



# Digital Tools for Creating and Analysing Corpora

Adam Mearns

## Outline

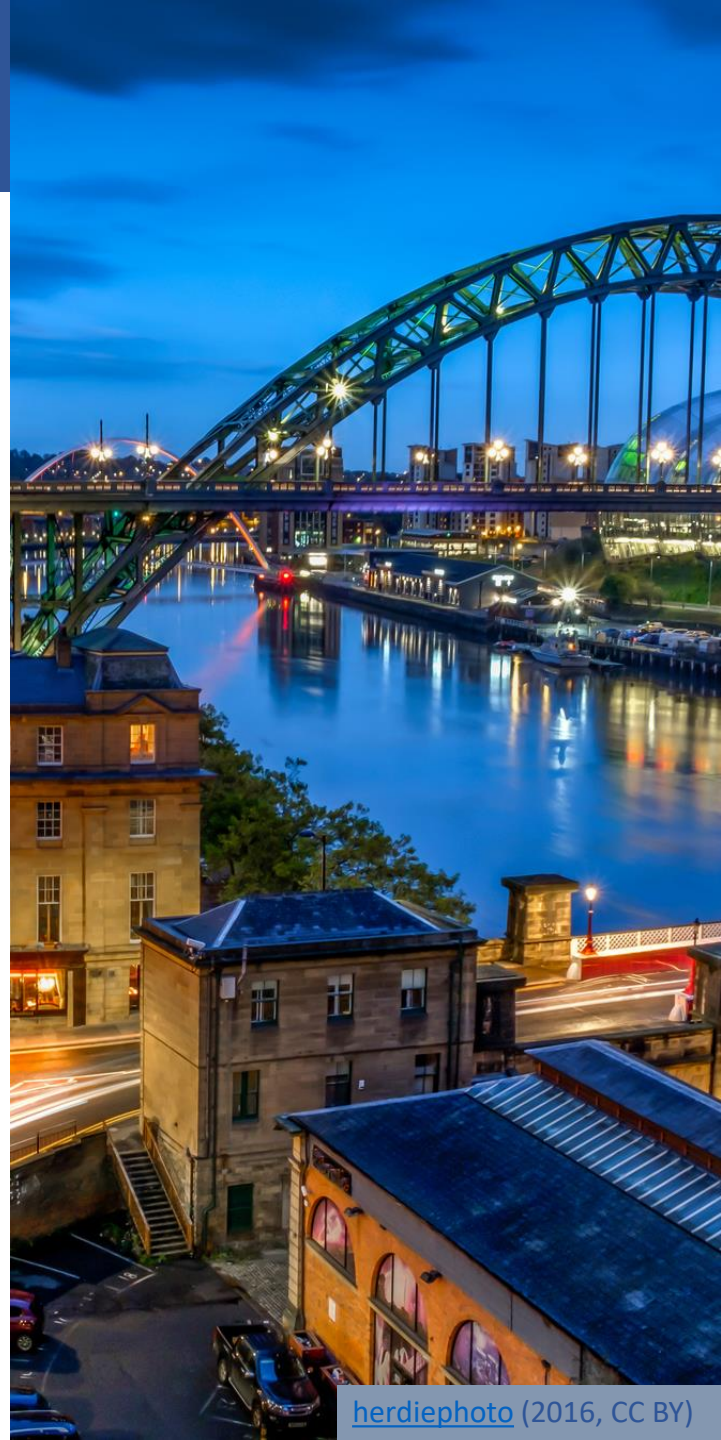
- DECTE: *The Diachronic Electronic Corpus of Tyneside English*  
<http://research.ncl.ac.uk/decte>
- Corpus Creation – Issues & Tools  
file formats, mark-up/tagging  
and text editing (regular expressions)  
*AntFileConverter, TEI XML,  
POS-Tagging, Notepad++*
- Corpus Analysis  
concordances, collocations &  
keywords  
*AntConc, Wordsmith, #LancsBox*



Combining interviews from 1970s & 1990s  
with interviews collected by  
Newcastle University students since 2007

### Current size

- *Processed*  
99 interviews  
c.800,000 words / c.72 hours
- *Full collection*  
c.890 interviews  
c.7m words / c.560 hours





# The Diachronic Electronic Corpus of Tyneside English

Combining interviews from 1970s & 1990s with interviews collected by Newcastle University students since 2007

## Format

- demographic files (speaker info)
- wav/mp3 audio files
- transcriptions:  
**TEI (P5)-conformant XML files**  
*a standard for the encoding of electronic texts*  
(and plain TXT versions)







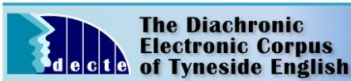
```
decten2y07i007.xml x
44 <person xml:id="informantY07i007a">
45 <age>
46 21-30
47 </age>
48 <sex>
49 Male
50 </sex>
51 <residence>
52 Wearside - Sunderland (born in Newcastle, Tyn
53 </residence>
54 <occupation>
55 Shop Worker
56 </occupation>
57 <education>
58 Higher Education
59 </education>
60 </person>
61 <person xml:id="informantY07i007b">
62 <age>
63 16-20
64 </age>
65 <sex>
66 Female
67 </sex>
68 <residence>
69 Tyneside - Newcastle
70 </residence>
71 <occupation>
72 University Student
73 </occupation>
74 <education>
75 Higher Education
```

```
decten2y07i007.xml x
141 <u who="#informantY07i007b"> No. <incident><desc>interruption
</desc></incident> Marmite. </u>
142 <u who="#informantY07i007a"> <incident><desc>interruption
</desc></incident> No. Ok, Marmite. A guy called, Mar- she nicknamed him
Marmite. Oh <anchor xml:id="decten2y07i007ortho0080"/> (NAME). Yeah.
Crazy (NAME). </u>
143 <u who="#informantY07i007b"> Yeah. </u>
144 <u who="#informantY07i007a"> Yeah. </u>
145 <u who="#interviewerY07i007"> Oh yeah. </u>
146 <u who="#informantY07i007a"> Yeah he's weird. </u>
147 <u who="#informantY07i007b"> You know about him too? <vocal><desc>
laughter</desc></vocal> </u>
148 <u who="#interviewerY07i007"> <vocal><desc>laughter</desc></vocal> I
think I might do. <vocal><desc>laughter</desc></vocal> </u>
149 <u who="#informantY07i007a"> Yeah. But no, (NAME) (NAME). She's
annoying, but she's left now. God help (NAME)! </u>
150 <u who="#informantY07i007b"> Still don't know who she is. </u>
151 <u who="#informantY07i007a"> She's like, our bosses' bosses' boss. </u>
152 <u who="#informantY07i007b"> <vocal><desc>laughter</desc></vocal> I
never see her. <vocal><desc>laughter</desc></vocal> </u>
153 <u who="#interviewerY07i007"> Erm <anchor xml:id=
"decten2y07i007ortho0100"/> so, do you see people from work socially? </u>
154 <u who="#informantY07i007a"> Er yeah. </u>
155 <u who="#informantY07i007b"> Yeah. </u>
```

