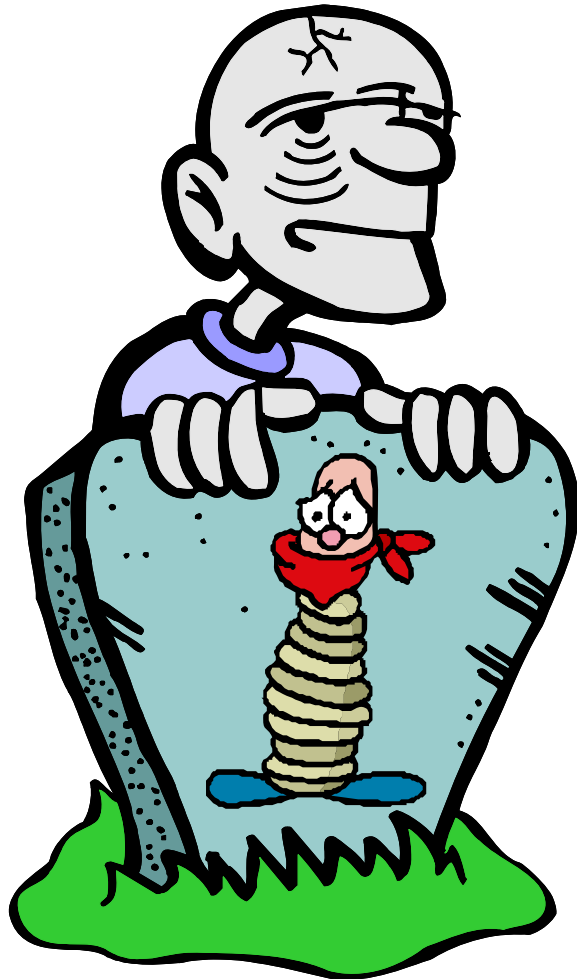


IVACS 2006

Letting the Corpus Speak



23/06/2006



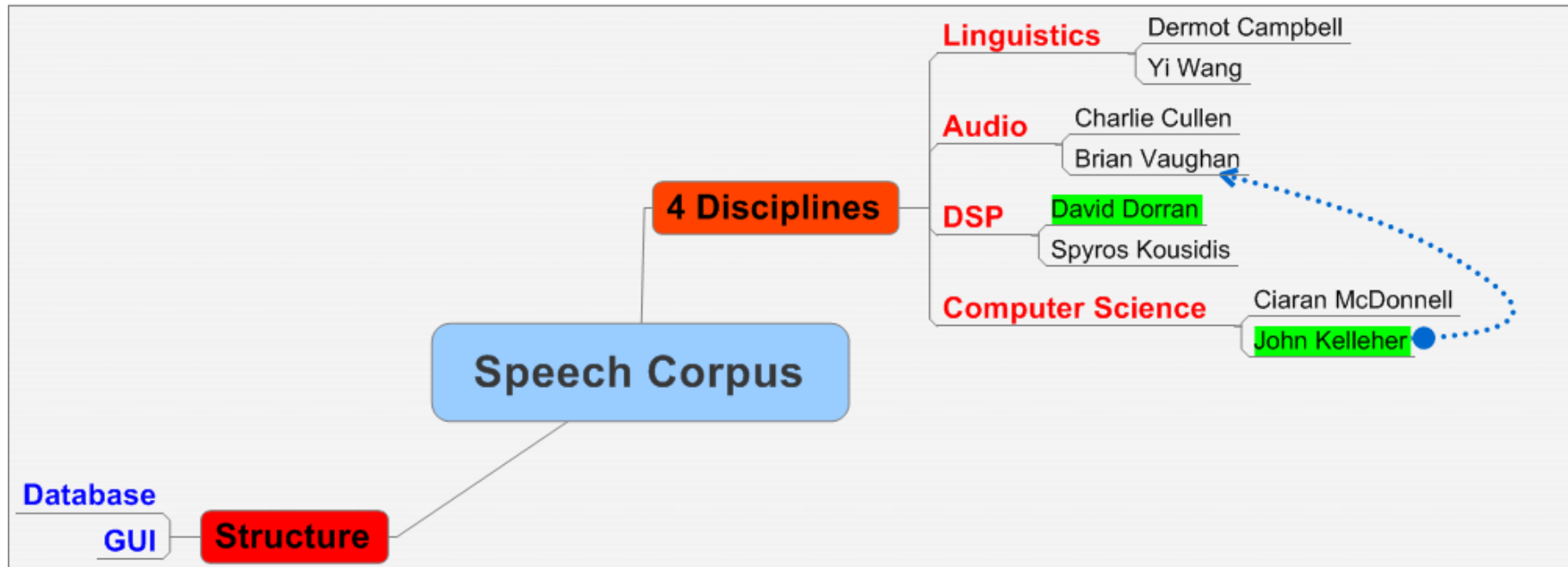
Structure of Presentation

- Route to the corpus + transcription (Dermot)
- Details of structure + Interface (Dermot/John)
- Problems of segmentation (Yi)
- What's missing ? (**YOU !**)

DITCall – the ‘Slow’ Way Forward

- **Slow-down** algorithm
 - slows speech
 - maintains natural pitch
- Importance of **speed** of delivery
- Need to work with **speech signal**
- Build speech corpus based on **dialogues**
- **Interfaces** for construction and use
- Search ► find ► sort ► access ► **play**

DIT *SALERO* Corpus Team



Why use a NS Speech Corpus?

Exposure to the target language and culture:

- **McCarthy (1998a):** ‘**Exposure to the spoken language**, without the pressure to imitate and produce it, could be one of the **most useful** elements [in language learning] ...’

Facilitates autonomous learning:

