Using forced alignment and HTML5 media syntax to share speech archive data

John Coleman

Phonetics Laboratory, Oxford

UNIVERSITY OF OXFORD

Outline

- Approaches to corpus dissemination
- The Audio British National Corpus
- Problem 1: Finding stuff
- Problem 2: Getting stuff
- Problem 3: Sharing stuff

Normal approach to corpus publication

- An institution or project collects and prepares a corpus.

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.

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.
- You log on and download the corpus. Fees and passwords may be required.

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.
- You log on and download the corpus. Fees and passwords may be required.
- Maybe, the corpus contains (some of) what you're looking for.

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.
- You log on and download the corpus. Fees and passwords may be required.
- Maybe, the corpus contains (some of) what you're looking for.

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.

Time and effort; other people's rules

- You log on and download the corpus. Fees and passwords may be required.
- Maybe, the corpus contains (some of) what you're looking for.

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.
- You log on and download the corpus. Fees and passwords may be required.
- Maybe, the corpus contains (some of) what you're looking for.

Time and effort;
other people's rules
The whole thing?

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.
- You log on and download the corpus. Fees and passwords may be required.
- Maybe, the corpus contains (some of) you're looking for.

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.
- You log on and download the corpus. Fees and passwords may be required.
- Maybe, the corpus contains (some of) what you're looking for.

Time and effort;
other people's rules
The whole thing?
What a hassle!
Or not! What is where?

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

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

UNIVERSITY OF OXFORD

Spoken texts

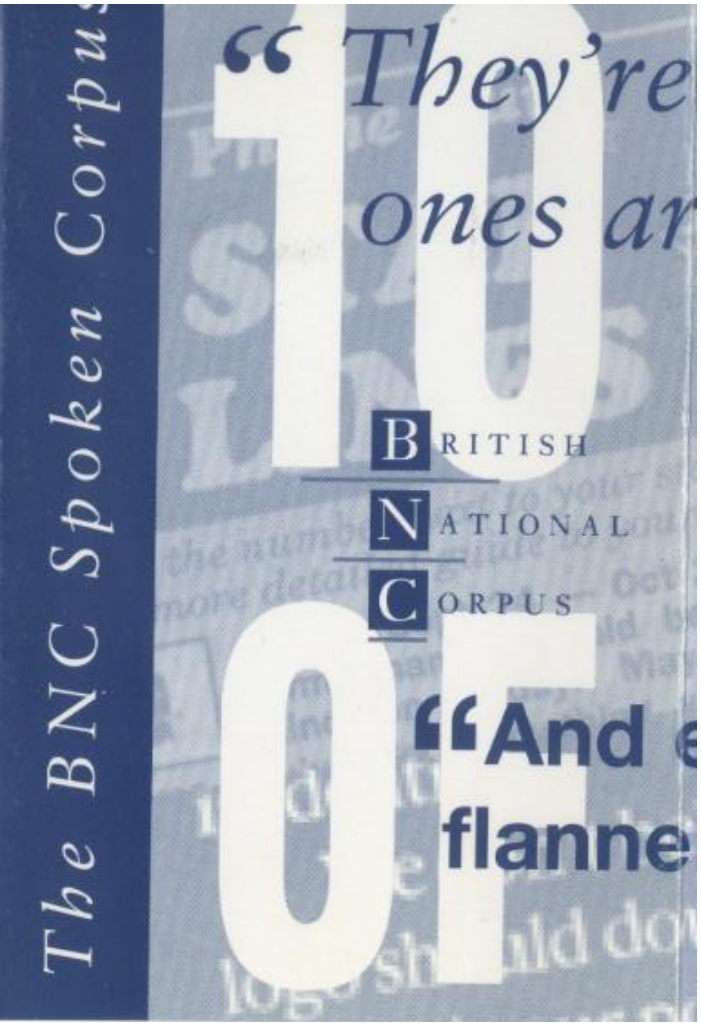
Demographic part: 4.2 million words

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, oral history
- *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

UNIVERSITY OF OXFORD

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, 7.5 million words)
- We anonymized the audio in accordance with the original transcription protocols

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?

