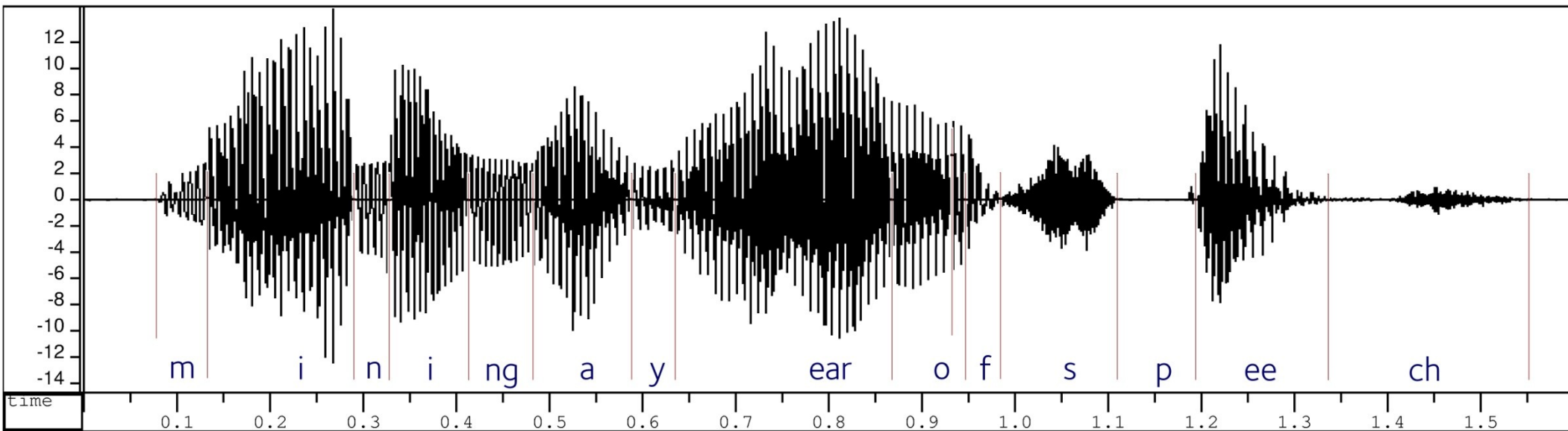


# Mining the Spoken BNC Audio

John Coleman, Oxford University Phonetics Lab



<http://www.phon.ox.ac.uk/mining/>

with support from our “Digging  
into Data” competition funders

JISC



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

# 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?

# 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

# 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

# Some large(ish) speech corpora

- 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 (OSU): c. 2 days.
- Wellington Corpus of Spoken New Zealand English, c. 3 days transcribed
- Digital Archive of Southern Speech (American)

# 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*

75% on 1/4" tape  
20% shellac and vinyl  
7% digital

# Analogue audio in libraries

World wide: ~100m hours (11,415 yrs) analogue  
*i.e. 4-5 Large Hadron Colliders!*

Cost of professional digitization: ~£20/\$32 per  
tape (e.g. C-90 cassette)

Using speech recognition and natural language  
technologies (e.g. summarization) could provide  
more detailed cataloguing/indexing without  
time-consuming human listening

# A rule of thumb

To catch most

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

# Main problem in large corpora

Finding needles in the haystack

To address that challenge, we think there are two “killer apps”

- Forced alignment
- Data linking: open exposure of digital material, coupled with cross-searching

# Collaboration, not collection

Search interface 1  
(e.g. Oxford)

LDC database -  
retrieve time  
stamps

Spoken LDC  
recordings -  
various locations

Search interface 2  
(e.g. BL)

BNC-XML  
database - retrieve  
time stamps

Spoken BNC  
recordings - BL  
sound server(s)

Search interface 3  
(e.g. Penn)

Search interface 4  
(e.g. Lancaster ?)

