Mining a Year of Speech
Exploiter un an de discours

John Coleman

Phonetics Laboratory, University of Oxford

http://www.phon.ox.ac.uk/SpokenBNC

TGE Adonis, Orleans, 5 April 2012

L'équipe

Ladan Ravary, Ros Temple,
*Greg Kochanski, Sergio Grau, John Pybus
Oxford University Phonetics Laboratory

Lou Burnard

Jonathan Robinson et al.
The British Library

Mark Liberman, Jiahong Yuan, Chris Cieri
UPenn Phonetics Laboratory
and Linguistic Data Consortium

with support from

and with thanks for pump-priming support from the Oxford University John Fell Fund and the British Library

UNIVERSITY OF OXFORD

The “Digging into Data” challenge

- “The creation of vast quantities of Internet accessible digital data and the development of techniques for large-scale data analysis and visualization have led to remarkable new discoveries in genetics, astronomy and other fields ...
- With books, newspapers, journals, films, artworks, and sound recordings being digitized on a massive scale, it is possible to apply data analysis techniques to large collections of diverse cultural heritage resources as well as scientific data.”

UNIVERSITY OF OXFORD

In “Mining a Year of Speech”

we addressed the challenges of working with very large audio collections of spoken language

UNIVERSITY OF OXFORD

Challenges of very large audio collections of spoken language

- How does a researcher find audio segments of interest?
- How do audio corpus providers mark them up to facilitate searching and browsing?
- How to make very large scale audio collections accessible?

UNIVERSITY OF OXFORD

The “Year of Speech”

- A grove *(bosquet)* of corpora, held at various sites with a common indexing scheme and search tools
- US English material: 2,240 hrs of telephone conversations
- 1,255 hrs of broadcast news
- Talk show conversations (1000 hrs), Supreme Court oral arguments (5000 hrs), political speeches and debates
- British English: Spoken part of the British National Corpus, >7.4 million words of transcribed speech

UNIVERSITY OF OXFORD

Corpora in the Year of Speech

- Spontaneous speech
  - Spoken BNC ~1400 hrs
- Conversational telephone speech
- Read text: audio books, broadcast news
- US Supreme Court oral arguments
- Political discourse
- Oral history interviews
- US vernacular dialects/Sociolinguistic interviews

UNIVERSITY OF OXFORD

Abbott v. United States - Oral Argument

Case: Abbott v. United States

ORAL ARGUMENT OF DAVID L. HORA...

sections

Chief Justice John G. Roberts 00:13

We will hear argument next in Case 09-479, Abbott v. United States and the consolidated case, 7073, Gould v. United States.

Mr. Horan.

David L. Horan

Mr. Chief Justice, and may it please the Court:

The statutory interpretation question here is what laws trigger section 924(c)(1)(A)'s except clause.

Mr. Gould offers an interpretation that gives meaning and effect to every word and phrase of section 924(c)(1)(A) and follows this Court's recent holdings regarding the broad scope of the phrase "any other provision of law".

The Government, on the other hand, advocates a narrow construction that is not supported by the text and defends it primarily on the basis that section 924(c) supposedly should always

Not just for linguists

Black et al. (forthcoming):

> “when the justices focus more unpleasant language toward one attorney, the side he represents is more likely to lose.”

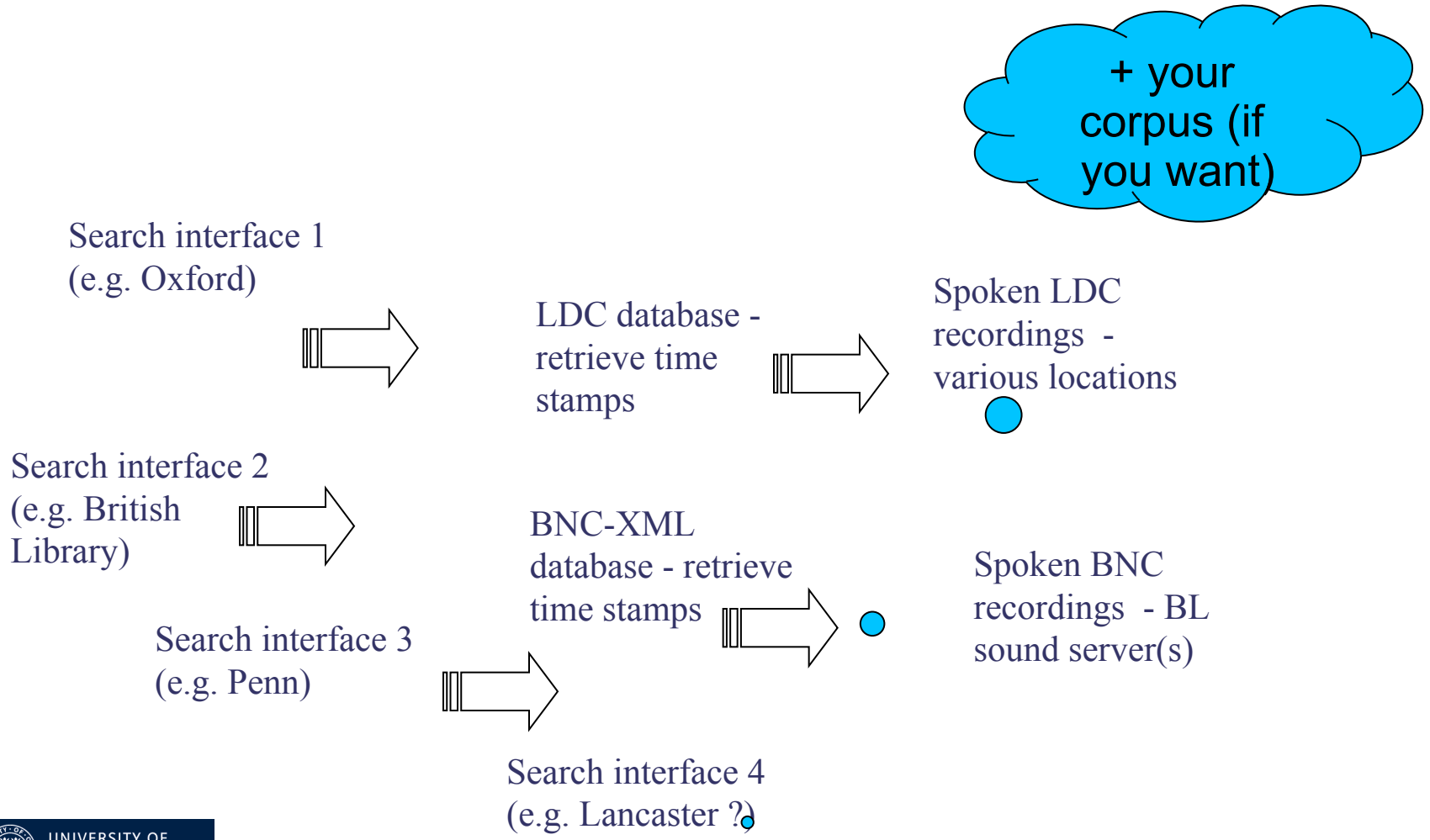
UNIVERSITY OF OXFORD

Not just for linguists

- Ireland et al. (2011):

Similarity in how people talk with one another in speed dating (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."

Cloud/crowd corpora: collaboration, not collection



UNIVERSITY OF OXFORD

Enabling other corpora to be added in future

- Negotiating with other speech corpus providers to join the federation
- Especially sociolinguistics collections
- Accumulating transcriptions (in ordinary spelling) by crowd-sourcing ??

UNIVERSITY OF OXFORD

II. What is the British National Corpus?

- a snapshot of British English at the end of the 20th century
- 100 million words in ~4000 different *text* samples of many kinds, spoken (10%) and written (90%)
- a synchronic (1990-4), sampled, general purpose language corpus
- freely available worldwide under licence since 1998; latest edition is BNC-XML
- various online portals

UNIVERSITY OF OXFORD

Production of the BNC, 1990-3

- Took >3 years and >£1.6 million
- Federation of
  - Dictionary publishers
  - Government (Department of Trade and Industry)
  - Science and Engineering Research Council
  - Linguistics research community
- Speech and Language Technology Club

UNIVERSITY OF OXFORD

Who produced the BNC and why?

- A consortium of dictionary publishers and academic researchers
  - OUP, Longman, Chambers
  - Oxford, Lancaster, British Library R&D
- government aim: to stimulate British industry
- expected users were lexicographers, NLP researchers, user-interface developers
  - (but not language teachers!)