decten2y07i007.xml



# Research Project Website and Public Website



DECTE is funded by the AHRC (Grant: AH/H037691/1)

- Home
- Acknowledgments
- Documentation
- Corpus Files
- People
- Bibliography
- Appendices
- Related Resources



Welcome to the *Diachronic Electronic Corpus of Tyneside English* (DECTE), a corpus of dialect speech from the Tyneside area of North-East England.

DECTE is an amalgamation of the existing *Newcastle Electronic Corpus of Tyneside English* (NECTE) created between 2001 and 2005 (<http://research.ncl.ac.uk/necte>), and NECTE2, a collection of interviews conducted in the Tyneside area since 2007. It thereby constitutes a rare example of a publicly available on-line corpus presenting dialect material spanning five decades.

The present website is designed for research use. DECTE also, however, includes an interactive website, [The Talk of the Toon](#), which integrates topics and narratives of regional cultural significance in the corpus with relevant still and moving images, and which is designed primarily for use in schools and museums and by the general public.

Graphics credits:

- River Tyne: [Jan Britton](#) (2008 CC BY NC 2.0)
- Angel of the North: [Taylor Dundee](#) (2007 CC BY NC-SA 2.0)
- Quayside Market: [Andrea 44](#) (1982 CC BY 2.0)

DECTE: <http://research.ncl.ac.uk/decte>

## The Talk of the Toon



*An archive of local language and stories*

the memories, thoughts and opinions of the people of Tyneside, past and present, in their own words



- Home
- About
- Interview Index
- Themes
- Quizzes
- Schools
- Top Stories
- Introduction to North East Dialects
- Links

### Themes

Family, Home, Work, Shopping, Sport, Entertainment, etc.



[Click Here]

### Word Search

Search the Archive



[Click Here]

### The Talk of the Toon

- [About](#)
- [Feedback | Get Involved](#)
- [Introduction to North East Dialects](#)
- [Interview Index](#)
- [Links](#)
- [Privacy Policy: Cookies](#)

### Quizzes

How well do you think you know the Geordie dialect?



[Click Here]

### Schools

Guidance for Key Stages 2 and 3, GCSE and A-Level



[Click Here]

## Top Stories — Memories of the War: cold beds, lino and igloos

1990s / Female 51-60 & Female 51-60

I can remember lying in my bed in the War and waking up so cold that I couldn't straighten my knees ... and then I used to get up in the morning, went to the lino ... eeh it was, eh, oh, they don't know they're born ...

[Hear the full story](#) — [See more Top Stories](#)



Picture: [A. Evans](#) (2010, CC BY-NC-ND 2.0)

- Home
- About
- Interview Index
- Themes
- Quizzes
- Schools
- Top Stories
- Introduction to North East Dialects
- Links

Toon: <http://research.ncl.ac.uk/decte/toon>



## Archive Interview: Y07i007

Return to: [Theme Results](#) | [Interview Index](#)

For a guide to the layout of this interview page and how to use it, [click here](#).



decten2y07i007audio

### Interview Transcript

**Speaker 3:** I'm only joking. No, Marmite.

**Speaker 2:** Yeah. Yeah erm, what (NAME)? Yeah.

**Speaker 3:** No. *(interruption)* Marmite.

**Speaker 2:** *(interruption)* No. Ok, Marmite. A guy called, Mar- she nicknamed him Marmite. Oh (NAME). Yeah. Crazy (NAME).

**Speaker 3:** Yeah.

**Speaker 2:** Yeah.

**Speaker 1:** Oh yeah.

**Speaker 2:** Yeah he's weird.

**Speaker 3:** You know about him too? *(laughter)*

**Speaker 1:** *(laughter)* I think I might do. *(laughter)*

**Speaker 2:** Yeah. But no, (NAME) (NAME). She's annoying, but she's left now. God help (NAME)!

**Speaker 1:** interviewerY07i007

**Speaker 2:** informantY07i007a

**Age Group:** 21-30

**Gender:** Male

**Residence:** Wearside - Sunderland  
(born in Newcastle,  
Tyneside)

**Education:** Higher Education

**Occupation:** Shop Worker

**Speaker 3:** informantY07i007b

**Age Group:** 16-20

**Gender:** Female

**Residence:** Tyneside - Newcastle

**Education:** Higher Education

**Occupation:** University Student

### Themes

Click a theme in the menu below to highlight related keywords in the transcript.



