g. some social psychologists) who need to search and retrieve based on features of pronunciation and sound. We expect that most potential users would already prefer to run searches on a server hosted by others -- corpus providers or research libraries or other third-party search systems -- and that the proportion will increase with the on-going exponential increase in available spoken material. Our plans for online access to the grove of corpora in our sample are discussed in section 5 below.

To give a sense of the scale of the Mining a Year of Speech corpora, let's make some comparisons with some previously-collected speech corpora, as well as with some famous "big science" datasets. The largest comparable transcribed and aligned spoken language dataset is the Switchboard corpus[3], published by LDC, which amounts to just 13 days of audio (36 GB). The Spoken Dutch Corpus[4] contains about one month of audio, but only a fraction is phonetically transcribed. There is a large (110 hours, 13 GB) corpus of orthographically transcribed spoken Spanish[5]. Other large speech datasets are the Buckeye Corpus from OSU, which is about 2 days of speech, and the Wellington Corpus of Spoken New Zealand English, which is about 3 days. As a rule of thumb, the audio files of a spoken audio corpus require over 250 times the storage needed for the corresponding transcriptions marked up in XML.

In fact, digitized audio data forms only a fraction of the currently available spoken audio material. Almost every sound archive and library holds analogue magnetic tape recordings; usually far more than their digital audio collections. For example, the British Library has over 1 million disks and tapes, of which only 46,200 are available in digitized form via the Archival Sound Recordings server[6]. The Library of Congress Recorded Sound Reference Center includes over 2 million items including one of the largest and most important archival collections of modern storytelling in the world (donated by the International Storytelling Foundation, Jonesborough, Tennessee), which includes eight thousand hours of audio and video recordings. The PRESTO Project (Preservation Technologies for European Broadcast Archives) estimates there are over 20 million hours - 2283 years - in European Broadcast archives alone (Wright and Williams 2001) and 100 million

[3] http://www.ldc.upenn.edu/Catalog/readme_files/switchboard.readme.html

[4] http://lands.let.ru.nl/cgn/ehome.htm

[5] http://lablita.dit.unifi.it/coralrom/uam_corpus.html

[6] http://sounds.bl.uk/

hours - 11,415 years - of analogue tape recordings in broadcast collections throughout the world (Bradley 2003). According to Wright and Williams (2001), 75% of the audio material in the 10 European broadcast archives in their survey are in ¼-inch tape format, 20% is on shellac and vinyl, and only 7% on digital media. If worldwide collections of analogue tape recordings were digitized according to the high fidelity standards employed in libraries and sound archives, they would require approximately 100 PB of storage. Ironically, some older material (e.g. vinyl disks), though in need of appropriate preservation standards, is more durable than ¼-inch tape; material in older digital formats (e.g. PCM on Betamax tape) is at greater risk of obsolescence as the equipment needed to play them is increasingly difficult to obtain or maintain.

Bringing these examples together, it is apparent that the scale of data storage resources needed for very large speech corpora (and some other “big humanities” projects) stands comparison with major international “big science” initiatives (Table 1).

Table 1. Relative scales required for data storage in sciences vs humanities

	“Big Science”		**“Big Humanities”**	
Smaller scale ...	Human genome	3 GB		
			Switchboard corpus	36 GB
			DASS audio sampler (Digital Archive of Southern Speech)	350 GB; more to come
Mid-scale ...	Hubble space telescope images	0.5 TB/year		
			Year of Speech	1 TB (9 TB archive format)
	Sloan digital sky survey	50 TB		
Large scale ...			CLAROS Web of Art	>25 TB
			C-SPAN archive	probably >20 TB
			USC Shoah Archive	200 TB
	Large Hadron Collider	13 PB/year		
			Worldwide analogue audio	12 - 100 PB

The cost to a research library for professionally digitizing analogue audio tapes and to create a corresponding library catalogue entry is ~£20 ($32) per tape.[7] Using speech recognition and associated natural language technologies (e.g. summarization) could provide much more detailed catalogue entries to be created without the need for so much time-consuming (and hence costly) human listening. This offers the prospect of greatly reducing the costs of digitization

[7] This was the actual 2009 cost to the British Library for digitizing and cataloguing one of the C-90 compact cassettes in the Spoken BNC. The audio consultant overseeing the work used a professional digitization system that worked on several tapes simultaneously.

projects.

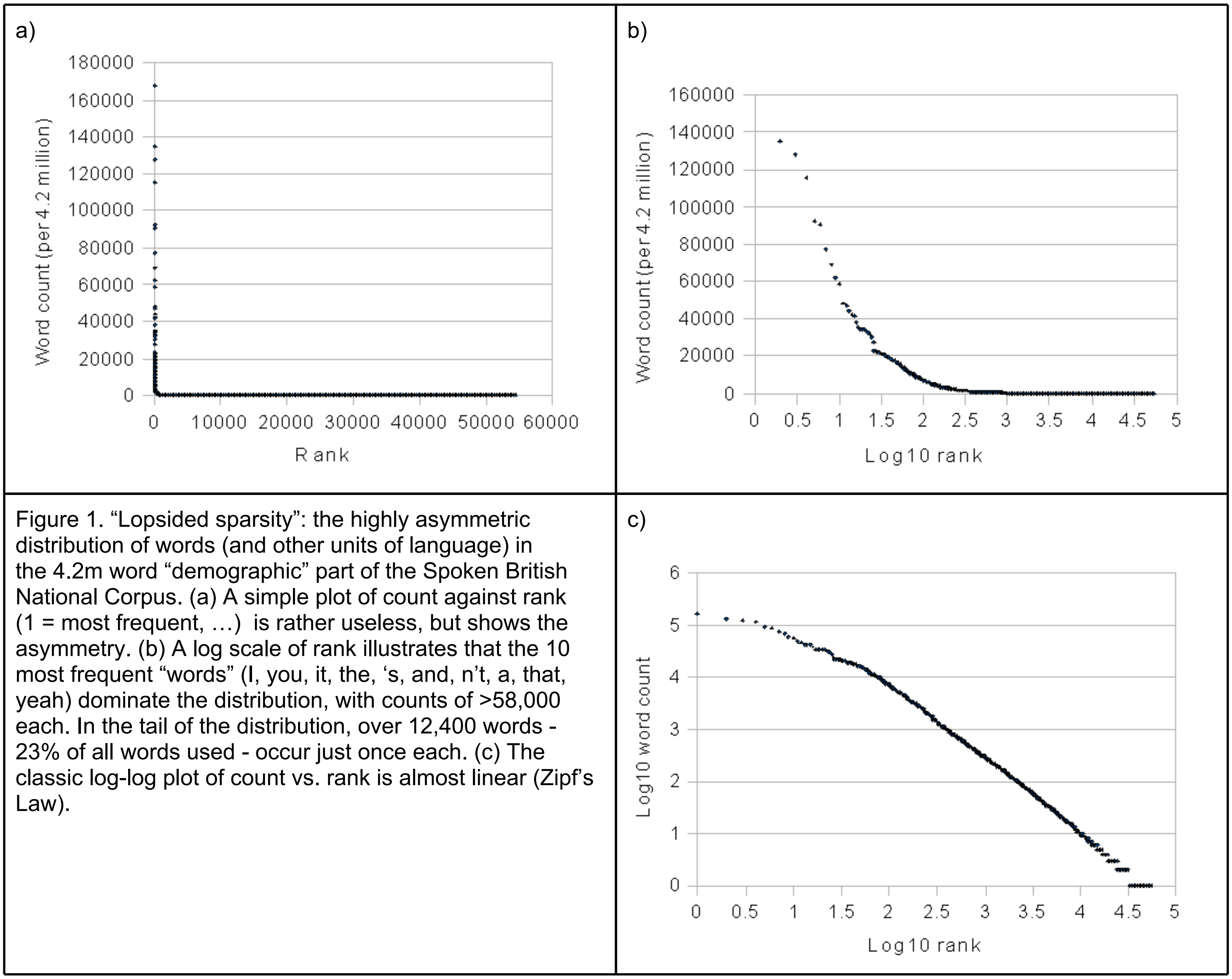


Figure 1. “Lopsided sparsity”: the highly asymmetric distribution of words (and other units of language) in the 4.2m word “demographic” part of the Spoken British National Corpus. (a) A simple plot of count against rank (1 = most frequent, …) is rather useless, but shows the asymmetry. (b) A log scale of rank illustrates that the 10 most frequent “words” (I, you, it, the, ‘s, and, n’t, a, that, yeah) dominate the distribution, with counts of >58,000 each. In the tail of the distribution, over 12,400 words - 23% of all words used - occur just once each. (c) The classic log-log plot of count vs. rank is almost linear (Zipf’s Law).

