Linked Speech in Crowd/Cloud Corpus Consortia

John Coleman

Phonetics Laboratory, University of Oxford

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

The Future of Language Resources:
LDC 20th Anniversary Workshop, 6th September 2012

Outline

- Approaches to corpus dissemination
- Digging into Data: Mining a Year of Speech
- The need for large corpora
- Problem 1: Finding stuff
- Problem 2: Getting stuff
- Problem 3: Sharing stuff

UNIVERSITY OF OXFORD

Normal approach to corpus publication

- An institution or project collects and prepares a corpus.

UNIVERSITY OF OXFORD

Normal approach to corpus publication

- An institution or project collects and prepares a corpus.
- They submit it to a data centre, and/or put it on their own website.

UNIVERSITY OF OXFORD

Normal approach to corpus publication

- An institution or project collects and prepares a corpus.
- They submit it to a data centre, and/or put it on their own website.
- Users log on and download the corpus. Fees and passwords may be required.

UNIVERSITY OF OXFORD

Normal approach to corpus publication

- An institution or project collects and prepares a corpus.
- They submit it to a data centre, and/or put it on their own website.
- Users log on and download the corpus. Fees and passwords may be required.
- Maybe, the corpus contains (some of) what they want.

UNIVERSITY OF OXFORD

Normal approach to corpus publication

Problems:

- An institution or project collects and prepares a corpus.
- They submit it to a data centre, and/or put it on their own website.
- Users log on and download the corpus. Fees and passwords may be required.
- Maybe, the corpus contains (some of) what they want.

UNIVERSITY OF OXFORD

Normal approach to corpus publication

Problems:

- An institution or project collects and prepares a corpus.
- They submit it to a data centre, and/or put it on their own website.
- Users log on and download the corpus. Fees and passwords may be required.
- Maybe, the corpus contains (some of) what they want.

Time and effort; other people's rules

UNIVERSITY OF OXFORD

Normal approach to corpus publication

Problems:

- An institution or project collects and prepares a corpus.
- They submit it to a data centre, and/or put it on their own website.
- Users log on and download the corpus. Fees and passwords may be required.
- Maybe, the corpus contains (some of) what they want.

Time and effort; other people's rules

The whole thing?

What a hassle!

UNIVERSITY OF OXFORD

Normal approach to corpus publication

Problems:

- An institution or project collects and prepares a corpus.
- They submit it to a data centre, and/or put it on their own website.
- Users log on and download the corpus. Fees and passwords may be required.
- Maybe, the corpus contains (some of) what they want.

Time and effort; other people's rules

The whole thing?

What a hassle!

Or not! What is where?

UNIVERSITY OF OXFORD

Cloud/crowd corpora: collaboration, not collection

Search interface 1
(e.g. Oxford)

LDC database - retrieve time stamps

Spoken LDC recordings

Search interface 2
(e.g. British Library)

BNC-XML database - retrieve time stamps

Spoken BNC recordings - various locations

Search interface 3
(e.g. Penn)

Search interface 4

The need for large corpora in the study of natural language

Lab speech	**Spoken corpora**
artificially elicited	spontaneous
balanced	reflect usage
few variables	many factors of variation
uncommunicative	communicative
controlled	out of control
small	(need to be very) large
rather odd	extremely odd, i.e. real

UNIVERSITY OF OXFORD

My example: AudioBNC

BRITISH NATIONAL CORPUS

- a snapshot of British English in the early 1990s
- 100 million words in ~4000 different *text* samples of many kinds, spoken (10%) and written (90%)
- freely available worldwide under licence since 1998; latest edition is BNC-XML
- various online portals
- no audio (until now)

UNIVERSITY OF OXFORD

Spoken part: 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
- includes London teenage talk, later published as COLT (Stenström et al.)