UNIVERSITY OF OXFORD

Project Goals

- A *synchronic* (1990-4) corpus of spoken and written samples from the full range of British English language production
- of *non-opportunistic* design, for generic applicability
- with *word class annotation*
- and *contextual information*
- for better ELT dictionaries
  - authoritative
  - both speech and writing
- A **REALLY BIG** corpus

UNIVERSITY OF OXFORD

How big is “really big”? Some quite large *transcribed* speech corpora

Spoken BNC: 2–3 months of audio

- *PAC (Phonologie de l’Anglais Contemporain)*
- SwitchBoard corpus: 13 days of audio.
- Spoken Dutch Corpus: 1 month, but only a fraction is phonetically transcribed.
- Spoken Spanish: 4.6 days, orthographically transcribed.
- Buckeye Corpus (Ohio State Univ.): ~ 2 days.
- Wellington Corpus of Spoken New Zealand English, ~ 3 days transcribed
- Digital Archive of Southern Speech (American)

UNIVERSITY OF OXFORD

How big is “big science”?

Human genome:	3 GB
Hubble space telescope:	0.5 TB/year
Sloan digital sky survey:	16 TB
Large Hadron Collider:	15 PB/year

UNIVERSITY OF OXFORD

How big is “big ~~science~~”? humanities

Human genome:	3 GB
Hubble space telescope:	0.5 TB/year
Sloan digital sky survey:	16 TB
Large Hadron Collider:	15 PB/year

UNIVERSITY OF OXFORD

How big is "big ~~science~~"? humanities

Human genome: 3 GB

Hubble space telescope: 0.5 TB/year

Year of Speech (Coleman): >1 TB

Sloan digital sky survey: 16 TB

Large Hadron Collider: 15 PB/year

UNIVERSITY OF OXFORD

How big is "big ~~science~~"? humanities

Human genome: 3 GB

Hubble space telescope: 0.5 TB/year

Year of Speech (Coleman): >1 TB

Sloan digital sky survey: 16 TB

CLAROS web of art (Kurtz): >25 TB

Large Hadron Collider: 15 PB/year

UNIVERSITY OF OXFORD

How big is "big ~~science~~"? humanities

Human genome: 3 GB

Hubble space telescope: 0.5 TB/year

Year of Speech (Coleman): >1 TB

Sloan digital sky survey: 16 TB

CLAROS web of art (Kurtz): >25 TB

Large Hadron Collider: 15 PB/year

= 1500 years of speech

UNIVERSITY OF OXFORD

How big is "big ~~science~~"? humanities

Human genome: 3 GB

Hubble space telescope: 0.5 TB/year

Year of Speech (Coleman): >1 TB

Sloan digital sky survey: 16 TB

CLAROS web of art (Kurtz): >25 TB

Large Hadron Collider: 15 PB/year

= 1500 years of speech

European broadcast archives: 2,283 years of speech, mostly *not* digitized yet

Worldwide analogue audio archives: (12-100 PB)

UNIVERSITY OF OXFORD

Analogue audio in libraries

British Library: >1m disks and tapes, 5% digitized

Library of Congress Recorded Sound Reference Center: >2m items, including …

International Storytelling Foundation: >8000 hrs of audio and video

European broadcast archives: >20m hrs (2,283 years) *cf. Large Hadron Collider*

World wide: ~100m hours (11,415 yrs) analogue

75% on ¼” (6mm) tape, 20% shellac and vinyl, 7% digital

UNIVERSITY OF OXFORD

How to build a ‘representative’ corpus?

- Speech production: demographic sampling
- Speech variety: context governed sampling
- Recording and transcribing speech “in the wild” is
  - Socially difficult (even dangerous)
  - Expensive
  - Technically challenging

Spoken texts: demographic

- 124 volunteers: male and females of a wide range of ages and social groupings, living in 38 different locations across the UK
- conversations recorded by volunteers over 2-3 days
- permissions obtained after each conversation
- participants' age, sex, accent, occupation, relationship recorded if possible as descriptive criteria
- includes a large amount of London teenage talk, later published as COLT (Stenström et al.)

Spoken texts: demographic

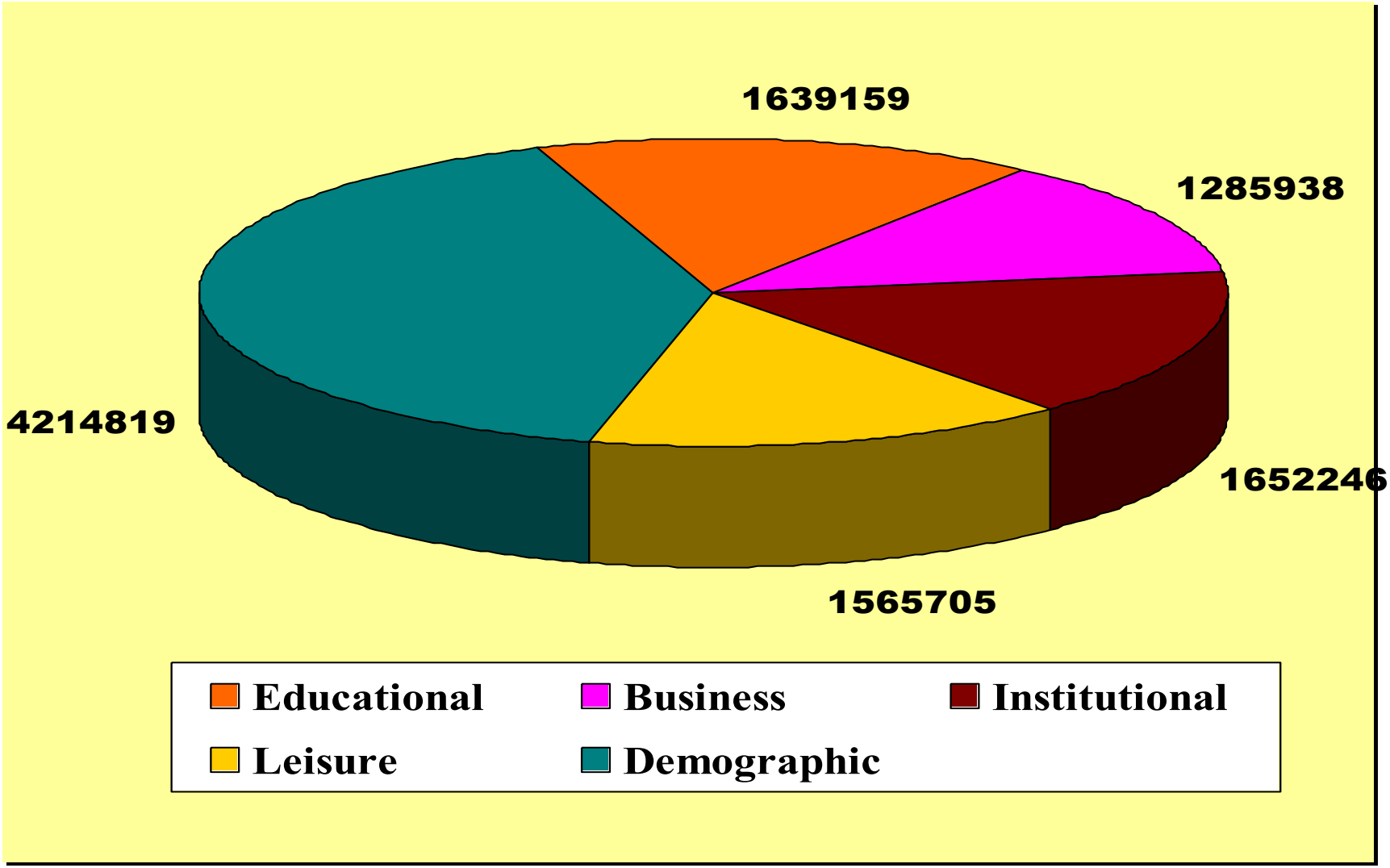
Antrim (Belfast)	Armagh	Berkshire	Birmingham	Bristol	Cambridgeshire	Cheshire
Clwyd	Devon	Dorset	Down	Durham	Dyfed	East Yorkshire
Essex	Greater London	Gwynedd	Hampshire	Hereford & Worcester	Hertfordshire	Kent
Lancashire	Leicestershire	Lincolnshire	London	Lothian (Edinburgh)	Manchester	Merseyside
Mid Glamorgan	Norfolk	North Yorkshire	Northamptonshire	Northumberland	Nottinghamshire	Orkney Islands
Pembrokeshire	Shropshire	South Yorkshire	Staffordshire	Strathclyde	Suffolk	Surrey
Tyne & Wear	Warwickshire	West Midlands	West Sussex	West Yorkshire	Wiltshire	

Spoken texts: context-governed

Four broad categories for social context, roughly equal quantities of speech in each:

- *Educational and informative* events, such as lectures, news broadcasts, classroom discussion, tutorials
- *Business* events such as sales demonstrations, trades union meetings, consultations, interviews
- *Institutional and public* events, such as religious sermons, political speeches, council meetings
- *Leisure* events, such as sports commentaries, after-dinner speeches, club meetings, radio phone-ins