But before we delve more deeply into the problems of time, storage and cost that such large collections present, and what we have learned about how to deal with them, let’s first spell out what various researchers stand to gain by having them. For research into observed (as opposed to elicited) language, **large size** is important because of the extremely lop-sided statistical distribution of the usage of words and other units of language (Figure 1). Even a small corpus will include numerous instances of the most frequent items; enlarging the scope of the corpus helps to capture more of the least frequent items - those occurring only once - which always account for a substantial fraction of the corpus. For example, here’s a small selection of words that occur only once in the 4.2m-word demographic sample of the Spoken BNC, hand-picked for their novelty value:

aeriated	aqualunging[8]	battyman	beaky	bodacious	bolshiness
boringest	bruv	bullshitter	canoodling	chambermaiding	chichi

[8] 'Aqualung' does not occur in the Spoken BNC except in this derived form!

chinesey
clangers
de-grandfathered
de-nailing
doofers
drived
drownded
europeaney
even-stevens
frazzled
gakky
gazump
geriatric-ing
gronnies
guesstimate
hap'orth
headbangers
hersen[9]
hoptastic
jimmy-jams
kiddiwinks
kiddy-fied
kiss-off
lawnmowing
lughole
mellies
mindblank
munchkins
mustardy
neaps
noggin
noseless
penn'orth
pickleey
plankers
potatoey
prickable
proposedly
punny
pythonish
raunchiest
re-putty
re-snogged[10]
reffing
regurgitate-arianism
relinoing
risqué
rissole
roofery
salady
sameyness
sarnies
sausagey
scootled
scraggle
scrobbling
scrooving
scruffle
scunny
semi-carnally
sequined
shebumkin
sheeps
shilly
shit-hot
shortstaffed
shoulder-less
shutted
shuttlings
sinked
skint
slaphead
slimmish
smackerooney
smartarses
spondoolies
spurtle
standed
steppy
stotties
stretchies
stripesey
stroppiness
stupidest
swimmed
swimmies
swolled
tatties
tike
toerag
tooked
toy-boy
tuitioning
typecasted
un-organised
velveting
verbals
vibes
wafted
waggling
waistcoaty
wasabe
watermanship
weppings
wiltering
wimpy
wowsers
yak-chucker
yamming
yattering
yourn
yousen[11]
yukkified
zombieness

For research looking at phrases or even just pairs of words, very large scale is essential. For example, in a study of the quotative "*It's like*" construction (e.g. "it's like 'mmmmmm' "), Fox and Robles (2010) lament that "*It's like*-enactments are not frequent in our American English data. In over 10 hours of interaction recorded since the mid-1980s we have found just 22 examples." Presumably, a year of speech - with almost 1000 times more data - would yield many more examples.

[9] A well-document Northern English form of 'herself'.

[10] "Re-snogged" occurs twice in Google, and one of those is a quotation of BNC http://www.cs.umd.edu/projects/metalanguage/BNC/KP6.html. The other instance is http://www.i-club.com/forums//showthread.php?p=1581713 and seems to have quite a different sense than its BNC usage.

[11] A well-document Northern English form of 'yourself'.

In a phonology project we have recently begun[12], we are investigating the pronunciation of Spoken BNC words ending in *m* or *ng* (IPA [ŋ]) to see whether they are affected by the following *t*, *p*, or *c* (IPA [k]) (compare Dilley and Pitt 2007). A few of the more frequent examples are given in (1).

(1)	Ending in -m	I'm trying	160 instances
		seem to	310
		alarm clock	18
	Ending in -ng	swimming pool	44
		getting paid	19
		wedding present	15

With such relatively rare combinations, it is clear that we shall need the 8-10m words of the Spoken BNC, and we would like more: if the corpus were 25 times smaller, like Switchboard, we might not expect to find any examples of some of these combinations. Even in the Spoken BNC, the use of such phrases is sometimes clumped into just a few conversations and spoken by just a few people. For example, 7 of the 44 tokens of “swimming pool” are found in recordings made by one family on a single occasion: a trip to a swimming pool! So, some research questions, especially those involving individual differences and the social and psychological factors that affect people’s use of language, may require much more than a year of speech.

Even for studies of single words, lopsided sparsity can be a problem. In our *Word Joins* project, we are also studying the incidence of “final -t/-d deletion” and the social and contextual factors that govern it in UK English (Tagliamonte and Temple 2005), in words such as:

(2)	just	19563 instances
	want	5221
	left	432
	slammed	6

Clearly, while some items provide abundant evidence, many rarer words will not be included in the dataset. Even in such a very large data set, there is the ever-present problem of having far more instances of some items than are needed, statistically, and not enough instances of other items.

As a rule of thumb (Kochanski et al. 2011), to catch a statistically reasonable sample of:

- most of the sounds of English, you need minutes of audio;
- the common words of English, you need a few hours;
- a typical person's vocabulary, you need 100+ hours;
- pairs of common words, you need 1000+ hours;
- arbitrary pairs of words, you need 100+ years.

So, large corpora open up new kinds of studies, using “speech in the wild”, as opposed to “speech in the lab”. This difference is expected to be very important (a) where we need informal speech, or (b) where we are looking at the word or sound choices people make. Of course, laboratory experiments can involve rare words or rare combinations of words. But in that case, it would be circular to study word *choices*, because they are not free choices any more: they are induced by the experiment.

[12] *Word joins in real-life speech: a large corpus-based study.* UK ESRC grant RES-062-23-2566 to Coleman, Kochanski, Temple and Yuan.

In experimental studies, the size of the sample must be adequate to make inferences about a larger population. But this is not a problem limited to experimental studies: even in non-experimental, more humanistic studies, the *rarity* of interesting forms is a key issue. If a word is so rare that its use use is not recorded, even the most qualitative study will treat it differently from one that has been observed. An example of this is illustrated by the selection of “new and interesting” words presented above. This is just a small, hand-picked sample from the >12,000 that occur only once in the 4.2m-word demographic part of the Spoken BNC, selected on the basis that they might merit inclusion in a new dictionary. The words are mostly new coinages, or obvious dialect words, some of which (e.g. “gronnies”) are so rare that they have not previously been recorded, it seems.

Furthermore, though experimental studies may artificially elicit items or phrases of interest that might be rare in natural speech, laboratory experiments are limited by being designed and controlled. Typically, they search for expected outcomes or plausible alternatives, and these are this limited by the imagination of the researcher. Factors that limit the imagination include: binary hypothesis tests; keeping experiments small (to be more affordable), and planning an experiment’s outcome and impact before starting. “Speech in the wild” can (and does) turn up the unexpected. Laboratory experiments on speech also suffer from a particular kind of observer effect: the contrast between how people naturally use language and what they believe to be “correct”. If a lab experiment, we can expect that people will consciously (or unconsciously) modify their behaviour because they would know the experimenter cares about their language performance. (In)formality can be important. So, large scale corpus research enables certain experiments to be carried out where the “laboratory observer effect” would otherwise be too large.

The only way of facing up to these problems is to collect corpora that are large enough to contain a sufficient number of examples of the phenomena of interest. The reliability of the results from an experiment aiming to identify generalisable aspects of linguistic behaviour is closely linked to the scale of the experiment. If you observe the same (range of) behaviour in 500 people, that’s much more persuasive than the range you might get from 50.

But spoken corpora are not just repositories of words: they also embody specific phrases or constructions, as well as revealing - through their audio aspect - particularities of people’s voices and habits of speaking (we present some case studies in section 5, below). For example, we have noticed various instances of people speaking with animals (pets) in the Spoken BNC demographic recordings. Without undertaking a detailed study of such behaviour, it is obvious to any listener that the speakers in question employ speech habits that are specific to “talking to animals”, and apparently different for different pet species. For example, dog-directed speech is similar in many respects to child-directed speech. But parrot-directed speech (there are at least two such “conversations” in the Spoken BNC) is something different, one speaker employing a quiet falsetto voice in addressing his parrot. In what circumstances do people adopt unusual voices, and how is the ‘voice’ in question selected for the context? Do men do it more than women? Young more than old? And how does the brain of the speaker (and human listeners) produce, interpret or store “odd voice” pronunciations, and strange duration/intonation patterns? These are questions that lie almost beyond the limits of previous linguistic research (but cf. Smith et al. 2002).