UNIVERSITY OF OXFORD

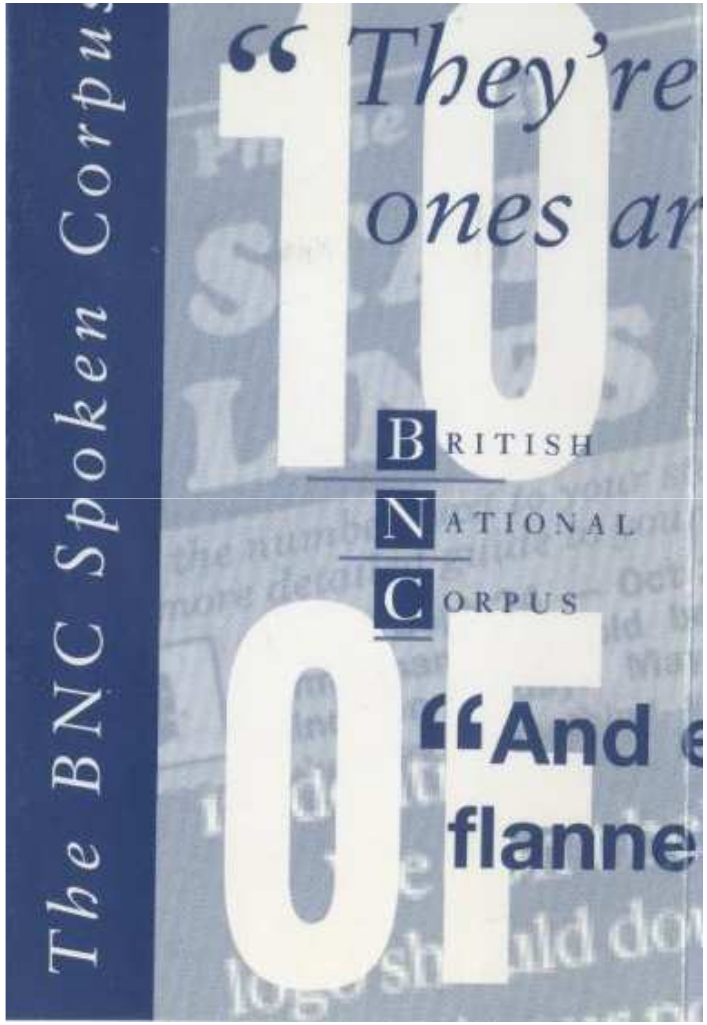
Spoken texts

Demographic part: 4.2 million words

Context-governed part: Four broad categories for social context, roughly 1.5 million words 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

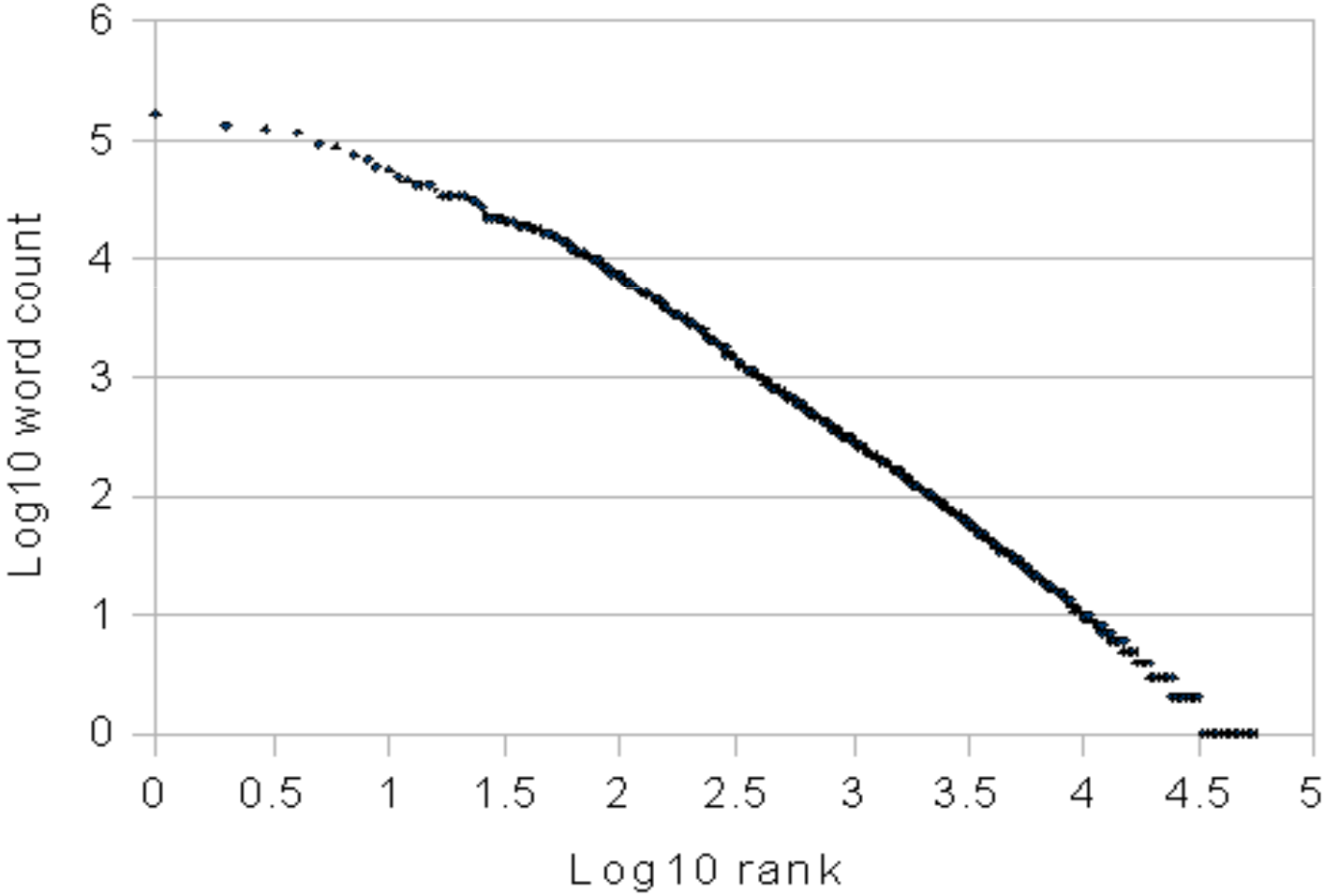
What happened to the audio?