Spoken domains

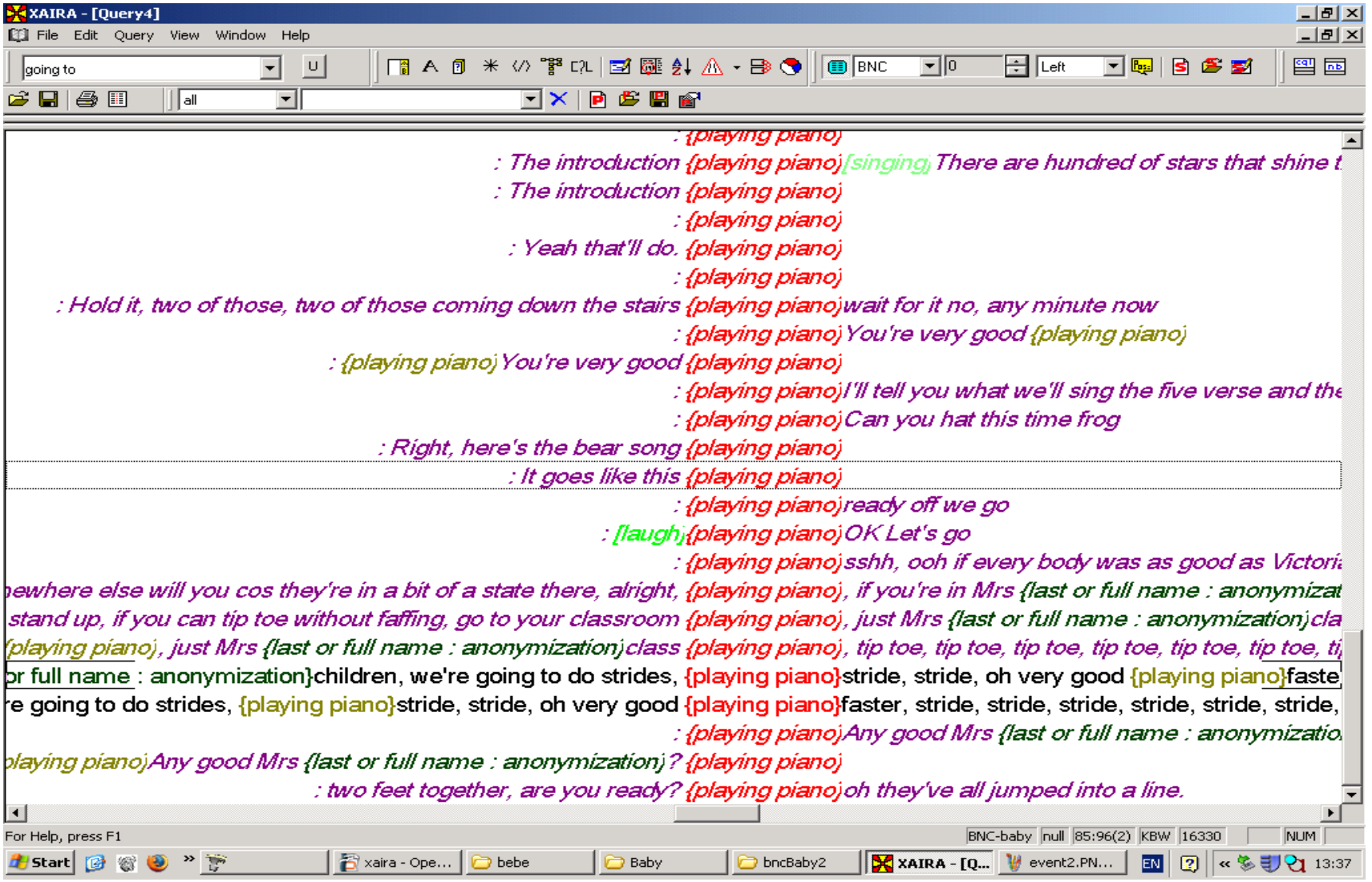


Explicit components of speech

Beyond transcription … markup makes explicit:

- changes of speaker and overlap
- words as perceived by transcriber
- indications of false starts, truncation, uncertainty
- some performance features e.g. pausing, dramatisation etc.
- speaker details where available (always for respondents, sometimes for others)

Performance features



Other features of spoken texts

- <shift> marks changes in voice quality

  e.g. whispering, laughing, etc., events and changes in voice quality.

- <vocal> marks non-verbal but vocalised sounds

  e.g. coughs, humming noises etc.

- <event> marks non-verbal and non-vocal events

  e.g. passing vehicles, animal noises, actions.

- <pause> marks significant pauses

  silence, longer than normal for the speaker(s).

- <unclear> marks unclear passages

  inaudible or incomprehensible

UNIVERSITY OF OXFORD

event description

baby
baby burped
baby cries
baby cry
baby crying
baby crying in background
baby gurgling
baby laughing
baby noise
baby noises
baby screaming
baby shouting
baby shouting over the top
baby shouts
baby speaking
baby squealing
baby talk
baby talking
background chatter
background chatter in pub
background chatter in pub
background chatting shuffling etcetera
background conversation

“Speech in the wild”

- Listen they were going [belch] that ain't a burp he said
- Like I'd be talking like this and suddenly it'll go [mimics microphone noises]
- He simply went [sound effect] through his nose
- Come on then shitbox

UNIVERSITY OF OXFORD

Vocal descriptions

<vocal desc="big breath"/>
<vocal desc="breathing out suddenly"/>
<vocal desc="drawing in breath"/>
<vocal desc="exhales"/>
<vocal desc="indrawn breath"/>
<vocal desc="inhales"/>
<vocal desc="intake of breath"/>
<vocal desc="sharp intake of breath"/>
<vocal desc="takes a deep breath"/>
<vocal desc="takes breath"/>
<vocal desc=“astonished snort"/>

Challenges

Amount of material; storage

- CD quality audio: 635 MB/hour
- Uncompressed .wav files: 115 MB/hour
- 2.8 GB/day
- 85 GB/month
- 1.02 TB/year
- Library/archive .wav files: 1 GB/hr, 9 TB/yr

Spoken audio = 250 times XML

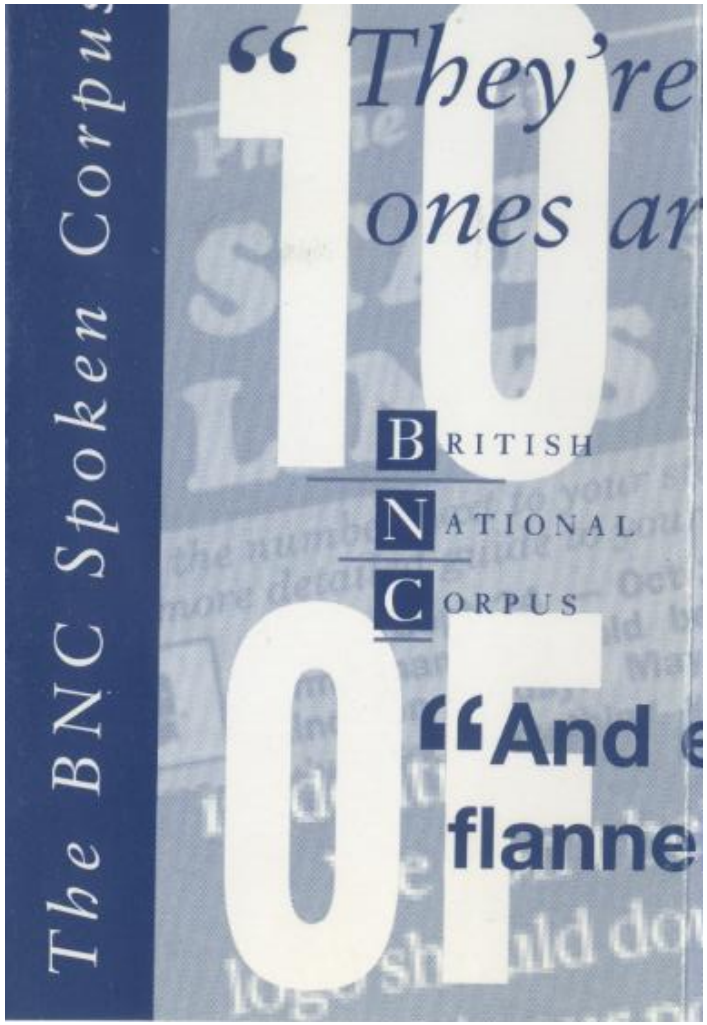
UNIVERSITY OF OXFORD

Information Technology, 1994

- WinWord or WordPerfect 5? the choice is yours
- On your desk … a 386 with 50 Mb disk space (just about enough to run Windows 3)
- 3½-inch floppy disks: 720 kB (later 1.44 MB)
- CD-R recorders cost $10-12,000
  - Wait until 1995 for Hewlett-Packard's $995 model
- In your lab (if you have a lab) a VAX, a Sparc or an SGI machine for serious work
- On the WWW (maybe) ... Mosaic for X

UNIVERSITY OF OXFORD

What happened to the audio?