What makes oral history and dialect corpora interesting to linguists?

- Unique and interesting words and expressions
- Regular differences, e.g. specifics of pronunciation

UNIVERSITY OF OXFORD

What makes oral history and dialect corpora interesting to linguists?

- Unique and interesting words and expressions — *needle in a haystack*
- Regular differences, e.g. specifics of pronunciation — many *needles in haystacks*

UNIVERSITY OF OXFORD

Searching *text* is easy ...

BNC Text {$docid} | BNCweb Query result

bnc.phon.ox.ac.uk/transcripts-html/GYS.

BNC Text GYS

Oral history project: interview. Sample containing about 4621 words speech recorded in leisure context

---

3 speakers recorded by respondent number C261

PS29U	Ag5	m	(William, age 72, farmer) unspecified
PS29V	X	m	(No name, age unknown) unspecified
GYSPS000	X	u	(No name, age unknown) unspecified

---

1 recordings

1. Tape 097801 recorded on unknown date. LocationStrathclyde: Kilmarnock () Activity: interview

Undivided text

(PS29V)	[1] Can we start off with your name?
William (PS29U)	
(PS29V)	[2] It's William isn't it?
William (PS29U)	[3] William aye.
(PS29V)	[4] And you're a retired farmer?

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

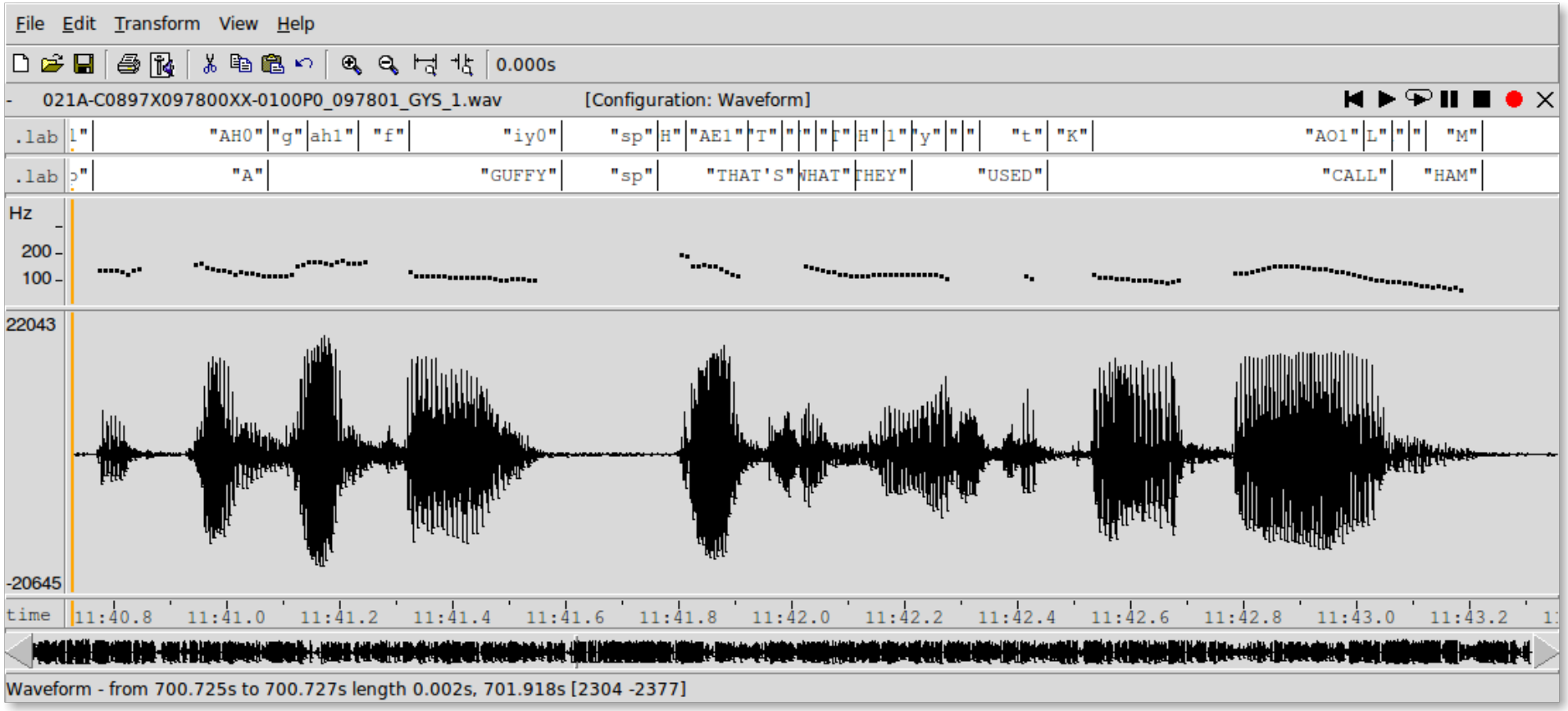
UNIVERSITY OF OXFORD

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



x 21 million

UNIVERSITY OF OXFORD

Forced alignment is *not* perfect

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

- In a pilot, 23% was accurately aligned within 20 ms
- In a phonetic study, 60% of 549 word-ends were well-aligned within 50 ms and 80% within 100 ms

UNIVERSITY OF OXFORD

AudioBNC publication