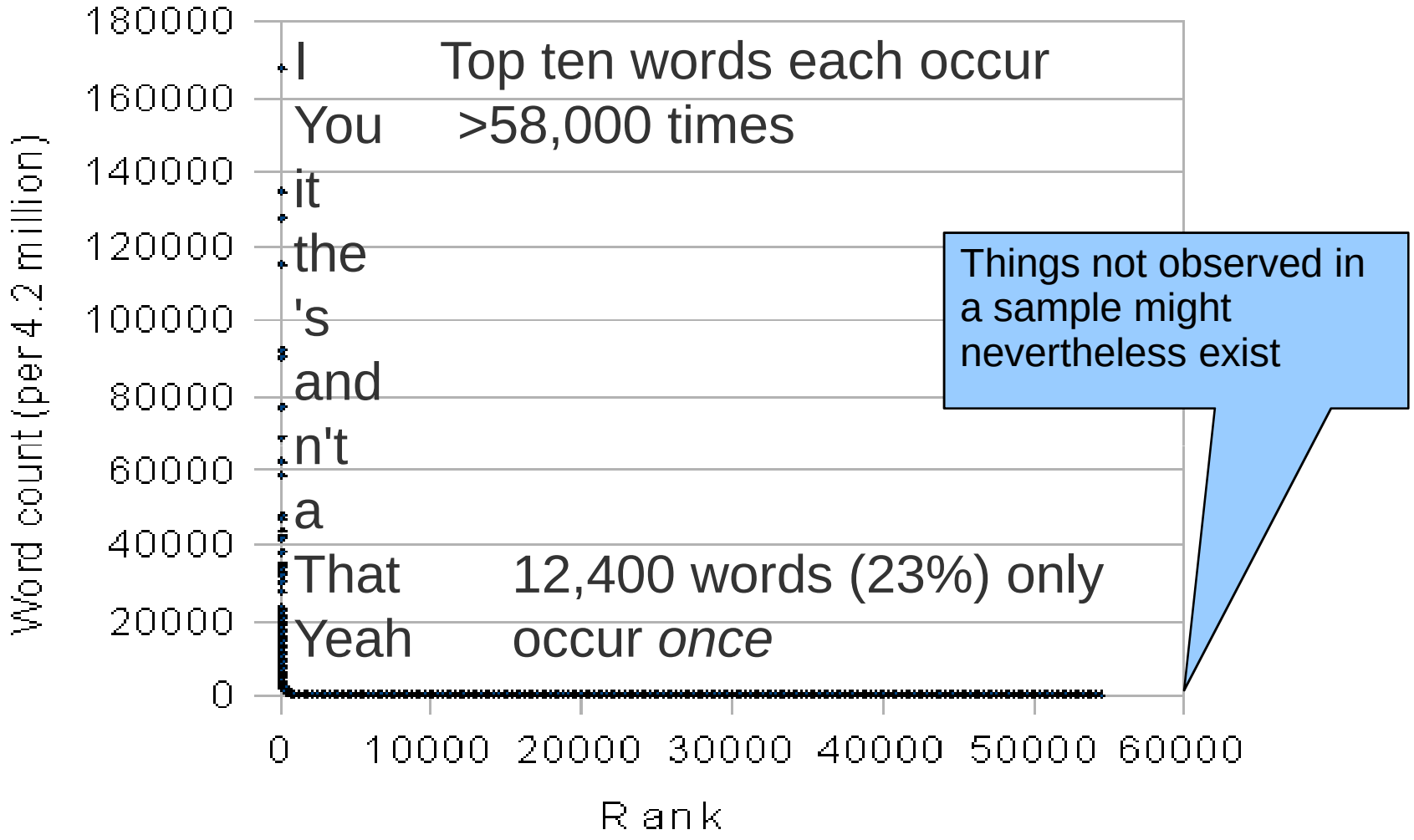
- All the tapes were transcribed in ordinary English spelling by audio typists
- Copies of the tapes were given to the National Sound Archive
- In 2009-10 we had a project with the British Library to digitize all the tapes (~1,400 hrs)
- We anonymized the audio in accordance with the original transcription protocols

UNIVERSITY OF OXFORD

The need for corpora to be *large*: lopsided sparsity (Zipf's law)



The need for *very* large corpora



UNIVERSITY OF OXFORD

Just listening and waiting, how long till items show up?

	For the 1st token, listen for	
[ʒ], the least frequent English phoneme (i.e. to get all English phonemes)	13 minutes	
"*twice*" (1000th most frequent word in the Audio BNC)	14 minutes	
"*from the*" (the most frequent word-pair in our current study)	17 minutes	
"*railways*" (10,000th most frequent word)	26 hours	
"*getting paid*" (the least frequent word-pair occurring >10 times in latest study)	95 hours (4 days)	

UNIVERSITY OF OXFORD

Just listening and waiting, how long till items show up?