In linguistic studies and as a resource for language teaching, one is typically only interested in specific instances of a word, a phrase, or a phonetic feature, insofar as they are examples of a more general category. In other fields, however, such as oral history, specific reports of events by individual narrators are the object of study. Here, size carries value: more is better. But for all these varied purposes, as well as for the “casually curious” enquirer, the main problem in navigating a large corpus is the problem of “finding a needle in a haystack”.

To address that challenge, we think there are two “killer app” technologies:

- Forced alignment
- Data linking (or at least exposure of digital material over the internet, coupled with cross-searching)

In Mining a Year of Speech, we have devoted most effort to the first of these, as it is a prerequisite to the second. In parallel projects that we shall pursue beyond the time-limited period of the Digging into Data challenge, we continue to promote data linking and work to establish common standards for open access to annotated audio corpora, including (of course) our own corpora. We discuss this further in section 4 below.

2. Mining a Year of Speech: What we did

Today's very large corpora are not collected overnight: they grow out of the corpora of former years, which were considered large in their day, and are themselves put together from earlier, smaller corpora. For example, the Spoken BNC includes COLT: the Bergen Corpus of London Teenage Language (Stenström et al. 2002). For "Mining a Year of Speech" we aggregated a number of corpora in our care, using common standards for the word- and phoneme-level mark-up. The corpora we are using were chosen on the basis that (a) each of them, alone, is already quite large; (b) we already have written transcriptions of the audio (though this may be inaccurate to some degree); (c) they may be published quite generally, at no charge. The corpora in the collection are listed in Table 2.

Table 2. Descriptions of the corpora in the Year of Speech, and their June 2011 alignment status

	Size	**Description**	**Status (June 2011)**
Spontaneous speech			
Spoken British National Corpus	1478 hrs	Broad sample of UK English; many dialects. Context-governed part includes meetings, radio phone-ins, oral history recordings. Demographic part: everyday conversations	About 2/3 aligned, remainder will soon be complete. Some of the demographic part is very hard to align accurately, due to poor audio quality.
Conversational and telephone speech			
Switchboard I	258 hours	~2,400 two-sided telephone conversations among 543 speakers (302 male, 241 female) from all areas of the United States	Aligned
Fisher Part 1	~1950 hrs	5,850 telephone conversations, each lasting up to 10 minutes	
Mixer-5	346 hrs	Three 1.5-hour interviews and up to 10 10-minute phone conversations per subject	Aligned
Mixer-6	329 hrs	Conversational speech from native speakers of American English from the Philadelphia area	

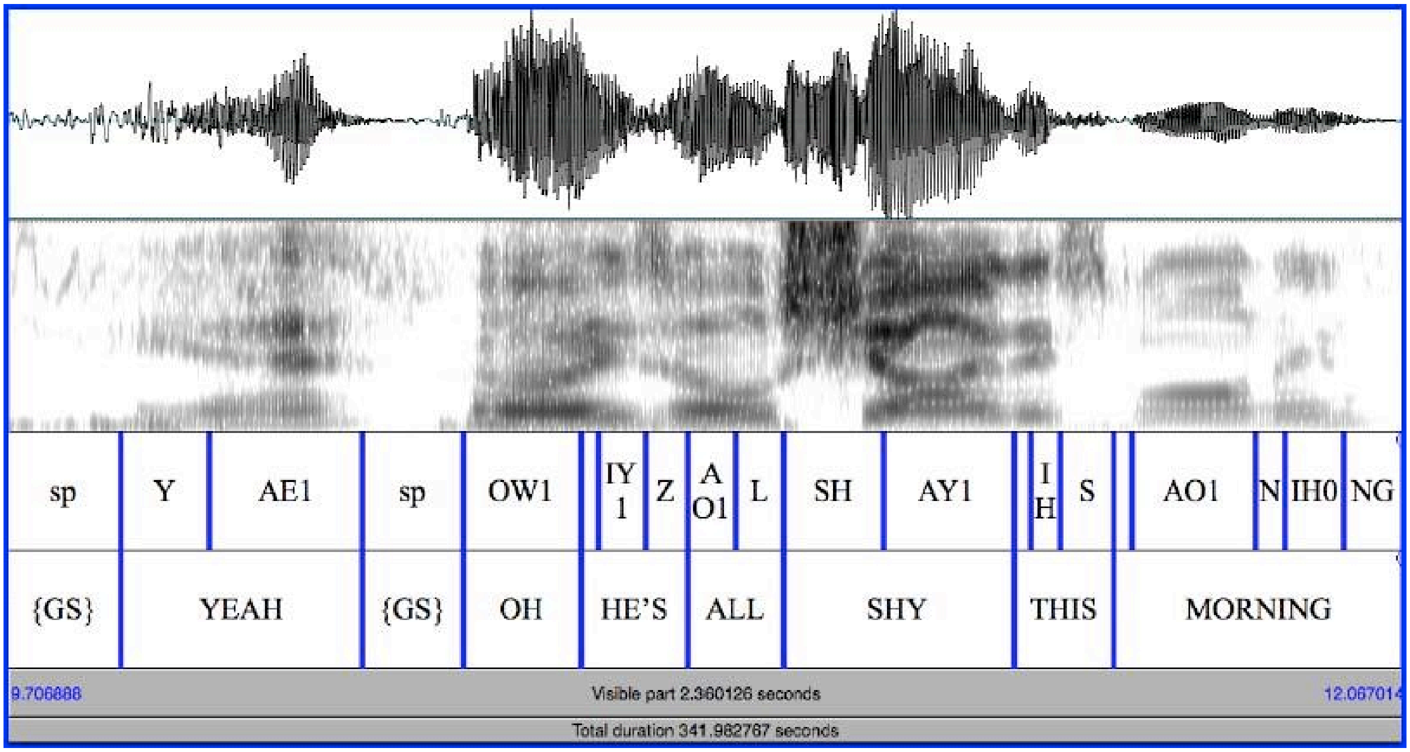
Read text			
LibriVox	>17,520 hrs	Audio books	Collected, partly aligned
Broadcast news			
TDT2	~633 hrs	CNN Headline News, ABC World News Tonight, PRI The World, VOA English News Service	Aligned
TDT3	~428 hrs	Same as TDT2 plus NBC Nightly News, MSNBC News with Brian Williams	Aligned
TDT4	~312 hrs	Same 6 sources as TDT3	Aligned
Hub4	~100 hrs	104 hours of broadcasts from ABC, CNN and CSPAN television networks and NPR and PRI radio networks	Aligned
Oyez.org	~9000 hrs	US Supreme Court oral arguments	About 6,000 hours digitized and partly aligned
Political Discourse	100 hours		Not yet aligned
Oral history interviews	100 hours		Not yet aligned
Sociolinguistic interviews	50-100 hours	From DASS (Digital Archive of Southern Speech) and Labov corpus	Transcriptions in progress
Brent	~240 hours	Part of CHILDES corpus	Aligned
Total so far	~5243 hours		

In total, our collections constitute c. 3.6 years of aligned speech, of which, at time of writing, 218 days has been aligned.

2.2 Forced alignment

For all of the corpora, UK and US, we are aligning the text transcriptions to the audio, using P2FA, the Penn Phonetics Lab Forced Aligner (Yuan and Liberman 2008). The alignment procedure yields a best-fitting phonemic transcription of the audio, together with detailed timing information: the start and end time of every vowel, consonant, and word. This data is encoded as Praat TextGrids, a data structure which can be rendered visually as in Figure 2. Such alignment data would enable the use of other software to search for specific vowels, consonants, words, or their combinations, in order to locate, navigate to or extract a copy of particular sections of the audio.

Figure 2. The output of a forced alignment of the word- and phoneme-level transcriptions from the Spoken BNC with the relevant portion of audio. It uses the PRAAT speech processing software (freeware) (http://www.fon.hum.uva.nl/praat/). This figure shows a 2.4-second extract; full alignment of a year of speech is 21 million times longer than this snippet.



We use the same set of (US English) acoustic models for all vowels and consonants of almost all varieties, even UK English. This meant that some acoustic models were used for quite different phonemes in UK and US varieties; for example, in view of their acoustic similarity, the same model is used for [ɑ] in US 'Bob' [bɑb] as in Southern British 'barb' [bɑɑb] (in which the 'r' is pronounced simply as prolongation of the vowel). Pronunciation differences between different varieties of English were dealt with by listing multiple variant phonetic transcriptions in the aligner's dictionary. Since American English lacks the UK English phoneme /OH1/ [ɒ], most spoken instances of UK [ɒ] are aligned with their US English transcription /AA1/. In ongoing work, we plan to build acoustic models trained to specific varieties, especially for the UK English data.)