- A few demo cassettes were circulated
- Some thoughts about CDs
- Copies of the original cassette tapes were deposited at the National Sound Archive … and ignored for over a decade
- A myth grew up that publication of the audio was not permitted

Speaker permissions form

"This is to confirm to the BNC that I agree to take part in the British National Corpus and that I give permission for all tape recordings and conversation details to be used as explained to me by the British Market Research Bureau and as confirmed in this letter, the accompanying letter, and Recording Guidelines, which I understand and accept.

I understand that all tapes and conversation details will be completely anonymous, and will be used for scientific study and publication by writers of dictionaries and educational material and language researchers."

UNIVERSITY OF OXFORD

What happened to the audio?

- All the tapes were transcribed in ordinary English spelling by audio typists
- The National Sound Archive moved into the new British Library building
- The 1000+ audiotapes could be audited if you went in to listen (‘digitized on demand’, in fact)
- In 2009-10 we set up a project with the British Library to have all the tapes digitized, at a cost of ~ £20k

UNIVERSITY OF OXFORD

Support from the Oxford University John Fell Fund and the British Library

Challenges

- Storing 1.02 TB/year: not really a problem in 21st century
- 1 TB (1000 GB) hard drive: c. ~~£65~~ *Now £39.95!*
- Computing (distance measures, alignments, labels etc): multiprocessor cluster

UNIVERSITY OF OXFORD

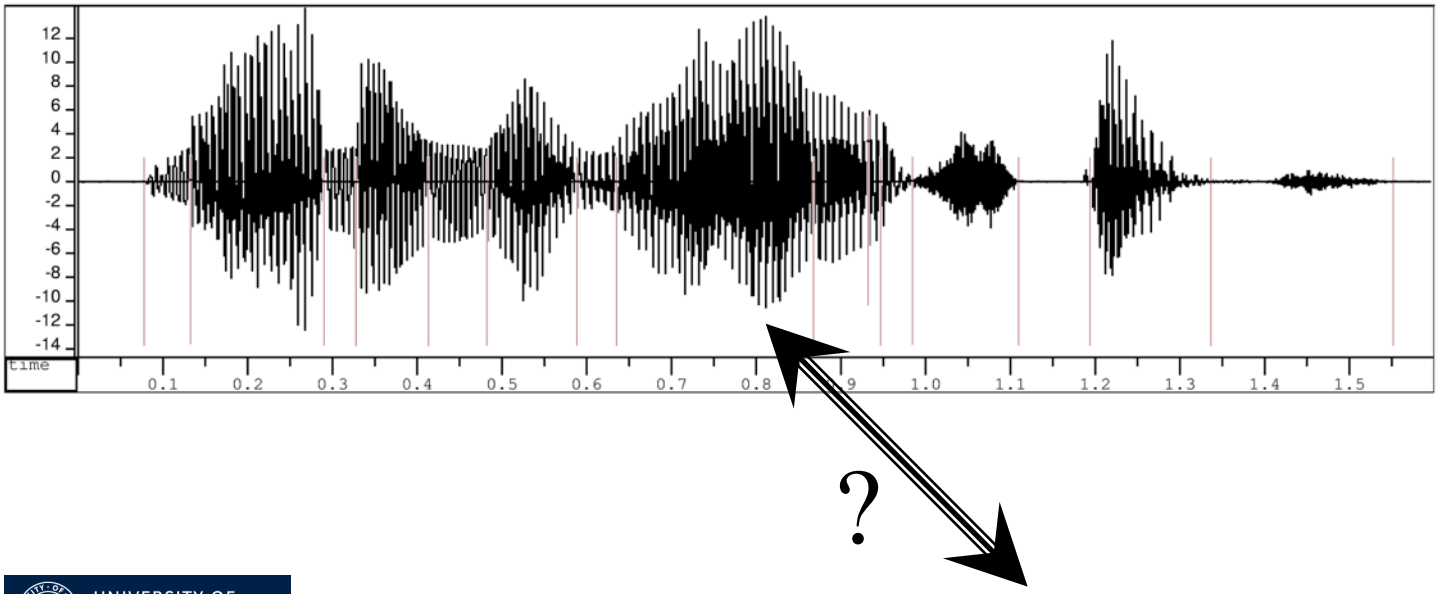
Challenges

- Amount of material
- Computing
  - distance measures, etc.
  - alignment of labels
  - searching and browsing
  - Just reading or copying 9 TB takes >1 day
  - Download time: days or weeks

Challenges

To make large spoken corpora practical, you need:

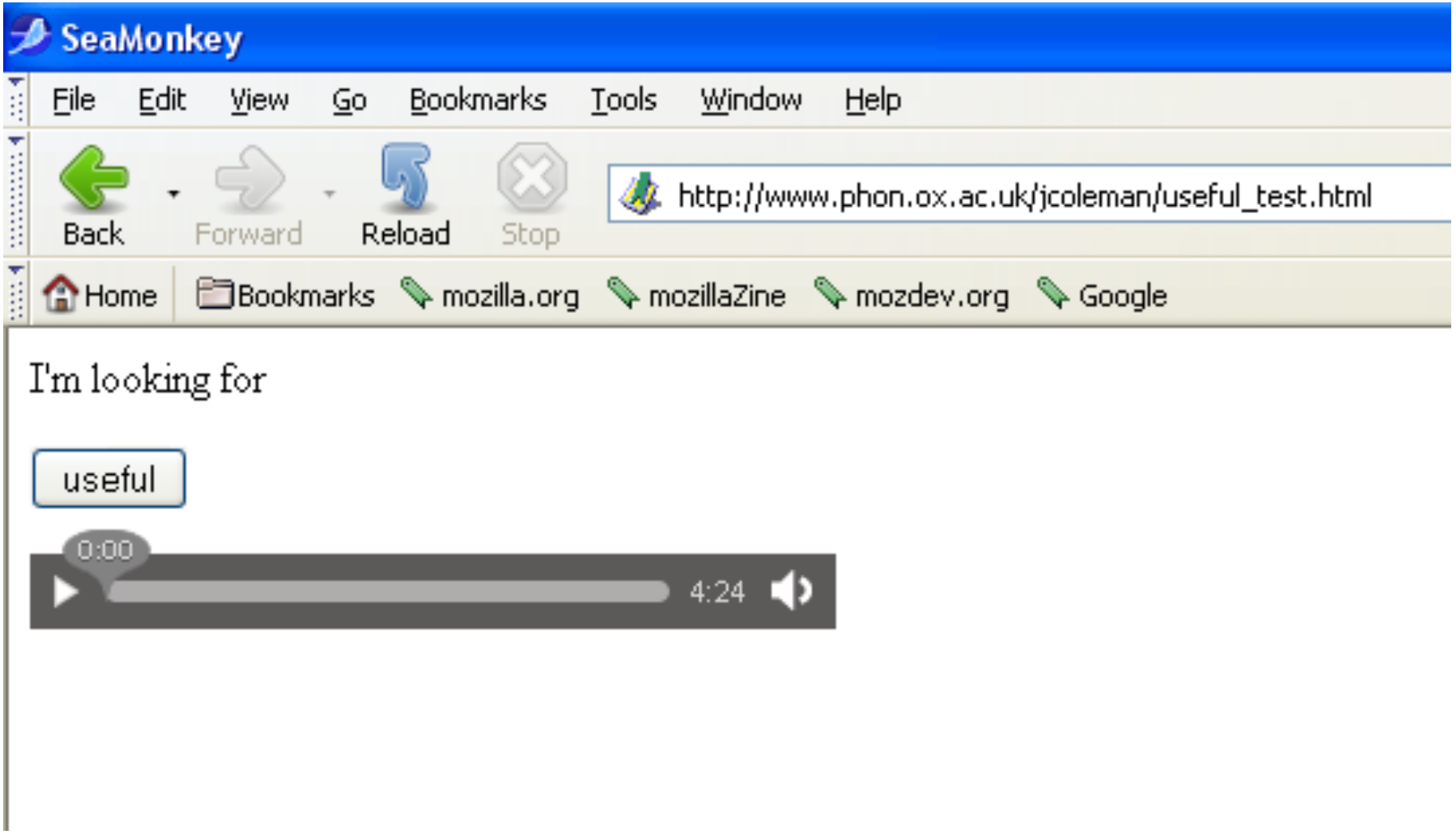
- A detailed index, so users can find the parts they need
- A way of using the index to access slices of the corpus