Education

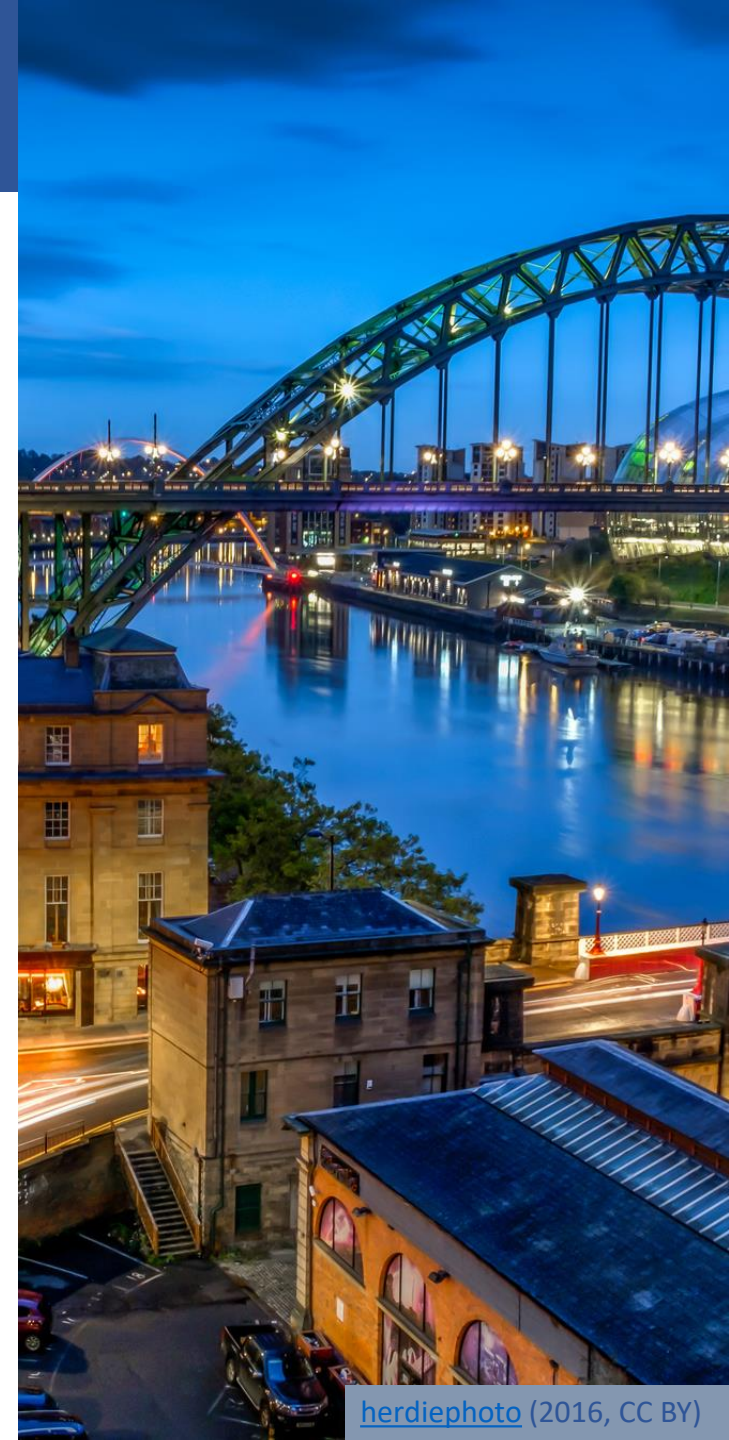
Work & Industry



## Processing

processing the interviews involves *e.g*

- correction of transcription  
*accuracy/regularity in trans. conventions*  
*accuracy as a record of what was said*
- time alignment  
*aligning points in audio with*  
*points in transcription*
- anonymization
- converting into TEI XML format





# Corpus Creation – Issues and Tools

2007\_SEL2091\_040

had like...places where you could put your coat and stuff and for s...I don-t know (reduced to dunno) what we were doing but I wasn-t involved in it I was just at the wrong place in the wrong time and I got chucked out of there and I wasn-t allowed to be with the girls I got segregated and I got put with the boys...and that..that was that. [07-08/N/YH/831]

[07-08/0/RL] what about when you went to senior school then...well behaved there or? [07-08/0/RL]

[07-08/N/BS/512] We got separated in{ em high school. [07-08/N/BS/512]

[07-08/N/YH/831] { Yes (rendered as 'yeah') [07-08/N/YH/831]

[07-08/N/BS/512] There was two sides of the year em..and like you got split into one half and so I didn-t get any em lessons with you until we were a bit older about fifteen or sixteen and then we got--we got Spanish didn-t we and me and Danielle got moved to the front for em..for em..@@ I don-t remem{ber what we were doing [07-08/N/BS/512]

[07-08/N/YH/831] {Talking. [07-08/N/YH/831]

[07-08/N/BS/512] but yes we had to sit like on the teacher'z desk..{..right at the very front. [07-08/N/BS/512]

[07-08/N/YH/831] {Yes (rendered as 'yeah') [07-08/N/YH/831]

[07-08/N/YH/831] But used to like you know how when you do listening tests in er Spanish we @ used to play with the tape @@ and like rewind it back to the wrong place and stuff and then like put the volume up and..I remember I used to draw p-you know when you used to draw pictures of your teacher...like in the back of your exercise book. I remember I drew a picture of em...our Spanish class Mrs. Wood and she was like always having babies like every other week she was having a baby and I drew a picture of her and like I drew attention to her like her feet and she came round and had a look like she was looking and I thought god she is going to see in the back of my book and it was just as well that I like shut it and like she was 'what are you doing in writing in the back of your exercise book?' and I was like oh my god please do not see that @. [07-08/N/YH/831]

[07-08/N/BS/512] @@...but..Mr. Jurich he was our Spanish teacher was looking in the the back of my book once because (reduced to 'cause') he was saying we shouldn-t be scribbling in our books and em..I-d written em..'M-Mr. Jurich is a freak' in the back of my book but he-d looked on the page before and seen 'Alex is a freak' and it was because (reduced to 'cause') I sat next to Alex and he was going 'Alex is a freak. Helen what are you talking about' and I was like and I got wronged for that but er..there was loads of stuff written ab..written about him on the page after it and he just didn-t..he just didn-t see it then his fl...