The standard dictionary used by the Penn Phonetics Lab Forced Aligner is CMUdict[13], the Carnegie Mellon University Pronouncing Dictionary, which was developed for North American English. Since the Spoken BNC contains 13223 words (including written word-like items such as "$700", "yaaa", or "PVC") that are neither listed in CMUdict not the electronic version of the (UK English) Oxford Advanced Learner's Dictionary (Mitton 1986), extension of the aligner's dictionary to deal with UK English was an anticipated problem, albeit on a scale that we had not foreseen (see section 3.5 below for discussion). Instead of manually writing additional phonetic transcriptions for the >13,000 "out of vocabulary" items, we used a combination of resources to guess a range of candidate transcriptions. The methods we used were: (a) find nearest-neighbor spellings in CMUdict and BEEP, a British English pronunciation dictionary[14]; (b) compute grapheme-to-phoneme conversion using Sequitur G2P software (Bisani and Ney 2008); (c) spell out "nonstandard words" (e.g. currency expressions like $700) as full written phrases (e.g. "seven hundred dollars"), which were then converted to phonemic transcription using the Festival text-to-speech system[15]. The candidate transcriptions were then all checked by two expert phonologists (J. Coleman and Ranjan Sen). Although such checking is somewhat time-consuming, it was found to be very much faster than manually writing transcriptions from scratch. The resulting additions to the dictionary all conformed to US

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

[14] ftp://svr-ftp.eng.cam.ac.uk/pub/comp.speech/dictionaries/beep.tar.gz

[15] http://www.cstr.ed.ac.uk/projects/festival/

phonological conventions, in the notation of CMUdict.

Then, to create transcriptions for the many regional varieties of UK English in the Spoken BNC, we semi-automatically converted CMUdict-style pronunciations for four (2×2) main dialect groups, according to two characteristics: ‘rhotic’ vs. ‘nonrhotic’ and ‘Northern’ vs. ‘Southern’. ‘Rhotic’ denotes dialects such as most US English varieties, Southwest UK, Northwest (Lancashire, Cumbria), Northumberland and East Anglia, in which ‘r’ is pronounced after certain vowels; non-rhotic dialects are those such as RP/Standard Southern British English, New England and Australian/New Zealand dialects in which ‘r’ is not pronounced after vowels, so that e.g. ‘park’ is [pɑɑk] or [paak], not [pɑrk]. ‘Northern’ UK here refers to varieties in which ‘but’ and ‘put’ have the same [ʊ] vowel as in ‘book’ or ‘pull’ (encoded UH in Table 3, ‘roadrunner’), and the same short, front [a] vowel in ‘bath’, ‘pass’ as in ‘bat’, ‘pat’, as opposed to the long, back [ɑɑ] of Southern UK English [pɑɑs] etc. While these dialect groupings would be far too broad and overly homogeneous for a serious dialectologist, we do not need to represent the pronunciation exactly in order to get a usable alignment. We let the aligner determine which pronunciation best fits the observed acoustics, on a word-by-word basis. Thus, in intermediate dialect regions, the aligner might end up picking (e.g.) rhotic pronunciations for some words and non-rhotic for others; we aim simply to label the data in the manner that desribes the major variations in the acoustics.

Table 3. Examples of some UK dialect-area differences, from a portion of the dictionary. Red text highlights pronunciation variations. (*Rioch* is transcribed R IY1 AA2 K for US English.)

	S Rhotic UK	**S Nonrhotic UK**	**N Rhotic UK**	**N Nonrhotic UK**
rioch	R IY1 OH2 K	R IY1 OH2 K	R IY1 OH2 K	R IY1 OH2 K
risecote	R AY1 Z K OW2 T	R AY1 Z K OW2 T	R AY1 Z K OW2 T	R AY1 Z K OW2 T
ritto	R IH1 T OW0	R IH1 T OW0	R IH1 T OW0	R IH1 T OW0
ritu	R IH1 T UW0	R IH1 T UW0	R IH1 T UW0	R IH1 T UW0
ritzi	R IH1 T S IY0	R IH1 T S IY0	R IH1 T S IY0	R IH1 T S IY0
rivermead	R IH1 V ER0 M IY2 D	R IH1 V AH0 M IY2 D	R IH1 V ER0 M IY2 D	R IH1 V AH0 M IY2 D
rivetus	R IH1 V AH0 T AH0 S	R IH1 V AH0 T AH0 S	R IH1 V AH0 T AH0 S	R IH1 V AH0 T AH0 S
roadrunner	R OW1 D R AH2 N ER0	R OW1 D R AH2 N AH0	R OW1 D R UH2 N ER0	R OW1 D R UH2 N AH0

2.3 Tools/user interfaces

We have developed and distributed tools for transcribing, for checking transcriptions, and for aligning transcripts with audio recordings[16]. We are developing and will distribute tools for interactive search and display of audio/transcript combinations, similar to those used by oyez.org, whitehousetapes.net (University of Virginia) and the Scottish Corpus of Texts and Speech[17]. We will develop and distribute tools for indexing and searching phonetic transcriptions and acoustic parameters,

[16] Transcriber, XTrans, Penn Phonetics Lab Forced Aligner

[17] E.g. http://www.scottishcorpus.ac.uk/corpus/search/document.php?documentid=800

with appropriate interfaces to other tools such as statistical analysis and visualization.

2.4. Dissemination-related activities

We omit presentations internal to UPenn or Oxford from the following list.

January 2010. J. Coleman interviewed about the project on BBC World Service *Digital Planet:* link

19 March 2010. Launch meeting in Oxford - outreach to academics from other UK speech corpus projects, language technology industries, and colleagues from the library and archive sector.

19 September 2010. J. Coleman gave an invited lecture, "Large-scale computational research in Arts and Humanities, using mostly unwritten (audio/visual) media", at the Universities UK-sponsored 'Future of Research' conference.

29-30 January 2011. Workshop on “New Tools and Methods for Very-Large-Scale Phonetics Research” organised by J. Yuan and M. Liberman at UPenn, preceded by a 1-day tutorial on forced alignment given by J. Yuan. J. Coleman presented our co-authored “Mining a Year of Speech” paper (Coleman et al. 2011), and G. Kochanski presented “Detecting gross alignment errors in the Spoken British National Corpus” (Baghai-Ravary et al. 2011).

31 May 2011. J. Coleman discussant, “The Federation of Corpora Galore: Problems and Prospects”, at Corpora Galore: Applications of Digital Corpora in Higher Education Contexts, a workshop for HE staff and postgraduate tutors, Newcastle University.

9-10 June 2011. ‘Digging into Data’ conference, Washington DC. Presentation by J. Coleman and M. Liberman.

3. Issues we grappled with

3.1. Funding logistics

The US side of the funding did not come through until nearly nine months into the project. As a result, our plan to hire a postdoc to work on the indexing and search methods had to be postponed and this part of our work will not be completed until 2012 (using a combination of remaining funds from this grant and other sources of funding). Instead, during the first period of the US side of the project, we focused on using less-expert labor to deal with the BNC alignment issues discussed below, and to assemble a larger and more diverse sample of spoken materials (nearly 4 times the originally projected amount).

3.2 Quality of transcriptions

It is well-known by Spoken BNC users that there are errors in its transcriptions: this is obvious from Adam Kilgarriff’s BNC word-lists[18], and has been studied by Mitton et al. (2007). We had anticipated that there would be typos, which are easily dealt with by listing them in the pronouncing dictionary alongside their correctly-spelled alternatives. We had also known in advance that hard-to-hear sections of the spoken audio was not transcribed by the audio typists who originally prepared the Spoken BNC. These sections are tagged as <unc> (for “unclear”), so we had expected to be able to model them as noise/burble intervals in between recognisable speech.