- **Stevens (1995):** ‘**Data Driven** Learning gives learners control over the learning process and exposes them to **authentic language**’

Exposure to ‘real’ language:

- **Carter (1998):** ‘A learner who has been **exposed to ‘real English’** has an advantage over a learner who has only been exposed to ‘disinfected textbook material’

A Corpus with **spoken** data can reveal:

- **features** of **connected speech**
- **difference in pronunciation** of a word depending on its occurrence in a sentence
- **prosodic** information
- preferred **pronunciation styles** of chosen socio-cultural model
- **kinetics** of speech **production** (imitation)

Features of Connected Speech

- **Elision:** omission of sounds
- **Assimilation:** replacing sounds to make them closer to neighbouring sounds
- **Weak Forms:** vowel quality modified; length reduced

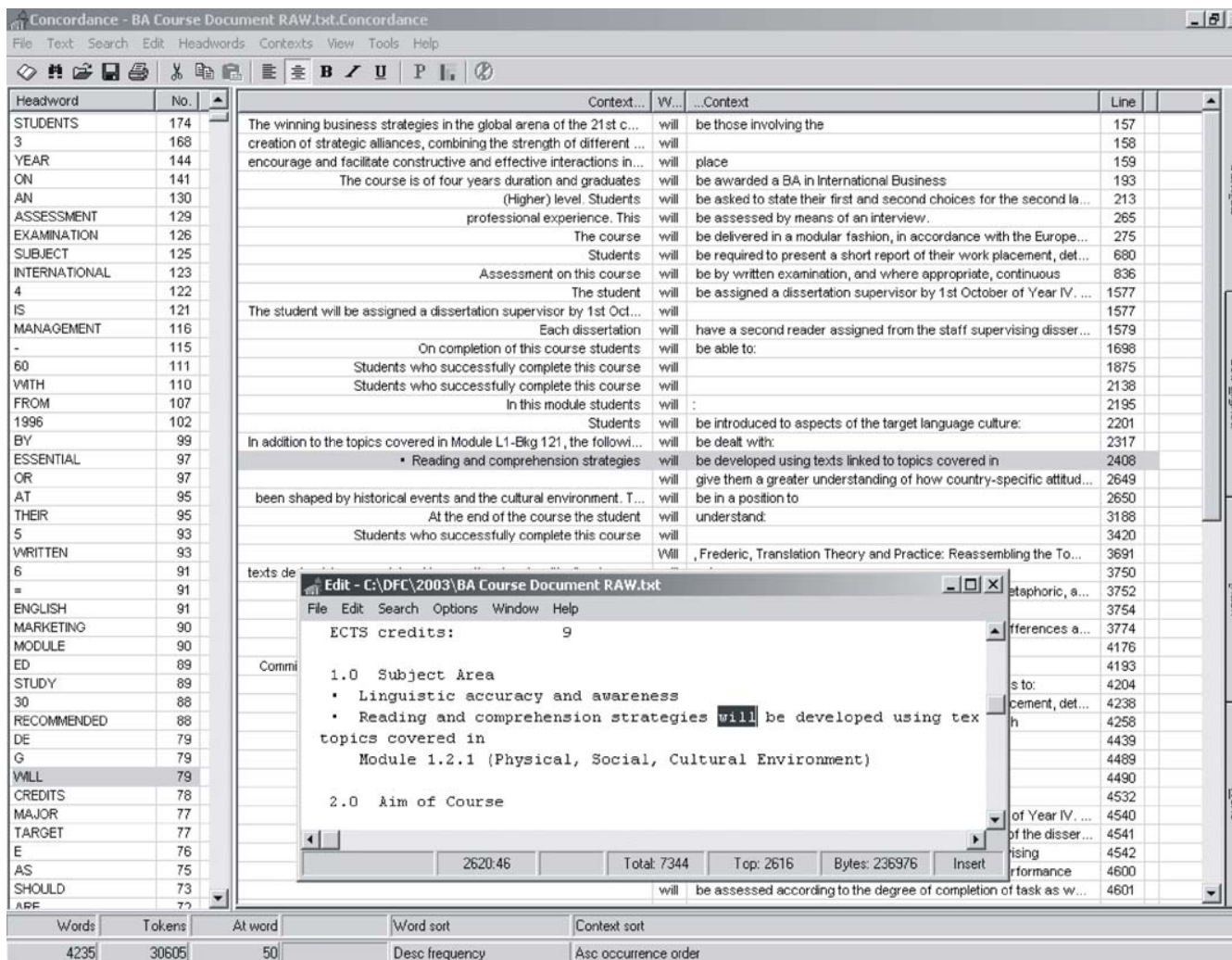
Why a **Speech** Corpus?