We released most of the aligned Audio BNC online:

- http://www.phon.ox.ac.uk/AudioBNC (webpage) and http://bnc.phon.ox.ac.uk (data)
- Includes .wav audio, Praat TextGrid alignments, HTML transcriptions, indices of word and sound time-stamps

UNIVERSITY OF OXFORD

Problem 2: Getting stuff

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

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

UNIVERSITY OF OXFORD

Browsing and searching

```
"GADGET" 2139.9725 2140.3925 021A-C0897X0143XX-AAZZP0_014307_KC9_28.result
"GADGET" 3057.3425 3057.7525 021A-C0897X103401XX-0100P0-2nd-0200P0_103401_HEM_1.result
"GADGET" 3065.6125 3066.0025 021A-C0897X103401XX-0100P0-2nd-0200P0_103401_HEM_1.result
"GADGET" 819.1025 819.2925 021A-C0897X0424XX-AAZZP0_042401_KST_9.result
"GADGET" 874.8925 875.2825 021A-C0897X0145XX-AAZZP0_014502_KC9_36.result
"GADGETS" 1025.2125 1025.6725 021A-C0897X0492XX-AAZZP0_049202_KBB_10.result
"GADGETS" 1051.5525 1052.0125 021A-C0897X0458XX-ABZZP0_045807_KDN_47.result
"GADGETS" 1175.2125 1175.7525 021A-C0897X104101XX-0100P0-2nd-0200P0_104101_HEV_1.result
"GADGETS" 1283.2925 1283.8025 021A-C0897X0141XX-ABZZP0_014104_KC9_15.result
"GADGETS" 1657.4325 1657.8825 021A-C0897X0145XX-ABZZP0_014506_KC9_39.result
"GADGETS" 814.7125 815.2325 021A-C0897X0424XX-AAZZP0_042401_KST_9.result
"GADGETS" 815.8925 816.2025 021A-C0897X0424XX-AAZZP0_042401_KST_9.result
"GADGY" 667.2125 667.7425 021A-C0897X097800XX-0100P0_097801_GYS_1.result
"GADGY" 838.7325 839.1525 021A-C0897X097800XX-0100P0_097801_GYS_1.result
"GADGY" 844.2325 844.6525 021A-C0897X097800XX-0100P0_097801_GYS_1.result
"GADGY" 850.5025 850.8125 021A-C0897X097800XX-0100P0_097801_GYS_1.result
```