2017\_SEL2091\_032

SPEAKER	UTTERANCE
[2017/IN/ED/0962]	not like student accommodation it's like it's me and five other students from here in Newcastle
[2017/CN/0962]	same course or
[2017/IN/ED/0962]	no well one of them's on the same course one other guy's been doing the same thing as me chm we from all over <INT> the <cough>
[2017/CN/0962]	yeah
[2017/IN/ED/0962]	we all know each other from last year so it's not bad not bad
[2017/MB/0962]	right
[2017/IN/ED/0962]	It's good for match days though 'cause you can always chm can always hear all the goals and all the cheering and stuff like that it's pretty decent
[2017/MB/0962]	yeah I was I was in Verde last year the green one
[2017/IN/ED/0962]	yeah yeah
[2017/MB/0962]	but I never heard it ever I prefer living ((+R+N)) in a house this year though <INT> more room for activities
[2017/IN/ED/0962]	is it better do you think <L> yeah
[2017/MB/0962]	and now I've got three other dentists with us like I don't miss things <INT> I'm never late they always just wake us up <XX>
[2017/CN/0962]	yes I was going to ((+R+gonna)) say it's
[2017/IN/ED/0962]	<L>
[2017/CN/0962]	yeah my course are melts so I can't really get a house with them
[2017/IN/ED/0962]	what do you study
[2017/CN/0962]	quantity surveying
[2017/IN/ED/0962]	oh <L> shit
[2017/MB/0962]	<XX>
[2017/IN/ED/0962]	what do you actually <INT> do
[2017/CN/0962]	some are alright there's a few decent lads but otherwise I just don't even don't even talk
[2017/IN/ED/0962]	what do you actually do for that
[2017/CN/0962]	just like stretch economics you could say but then er you just look at buildings and count bricks it's quite easy to be fair
[2017/IN/ED/0962]	<L>
[2017/CN/0962]	then there's another more law fuelled one which is sort of just loads of essays and referencing ((+R+N))
[2017/MB/0962]	my ((+R-me)) dad was saying ((+R+N)) he knows a couple of quantity surveyors 'cause he started off on a building site
[2017/CN/0962]	yeah <INT> earn some serious money
[2017/IN/ED/0962]	yeah I've got some serious money

c. 779 interviews (text & audio) collected between 2007 and 2017  
Transcribed by students as DOCX files





## File Formats

Batch conversion of files (docx, pdf) to txt [xml]

- **AntFileConverter** (Laurence Anthony)  
<http://www.laurenceanthony.net/software/antfileconverter>

*docx or pdf → txt*

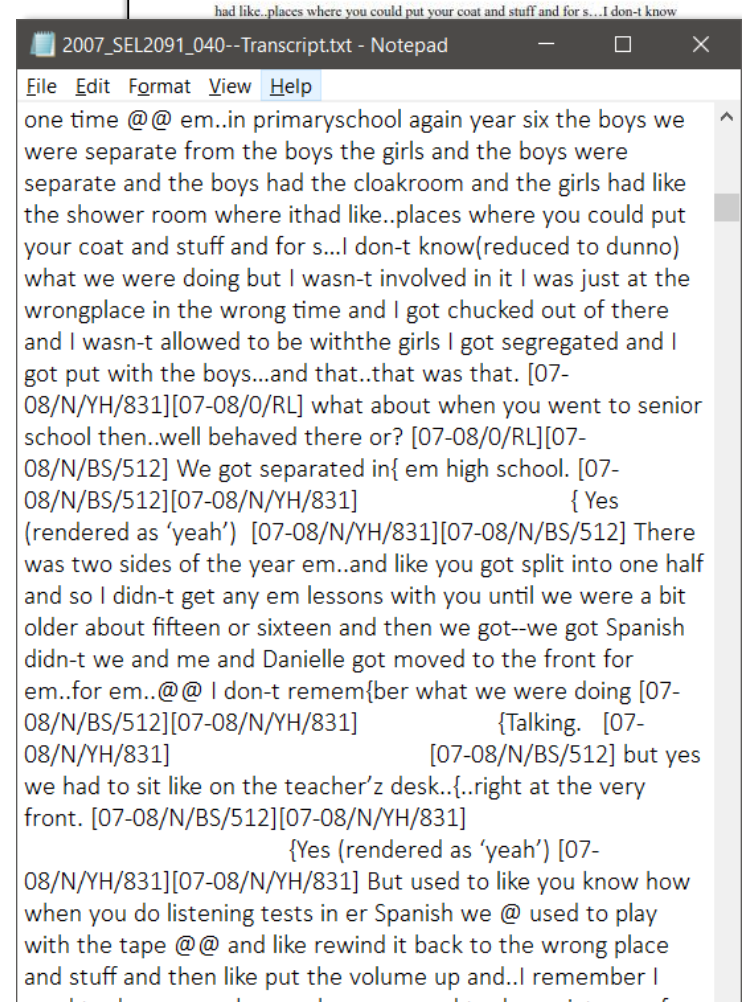
Windows, Mac, Linux; freeware

- **MultiDoc Converter** (Pawel Idzikowski)  
<http://www.multidoc-converter.com/en/index.html>

*↔ doc, docx, rtf, odt, epub, htm, html, mht, xml, txt*

Windows; freeware

2007\_SEL2091\_040



Converted from DOCX to TXT with AntFileConverter



## File Formats

Batch conversion of files (docx, pdf) to txt [xml]

- *AntFileConverter* (Laurence Anthony)  
<http://www.laurenceanthony.net/software/antfileconverter>