UNIVERSITY OF OXFORD

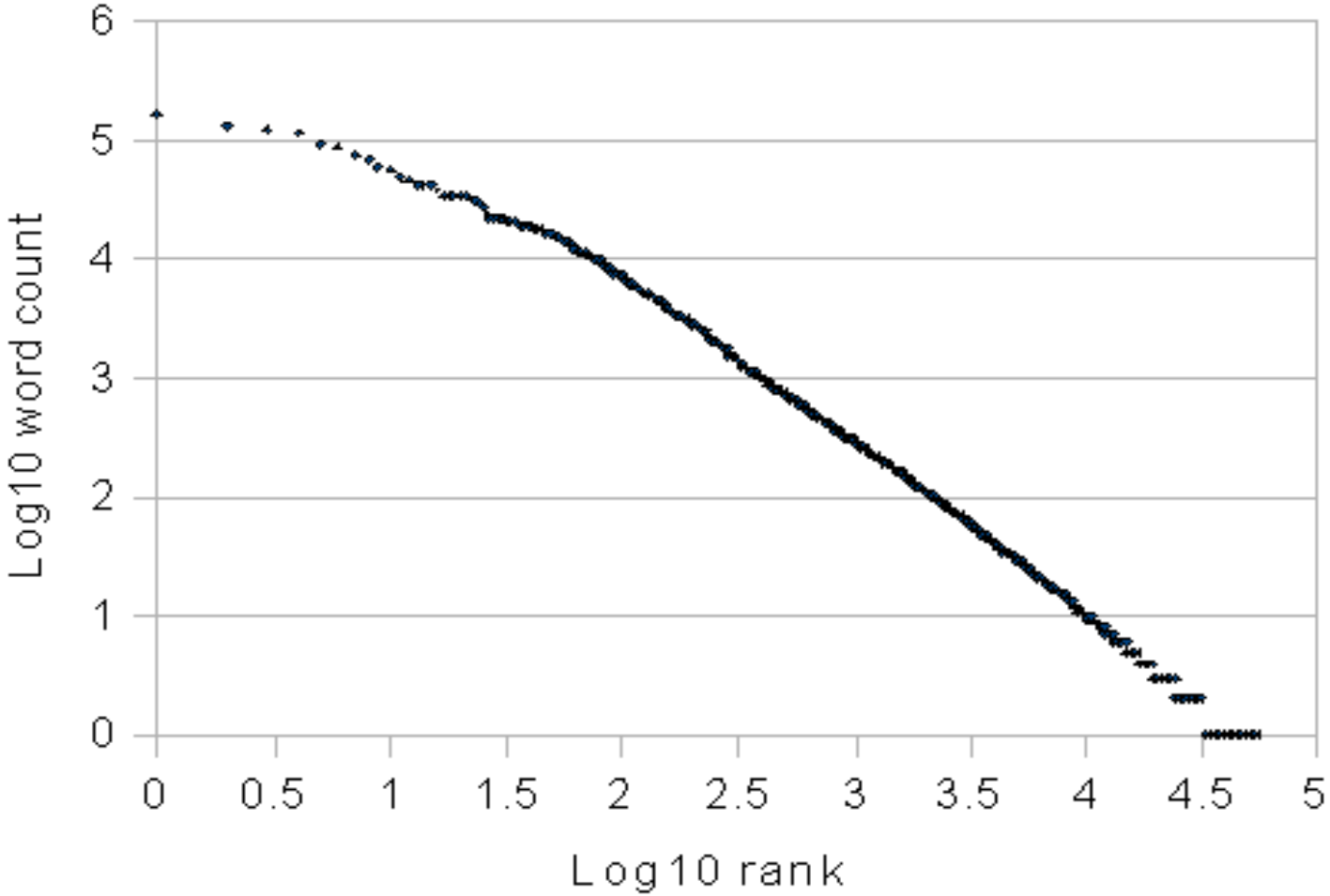
<w c5="AV0" hw="well" pos="ADV" >Well </w>

Streaming audio fragments

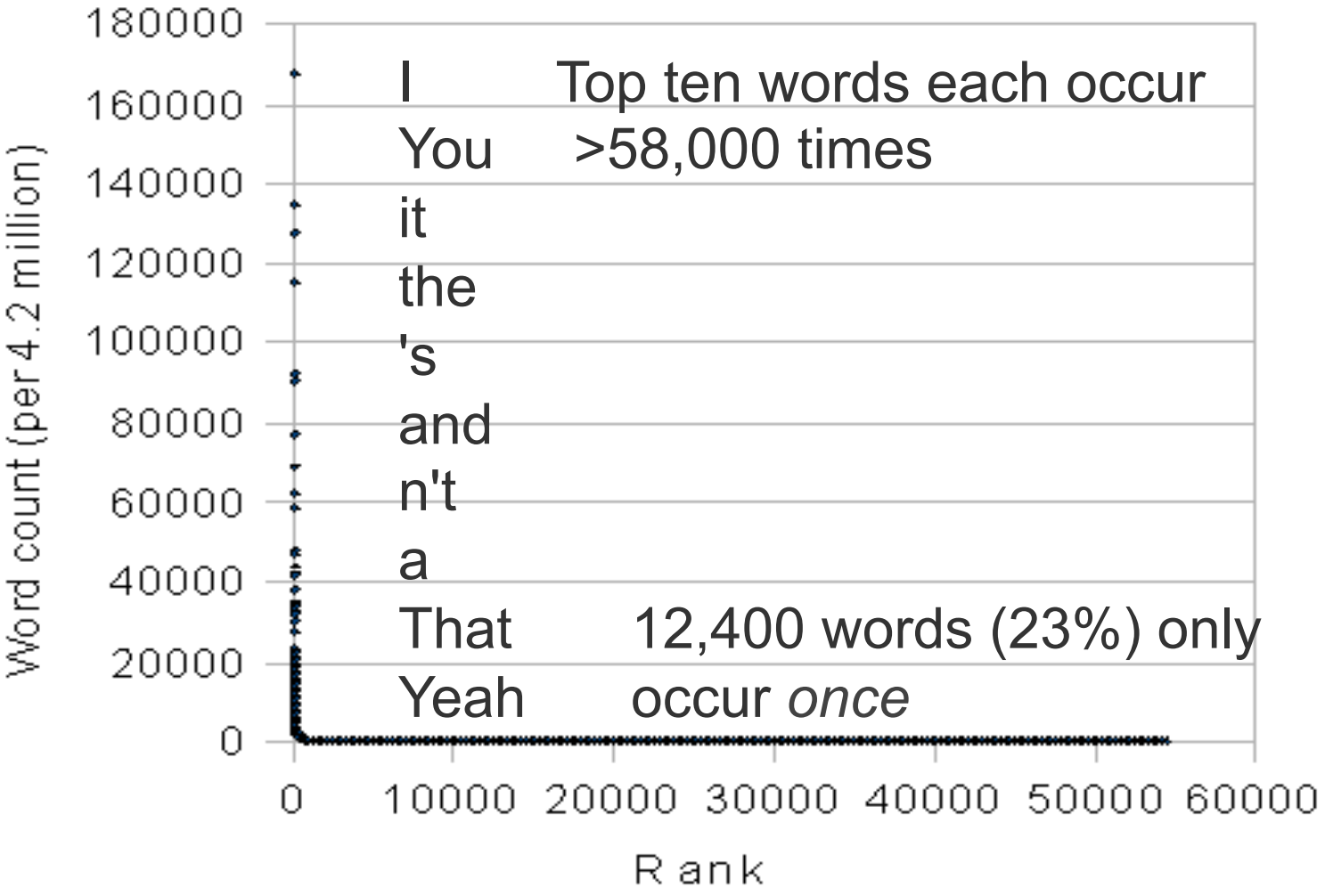



SeaMonkey
File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
https://aloe.phon.ox.ac.uk/BNC/test2.html
Search
Print
Home Bookmarks mozilla.org mozillaZine mozdev.org Google
File 021A-C0897X123400XX-0100P0.wav
SO NOW WE'RE BEING RECORDED ALL VERY EXCITING I HOPE
THEY CAN HEAR US SO IF WE CAN HAVE SOME
GOOD EXAMPLES OF THE ENGLISH LANGUAGE PLEASE {LG} OKAY UP
TO NOW WE'VE BEEN COVERING ONE PART OF THE NEURAL
NETWORKS ERM IN FACT WE ONLY COV COVERED ONE PART
OF ONE NETWORK IN FACT {OOV} ER THAT A FORM
OF CONTINUOUS NETWORK ER BECAUSE IT HAS CONTINUOUS WEIGHTS VARI
LOT OF VARIATIONS IN DIFFERENT SORTS OF NEURAL NETWORKS WE'VE
HAD DIFFERENT SORTS OF WEIGHTS DIFFERENCE SORTS OF INPUTS AND
LEARNING RULES AS YOU'LL SEE WHAT I WANT TO COVER
NOW IS A TYPE OF NETWORK THAT WE STUDY A
LOT IN THE COMPUTER ARCHITECTURE GROUP AND I'VE BEEN WORKING
WITH FOR YEARS AND YEARS AND YEARS ERM CALLED THE
N TUPLE NETWORK N TUPLE METHOD IT IN NEURAL NETWORK
TERMS IS A BINARY WEIGHTED IT MEANS THAT THE WEIGHTS
ARE TYPICALLY BINARY IN THE NETWORK SO YOU CAN ALSO
USE LOTS OF DIFFERENT LEARNING RULES COMPAR COMPARED TO THE
ER NETWORKS YOU SEE QUITE DIFFERENT HOWEVER EVEN THOUGH IT
IS A NEURAL NETWORK HOWEVER ERM YOU CAN SEE IT
IS A PACK RECOGNITION TECHNIQUE AND IN FACT IT WAS
FIRST DEVELOPED IN NINETEEN FIFTY NINE AS THE N TUPLE
METHOD ER BY SOME CHAP {GAP_ANONYMIZATION_NAME} BROWNING ERM I THINK
{GAP_ANONYMIZATION_NAME} {OOV} COS I'VE NEVER HEARD OF BROWNING SINCE THAT
PAPER SO WHAT I WANT TO DO IS TO INTRODUCE
THE BASIC IDEA OF WHAT THE LEARNING {OOV} DOES AND
https://aloe.phon.ox.ac.uk/BNC/test2.html
start
New BNC audio ...
SeaMonkey
Work
Digging into Dat...
Coleman_Orlean...
MiningYearOfSp...
13:44