+7,931,695 more lines

UNIVERSITY OF OXFORD

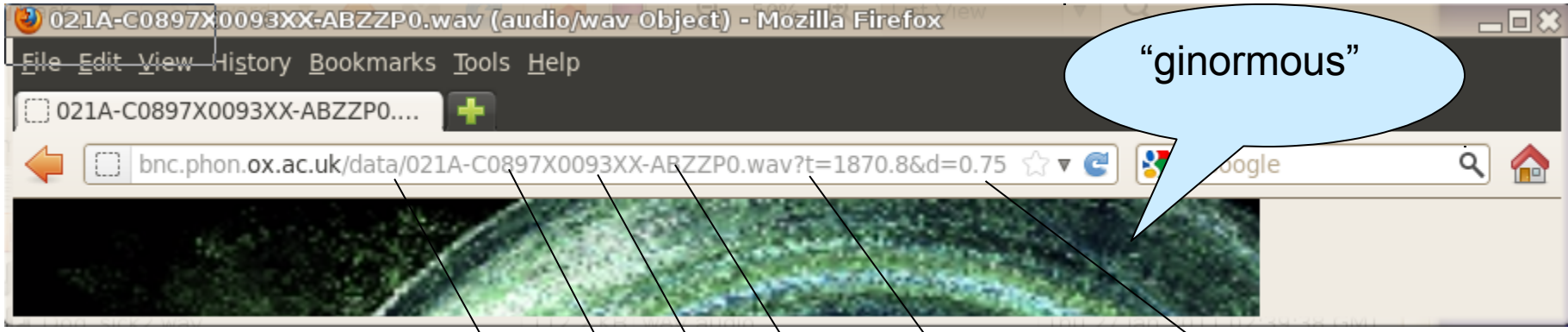
Browsing and searching

```
"GADGET" http://bnc.phon.ox.ac.uk/data/021A-C0897X0143XX-AAZZP0.wav?t=2139.9725,2140.3925
"GADGET" http://bnc.phon.ox.ac.uk/data/021A-C0897X103401XX-0100P0.wav?t=3057.3425,3057.7525
"GADGET" http://bnc.phon.ox.ac.uk/data/021A-C0897X103401XX-0100P0.wav?t=3065.6125,3066.0025
"GADGET" http://bnc.phon.ox.ac.uk/data/021A-C0897X0424XX-AAZZP0.wav?t=819.1025,819.2925
"GADGET" http://bnc.phon.ox.ac.uk/data/021A-C0897X0145XX-AAZZP0.wav?t=874.8925,875.2825
"GADGETS" http://bnc.phon.ox.ac.uk/data/021A-C0897X0492XX-AAZZP0.wav?t=1025.2125,1025.6725
"GADGETS" http://bnc.phon.ox.ac.uk/data/021A-C0897X0458XX-ABZZP0.wav?t=1051.5525,1052.0125
"GADGETS" http://bnc.phon.ox.ac.uk/data/021A-C0897X104101XX-0100P0.wav?t=1175.2125,1175.7525
"GADGETS" http://bnc.phon.ox.ac.uk/data/021A-C0897X0141XX-ABZZP0.wav?t=1283.2925,1283.8025
"GADGETS" http://bnc.phon.ox.ac.uk/data/021A-C0897X0145XX-ABZZP0.wav?t=1657.4325,1657.8825
"GADGETS" http://bnc.phon.ox.ac.uk/data/021A-C0897X0424XX-AAZZP0.wav?t=814.7125,815.2325
"GADGETS" http://bnc.phon.ox.ac.uk/data/021A-C0897X0424XX-AAZZP0.wav?t=815.8925,816.2025
"GADGY" http://bnc.phon.ox.ac.uk/data/021A-C0897X097800XX-0100P0.wav?t=667.2125,667.7425
"GADGY" http://bnc.phon.ox.ac.uk/data/021A-C0897X097800XX-0100P0.wav?t=838.7325,839.1525
"GADGY" http://bnc.phon.ox.ac.uk/data/021A-C0897X097800XX-0100P0.wav?t=844.2325,844.6525
"GADGY" http://bnc.phon.ox.ac.uk/data/021A-C0897X097800XX-0100P0.wav?t=850.5025,850.8125
```