- Written corpus useful for e.g. collocations and discourse analysis
- Access to **actual speech** is necessary if we wish to study speech **production**.
- Once accessed, this can be slowed down for detailed study (**DITCall** Project)

Riding the WAV

- Need to study **original** sound files
- Access to WAV files by means of concordancing interface ...
- ... via '**citation transcription**'
- Allows comparison of:
 - sound file with transcript
 - sound files of different speeds

'Discourse' Concordancer Collocations



| Headword | No. | Context | VV... | ...Context | Line |
|---------------|-----|--|-------|---|------|
| STUDENTS | 174 | The winning business strategies in the global arena of the 21st c... | will | be those involving the | 157 |
| 3 | 168 | creation of strategic alliances, combining the strength of different ... | will | | 158 |
| YEAR | 144 | encourage and facilitate constructive and effective interactions in... | will | place | 159 |
| ON | 141 | The course is of four years duration and graduates | will | be awarded a BA in International Business | 193 |
| AN | 130 | (Higher) level. Students | will | be asked to state their first and second choices for the second la... | 213 |
| ASSESSMENT | 129 | professional experience. This | will | be assessed by means of an interview. | 265 |
| EXAMINATION | 126 | The course | will | be delivered in a modular fashion, in accordance with the Europe... | 275 |
| SUBJECT | 125 | Students | will | be required to present a short report of their work placement, det... | 680 |
| INTERNATIONAL | 123 | Assessment on this course | will | be by written examination, and where appropriate, continuous | 836 |
| 4 | 122 | The student | will | be assigned a dissertation supervisor by 1st October of Year IV. ... | 1577 |
| IS | 121 | The student will be assigned a dissertation supervisor by 1st Oct... | will | | 1577 |
| MANAGEMENT | 116 | Each dissertation | will | have a second reader assigned from the staff supervising disser... | 1579 |
| - | 115 | On completion of this course students | will | be able to: | 1698 |
| 60 | 111 | Students who successfully complete this course | will | | 1875 |
| WITH | 110 | Students who successfully complete this course | will | | 2138 |
| FROM | 107 | In this module students | will | : | 2195 |
| 1996 | 102 | Students | will | be introduced to aspects of the target language culture: | 2201 |
| BY | 99 | In addition to the topics covered in Module L1-Bkg 121, the followi... | will | be dealt with: | 2317 |
| ESSENTIAL | 97 | • Reading and comprehension strategies | will | be developed using texts linked to topics covered in | 2408 |
| OR | 97 | | will | give them a greater understanding of how country-specific attitud... | 2649 |
| AT | 95 | been shaped by historical events and the cultural environment. T... | will | be in a position to | 2650 |
| THEIR | 95 | At the end of the course the student | will | understand: | 3188 |
| 5 | 93 | Students who successfully complete this course | will | | 3420 |
| WRITTEN | 93 | | Will | , Frederic, Translation Theory and Practice: Reassembling the To... | 3691 |
| 6 | 91 | texts de | | | 3750 |
| = | 91 | | | | 3752 |
| ENGLISH | 91 | | | staphoric, a... | 3754 |
| MARKETING | 90 | | | ferences a... | 3774 |
| MODULE | 90 | | | | 4176 |
| ED | 89 | | | s to: | 4193 |
| STUDY | 89 | | | | 4204 |
| 30 | 88 | | | cement, det... | 4238 |
| RECOMMENDED | 88 | | | h | 4258 |
| DE | 79 | | | | 4439 |
| G | 79 | | | | 4489 |
| WILL | 79 | | | | 4490 |
| CREDITS | 78 | | | | 4532 |
| MAJOR | 77 | | | of Year IV. ... | 4540 |
| TARGET | 77 | | | of the disser... | 4541 |
| E | 76 | | | ising | 4542 |
| AS | 75 | | | formance | 4600 |
| SHOULD | 73 | | will | be assessed according to the degree of completion of task as w... | 4601 |
| AFF | 72 | | | | |