[18] http://www.kilgarriff.co.uk/bnc-readme.htm

What we had not adequately appreciated at the outset of this project was that (a) there were long untranscribed portions in between transcribed <div>'s (= conversational units), and (b) we didn't know that there were large transcribed regions where the audio had been lost. The former is due to the fact that the audio typists mainly attended to coherent sections of conversation. In between these, short isolated utterances of speakers made in between conversations, as they went about their daily lives, were simply ignored. The latter problem was only revealed when large-scale alignment runs were under way. The fact that we have transcriptions for portions of conversations that are not on the audio tapes, and which lie in between the end of one side of the tape and the start of the other side of the tape, strongly suggests that the Spoken BNC audiotapes deposited in the British Library Sound Archive are neither the originals nor faithful copies of the originals; they appear to be slightly incautious copies of the - now presumed lost - originals, in which the original recording was left running while the deposited tapes were turned over. This illustrates a common difficulty when attempting to rejuvenate old data: the documentation and records of how the corpus was made are never as complete as you would like, because those who originally produced the data didn't imagine quite what we might want to do in later years, nor what technologies might become available to us.

Some of the other data sources have (minor) versions of the issues discussed above. Thus broadcast recordings may have commercial or other interpolations that are not transcribed; transcripts generally edit out many disfluencies; political speeches may include extemporized pieces that are different from the "as prepared for delivery" version that may be the only textual version available. We have developed several different ways of dealing with these issues. Our automatic alignment techniques can deal fairly well with untranscribed disfluencies, and reasonably well with commercial and similar interruptions. We are working on improving the state of the art with respect to "knowing what we don't know", that is, accurately identifying (small or large) regions where the transcript does not correspond to the recorded audio.

3.3 Quality of alignments

There are a number of fairly obvious difficulties in forced alignment of spontaneous speech in natural settings:

- overlapping speakers
- background noise, music or babble
- variable signal loudness
- reverberation
- distortion
- poor speaker vocal health or voice quality

Even though we must simply accept that the most "difficult" segments cannot yet be aligned well, most of the time forced alignment is good enough to be quite useful. In an earlier pilot project, in which we tested the P2FA alignment software on a sample of BNC audio files, we found that 83% of the phoneme boundaries were within 2s of their correct position - accurate enough to take a user to the right portion of the recording. 24% of the phoneme boundaries were within 20 ms of expert human labels. This is not a large fraction, but when one boundary is accurate, almost all the boundaries within the neighboring 2 seconds are also accurate. Practically speaking, it is not sufficient to know that one quarter of the data is very accurately aligned; a user wants to know whether a search for particular words or phonemes yields results that are well-aligned or poorly aligned. As part of our work towards developing a running confidence measure of alignment accuracy across large corpora, Baghai-Ravary, Grau and Kochanski (2011) presented methods for evaluating the accuracy of alignment.

3.4 Intellectual property responsibilities

We had ascertained before applying for funding that as participants in the BNC Consortium - as its curators, in fact - the University of Oxford and the British Library Sound Archive are permitted to publish the audio data provided that participants' anonymity is respected, by muting those portions of audio corresponding to the <gap> tags in the TEI-XML transcriptions. However, as there are over 18,000 such <gap> tags to check - we had not anticipated there being so many! - it has not been possible to manually check them. The process is made slower by limitations on alignment accuracy; as this is improved, listening to segments in need of anonymization will become easier and faster. Because of this difficulty alone, we have had to revise our public release plans (see section 4.1 below).

3.5 Problems of “scaling up”

The tasks of (a) adding extra dictionary entries for out-of-vocabulary words (described in section 2.2 above), and (b) checking the alignment <gap> tags, illustrate a more general problem of scale that afflicts many large-scale data-processing projects: a small difficulty, which by itself or for a few instances would be quick and easy to fix by hand, blows up into a major time-sink when it has to be repeated tens of thousands of times, or more. For example, collecting and sifting through the candidate pronunciations for out-of-vocabulary items was several person-months of unanticipated extra work. By contrast, automatically generating four variant UK dialect pronunciations for the entire dictionary only took a couple of days’ work. As a more general observation, any large-scale data activity with an error rate of, say, 1%, amounts to 10,000 errors per million items (e.g. 100,000 word errors in a 10-million word corpus). In such circumstances, manual error correction or checking quickly becomes impractical, and perfection may be impossible to attain.

4. Some Results and Prospects

4.1 Publication/release plans

a) LDC data

Nearly all of the relevant LDC data has already been published, or is queued up for publication. The exceptions are the small political-speech and oral-history samples collected for this project, which will be published, pending IPR arrangements. The word-level and segment-level alignments for all of the material will also be published.

About 6,000 hours of the SCOTUS corpus is already available to the public for interactive search and display on the oyez.org web site. As we develop web-based interfaces for phonetic search and retrieval, we’ll make this capability available on an experimental basis for different corpora and combinations of corpora.

b) BNC Spoken Audio

Until we are able to check or be confident of the alignment of the anonymization tags, for muting the corresponding audio, we will are not able to publicly release those recordings that require anonymization. This is a task that could not be completed by the end of the project, and on which we continue to work in follow-on projects. In the mean time, therefore, we have created a BNC Spoken Audio Sampler which we have published on the web[19] and also on DVD’s that we shall distribute gratis. We still intend in the near future to make a full release as linked data via the British Library Archival Sound Recordings server, through other, aligned projects.

4.2. Prospects

To enable other corpora to be added to the collection in future, we have been studying and discussing with other corpus curators common standards for spoken audio with linked transcriptions or other forms of annotation. As part of their work on Mining a Year of Speech, Sergio Grau and Lou Burnard developed three alternative proposals for adding audio timing information to linguistic transcriptions in TEI. We are currently experimenting with these to see which one of them meets our needs best without breaking older software that works with the TEI-XML BNC, such as Xaira. We are also studying

[19] http://www.phon.ox.ac.uk/SpokenBNC

similar markup proposals from other tool developers. In due course, we intend to submit a settled proposal to the TEI Consortium for adoption as a standard. In parallel, we are also keeping a watch on emerging linked data standards for audio segments.

All of the corpora in Mining a Year of Speech were chosen because they already had some kind of text transcription of the audio. One avenue for adding transcriptions (in ordinary spelling) to audio corpora that currently lack them is to accumulate them by crowdsourcing. For example, dotsub.com provides a Flash app that enables a transcriber to add a time-aligned subtitles to a submitted web video. You can use their time marks even if you change the text. LDC proposes to make an open-source version, for certain communities to do their own commentary/transcription of e.g. language teaching classes Oyez.org includes a similar provision for e.g. instructors to add marginal annotations. Some untranscribed audio material is clean enough that it is suitable for automatic transcription using automatic speech recognition, a prospect which is advanced by having large amounts of previously-aligned data on which to train new acoustic models.

5 New insights?

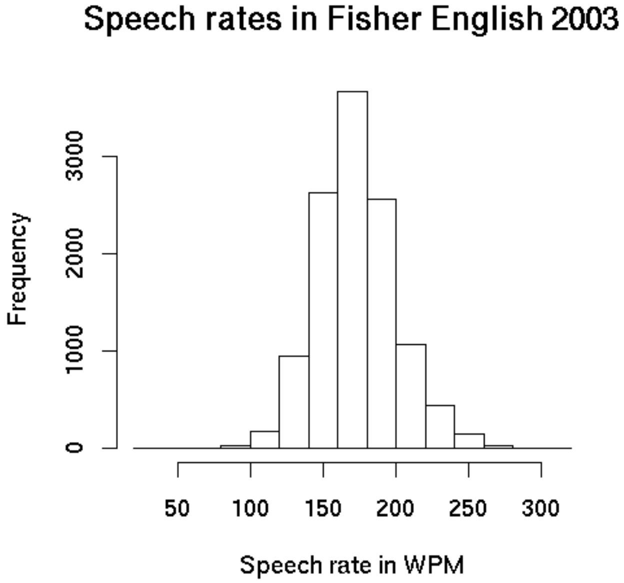
Our primary aims in this project were to assess the challenges of working with very large digital audio collections of spoken language, and to explore methods to address those challenges. In ongoing research supported from other sources, we are also beginning to make use of these large-scale spoken resources to address some specific research questions. To conclude, we present an illustrative selection of pilot studies that illustrate the value and great potential of such work. The first three pilot studies below are “breakfast experiments” which took no more than about 1 hour from conception to completed blog post, and show how accessible data can profoundly change speech research. No new data collection is required for any of these, and “experimental design” boils down to just framing a query.

Case study 1. Sex differences in conversational speaking rates
Question: Do women talk faster than men, as has sometimes been asserted[20]?
Method: Words and speaking times in conversational turns, in the Fisher 2003 corpus (Cieri, Miller and Walker 2004). This is a collection of 5,850 ten-minute conversations, collected by LDC in 2003 and published in 2004-5[21]. 2,368 of these conversations were between two women; 1,572 between two men; 1,910 between one man and one woman.