Why so large? Lopsided sparsity (Zipf's law)



Why so large? Lopsided sparsity



Lopsided sparsity and size

Final -t/-d ‘deletion’:

- just 19563 tokens
- want 5221
- left 432
- slammed 6

UNIVERSITY OF OXFORD

A rule of thumb

To catch most

- English phonemes, you need minutes of audio
- common words … a few hours
- a typical person's vocabulary … >100 hrs
- pairs of common words … >1000 hrs
- arbitrary word-pairs … >100 years

UNIVERSITY OF OXFORD

Lopsided sparsity and size

Fox and Robles (2010): 22 examples of *It's like*-enactments [e.g. it's like 'mmmmmm'] in 10 hours of data

Rare and unique wonders

aqualunging boringest chambermaiding

de-grandfathered europeaney gronnies

hoptastic lawnmowing mellies noseless

punny regurgitate-arianism scunny

smackerooney tooked weppings

yak-chucker zombieness

UNIVERSITY OF OXFORD

Not just repositories of words

Specific phrases or constructions

Particularities of people's voices and speaking habits

Dog-directed speech

Parrot-directed speech

Language in the wild

- Talking to George (a bird)
- Talking to dogs
- Try transcribing this!
- There’s gronnies lurking about



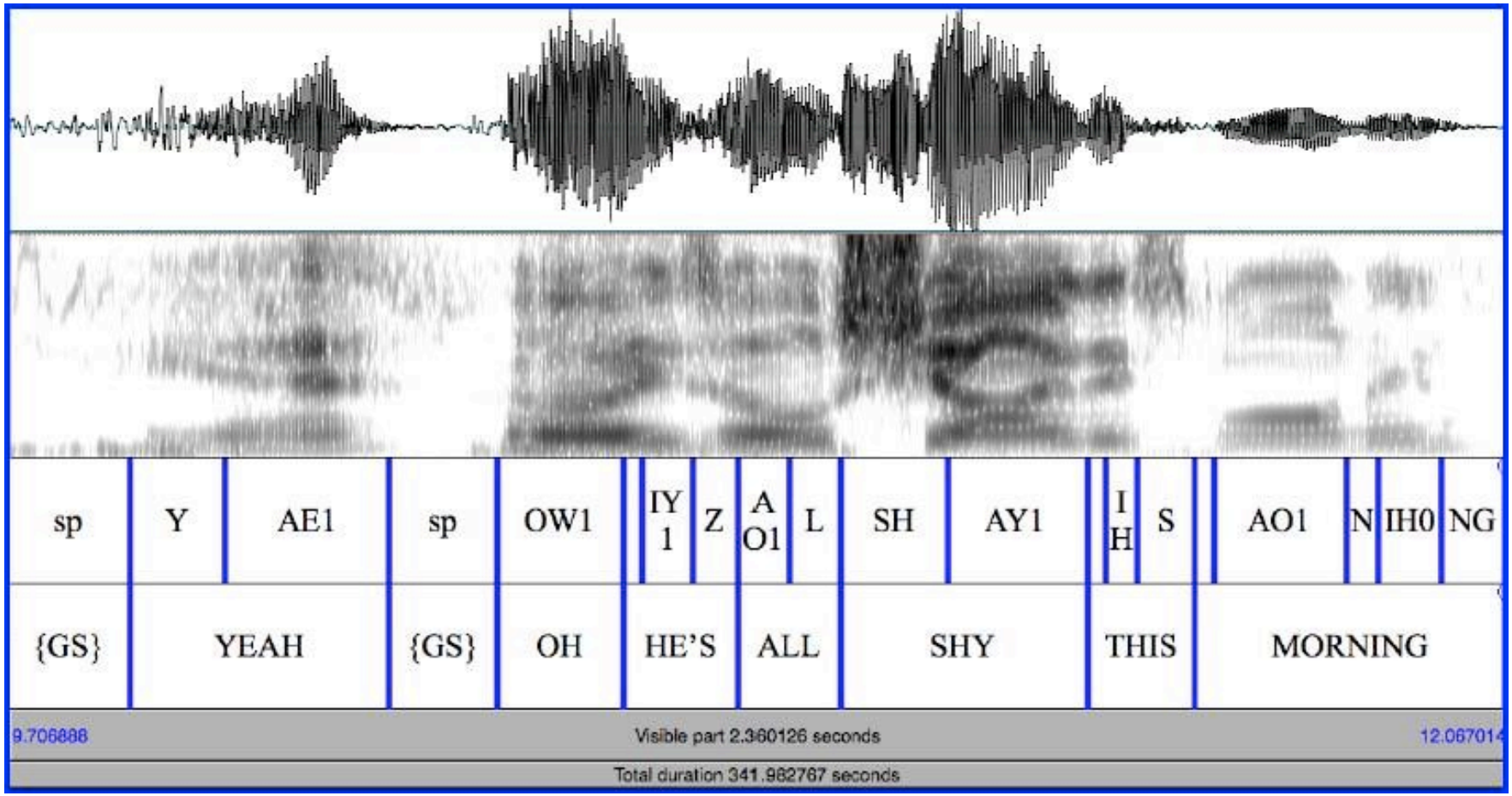



UNIVERSITY OF OXFORD

Practicalities

- In order to be useful, such very large corpora must be indexed at word and segment level
- All included speech corpora must therefore have associated text transcriptions
- We use the Penn Phonetics Laboratory Forced Aligner to associate each word and segment with the corresponding start and end points in the sound files