*docx or pdf → txt*

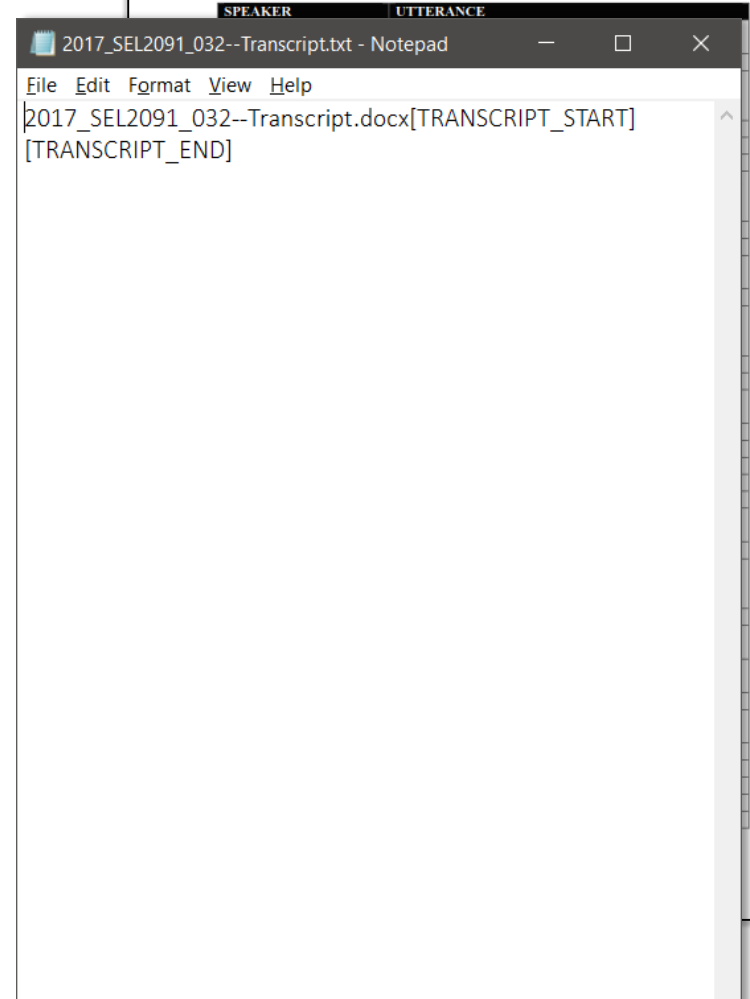
Windows, Mac, Linux; freeware

- *MultiDoc Converter* (Pawel Idzikowski)  
<http://www.multidoc-converter.com/en/index.html>

*↔ doc, docx, rtf, odt, epub, htm, html, mht, xml, txt*

Windows; freeware

2017\_SEL2091\_032



Converted from DOCX to TXT with *AntFileConverter*



## File Formats

Batch conversion of files (docx, pdf)  
to txt [xml]

- **AntFileConverter** (Laurence Anthony)  
<http://www.laurenceanthony.net/software/antfileconverter>

*docx or pdf* → *txt*

Windows, Mac, Linux; freeware

- **MultiDoc Converter** (Pawel Idzikowski)  
<http://www.multidoc-converter.com/en/index.html>

↔ *doc, docx, rtf, odt, epub, htm, html, mht, xml, txt*

Windows; freeware

2017\_SEL2091\_032

```
File Edit Format View Help
[2017/IN/ED/0962]
hot like student accommodation it's like it's me and five other
students from here in Newcastle
[2017/CN/0962]
same course or
[2017/IN/ED/0962]
no well one of them's on the same course one other guy's been
doing the same thing as me ehm we from all over <INT> the
<cough>
[2017/CN/0962]
yeah
[2017/IN/ED/0962]
we all know each other from last year so it's not bad not bad
[2017/MB/0962]
right
[2017/IN/ED/0962]
It's good for match days though 'cause you can always ehm can
always hear all the goals and all the cheering and stuff like that
it's pretty decent
[2017/MB/0962]
yeah I was I was in Verde last year the green one
[2017/IN/ED/0962]
yeah yeah
[2017/MB/0962]
but I never heard it ever I prefer living ((+R+N)) in a house this
year though <INT> more room for activities
[2017/IN/ED/0962]
is it better do you think <L> yeah
[2017/MB/0962]
```

Converted from DOCX to TXT with *MultiDoc Converter*





## File Formats

Batch conversion of files (docx, pdf) to txt [xml]

- *AntFileConverter* (Laurence Anthony)  
<http://www.laurenceanthony.net/software/antfileconverter>

*docx* or *pdf* → *txt*

Windows, Mac, Linux; freeware

- *MultiDoc Converter* (Pawel Idzikowski)  
<http://www.multidoc-converter.com/en/index.html>

↔ *doc*, *docx*, *rtf*, *odt*, *epub*, *htm*, *html*,  
*mht*, *xml*, *txt*

Windows; freeware

Newcastle University  
School of English  
Literature, Language & Linguistics

The Diachronic Electronic Corpus of Tyneside English  
DECTE is funded

Home  
Acknowledgements  
Documentation  
Corpus Files  
People  
Bibliography  
Appendices

**Acknowledgements**

The DECTE project is of course indebted to the researchers who created the legacy materials that it incorporates:

- The recordings from the 1960s-1970s were collected as part of the *Tyneside Linguistic Survey* by Joan Beal, Anthea Fraser Gupta, Val Jones, John Local, Vince McNeaney, Graham Nixon, John Pellowe and Barbara Strang.
- The 1990s recordings were gathered for the *Phonological Variation and Change in Contemporary Spoken English* project by Gerry Docherty, Paul Foulkes, Jim Milroy, Lesley Milroy, Penny Oxley, David Walshaw and Dominic Watt.
- Between 2000 and 2005, these two collections were digitized and amalgamated as the *Newcastle Electronic Corpus of Tyneside English* by Will Allen, Joan Beal, Karen Corrigan, Warren Maguire and Hermann Moisl. Full acknowledgements for that project can be found here: <http://research.ncl.ac.uk/necte/acknowledgements.htm>

```
<tr>
<td><font face="Arial" size="3"><b>Acknowledgements</b></td></tr>
<p style="line-height: normal; margin-top: 0; margin-bottom: 0">
The DECTE project
is of course indebted to the researchers who created
the legacy materials that it incorporates:</p>
<p style="line-height: normal; margin-top: 0; margin-bottom: 0">
</p>
<ul>
<li>
<p style="line-height: normal; margin-top: 0; margin-bottom: 0">
The recordings
from the 1960s-1970s were collected as part of
the <i>Tyneside Linguistic Survey</i> by
Joan Beal, Anthea Fraser
Gupta, Val Jones, John Local, Vince McNeaney,
Graham Nixon, John Pellowe and Barbara Strang.</li>
<li>
<p style="line-height: normal; margin-top: 10px; margin-bottom: 0">
The 1990s
recordings were gathered for the <i>Phonological
Variation and Change in Contemporary Spoken
English</i> project by Gerry Docherty, Paul Foulkes, Jim Milroy,
Lesley Milroy, Penny Oxley,
David Walshaw and Dominic Watt.</li>
<li>
<p style="line-height: normal; margin-top: 10px; margin-bottom: 0">
Between 2000
and 2005, these two collections were digitized
and amalgamated as the <i>Newcastle Electronic
Corpus of Tyneside English</i> by
Will Allen, Joan Beal, Karen Corrigan, Warren
Maguire and Hermann Moisl. Full acknowledgements
```