	For the 1st token, listen for	**For 10 tokens, listen for**
[ʒ], the least frequent English phoneme (i.e. to get all English phonemes)	13 minutes	5 hours
"*twice*" (1000th most frequent word in the Audio BNC)	14 minutes	44 hours
"*from the*" (the most frequent word-pair in our current study)	17 minutes	22 hours
"*railways*" (10,000th most frequent word)	26 hours	41 days without sleep
"*getting paid*" (the least frequent word-pair occurring >10 times in latest study)	95 hours (4 days)	37 days

UNIVERSITY OF OXFORD

Problem 1: Finding stuff

- 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?

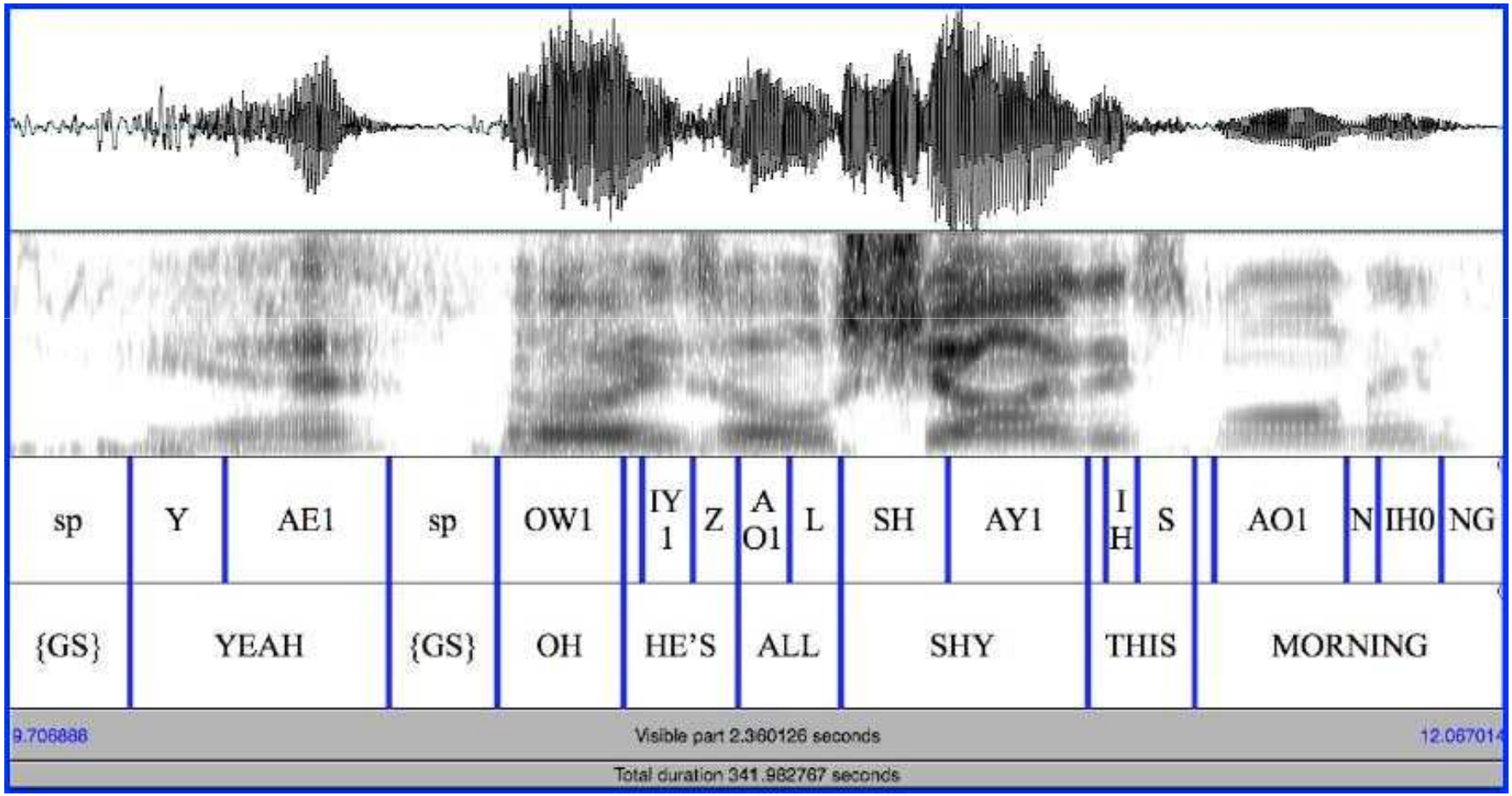
UNIVERSITY OF OXFORD

Practicalities

- To be useful, large speech corpora must be indexed at word and segment level
- We used a forced aligner* to associate each word and segment with their start and end points in the sound files
- Pronunciation differences between varieties are dealt with by listing multiple phonetic transcriptions in the lexicon, and letting the aligner choose for each word which sequence of models is best
  - * HTK, with HMM topology to match P2FA, with a combination of P2FA American English + our UK English acoustic models

UNIVERSITY OF OXFORD

Indexing by forced alignment



UNIVERSITY OF OXFORD

x 21 million

Forced alignment is *not* perfect

- Overlapping speakers
- Variable signal loudness
- Transcription errors
- Unexpected accents
- In a pilot:
  - ~23% was accurately aligned (20 ms)
  - ~80% was aligned within 2 seconds
- In our nasals study, ~67% of word-pairs of interest are well-aligned within 100 ms

- Background noise/music/babble
- Reverberation, distortion
- Poor speaker vocal health/voice quality

UNIVERSITY OF OXFORD

AudioBNC publication

- We have now released most of the aligned Audio BNC
  - http://www.phon.ox.ac.uk/AudioBNC (webpage) and

    http://bnc.phon.ox.ac.uk (data)
  - Includes .wav audio, Praat TextGrid alignments, HTML texts, and in future will include TEI-XML for speech
- Later: permanent release via the British Library
- Experiments in search tools, linked data etc.