+7,931,695 more lines

UNIVERSITY OF OXFORD

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

Search for media fragments

BNC Text {$docid} | BNCweb (CQP-Edition)

https://bncweb.lancs.ac.uk/cgi-binbncXML/dlogs.pl?selected

Main menu

Query options
- Standard query
- Written restrictions
- Spoken restrictions

User-specific functions
- User settings
- Query history
- Saved queries
- Categorized queries
- Make/edit subcorpora
- Upload external data file

Additional functions
- Browse a text
- Scan keywords/titles
- Explore genre labels
- Frequency lists
- Keywords

About BNCweb
- BNCweb book
- The BNCweb team
- New features
- Bug reports
- The CLAWS-5 tagset
- Oxford BNC homepage

Version 4.3, November 2013

BNC*web* (CQP-Edition)

Standard Query

chappens

Query mode: Simple query (ignore case) — Simple Query Syntax help

Number of hits per page: 50

Restriction: None (search whole corpus)

Start Query | Reset Query

News

24.3.15: Apologies for the interruption in (reliable) service today - things should be back to normal now. Sebastian Hoffmann

All users of BNCweb now have full access to the corpus, i.e. they can see the larger context of a query hit and browse a text.

The audio data for a sizeable proportion of the spoken component of the BNC has recently been made public on the Internet - see Audio BNC: the audio edition of the Spoken British National Corpus. Access to these recordings is now available straight from a query result in BNCweb. Users who have access to

Search for media fragments

BNC Text {$docid} | BNCweb Query result | BNC Audio Data | BNC Audio Data

https://bncweb.lancs.ac.uk/cgi-binbncXML/processQuery.pl?thel

Search

Your query "chappens" returned 2 hits in 1 text (98,313,429 words [4,048 texts]; frequency: 0.02 instances per million words) [0.128 seconds]

|< << >> >| Show Page: 1 Show KWIC View Show in random order New Query Go!

No	Filename	Hits 1 to 2 Page 1 / 1
1	GYS 51	And they used to make what they called **chappens** That was, that was a, a tin jug.
2	GYS 54	And they called them **chappens**