The overall distribution of speaking rates across the 11,700 conversational sides had the expected bell-shaped distribution, with a mean of 173 words per minute, and a standard deviation of 27:

[20] e.g. Louann Brizendine, *The Female Brain*, 2006.

[21] LDC2004S13, LDC2004T19, LDC2005S13, LDC2005T19



For the male speakers, the mean value was 174.3; for the female speakers, the mean value was 172.6. The difference of 1.7 words per minute is statistically significant, due to the large sample size, but is clearly of no functional significance.

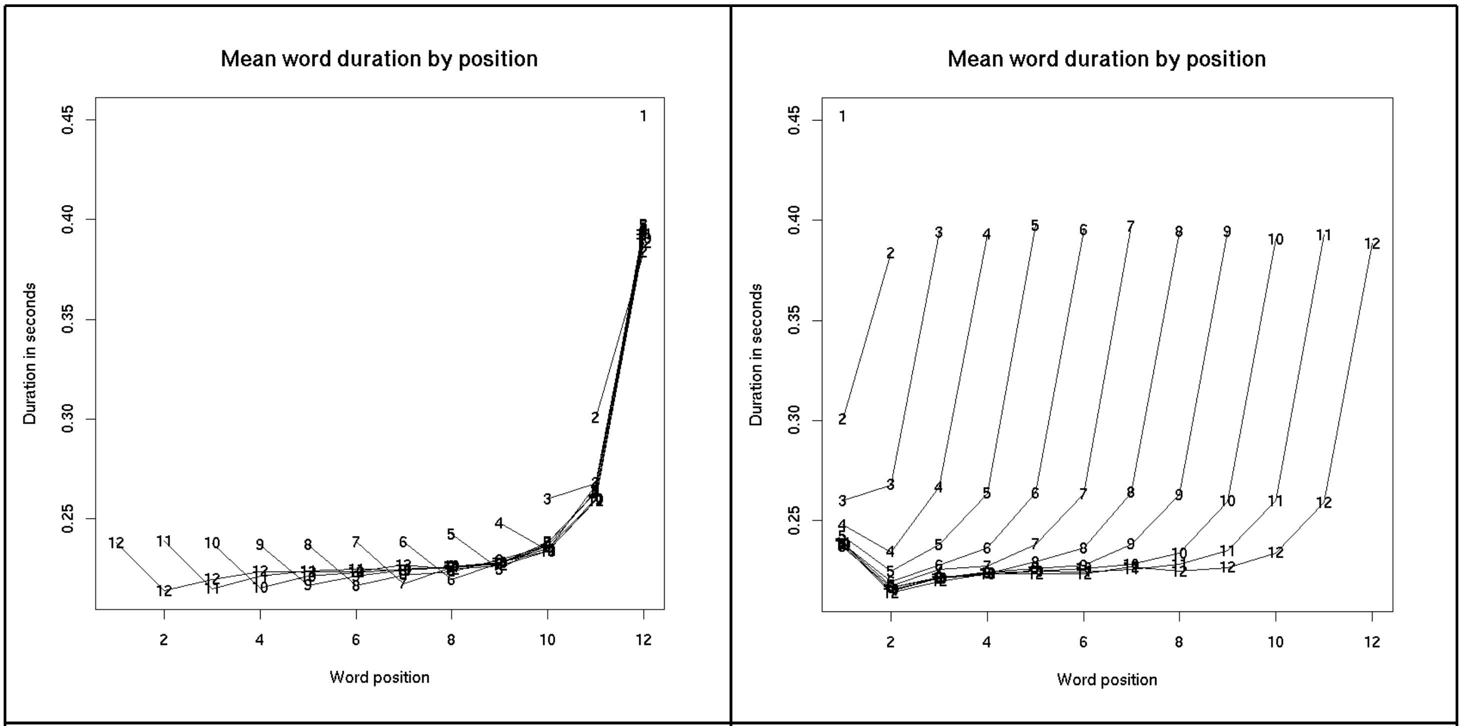
Case study 2. Phrasal modulation of speaking rate (Yuan, Liberman & Cieri, ICSLP 2006)
"Phrase final lengthening" (similar to rallentando in music) is a well-established effect, first observed by Abbé J.-P. Rousselot for French around 1870, and documented by many researchers in many languages since then. It is clear that unit-final syllables tend to be lengthened, but effects in other phrasal positions are less clear. The question of what should count as a "phrase" for this purpose is also an open one: is it a syntactic unit? a unit of information structure? a unit of speech production? Finally, the consistency of this effect across languages is unclear.

There are many other factors that influence timing in speech: the intrinsic phonetic duration of different speech sounds, the arrangement of sounds in syllables and words, the effect of emphasis, the effect of the process of deciding what to say, and so on. The usual method for studying this problem has been to have subjects in the laboratory read sentences that have been designed to hold these other factors constant or to vary them orthogonally. The result is material that is unnatural at best.

In a large collection of natural speech, we can expect that many of these factors will balance out; and we can use statistical modeling techniques to control after the fact for those that don't. So as a simple first step, we decided to define a phrase as a "pause group", i.e. any stretch of speech without internal silence >100 ms, and to look at word durations as a function of position in these "phrases" in the Switchboard[22] corpus. Switchboard comprises about 2,400 telephone conversations, recorded in 1990-91, published 1992-93, and republished with some corrections 1997. It was originally designed for research in speaker recognition.

The result is a strikingly regular pattern:

[22] LDC97S62



a) Phrases defined by silent pauses, all aligned to the end of the phrase

b) Same data, all phrases aligned by their first word

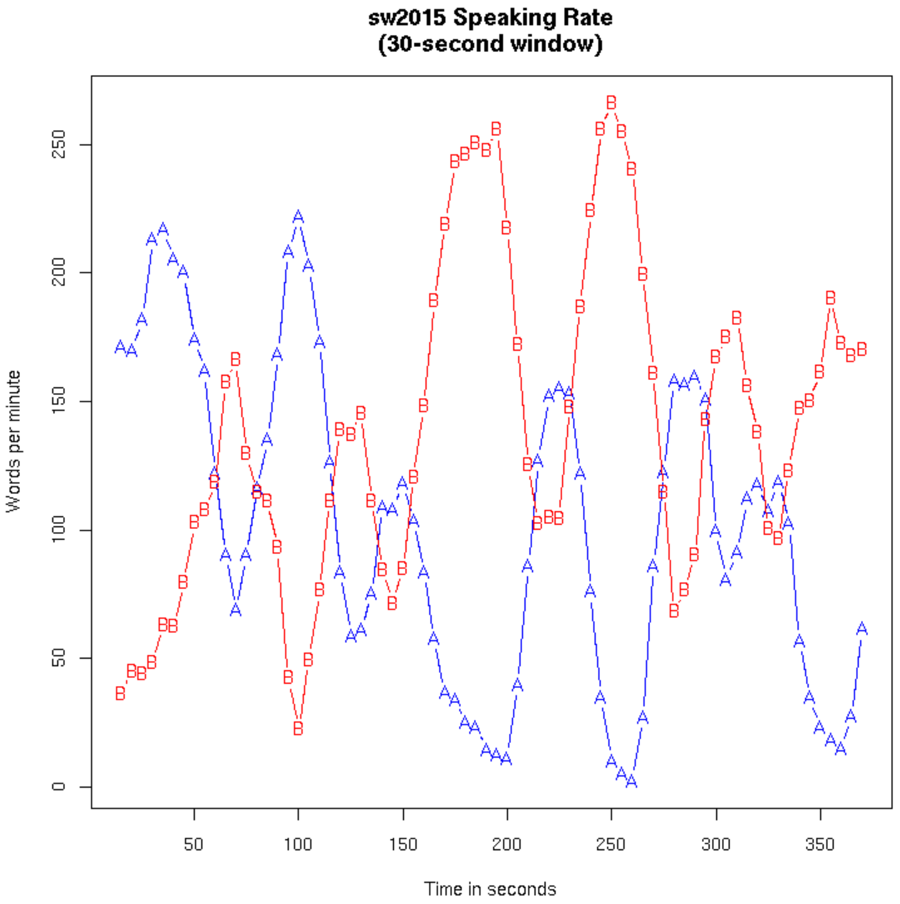
This is the beginning of the story, not the end: many variations in measurement and modeling follow. But once the data is collected, transcribed, aligned, and organized, most new questions are relatively easy to answer. And the success of this relatively superficial method means that it will not be difficult to make comparisons across genres, dialects, languages, and so on.