## File Formats

Batch conversion of files (docx, pdf)  
to txt [xml]

- **AntFileConverter** (Laurence Anthony)  
<http://www.laurenceanthony.net/software/antfileconverter>

*docx or pdf → txt*

Windows, Mac, Linux; freeware

- **MultiDoc Converter** (Pawel Idzikowski)  
<http://www.multidoc-converter.com/en/index.html>

*↔ doc, docx, rtf, odt, epub, htm, html,  
mht, xml, txt*

Windows; freeware



acknowledgements.txt - Notepad

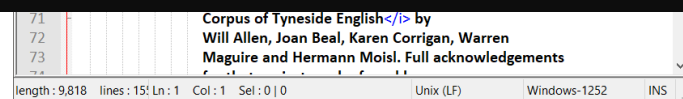
File Edit Format View Help

Home  
Acknowledgments  
Documentation  
Corpus Files  
People  
Bibliography  
Appendices  
Related Resources

Acknowledgements  
The DECTE project is of course indebted to the researchers who created the legacy materials that it incorporates:

- The recordings from the 1960s-1970s were collected as part of the Tyneside Linguistic Survey by Joan Beal, Anthea Fraser Gupta, Val Jones, John Local, Vince McNeaney, Graham Nixon, John Pellowe and Barbara Strang.
- The 1990s recordings were gathered for the Phonological Variation and Change in Contemporary Spoken English project by Gerry Docherty, Paul Foulkes, Jim Milroy, Lesley Milroy, Penny Oxley, David Walshaw and Dominic Watt.
- Between 2000 and 2005, these two collections were digitized and amalgamated as the Newcastle Electronic Corpus of Tyneside English by Will Allen, Joan Beal, Karen Corrigan, Warren Maguire and Hermann Moisl. Full acknowledgements for that project can be found here:  
<http://research.ncl.ac.uk/necte/acknowledgements.htm>
- Since 2007, a great number of undergraduate and postgraduate students have recorded and transcribed interviews for the current stage of the corpus (NECTE2), as part of their studies in the School of English Literature, Language and Linguistics at Newcastle University.

Converted from HTM to TXT with MultiDoc Converter





## File Formats

Batch conversion of files (docx, pdf)  
to txt [xml]

*Alternatively*

- *WordSmith Tools* (Mike Scott)

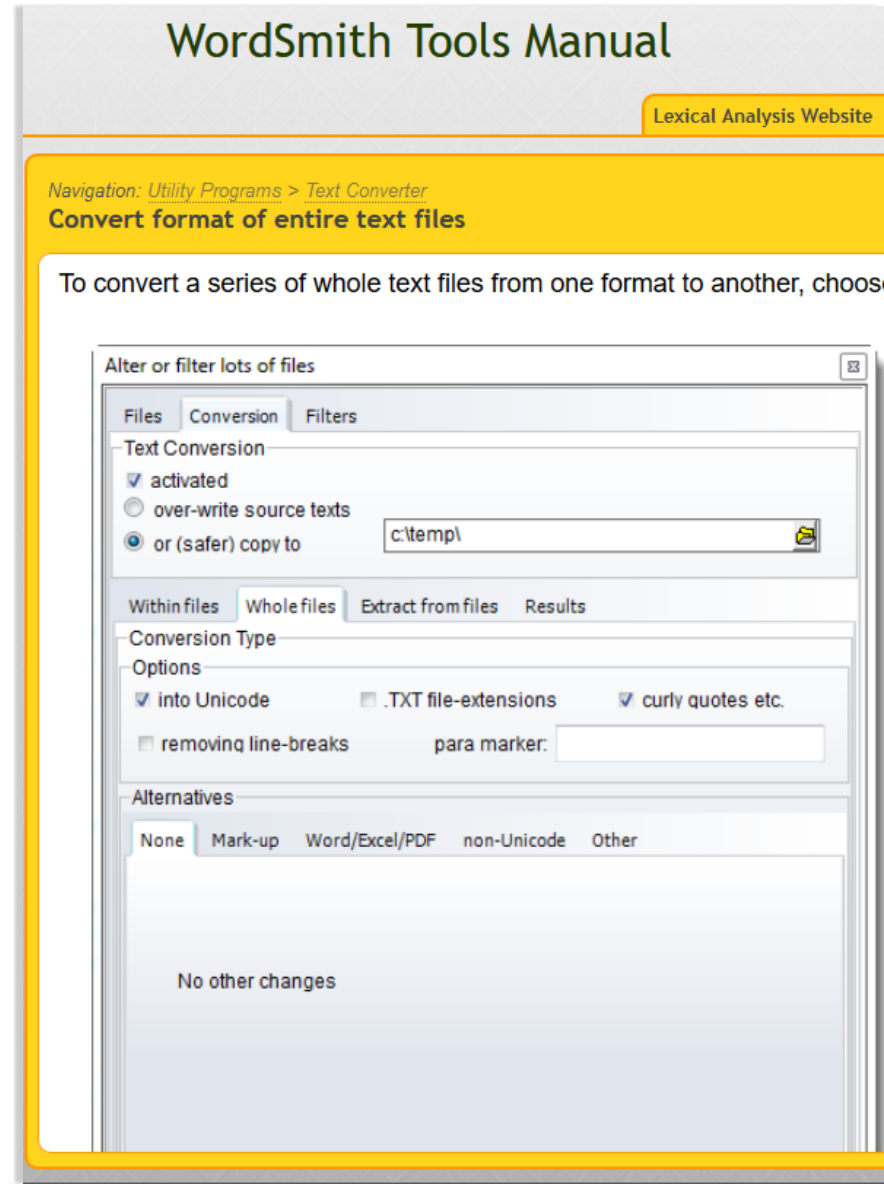
<http://lexically.net/wordsmith>

*utility programs inc. Text Converter*

single user licence £50

Windows

- *Online HTML scrapers / converters*







## Mark-up/Tagging



< Text Encoding Initiative >

**TEI (Text Encoding Initiative)** – guidelines for XML markup

- <http://www.tei-c.org/index.xml>

‘The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics. Since 1994, the TEI Guidelines have been widely used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation.’