UNIVERSITY OF OXFORD

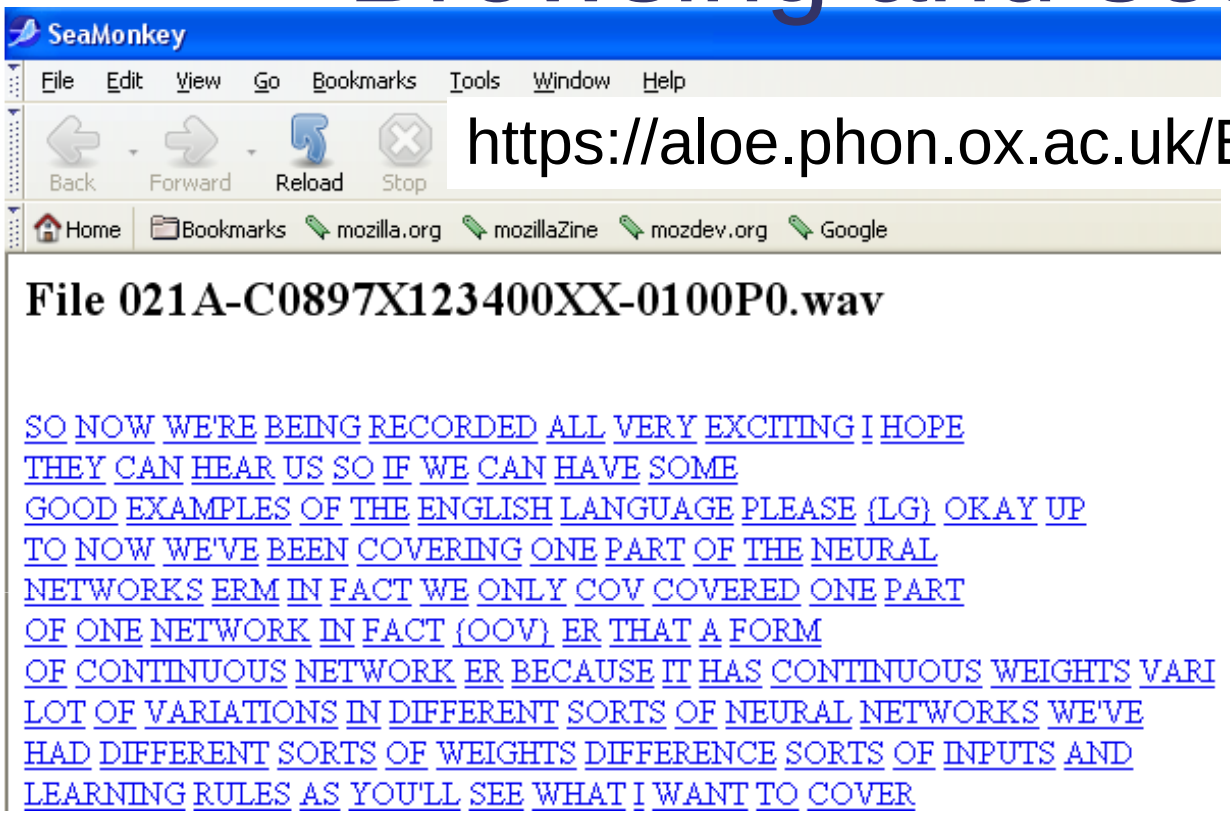
Problem 2: Getting stuff

- just reading or copying a year of audio takes >1 day
- download time: days or weeks

- browsing
- searching
- saving
- *linking* to stable clips

UNIVERSITY OF OXFORD

Browsing and searching



Browsing and searching

SeaMonkey

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop

https://aloe.phon.ox.ac.uk/BNC/test2.html

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
NETWORKS ERM IN FACT WE ONL
OF ONE NETWORK IN FACT {OOV
OF CONTINUOUS NETWORK ER B
LOT OF VARIATIONS IN DIFFEREN
HAD DIFFERENT SORTS OF WEIGH
LEARNING RULES AS YOU'LL SEE

http://www.phon.ox.ac.uk/jcoleman/useful_test.html

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop http://www.phon.ox.ac.uk/jcoleman/useful_test.html