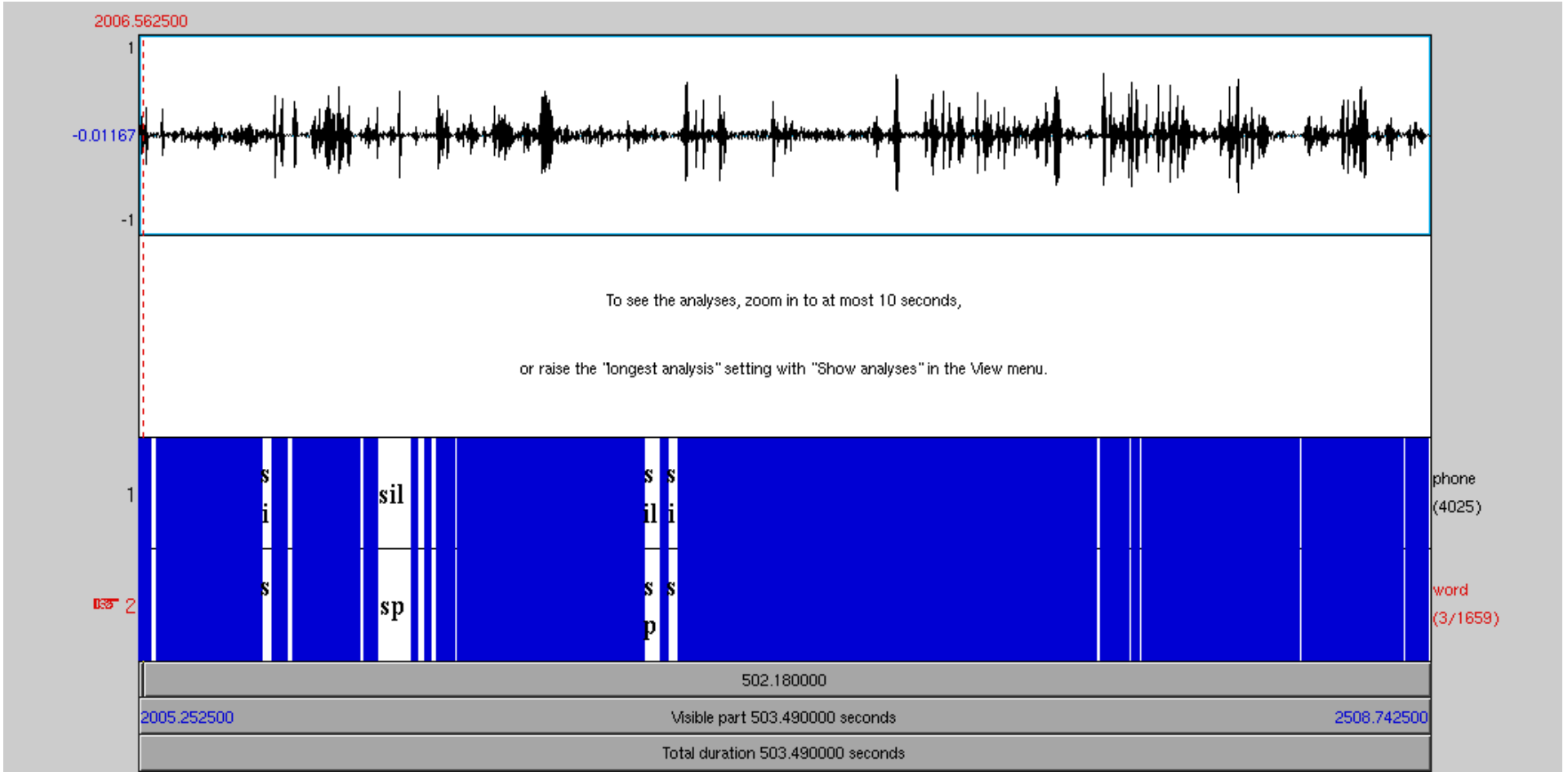
Indexing by forced alignment



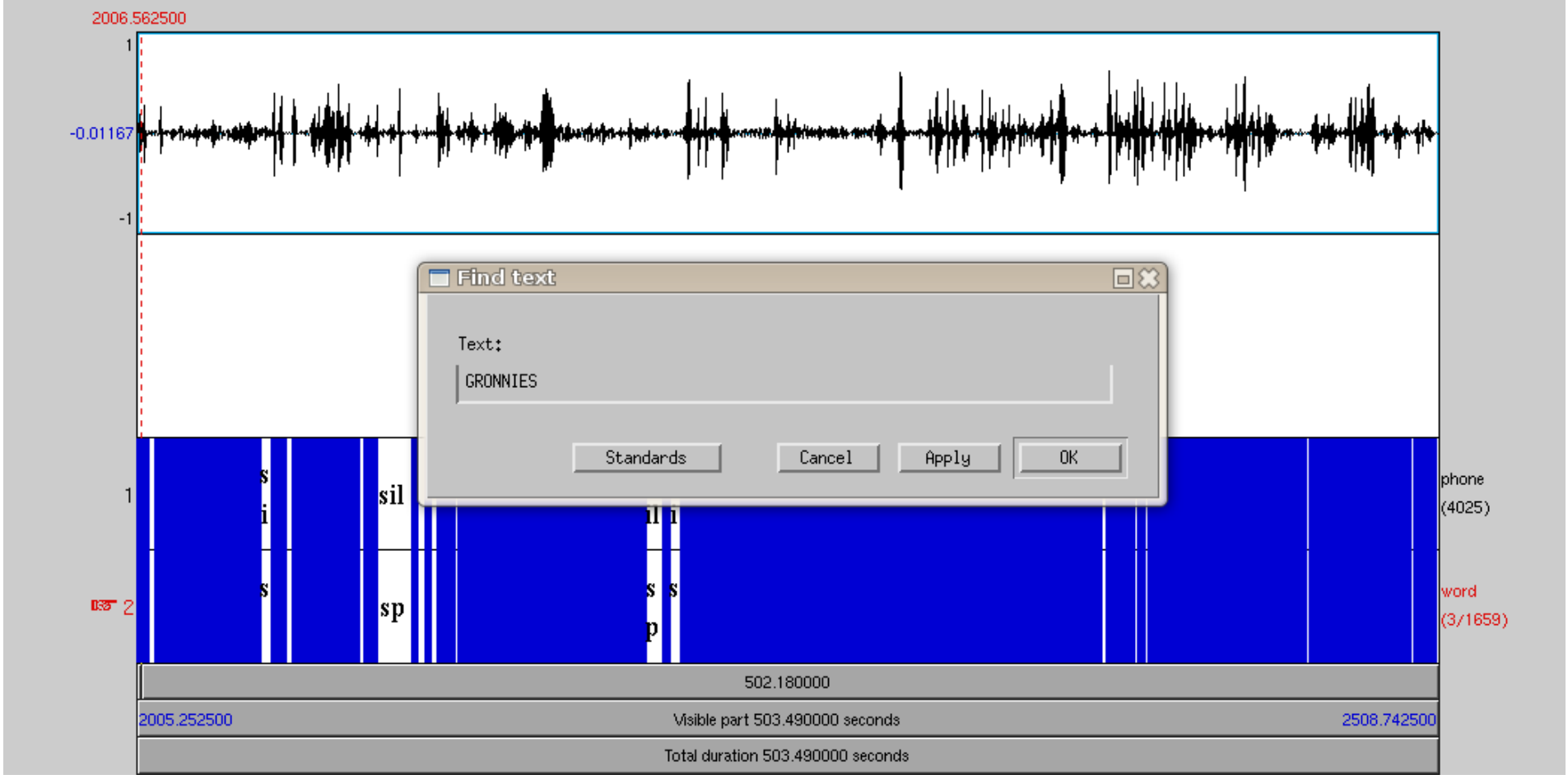
UNIVERSITY OF OXFORD

x 21 million

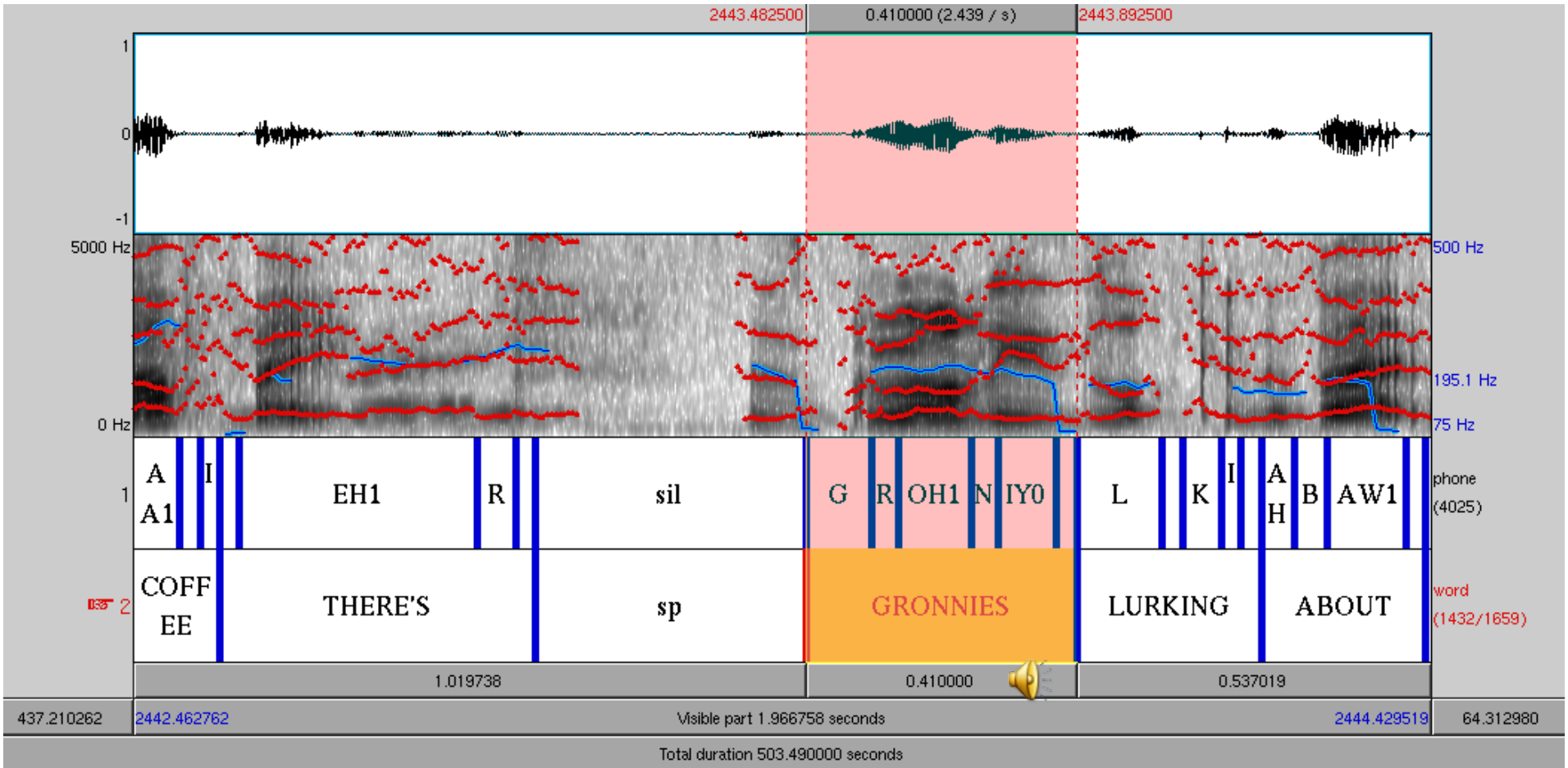
Indexing by forced alignment



Digging for treasure



Digging for treasure



Using an American aligner with British English recordings

Same set of acoustic models

e.g. same [ɑ] model for US “Bob” and UK “Ba(r)b”