Software for creating/editing XML files, e.g:

**Notepad++** [freeware; see below]

**Oxygen:** <https://www.oxygenxml.com>



## Mark-up/Tagging



## TEI (Text Encoding Initiative) – guidelines for XML markup

```

• decten2y07i007.xml x
TEI text group text body u
190 <u who="#informantY07i007b"> <vocal><desc>laughter</desc></vocal> </u>
191 <u who="#interviewerY07i007"> So erm, when did you each of you last go on holiday? </u>
192 <u who="#informantY07i007a"> Er </u>
193 <u who="#informantY07i007b"> Wha- wha- what's a holiday? Like abroad or not? </u>
194 <u who="#interviewerY07i007"> Well it doesn't matter. Anytime you got away,
194 <incident><desc>interruption</desc></incident> like. </u>
195 <u who="#informantY07i007b"> <incident><desc>interruption</desc></incident> That was, erm
195 <anchor xml:id="decten2y07i007ortho0180"/> September or something. Something like that.
195 </u>
196 <u who="#interviewerY07i007"> Yeah? Where did you go? </u>
197 <u who="#informantY07i007b"> I went to Scotland. </u>
198 <u who="#interviewerY07i007"> Ah. That's all right. </u>
199 <u who="#informantY07i007b"> And I saw a goat on a podium. </u>
200 <u who="#interviewerY07i007"> <vocal><desc>laughter</desc></vocal> You what? </u>
201 <u who="#informantY07i007b"> It's like, there was this -- it was like, this little stand thing and
201 the goat was on top of it, and it was like it was on a podium. </u>
202 <u who="#interviewerY07i007"> <vocal><desc>laughter</desc></vocal> Why was it up there
202 start with? <vocal><desc>laughter</desc></vocal> </u>
203 <u who="#informantY07i007b"> <anchor xml:id="decten2y07i007ortho0200"/> Because it was
203 in like a little children's zoo bit. </u>

```

decten2y07i007.xml



## Mark-up/Tagging



< Text Encoding Initiative >

## TEI (Text Encoding Initiative) – guidelines for XML markup

### Jane Loraine’s Recipe Book (c.1684)

*Manuscript, Print, Digital* module

MA students, SELLL Newcastle 2017

<http://janelorraine-recipe.ncl.ac.uk>

```

• RECIPE_BOOK_2017_Final.xml x
TEI text group text body div div p
1355 reason="illegible" resp="#AKMH2521">call</unclear> as will
1355 <choice><orig>coler</orig><reg>colour</reg></choice> it deep it must be <unclear
1355 reason="illegible" resp="#AKMH2521">brused</unclear> small</p>
1356 </div>
1357 <div type="recipe" n="11" resp="#AKMH2521">
1358 <head><ref xml:id="recipe-11-text">11</ref>
1358 <choice><orig>Egge</orig><reg>Egg</reg></choice>
1358 <choice><orig>creame</orig><reg>cream</reg></choice></head>
1359 <p>Take a quart<ptr target="#JWM6885-n4"/> of
1359 <choice><orig>creame</orig><reg>cream</reg></choice> and
1359 <choice><orig>boyle</orig><reg>boil</reg></choice> it up: that
1359 <choice><orig>haue</orig><reg>have</reg></choice> 4 whites of<lb
1359 type="word_end"/><choice><orig>Eggs</orig><reg>eggs</reg></choice> well beaten: with 3
1359 <choice><orig>spounfulls</orig><reg>spoonfuls</reg></choice> of rose water<ptr
1359 target="#LCD7990-n2"/>; when <expan><ex>th</ex>e</expan> cream<lb type="word_end"/>is
1359 <choice><orig>boyled</orig><reg>boiled</reg></choice> take it <expan>of<ex>f</ex></expan>
1359 the fire: when it is a little <choice><orig>coule</orig><reg>cool</reg></choice>
1359 <choice><orig>stirre</orig><reg>stir</reg></choice><lb type="word_end"/>in your
1359 <choice><orig>Eggs</orig><reg>eggs</reg></choice> so
1359 <choice><orig>serue</orig><reg>serve</reg></choice> it up</p>
1360 </div>

```





## Mark-up/Tagging



< Text Encoding Initiative >

### **TEI (Text Encoding Initiative)** – guidelines for XML markup

The TEI P5 guidelines define XML markup tags that allow us to label / encode various features of a text (i.e. embed information about them within the text), for example:

- **Metadata** (authorship, provenance, encoding rationale...)
- **Physical features** (paper, condition, binding...)
- **Structure** (title, chapter, paragraph, line...)
- **Presentation** (italics, underlined, centered...)
- **Linguistic** (parts of speech)
- **Editorial** (additions, deletions, marginalia, corrections...)
- **Context** (named entities, dates, references...)



## Mark-up/Tagging



< Text Encoding Initiative >

## TEI (Text Encoding Initiative) – guidelines for XML markup

Features of TEI XML tags ('elements'):

- Tags come in pairs: a **<start tag>** and a **</closing tag>**
  - The Old English poem **<title>Beowulf</title>**
  - **<p>**This is a very short paragraph.**</p>**
- Tags belong to the words/text they are marking
  - an example of **<hi>tmesis</hi>**, such as absobloodylutely
- Tags can nest within other tags: embedding
  - **<head><q>**A week is a long time in politics**</q></head>**
- Tags can also encode **attributes** (and **values** for those attributes)
  - **<hi rend="italic">**tmesis**</hi>**



**TEI (Text Encoding Initiative)** – guidelines for XML markup

A well-formed TEI XML file consists of: a **<teiHeader>** section and a **<text>** section:

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <!--metadata about the text--!>
  </teiHeader>
  <text>
    <!--the text--!>
  </text>
</TEI>
```



## Mark-up/Tagging



< Text Encoding Initiative >

## TEI (Text Encoding Initiative) – guidelines for XML markup

Example of a minimal <teiHeader> from the TEI P5 Guidelines:

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>Thomas Paine: Common sense, a machine-readable transcript</title>
      <respStmt>
        <resp>compiled by</resp> <name>Jon K Adams</name>
      </respStmt>
    </titleStmt>
    <publicationStmt><distributor>Oxford Text Archive</distributor></publicationStmt>
    <sourceDesc>
      <bibl>The complete writings of Thomas Paine, collected and edited
        by Phillip S. Foner (New York, Citadel Press, 1945)</bibl>
    </sourceDesc>
  </fileDesc>
</teiHeader>
```