# Corpora in the Year of Speech

Spontaneous speech

Spoken BNC ~1400 hrs

Conversational telephone speech

Read text

LibriVox audio books

Broadcast news

US Supreme Court oral arguments

Political discourse

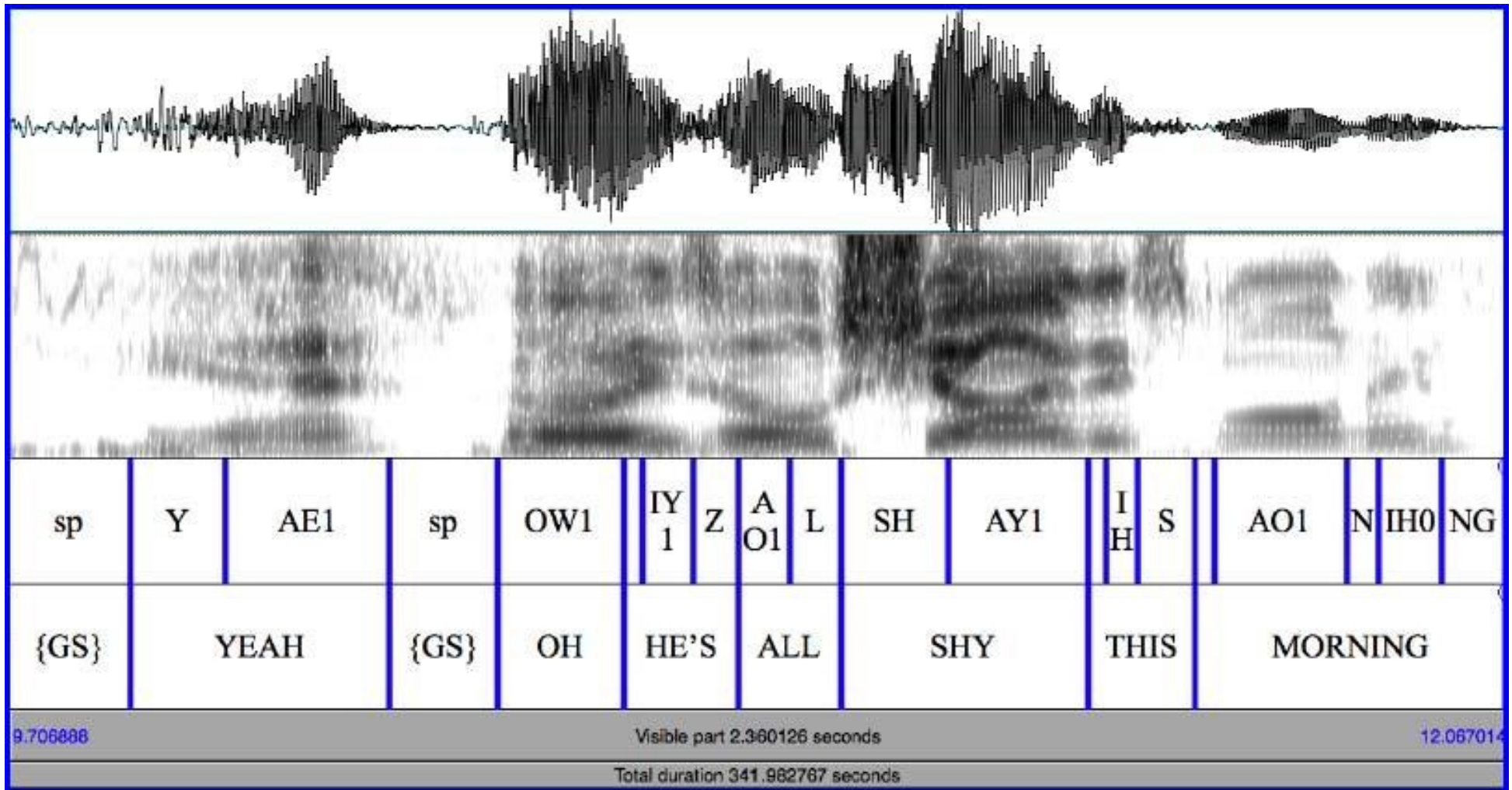
Oral history interviews

US vernacular dialects/Sociolinguistic interviews

# Practicalities

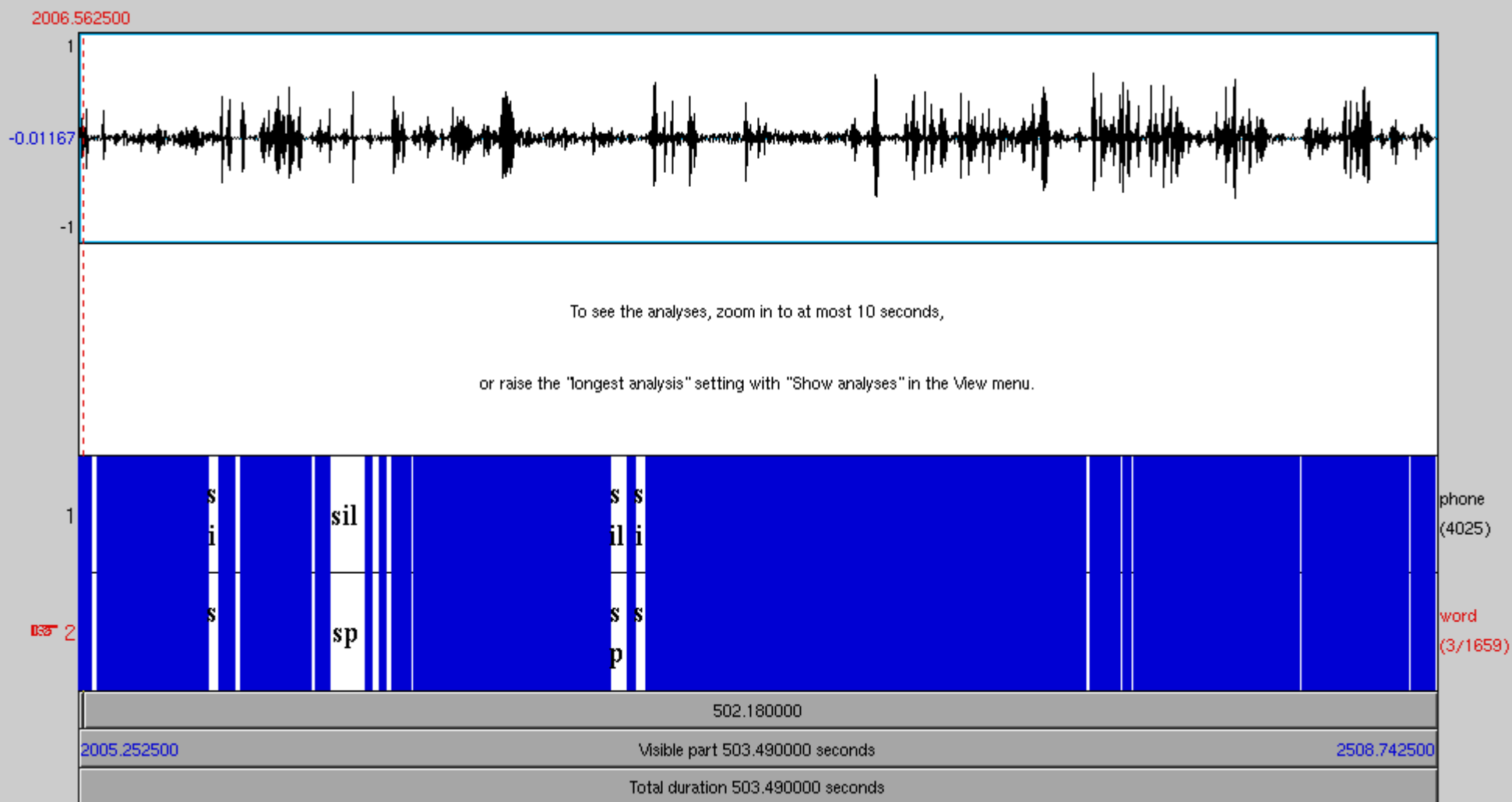
- In order to be of much practical use, such very large corpora must be indexed at word and segment level
- All included speech corpora must therefore have associated text transcriptions
- We're using the Penn Phonetics Laboratory Forced Aligner to associate each word and segment with the corresponding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