Case study 3. How does speaking rate reflect the ebb and flow of a conversation?

Speaking rate is not constant. Speakers may slow down to weigh their words, or to make things easier for their hearers; they may speed up to get past familiar or backgrounded material; and they join in the natural rises and falls of interpersonal arousal. We were interested in the distribution of local speaking rates in certain kinds of recordings, in order to provide background information for setting the criteria for certification of court reporters. But we also found that plots of local rates as time functions offer a suggestive picture of conversational rhythms.

In order to make these plots, we simply moved a fixed-length window -- 10 seconds or 20 seconds or 30 seconds -- smoothly through the timeline of a conversation, counting the number of words or syllables within the window at each step. This is easy to do with a word-aligned audio corpus -- we also note that there have been fairly good language-independent syllable detectors for many decades (e.g. Mermelstein 1975), so that a very similar plot could be created even if the language of the conversation is unknown.

The following plot shows a six-minute segment from a conversation in the Switchboard corpus. We can see the conversational initiative switching back and forth, on scales typically larger than a sentence: first speaker A takes the floor, while speaker B responds, and then they switch roles:



Case study 4. Previously unattested ("impossible") assimilations of word-final consonants

It has long been reported in works on English phonetics that the pronunciation of certain consonants ([t], [d], [n], [s] and [z]) is apt to be altered when the following word begins with a consonant at a different place of articulation. For example:

(4) that case ⇒ tha[k] case
bad case ⇒ ba[g] case
ran quickly ⇒ ra[ŋ] quickly
this shop ⇒ thi[ʃ] shop
hi[z] shop ⇒ hi[ʒ] shop

This generalization is usually formalized by phonologists as a statement about alveolar consonants: alveolar consonants may assimilate to following labial ([p], [b], [m], [f], [v]) and/or velar consonants ([k], [g], [ŋ]), but labials or velars do not themselves assimilate. Thus, one does not normally hear "u[t] town" (for "uptown"), nor "do[d] leash" (for "dog leash"). This pattern has been so often repeated in phonological descriptions and textbooks that it is hard to imagine it could be factually incorrect. Nevertheless, there are several scattered reports in the phonetics literature of such "prohibited" or "impossible" assimilations, in which a labial or a velar consonant is pronounced as an alveolar. For example, Barry (1985), a study that used electropalatography to observe regions of contact between the tongue and the palate, found "like that" articulated as ” li[t]e that“. Ogden (1999) reports that “I'm going” can be pronounced “I'[ŋ] going”, and we have heard “sometimes” pronounced like “suntimes” and “timetable” pronounced as if “tinetable” in informal conversational speech. In written English, there is abundant evidence from frequent typos: “sometimes” spelled as “sonetimes” (52,700 hits = 20 ppm in Google search data), “timetable” spelled as “tinetable” (over 32 million hits, or 23 ppm), “Washington” spelled as “Washinton” (2093 ppm), and Wellington” spelled as “Wellinton” (1583 ppm). While these typos could be because *n* and *m* are adjacent on the keyboard, or because the writer just accidentally missed a keystroke, they might be triggered by the corresponding spoken forms, if such pronunciations are normal variants, rather than random speech errors. Clearly, even if assimilated labials or velars are permitted, natural pronunciations variants, they are unlikely to be frequent. But, as we mentioned in the Introduction, the Spoken BNC provides sufficient data to begin to seek out such “impossible” or at least rare combinations of words. In fact, the first several instances of “swimming pool” that we were able to mine out of the Spoken BNC were all pronounced as “swimmi[m] pool”.[23]

Case study 5. Integration of language and other aspects of human behavior

Among people who are professionally interested in language, there has for a very long time been two somewhat different perspectives. On the one hand, there are those, such as Aristotle, Bertrand Russell or Noam Chomsky, who emphasise the formal/logical structure of language, and its cognitive instantiation, and who de-emphasise or are at least not very interested in how language is *used* in everyday life. And on the other hand, there are those who are more interested in, and give greater emphasis to, “real” language - that is, normal, spontaneous language, with all its errors, hesitations, re-starts, interruptions, and fragmentary phrase - and how its patterns and usage are shaped by the extra-linguistic context of its use. Studying language from this perspective is beset with difficulties, however, not the least of which is the great time and expense required to collect and analyse such recordings. Spoken corpora form a rich mine of data for studies of language as human behavior, and are replete with expressions that can hardly be understood (or even represented) in a decontextualised manner.

In the following examples from the Spoken BNC, linguistic expressions and non-linguistic events together constitute the action of the occasion. Examples (5) and (6) are from an elementary school “music and movement” class, in which a teacher is playing a tune on the piano while giving verbal directions to the children, who are dancing/acting:

(5) Hold it, two of those, two of those coming down the stairs [playing piano] wait for it no, any minute now

(6) Right, here’s the bear song [playing piano] It goes like this [playing piano] ready off we go
[playing piano] OK let’s go

Here, the annotator has noted the “background sound” of the piano playing at a number of chosen points in the transcription. In the audio recordings, the teacher is speaking and playing the piano simultaneously, and we can infer that the children are also dancing/moving simultaneously: the linear form of a written transcription does not do justice to this temporal complexity. The expression “coming down the stairs” refers to both the children’s action and a certain phrase in the music. “Wait for it” and “any minute now” are instructions to be still, but are also synchronised with some part of the music; “off we go” and “let’s go” are instructions for the children to begin a certain kind of movement, and are synchronised with another point in the music. “Here’s the bear song” and “it goes like this” are straightforward referential expressions.

In examples (7) to (9), various forms of the verb *to go* are used with non-verbal sound actions in direct object position. The sequencing of the non-verbal direct object is in accordance with the usual rules of syntax for more conventional, entirely verbal utterances. This form of *go*, and “quotative *like*” (as in “I’m like [speaker shrugs]), are quite frequent and familiar conversational events; though they defy the conventional clean categorisation of linguistic vs. nonlinguistic sounds and other actions, such constructions have been rather little studied (but see e.g. Postal 2002, Fox and Robles 2010), formal

[23] Have a listen to http://www.phon.ox.ac.uk/jcoleman/swimming_pool1_KBF_0156XX-AB.wav

theories of natural language syntax usually have no means for representing them.

(7) Listen they were all going [belch] that ain't a burp he said

(8) Like I'd be talking like this and suddenly it'll go [mimics microphone noises]

(9) He simply went [sound effect] through his nose

In other examples that have turned up in our working through the BNC Spoken Audio, we have impressionistically noted the various and different speaking styles and voices in which people address different species of animals (e.g. dogs vs. parrots), and inanimate objects (e.g. 10, uttered as the speaker removes the litter box from a parrot's cage, apparently).

(10) Come on then shitbox

In addressing pets, speakers employ unusual patterns of intonation, tempo, dynamics and voice quality (e.g. a man who uses falsetto voice in talking with his parrot). Another recording presents us with the dual challenge of (i) a man employing what can only presently be described as a "disgusted" voice quality (a certain combination of creak voice with long, drawn-out vowels), in response to the family dog being sick on the floor; while (ii) another member of the family colorfully imitates the sound of vomiting.[24]

Among the repertoire of voice qualities marked in the Spoken BNC is "laughing speech"; very little attention has yet been given to the acoustics of different kinds of laughter (e.g. Szameitat et al. 2009) or the altered acoustic characteristics of speech in talkers who are simultaneously smiling or laughing (Ford and Fox 2010). The fusion of linguistic with non-linguistic action urges us to rethink commonly-held assumptions about the separation of language, society and behaviour, and perhaps guides linguistics in the direction of a *rapprochement* with sociolinguistics, ethnomethodology, social psychology, etc.

The value of working with large spoken language corpora is certainly not restricted to linguists; in several other fields, the same excitement about the prospects opened up by such resources is evident. For example, social psychologists Tausczik and Pennebaker (2010) observe:

> We are in the midst of a technological revolution whereby, for the first time, researchers can link daily word use to a broad array of real-world behaviors. … Empirical results using LIWC [Linguistic Inquiry and Word Count - their text analysis software] demonstrate its ability to detect meaning in a wide variety of experimental settings, including to show attentional focus, emotionality, social relationships, thinking styles, and individual differences.

In another example of such work, Ireland et al. (2011) show that similarity in how people talk with one another on speed dates (measured by their usage of function words) predicts "increased likelihood of mutual romantic interest", "mutually desired future contact" and "relationship stability at a 3-month follow-up".