Words: 4235 | Tokens: 30605 | At word: 50 | Word sort: Desc frequency | Context sort: Asc occurrence order

23/06/2006

IVACS 2006



SPEECH Concordancer

| | A | B | C | D | E | F | G | H |
|----|---------|------|-----------------------------|---------|-----------|----------|-----------|-------|
| 1 | Speaker | Line | Orthotext | SigSyll | Start | Duration | OrthoSyll | Speed |
| 2 | Aelish | 185 | probably | 1 | 6:38.162 | 0.185 | 3 | 973 |
| 3 | Aelish | 151 | I do not know | 3 | 5:33.390 | 0.259 | 4 | 927 |
| 4 | Marc | 307 | yeah | 1 | 10:37.009 | 0.065 | 1 | 923 |
| 5 | Marc | 15 | I am not giving | 4 | 0:28.801 | 0.359 | 5 | 836 |
| 6 | Marc | 254 | yeah | 1 | 8:59.738 | 0.072 | 1 | 833 |
| 7 | Marc | | Yeah | 1 | 3:14.785 | 0.075 | 1 | 800 |
| 8 | Marc | | Yeah | 1 | 6:22.970 | 0.075 | 1 | 800 |
| 9 | Aelish | 317 | and | 1 | 10:55.488 | 0.076 | 1 | 789 |
| 10 | Marc | | yeah | 1 | 8:59.958 | 0.079 | 1 | 759 |
| 11 | Marc | 166 | That is right | 2 | 6:0.976 | 0.243 | 3 | 741 |
| 12 | Marc | 109 | that is right | 2 | 4:12.225 | 0.250 | 3 | 720 |
| 13 | Marc | 256 | yeah | 1 | 9:3.070 | 0.084 | 1 | 714 |
| 14 | Marc | 10 | originally | 4 | 0:19.304 | 0.424 | 5 | 708 |
| 15 | Marc | | very good | 3 | 9:33.502 | 0.258 | 3 | 698 |
| 16 | Marc | 65 | Right | 1 | 2:33.489 | 0.087 | 1 | 690 |
| 17 | Marc | 208 | That is right | 2 | 7:32.976 | 0.266 | 3 | 677 |
| 18 | Marc | | Yeah | 1 | 8:1.450 | 0.092 | 1 | 652 |
| 19 | Marc | 240 | engineer | 3 | 8:32.314 | 0.281 | 3 | 641 |
| 20 | Marc | 65 | right | 1 | 2:33.749 | 0.095 | 1 | 632 |
| 21 | Marc | 111 | that is right | 2 | 4:15.730 | 0.287 | 3 | 627 |
| 22 | Marc | 22 | how are you doing? | 4 | 0:40.679 | 0.482 | 5 | 622 |
| 23 | Marc | 252 | That is correct | 3 | 8:55.194 | 0.387 | 4 | 620 |
| 24 | Marc | | right | 1 | 5:5.544 | 0.097 | 1 | 619 |
| 25 | Marc | 81 | Civil War? | 3 | 3:14.307 | 0.294 | 3 | 612 |
| 26 | Marc | | isn't that right? | 4 | 6:37.768 | 0.393 | 4 | 611 |
| 27 | Aelish | 206 | he was in the | 4 | 7:27.517 | 0.393 | 4 | 611 |
| 28 | Marc | 153 | That is where Daddy is from | 5 | 5:37.579 | 0.692 | 7 | 607 |
| 29 | Marc | 127 | He was a Taylor | 5 | 4:41.263 | 0.498 | 5 | 602 |
| 30 | Marc | | the | 1 | 6:24.456 | 0.100 | 1 | 600 |
| 31 | Marc | 130 | Israel | 3 | 4:45.854 | 0.310 | 3 | 581 |
| 32 | Aelish | 66 | but it does not mean | 4 | 2:37.791 | 0.520 | 5 | 577 |
| 33 | Marc | 153 | is not it | 2 | 5:38.272 | 0.312 | 3 | 577 |
| 34 | Aelish | 271 | sure | 1 | 9:31.063 | 0.105 | 1 | 571 |