# Mining (indexing by forced alignment)

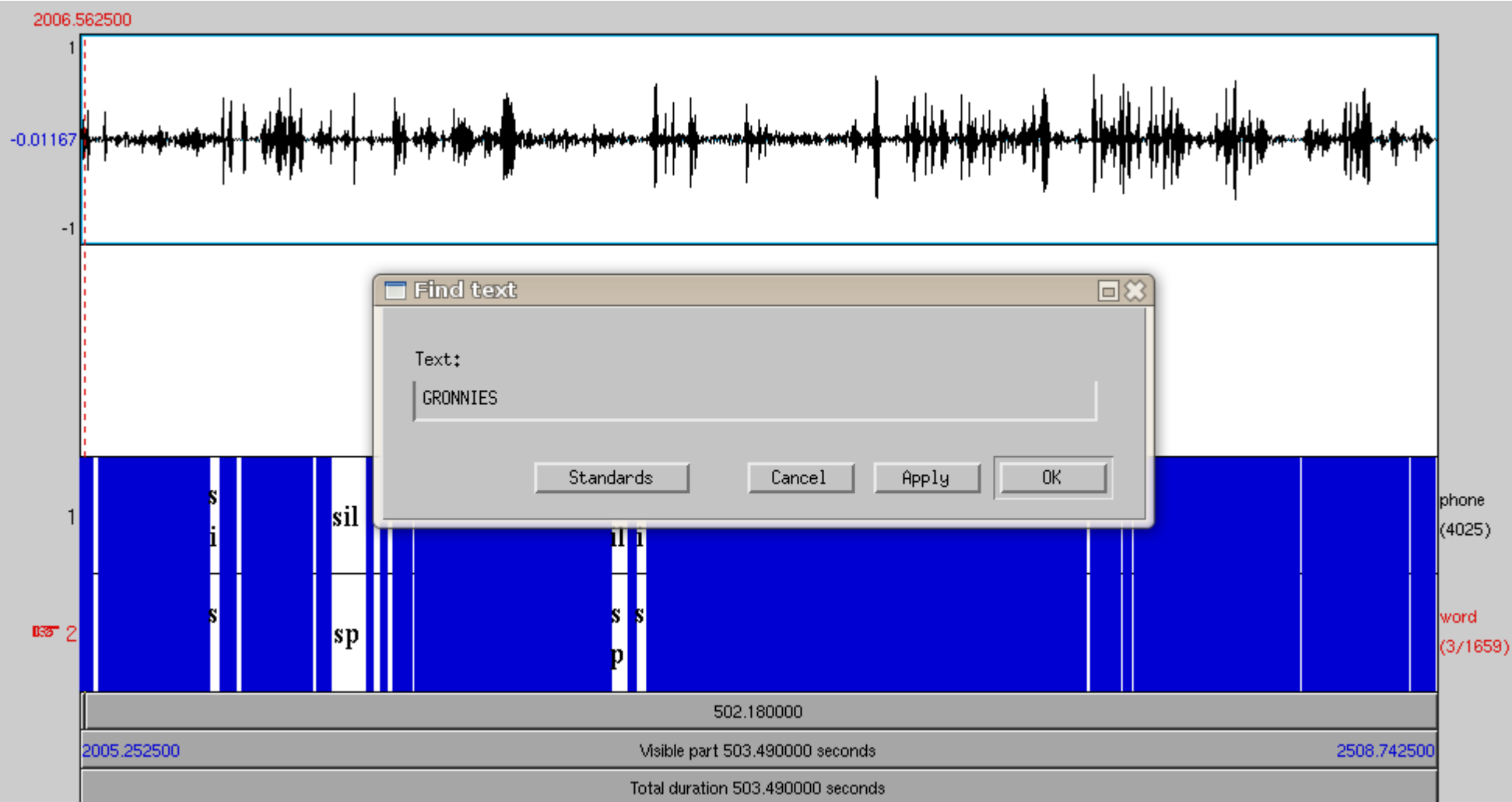