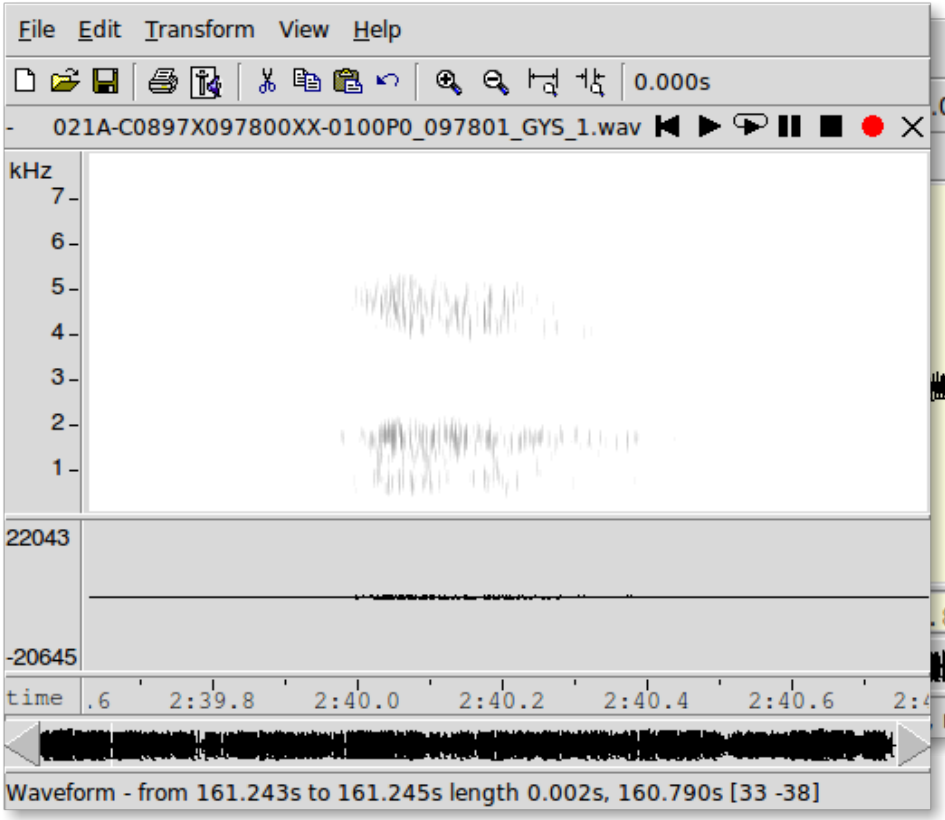
BNCweb (CQP-edition) © 1996-2013

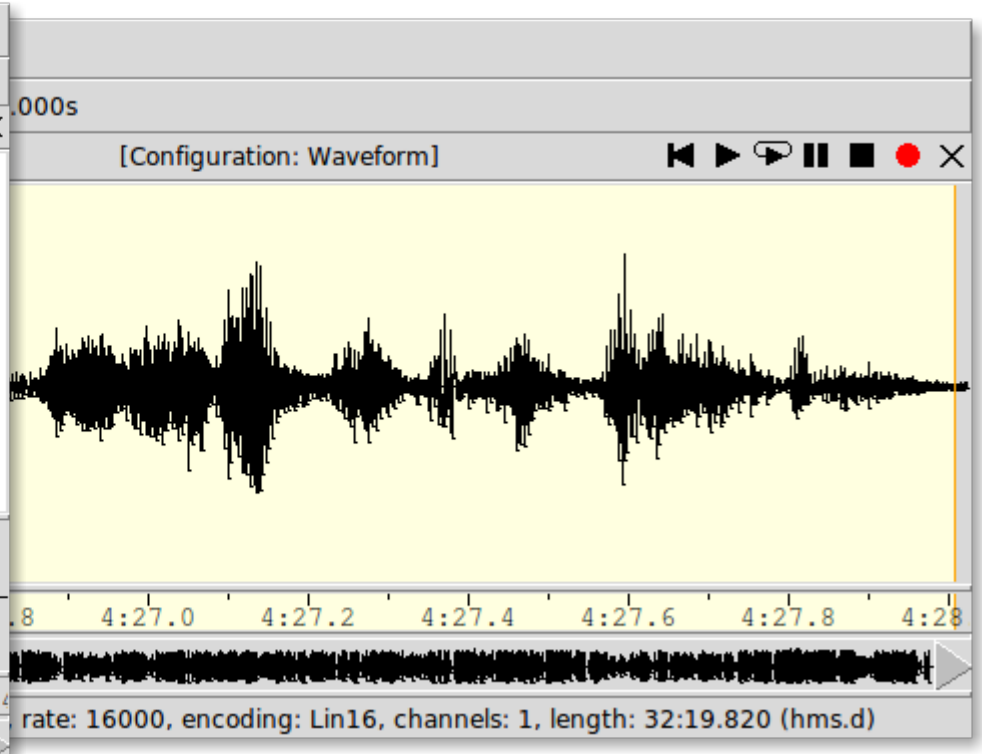
You are logged in as user "jcoleman"

Unsearchable media fragments

It's important to be able to access parts of the audio that *aren't* indexed, e.g.