Marc
average
388 syll/min

Citation Transcription

| Spoken | Sec. | Syl. | Citation | Syl. | Syl./min |
|------------------|-------|------|----------------------|------|--------------|
| Whydjedotha? | 0.863 | 4 | Why did you do that? | 5 | 348 (278) |
| I'll be 5 mints. | 0.986 | 4 | I will be 5 minutes. | 6 | 365 (243) |
| 'Mon quick! | 0.497 | 2 | Come on quick! | 3 | 362 (241) |

Why use Citation Transcription?

- Uses **idealised** forms of ‘reduced’ NS flow
- Ability to find what is **barely** in speech signal
- Easy to find multiple instances via **search string**
- Speech records can be **compared** via transcript
- Can be **tagged** for multiple features: variety, emotion etc.

GUI Interface - Key Features

Use various DSP functions

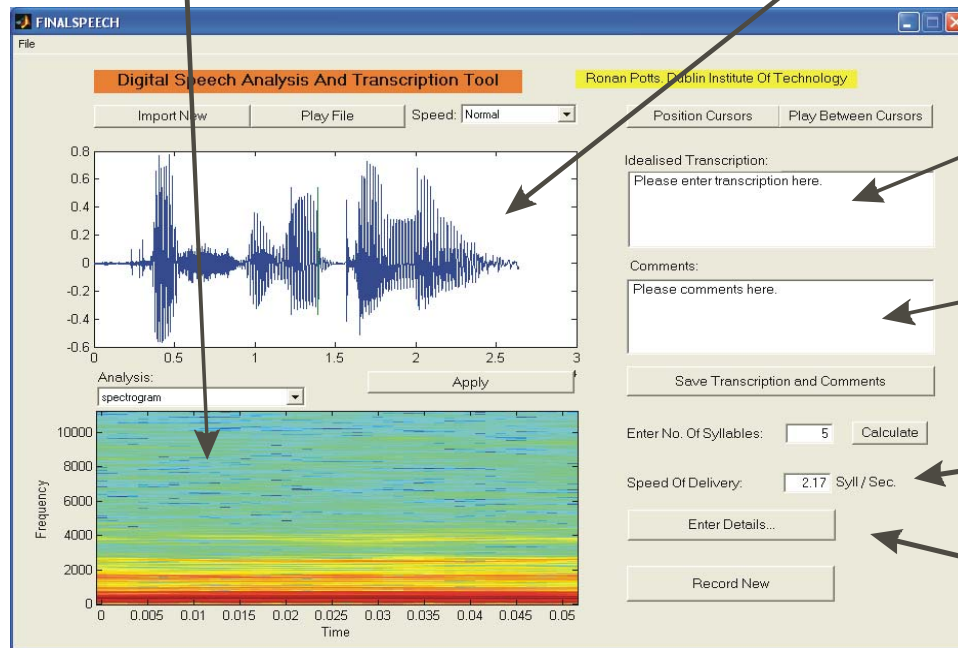
Assist in segmentation

Listen to recording and produce CITATION transcription

Tag for various features?