*x 21 million*

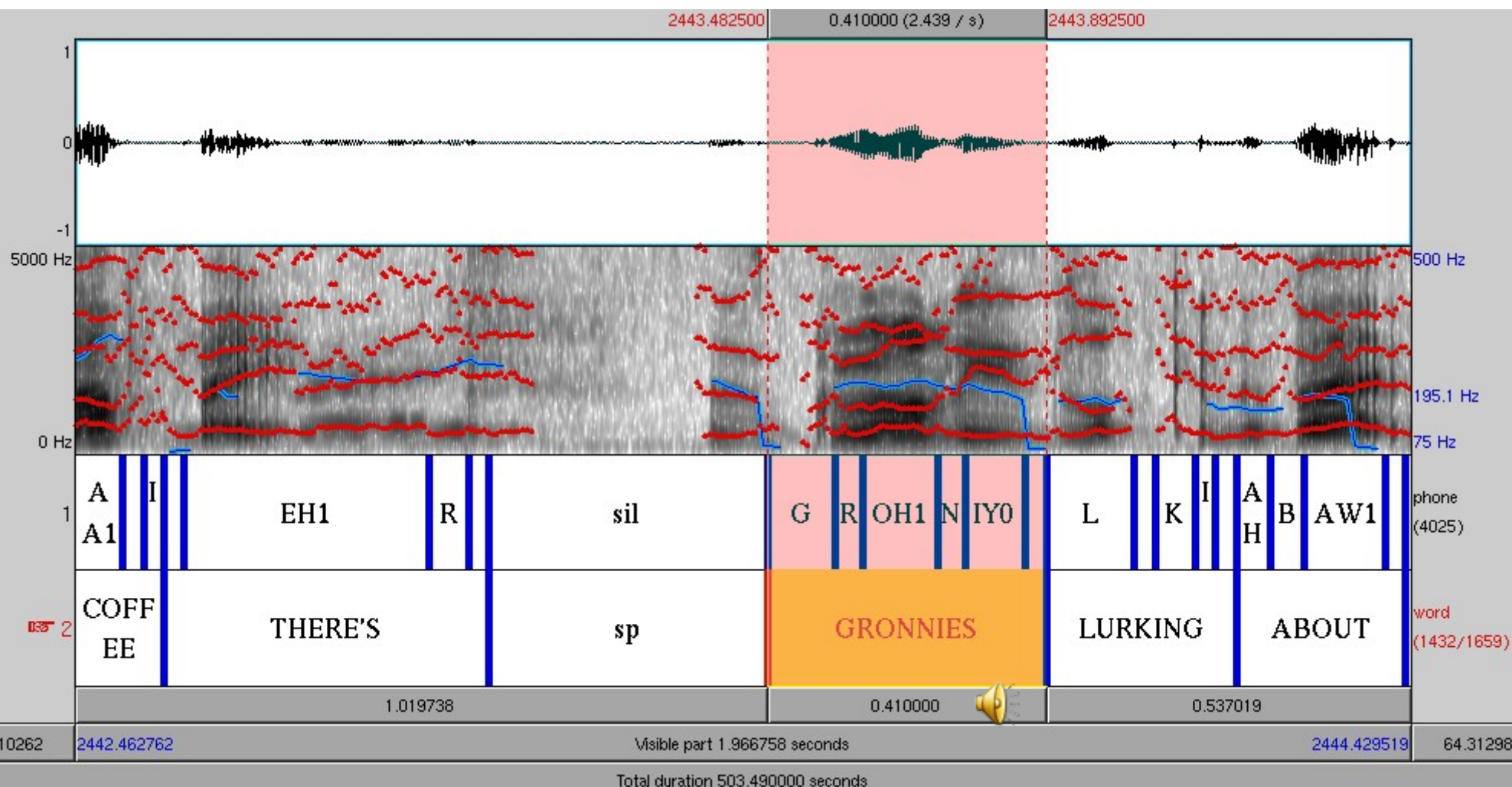
# Mining (indexing by forced alignment)