Pronunciation differences between different varieties are dealt with by listing multiple phonetic transcriptions

UNIVERSITY OF OXFORD

Building a multi-dialect dictionary (1): diagnosis of problems

		Problem	corrected form	American pron Variants
9766	id= 375s1	?		
5984	upa	??		
4550	mm.mm	(Two filled hesitations)		M M
979	banw	abbreviation or word?		B AE1 N UW0
1765	ceau?escu	Accented lette	Ceausescu	CH AW0 CH EH1 S K Y UW0
13219	À	Accented letters		AA1 AE1
13220	École	Accented letters		EH2 K OW1 L
13221	Époque	Accented letters		EH0 P OW1 K
13222	élite	Accented letters		EY0 L IY1 T
13223	émigré	Accented letters		EH1 M IH0 G R EY2
6867	verus	deliberate mispronunciation		V IH1 ER0 AH0 V EH1 R AH0 S
10656	oia_011207.tm	Filename?		
691	aldate's.	Final .		AO1 L D EY2 T AO1 L D EY2 T S
792	irish.	Final .	Irish	AY1 R IH2 S H
926	attaboy.	Final .		AE1 T AH0 B OY2
934	aubergines.	Final .		OW1 B ER0 Z H OW1 B AH0 Z H IY0 N Z

UNIVERSITY OF OXFORD

Building a multi-dialect dictionary (2): generation of transcriptions

- BEEP dictionary
- g2p (grapheme-to-phoneme) algorithm
- Orthographic nearest neighbours
- Expert phonologists selected correct candidates

  (checking is far quicker than transcription)

Transcriptions needed to be manually created only for ~10,000 items

UNIVERSITY OF OXFORD

Building a multi-dialect dictionary (3): extend to 4 main dialect regions by rule

- Southern vs. Northern × Rhotic vs. Nonrhotic
- “Southern” = /ʌ/, /baθ/ “Northern” = /ʊ/, /bɑθ/
- “Southern Rhotic” taken as a basis (~ American)

	S Rhotic British	S Nonrhotic British	N Rhotic British	N Nonrhotic British
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
roadrunners	R OW1 D R AH2 N ER0 Z	R OW1 D R AH2 N AH0 Z	R OW1 D R UH2 N ER0 Z	R OW1 D R UH2 N AH0 Z

UNIVERSITY OF OXFORD

Forced alignment is *not* perfect

- “un alignement parfait entre l’enregistrement sonore et sa transcription phonétique”?
- Non, hélas!
- ~23% is accurately aligned (20 ms)
- ~80% is aligned within 2 seconds

UNIVERSITY OF OXFORD

Some causes of difficulty

- Overlapping speakers
- Background noise/music/babble
- Transcription errors
- Variable signal loudness
- Reverberation, distortion
- Poor speaker vocal health/voice quality
- Unexpected accents

Anonymization

- The text transcriptions in the published BNC have already been anonymized
- Some parts of the audio have also been published (e.g. COLT)
- Full names, personal addresses and telephone numbers were replaced by <gap> tags
- We use the location of all such tags to mute (silence) the corresponding portions of audio

UNIVERSITY OF OXFORD

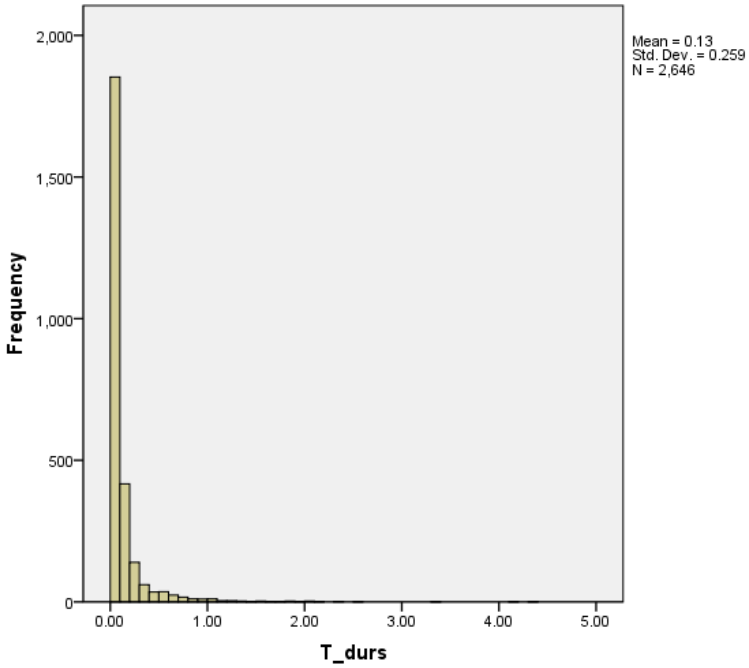
Publication/release plans: BNC

- When we finish checking the alignment of anonymization gaps, we will release the whole BNC Spoken Audio corpus
- In the mean time, there is a small sampler
  - http://www.phon.ox.ac.uk/SpokenBNC
  - Including the audio, alignments, and HTML texts
- We'll soon release the well-aligned *half* of the corpus
- Later: full release as linked data via the British Library Archival Sound Recordings server

UNIVERSITY OF OXFORD

Final -t/-d variation: deletion or "continuous speech process"?

- E.g. jus(t), wan(t), lef(t), slamm(ed)
- Distribution of durations
- Acoustic differences from unreduced standards
- Correlations with social and linguistic factors



UNIVERSITY OF OXFORD

Previously unattested ("impossible") assimilations of word-final consonants

I'm gonna 
seem/n to
alarng clock

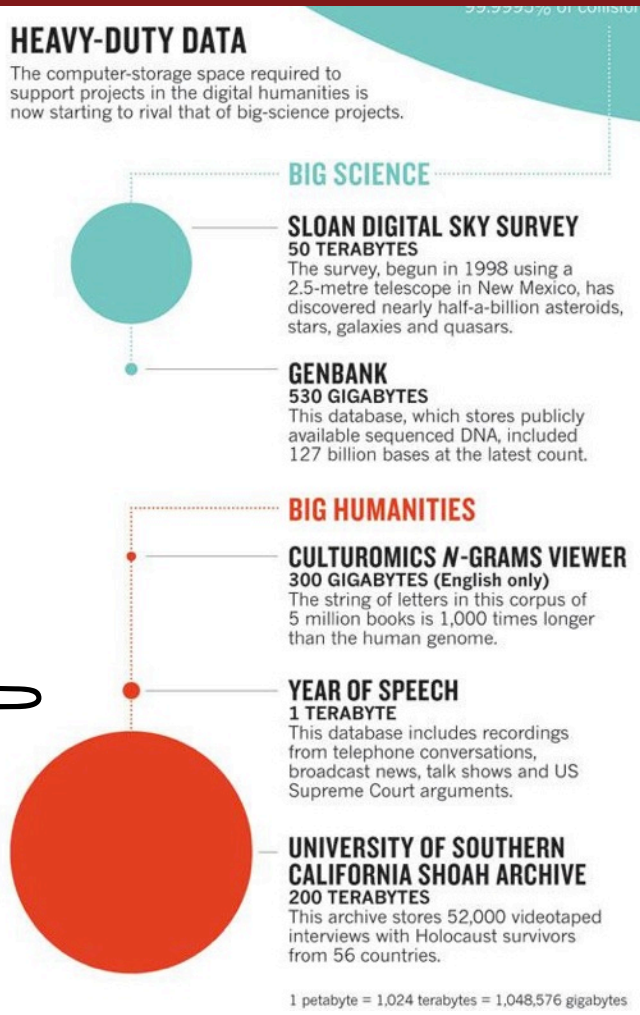
swimmim pool 
gettim paid
weddim present

UNIVERSITY OF OXFORD

nature International weekly journal of science vol. 474, 436-440 (17 Jun 2011)

Merci beaucoup!

Thank you very much!



Questions?

UNIVERSITY OF OXFORD