Home Bookmarks mozilla.org mozillaZine mozdev.org Google

I'm looking for

useful

0:00 4:24

"USEFUL" 271.2325 271.5925 021A-C0897X0020XX-AAZZP0_002011_KBK_11.result
"USEFUL" 733.5825 733.7625 021A-C0897X0020XX-AAZZP0_002011_KBK_11.result
"USEFUL" 670.2725 670.4525 021A-C0897X0020XX-ABZZP0_002010_KBK_10.result
"USEFULLY" 1513.2225 1513.9625 021A-C0897X0023XX-AAZZP0_002306_KBK_72.result
"USEFUL" 519.3125 519.5325 021A-C0897X0023XX-ABZZP0_002317_KBK_83.result

W3C media fragments protocol

"GINORMOUS" 1870.8425 1871.4425 021A-C0897X0093XX-ABZZP0_009304_KBE_18.wav
"GINORMOUS" 1360.7725 1361.5625 021A-C0897X0097XX-ABZZP0_009707_KC5_7.wav
"GINORMOUS" 917.8625 918.3825 021A-C0897X0102XX-AAZZP0_010203_KE3_3.wav
"GINORMOUS" 838.7625 839.1725 021A-C0897X0103XX-AAZZP0_010305_KE3_19.wav
"GINORMOUS" 840.1925 840.6525 021A-C0897X0103XX-AAZZP0_010305_KE3_19.wav

duration (or t = end time)