# Mining (a needle in a haystack)



# Mining (a diamond in the rough)

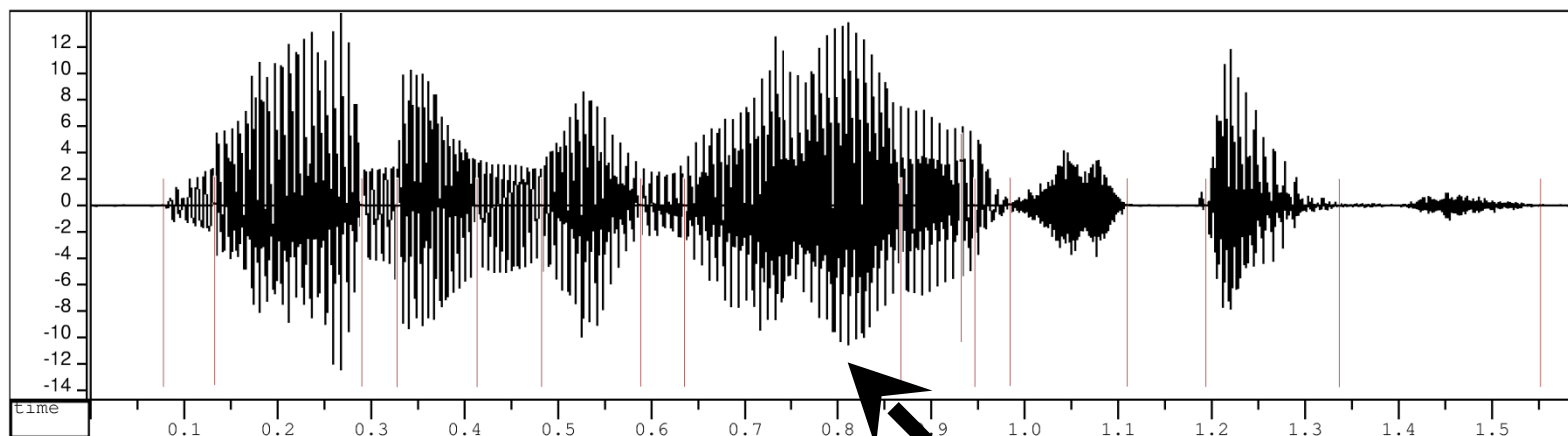


# Issues we're still grappling with

- No standards for adding phonemic transcriptions and timing information to XML transcriptions
- Many different possible schemes
- How to decide?

# Enabling other corpora to be brought in in future

Promoting common standards for audio with linked transcription



?

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

<w c5="AV0" hw="well" pos="ADV" synch="W1">Well  
</w>

```
<w c5="AV0" hw="well" pos="ADV" synch="W1">Well  
</w>
```

```
<fs type="word">  
  <f name="transcription">  
    <string>Well </string>  
  </f>  
  <f name="phonemes">  
    <fs type="phoneme">  
      <f name="W" fVal="#W" synch="P1.1"/>  
      <f name="EH1" fVal="#EH1" synch="P1.2"/>  
      <f name="L" fVal="#L" synch="P1.3"/>  
    </fs>  
  </f>  
</fs>
```

```
<timeline origin="0" unit="s" xml:id="TL0">
  <when xml:id="W1"    from="1.6925" to="2.1125"/>
  <when xml:id="W2"    from="2.1125" to="2.3125"/>
  <when xml:id="P1.1"  from="1.6925" to="1.8225"/>
  <when xml:id="P1.2"  from="1.8225" to="1.9225"/>
  <when xml:id="P1.3"  from="1.9225" to="2.1125"/>
  <when xml:id="P2.1"  from="2.1125" to="2.1825"/>
  <when xml:id="P2.2"  from="2.1825" to="2.3125"/>
  ...
</timeline>
```

Thank you very much!