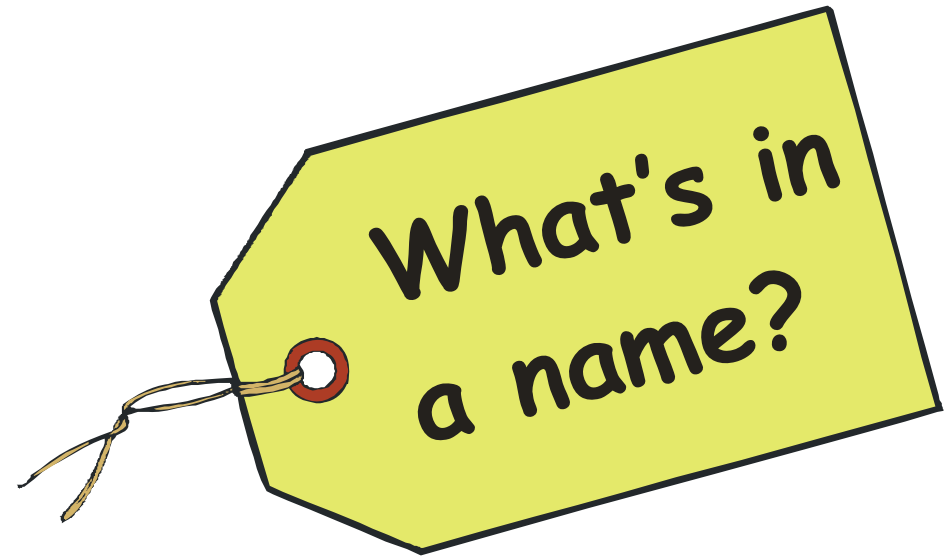
Indicate speed: syll/min

Further screens/views

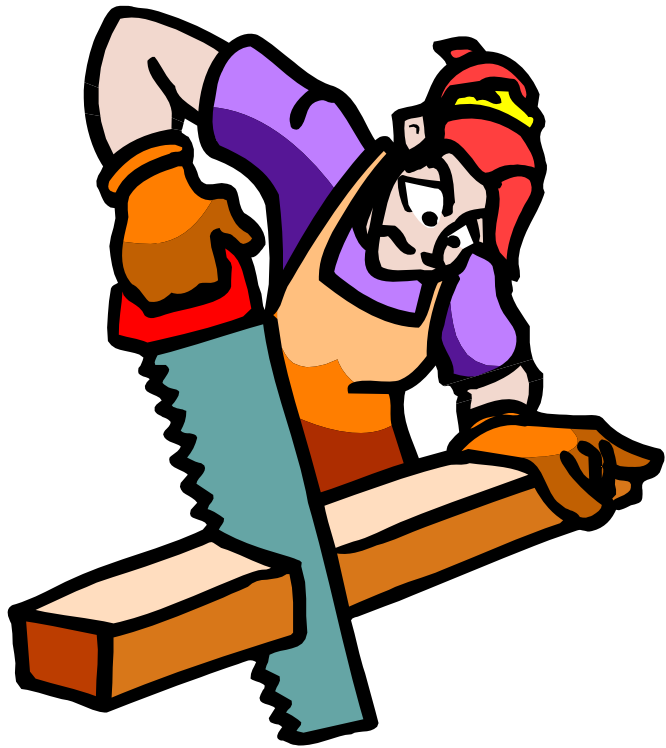


'Segmentation' = Tagging

- Speed
- Emotion
- Speech variety
- NNS English



Problems of 'Segmentation'



DIY Rule:

- **Measure twice ...**
- **Cut once !**

Possible 'Segmentation' Units

- Clip
- Tone unit
- Formulaic sequence
- Chunk
- Paradigm
- Collocation

Enter ... the ‘Timed Unit’

Draft definition:


“A segment which has its own flow and **pragmatic integrity**. It is speaker determined and can be timed, and is bounded by pauses— no matter how brief— or marked by a sudden change of tack/thought.”