start time

B side

Tape No

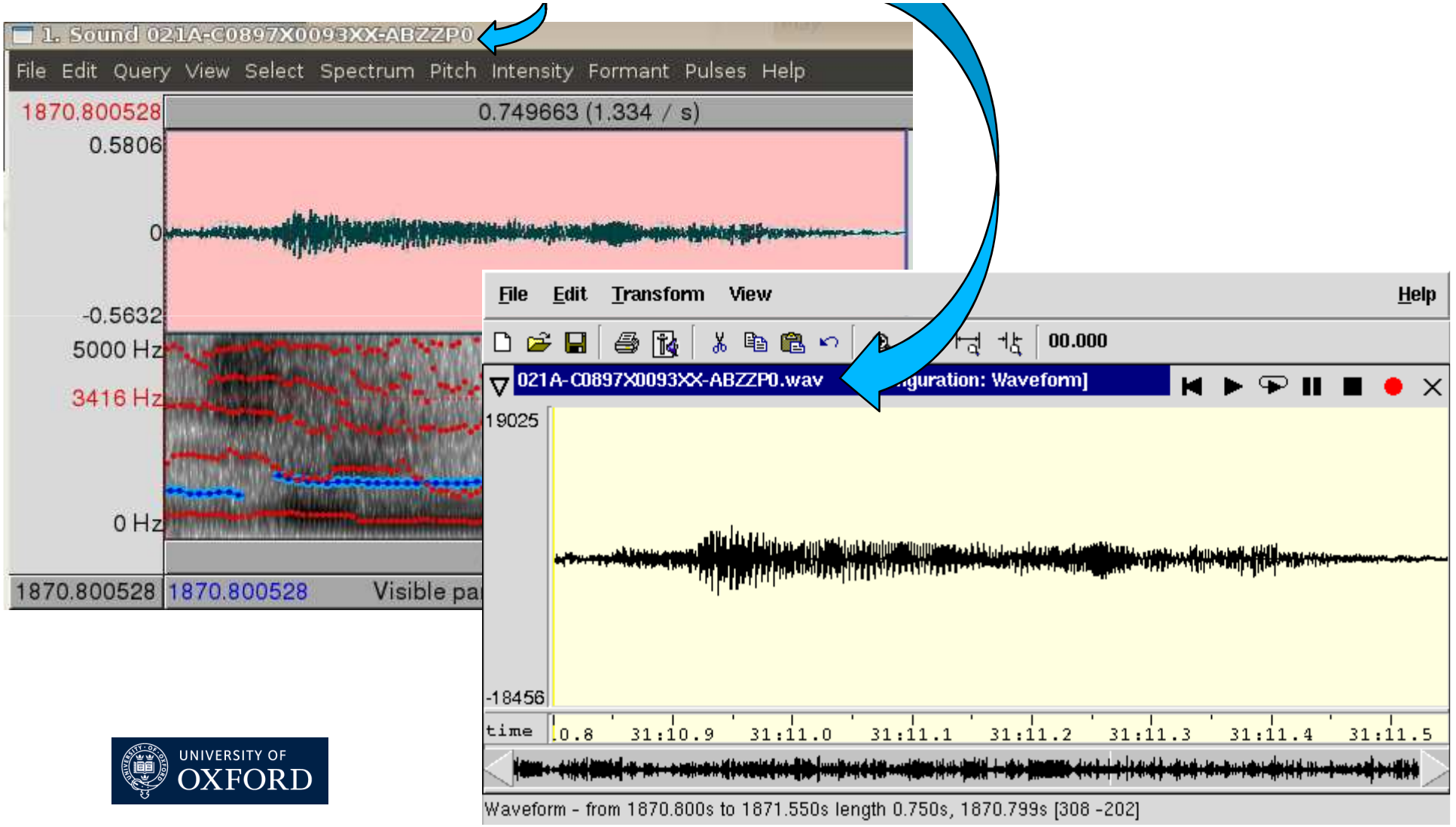
BL Cat No

Server URL

bnc.phon.ox.ac.uk/data/021A-C0897X0093XX-ABZZP0.wav?t=1870.8&d=0.75

W3C media fragments protocol

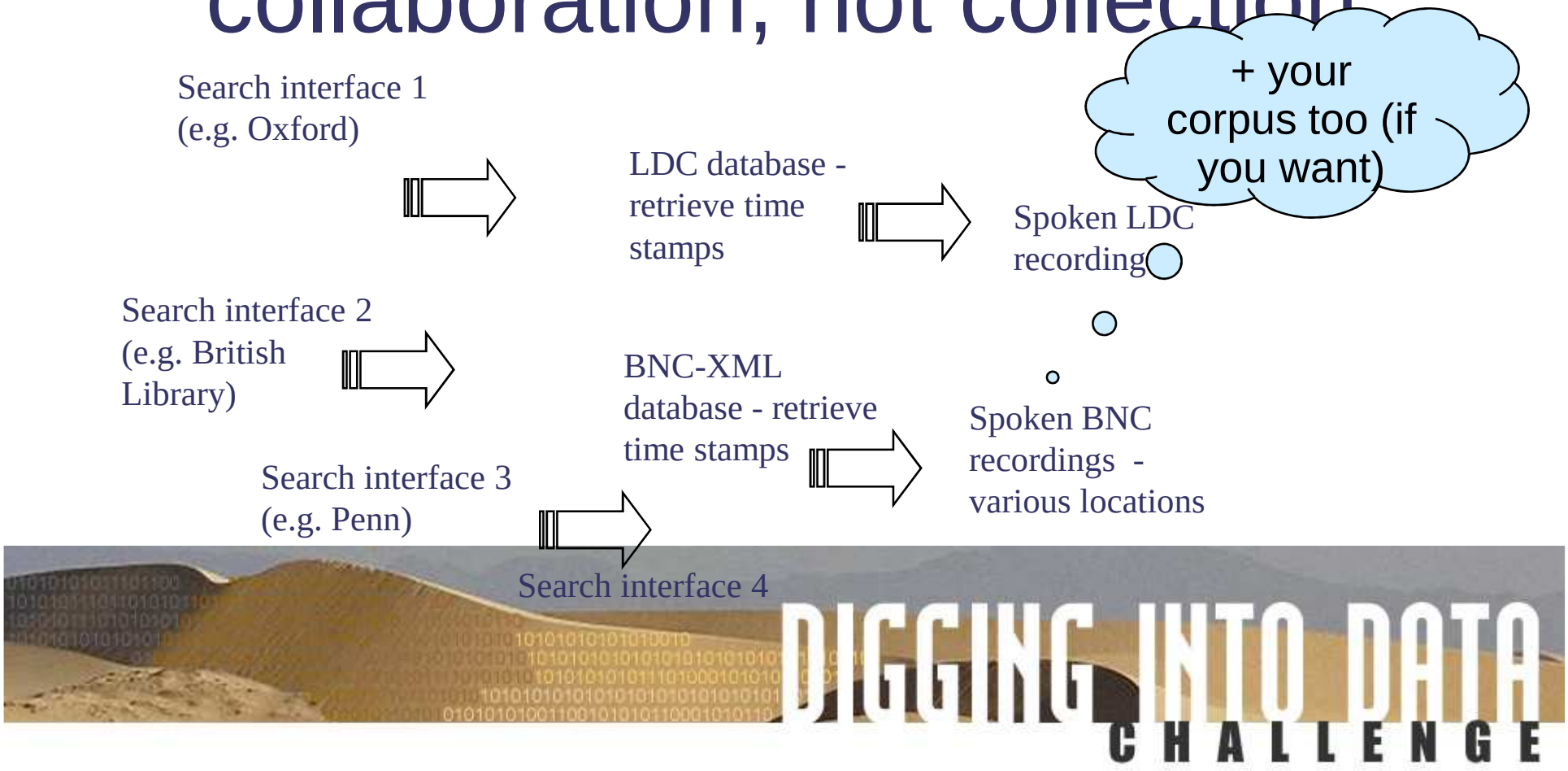
bnc.phon.ox.ac.uk/data/021A-C0897X0093XX-ABZZP0.wav?t=1870.8,1871.55



UNIVERSITY OF OXFORD

Problem 3: Sharing stuff

Cloud/crowd corpora: collaboration, not collection



Need to agree, and to follow, some data *standards*

Open access: passwords kill federated search

Corpus	File format	Transcription convention
SBCSAE (Am English)	SBCSAE text format	DT1
BNC Spoken + Audio (UK English)	BNC XML (TEI 1) + Praat TextGrids	BNC Guidelines
IViE (UK English)	Xlabel files	IViE guidelines (modified ToBI)
CallFriend (AmEng)	CHAT text format	CA-CHAT
METU Spoken Turkish	EXMARaLDA (XML)	HIAT
CGN (Dutch)	Praat TextGrids	CGN conventions
FOLK (German)	FOLKER (XML)	cGAT
CLAPI (French)	CLAPI XML (TEI 2)	ICOR
Swedish Spoken Language Corpus	Göteborg text format	GTS