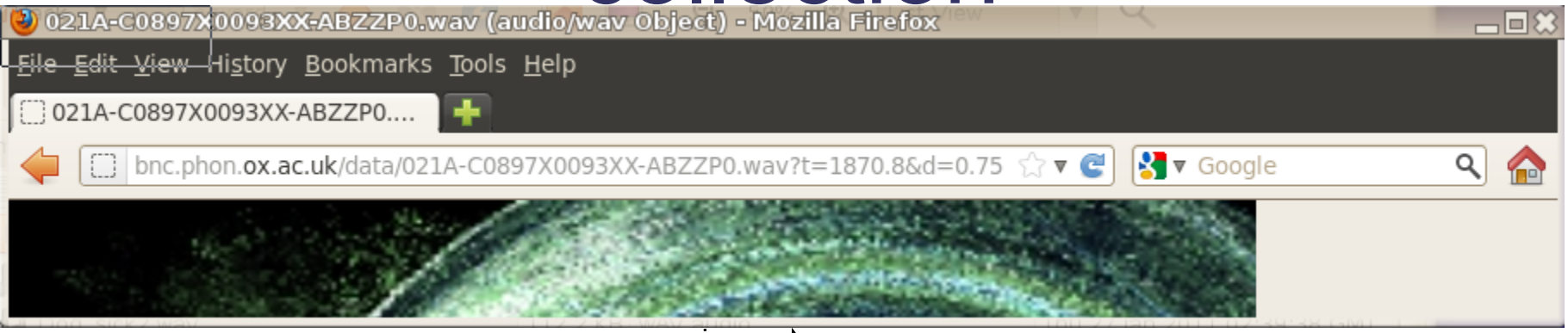
- a sigh
- untranscribed material



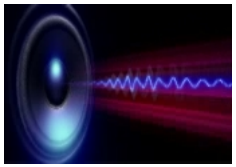


Problem 3: Sharing stuff

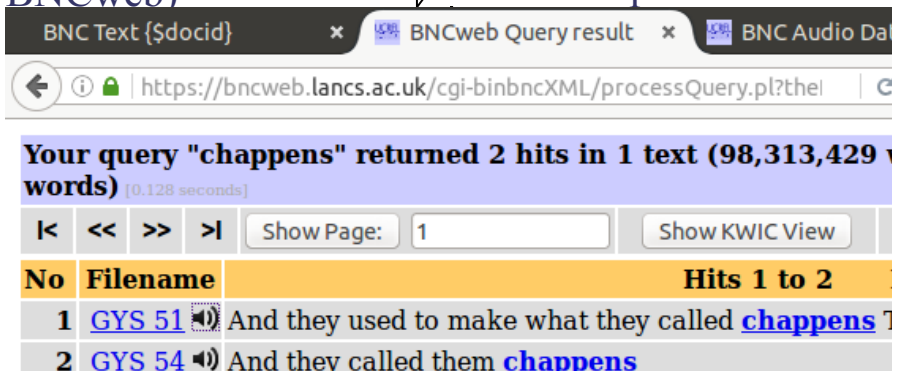
Cloud corpora: federation not collection



User interface 1 (e.g. Oxford) → retrieve time stamps → AudioBNC recordings



User interface 2 (e.g. Lancaster BNCweb) → database - retrieve time stamps → AudioBNC recordings



UNIVERSITY OF OXFORD

Cloud corpora: federation not collection

Need to agree, and to follow, some data standards

Open access: passwords kill federated search

UNIVERSITY OF OXFORD

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

Proposal by Saul Albert for extending BNC markup for conversation analysis

```
<setting xml:id="KB0SE00D" n="029103" who="PS000 PS001 PS006 PS007">
          <audioFile>021A-C0897X0291XX-ABZZP0.wav</audioFile>
          <placeName>Clwyd:  Holywell </placeName>
          <locale> synod meeting </locale>
          <activity spont="H"> end of meeting </activity>
        </setting>

…

<u who="PS006">
        <s n="6" nbw="18" startTime="6.7525" endTime="8.3325">
          <w c5="VVD" hw="see" pos="VERB" startTime="6.7525" endTime="7.0025">saw </w>
          <w c5="NP0" hw="mary" pos="SUBST" startTime="7.1625" endTime="7.5925">Mary </w>
          <w c5="CJC" hw="and" pos="CONJ" startTime="7.5925" endTime="7.6625">and </w>
          <w c5="NP0" hw="andrew" pos="SUBST" startTime="7.7425" endTime="8.2725">Andrew </w>
          <w c5="CJC" hw="and" pos="CONJ" startTime="8.2725" endTime="8.3325">and</w>
        </s>
      </u>
```

UNIVERSITY OF OXFORD

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)

= words and sounds

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 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), for the catalogue ?

UNIVERSITY OF OXFORD

Cloud 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), for the catalogue ?

Important role for data centres

UNIVERSITY OF OXFORD