SPEECH Concordancer

| | A | B | C | D | E | F | G | H |
|----|---------|------|-----------------------------|---------|-----------|----------|-----------|-------|
| 1 | Speaker | Line | Orthotext | SigSyll | Start | Duration | OrthoSyll | Speed |
| 2 | Aelish | 185 | probably | 1 | 6:38.162 | 0.185 | 3 | 973 |
| 3 | Aelish | 151 | I do not know | 3 | 5:33.390 | 0.259 | 4 | 927 |
| 4 | Marc | 307 | yeah | 1 | 10:37.009 | 0.065 | 1 | 923 |
| 5 | Marc | 15 | I am not giving | 4 | 0:28.801 | 0.359 | 5 | 836 |
| 6 | Marc | 254 | yeah | 1 | 8:59.738 | 0.072 | 1 | 833 |
| 7 | Marc | | Yeah | 1 | 3:14.785 | 0.075 | 1 | 800 |
| 8 | Marc | | Yeah | 1 | 6:22.970 | 0.075 | 1 | 800 |
| 9 | Aelish | 317 | and | 1 | 10:55.488 | 0.076 | 1 | 789 |
| 10 | Marc | | yeah | 1 | 8:59.958 | 0.079 | 1 | 759 |
| 11 | Marc | 166 | That is right | 2 | 6:0.976 | 0.243 | 3 | 741 |
| 12 | Marc | 109 | that is right | 2 | 4:12.225 | 0.250 | 3 | 720 |
| 13 | Marc | 256 | yeah | 1 | 9:3.070 | 0.084 | 1 | 714 |
| 14 | Marc | 10 | originally | 4 | 0:19.304 | 0.424 | 5 | 708 |
| 15 | Marc | | very good | 3 | 9:33.502 | 0.258 | 3 | 698 |
| 16 | Marc | 65 | Right | 1 | 2:33.489 | 0.087 | 1 | 690 |
| 17 | Marc | 208 | That is right | 2 | 7:32.976 | 0.266 | 3 | 677 |
| 18 | Marc | | Yeah | 1 | 8:1.450 | 0.092 | 1 | 652 |
| 19 | Marc | 240 | engineer | 3 | 8:32.314 | 0.281 | 3 | 641 |
| 20 | Marc | 65 | right | 1 | 2:33.749 | 0.095 | 1 | 632 |
| 21 | Marc | 111 | that is right | 2 | 4:15.730 | 0.287 | 3 | 627 |
| 22 | Marc | 22 | how are you doing? | 4 | 0:40.679 | 0.482 | 5 | 622 |
| 23 | Marc | 252 | That is correct | 3 | 8:55.194 | 0.387 | 4 | 620 |
| 24 | Marc | | right | 1 | 5:5.544 | 0.097 | 1 | 619 |
| 25 | Marc | 81 | Civil War? | 3 | 3:14.307 | 0.294 | 3 | 612 |
| 26 | Marc | | isn't that right? | 4 | 6:37.768 | 0.393 | 4 | 611 |
| 27 | Aelish | 206 | he was in the | 4 | 7:27.517 | 0.393 | 4 | 611 |
| 28 | Marc | 153 | That is where Daddy is from | 5 | 5:37.579 | 0.692 | 7 | 607 |
| 29 | Marc | 127 | He was a Taylor | 5 | 4:41.263 | 0.498 | 5 | 602 |
| 30 | Marc | | the | 1 | 6:24.456 | 0.100 | 1 | 600 |
| 31 | Marc | 130 | Israel | 3 | 4:45.854 | 0.310 | 3 | 581 |
| 32 | Aelish | 66 | but it does not mean | 4 | 2:37.791 | 0.520 | 5 | 577 |
| 33 | Marc | 153 | is not it | 2 | 5:38.272 | 0.312 | 3 | 577 |
| 34 | Aelish | 271 | sure | 1 | 9:31.063 | 0.105 | 1 | 571 |

Marc
average
388 syll/min

Grouping into Tone Units

| | A | B | C | D | E | F | G | H | I |
|----|---------|------|---|---------|---------|----------|-----------|-------|----------|
| 1 | Speaker | Unit | Orthotext | SigSyll | Start | Duration | OrthoSyll | Speed | Emotion? |
| 2 | Marc | 1a | The  | 1 | 0:0.000 | 0.363 | 1 | 165 | |
| 3 | Marc | 1b | lapel mike is on | 5 | 0:0.419 | 1.041 | 5 | 288 | |
| 4 | Marc | 2 | The front mike is on | 5 | 0:1.916 | 1.028 | 5 | 292 | |
| 5 | Marc | 3 | #erm# | | | | | | |
| 6 | Marc | 4 | I shall leave that for the moment | 7 | 0:3.870 | 0.863 | 8 | 556 | |
| 7 | Marc | 5 | #erm# | | | | | | |
| 8 | Marc | 6a | Pat | 1 | 0:6.517 | 0.327 | 1 | 183 | |
| 9 | Marc | 6b | Patrick | 2 | 0:6.873 | 0.366 | 2 | 328 | |
| 10 | Marc | 6c | Paddy | 2 | 0:7.234 | 0.366 | 2 | 328 | |
| 11 | Pat | 7 | P | 1 | 0:8.104 | 0.359 | 1 | 167 | |

... and what else ?

**Other functions for
the speech corpus:**

- **Learner ?**
- **Researcher ?**