UNIVERSITY OF OXFORD

(after Schmidt 2011, *JTEI*)

Towards TEI-XML standards for sound

```
<u who="D94PSUNK">
 <s n="3">
  <w
   ana="#D94:0083:11"
   c5="VVD"
   hw="want"
   pos="VERB">Wanted </w>
  <w
   ana="#D94:0083:12"
   c5="PNP"
   hw="i"
   pos="PRON">me </w>
  <w
   ana="#D94:0083:13"
   c5="TO0"
   hw="to"
   pos="PREP">to</w>
  <c c5="PUN">.</c>
```

UNIVERSITY OF OXFORD

Towards TEI-XML standards for sound

```
<u who="D94PSUNK">
 <s n="3">
  <w
   ana="#D94:0083:11"
   c5="VVD"
   hw="want"
   pos="VERB">Wanted </w>
  <w
   ana="#D94:0083:12"
   c5="PNP"
   hw="i"
   pos="PRON">me </w>
  <w
   ana="#D94:0083:13"
   c5="TO0"
   hw="to"
   pos="PREP">to</w>
  <c c5="PUN">.</c>
```

```
<fs xml:id="D94:0083:11">
 <f name="orth">wanted</f>
 <f name="phon_ana">
  <vcoll type="lst">
   <symbol synch="#D94:0083:11:0" value="W"/>
   <symbol synch="#D94:0083:11:1" value="AO1"/>
   <symbol synch="#D94:0083:11:2" value="N"/>
   <symbol synch="#D94:0083:11:3" value="AH0"/>
   <symbol synch="#D94:0083:11:4" value="D"/>
  </vcoll>
 </f>
</fs>
```

UNIVERSITY OF OXFORD

Towards TEI-XML standards for sound

```
<u who="D94PSUNK">
 <s n="3">
  <w
   ana="#D94:0083:11"
   c5="VVD"
   hw="want"
   pos="VERB">Wanted </w>
  <w
   ana="#D94:0083:12"
   c5="PNP"
   hw="i"
   pos="PRON">me </w>
  <w
   ana="#
   c5="TO
   hw="to
   pos="P
  <c c5="P
```

```
<fs xml:id="D94:0083:11">
 <f name="orth">wanted</f>
 <f name="phon_ana">
  <vcoll type="lst">
   <symbol synch="#D94:0083:11:0" value="W"/>
   <symbol synch="#D94:0083:11:1" value="AO1"/>
   <symbol synch="#D94:0083:11:2" value="N"/>
   <symbol synch="#D94:0083:11:3" value="AH0"/>
   <symbol synch="#D94:0083:11:4" value="D"/>
```

```
<timeline origin="0" unit="s" xml:id="TL0">
    ...
    <when xml:id="#D94:0083:11:0" from="1.6925" to="1.8225"/>
    <when xml:id="#D94:0083:11:1" from="1.8225" to="1.9225"/>
    <when xml:id="#D94:0083:11:2" from="1.9225" to="2.1125"/>
    <when xml:id="#D94:0083:11:3" from="2.1125" to="2.1825"/>
    <when xml:id="#D94:0083:11:4" from="2.1825" to="2.3125"/>
    ...
</timeline>
```

Linked Data Principles (Berners-Lee 2006)

1. All resources should be identified using URI's

2. All URI's should be dereferenceable, that is HTTP URI's, as it allows looking up the resources identified

3. When looking up a URI, it leads to more (useful) data about that resource

4. Links to other URI's should be included in order to enable the discovery of more data

UNIVERSITY OF OXFORD

Linked Data Principles (Berners-Lee 2006)

1. All resources should be identified using URI's

   http://bnc.phon.ox.ac.uk/data/021A-C0897X0093XX-ABZZP0.wav?t=1870.8,1871.55

2. All URI's should be dereferenceable, that is HTTP URI's, as it allows looking up the resources identified

   Yup! (requires server-side capability, but this is not difficult)

3. When looking up a URI, it leads to more (useful) data about that resource

   Hmm. Audio clip references ↔ metadata, e.g. labels, place in transcript ?

4. Links to other URI's should be included in order to enable the discovery of more data

   Links to similarly-labelled items in other corpora would be useful

UNIVERSITY OF OXFORD

Cloud/crowd corpus consortia

Old model	*New approach*
Distributed user base	Distributed user base
Centralized catalogue	Central catalogues
Centralized data	*Data is distributed*
Subscribers pay	Providers pay (like open-access journals), to be in the catalogue ?

UNIVERSITY OF OXFORD

Cloud/crowd corpus consortia

Old model	*New approach*
Distributed user base	Distributed user base
Centralized catalogue	Central catalogues
Centralized data	*Data is distributed*
Subscribers pay	Providers pay (like open-access journals), to be in the catalogue ?

Both are OK!

Important role for data centres

UNIVERSITY OF OXFORD

Team & collaborators

Ladan Baghai-Ravary, Ros Temple,
Margaret Renwick, John Pybus

previously, Greg Kochanski and Sergio Grau
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