Legal scholars, similarly, are also interested in spoken discourses (judicial rather than romantic, of course!) For example, there is a growing body of work by legal scholars on modelling Supreme Court decisions, and in particular predicting justices' votes on the basis of the oral arguments. Black et al. (forthcoming) examines a corpus which they assembled of more than 8 million words spoken by the justices during oral arguments over the past 30 years. They argue that "when the justices focus more unpleasant language toward one attorney, the side he represents is more likely to lose. The same relationship holds between an individual justice's questioning patterns and her final vote on the merits. … [they] extend to

[24] Audio clips http://www.phon.ox.ac.uk/jcoleman/Dog_sick1.wav and http://www.phon.ox.ac.uk/jcoleman/Dog_sick2.wav

Supreme Court justices the burgeoning focus on how the linguistic nature of language used by political actors – presidents (Sigelman and Whissell 2002a, 2002b) and members of Congress (Sigelman, Deering and Loomis 2000; Monroe et al. 2009) – affects decisions they make. As scholars continue to build models of the Court's decision-making process they must account for what transpires during these proceedings, including the justices' emotional state as they move toward decisions." Their work builds on, *inter alia*, Epstein, Landes and Posner (2009), who found that measures as simple as the number of words and utterances in justices' questions are predictive of judicial decisions.

Conclusion

The spoken material in the Year of Speech corpora includes some parts which are extremely challenging to automatically align with transcriptions, and other parts which are relatively easy to mark-up using forced alignment. For speech recordings that are relatively "clean" (scripted or captioned, in some cases, without background noise or overlapping speech), it is clear that forced alignment of far larger corpora is already possible, in principle. For example, there is now a 20-year archive of publicly open video footage from the C-SPAN cable TV channel covering US Senate/House proceedings, Canadian and UK Parliamentary proceedings, committees and hearings, as well as current affairs discussion shows. Since large parts of the proceedings are supported by officially published transcriptions, there is plenty of material for a future "Decades of Speech" project to get stuck into. As more and more audiovisual material becomes digitized, automatic alignment of transcriptions will be essential for users to navigate through these libraries of the future.

References

Baghai-Ravary, L., S. Grau and G. Kochanski (2011) Detecting gross alignment errors in the Spoken British National Corpus. In *VLSP 2011: New Tools and Methods for Very-Large-Scale Phonetics Researc*h. University of Pennsylvania. 103-106. http://ora.ouls.ox.ac.uk/objects/uuid%3Ab6438388-68bb-434e-9d73-7c2d32f04557/datastreams/ATTACHMENT02.pdf

Barry, Martin C. (1985) A palatographic study of connected speech processes. *Cambridge Papers in Phonetics and Experimental Linguistics* **4.**

Bisani, M. and H. Ney (2008) Joint-Sequence Models for Grapheme-to-Phoneme Conversion. *Speech Communication*, **50 (5)**, 434-451. http://dx.doi.org/10.1016/j.specom.2008.01.002

Black, Ryan C., Sarah A. Treul, Timothy R. Johnson, Jerry Goldman (forthcoming) Emotions, Oral Arguments, and Supreme Court Decision Making. To appear in *Journal of Politics.* Preprint: www.polisci.umn.edu/~tjohnson/black-treul-johnson-goldman-nd.pdf

Bradley, K. (2003) Critical Choices, Critical Decisions: Sound Archiving and Changing Technology. Paper presented at Researchers, Communities, Institutions, Sound Recordings, workshop held at University of Sydney, School of Society Culture and Performance, September 30 - October 1, 2003. http://conferences.arts.usyd.edu.au/viewpaper.php?id=57&cf=2) [Accessed 28/3/11]

Cassidy, Steve. DADA-HCS project (Distributed Access and Data Annotation for the Human Communication Sciences http://www.clt.mq.edu.au/research/projects/dada/

Cieri, C., D. Miller and K. Walker (2004) The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text. In Proceedings of LREC 2004. http://papers.ldc.upenn.edu/LREC2004/LREC2004_Fisher_Paper.pdf

Coleman, J., M. Liberman, G. Kochanski, L. Burnard and J. Yuan (2011) Mining a Year of Speech. In *VLSP 2011: New Tools and Methods for Very-Large-Scale Phonetics Research*. University of Pennsylvania. 16-19. http://www.phon.ox.ac.uk/jcoleman/MiningVLSP.pdf

Crowdy, Steve (1993) Spoken Corpus Design. *Literary and Linguistic Computing* **8** (4), 259-265.

Dilley, Laura C. and Mark A. Pitt (2007) A study of regressive place assimilation in spontaneous speech and its implications for spoken word recognition. *Journal of the Acoustical Society of America* **122 (4),** 2340-2354.

Epstein, Lee, William M. Landes, and Richard A. Posner (2009) Inferring the winning party in the Supreme Court from the pattern of questioning at oral argument. *John M. Olin Law & Economics Working Paper No. 466*, The Law School, The University of Chicago. http://www.law.uchicago.edu/files/files/466-wml-rap-inferring.pdf

Ford, C. and B. Fox (2010) Multiple Practices for Constructing Laughables. In D. Barth-Weingarten, E. Reber and M. Selting (eds) *Prosody in Interaction*. Amsterdam: John Benjamins. 339–368.

Fox, B. A. and J. Robles (2010) It's like mmm: Enactments with it's like. *Discourse Studies* **12,** 715-738. DOI: 10.1177/1461445610381862

Ireland. M. E., R. B. Slatcher, P. W. Eastwick, L. E. Scissors, E. J. Finkel, and J. W. Pennebaker (2011) Language style matching predicts relationship initiation and stability. *Psychological Science* 22 (1), 39-44. DOI: 10.1177/0956797610392928

Kochanski, G. P., C. Shih, R. Shosted (2011) Should corpora be big, rich, or dense? In *VLSP 2011: New Tools and Methods for Very-Large-Scale Phonetics Researc*h. University of Pennsylvania. 28-31. http://arxiv.org/abs/1012.2797

Mermelstein, P. (1975) “Automatic Segmentation of Speech into Syllabic Units”, *Journal of the Acoustical Society of America* 58 (4).

Mitton, Roger (1986) A partial dictionary of English in computer-usable form. *Literary and Linguistic Computing* **1 (4),** 214-5.

Mitton, Roger, David Hardcastle, and Jenny Pedler (2007) BNC! Handle with care! Spelling and tagging errors in the BNC. In *Proceedings of the Fourth Corpus Linguistics Conference,* 27-30 July 2007, Birmingham, U.K.. http://www.corpus.bham.ac.uk/corplingproceedings07/paper/142_Paper.pdf

Ogden, Richard (1999) A declarative account of strong and weak auxiliaries in English. *Phonology* **16,** 55-92.

Postal, Paul M. (2002) The openness of natural languages. Linguistics in the Big Apple (CUNY/NYU Working Papers in Linguistics). http://web.gc.cuny.edu/dept/lingu/liba/papers/Postal2002.pdf

Smith, Z.M., B. Delgutte and A. J. Oxenham (2002) Chimaeric sounds reveal dichotomies in auditory perception. *Nature* 416, 87-90.

Stenström, Anna-Brita, Gisle Andersen and Ingrid Kristine Hasund (2002) *Trends in Teenage Talk: Corpus compilation, analysis and findings.* John Benjamins.

Szameitat, D. P., Alter, K., Szameitat, A. J., Wildgruber, D., Sterr, A., and Darwin, C. J. (2009). "Acoustic profiles of distinct emotional expressions in laughter" *Journal of the Acoustical Society of America*, 126, 354-366

Tagliamonte, Sali and Rosalind Temple (2005) New perspectives on an ol' variable: (t,d) in British English. *Language Variation and Change* **17,** 281-302.

Tausczik, Y., & Pennebaker, J.W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29 (1), 24-54. doi: 10.1177/0261927X09351676

Wright, Richard and Adrian Williams (2001) *Archive Preservation and Exploitation Requirements*. PRESTO-W2-BBC-001218, PRESTO Preservation Technologies for European Broadcast Archives, 2001. http://presto.joanneum.ac.at/Public/D2.pdf [accessed 28/3/11].

Yuan, J. and M. Liberman (2008) Speaker identification on the SCOTUS corpus. *Proceedings of Acoustics '08*, 5687–5690. http://www.ling.upenn.edu/~jiahong/publications/c09.pdf
Software download page: http://www.ling.upenn.edu/phonetics/p2fa/