## Mark-up/Tagging



< Text Encoding Initiative >

### TEI (Text Encoding Initiative) – guidelines for XML markup

A well-formed `<text>` element will have the following structure  
(*necessary* and *optional elements*):

```
<text>
  <front>
    <!--e.g: title page, preface, contents, etc--!>
  </front>
  <body>
    <!--main body of the text, e.g: chapters, stanzas, scenes ... recipes, etc--!>
  </body>
  <back>
    <!--e.g: index, appendix, glossary, bibliography, etc--!>
  </back>
</text>
```



## Mark-up/Tagging



< Text Encoding Initiative >

### **TEI (Text Encoding Initiative)** – guidelines for XML markup

The elements that can be used to encode the text in the <text> section can be divided into the following four categories:

- **Structural**  
e.g. divisions, chapters, lists, headings, paragraphs, tables, line groups, lines, etc.
- **Presentational**  
e.g. typographic features like bold, italics, small case, indentations, etc.
- **Contextual**  
e.g. identification of names, titles, places, languages, emphasis, etc.
- **Editorial/Analytic**  
e.g. annotation, explication, correction, normalization, etc.



### TEI (Text Encoding Initiative) – guidelines for XML markup

The elements that can be used to encode the text in the <text> section can be divided into the following four categories:

- **Structural**

e.g. divisions, chapters, lists, headings, paragraphs, tables, line groups, lines, etc.

- Some elements have their own dedicated tags  
e.g. <p> paragraph, <l> line, <head> heading, <table>, <list>
- Others can be encoded using **attributes** with appropriate **value** terms  
e.g. <div type="chapter">, <div type="scene">, <div type="letter">
- Tags can contain multiple **attributes**  
e.g. <div type="chapter" n="7">



### TEI (Text Encoding Initiative) – guidelines for XML markup

#### Contextual markup:

‘Just as we can use TEI to represent the structure of our document, we can use it to define and provide contextual information for things mentioned in the text, like **named entities**, **dates**, **geographic features**, and interpretive information like **themes** or **keywords**.’

- *Details of any named entity can be recorded in a list (an ‘ography’) e.g. in the <back> of the <text type="editorial\_introduction">*
  - `<person xml:id="JL"><!--details about Jane Loraine--!></person>`
- *References in the main text can then be linked to the information in the ‘ography’.*
  - `<persName ref="#JL">Jane Loraine</persName>`





## Mark-up/Tagging

### Part-of-Speech (POS) Tagging

- **TagAnt** (Laurence Anthony)  
<http://www.laurenceanthony.net/software/tagant>

uses *TreeTagger* (Schmid – [link](#))

Dutch, English, French, German,  
Italian, Spanish

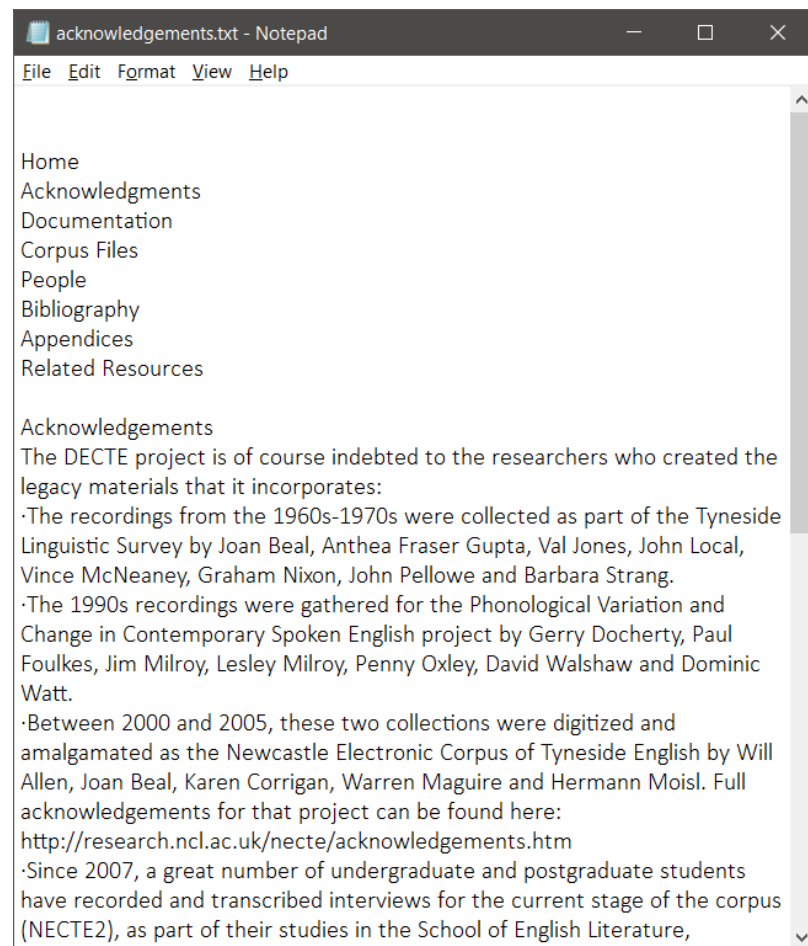
Windows, Mac, Linux; freeware

- **#LancsBox**  
<http://corpora.lancs.ac.uk/lancsbox>

also uses *TreeTagger*

English, Chinese; Arabic, Catalan, Czech, Danish, Dutch, Finnish, French, German, Italian, Korean,  
Latin, Mongolian, Portuguese, Romanian, Russian, Slovak, Slovenian, Spanish, Swahili

Windows, Mac, Linux; freeware [automatically POS-tags imported corpora]





## Mark-up/Tagging

### Part-of-Speech (POS) Tagging

- **TagAnt** (Laurence Anthony)  
<http://www.laurenceanthony.net/software/tagant>

uses *TreeTagger* (Schmid – [link](#))

Dutch, English, French, German, Italian, Spanish

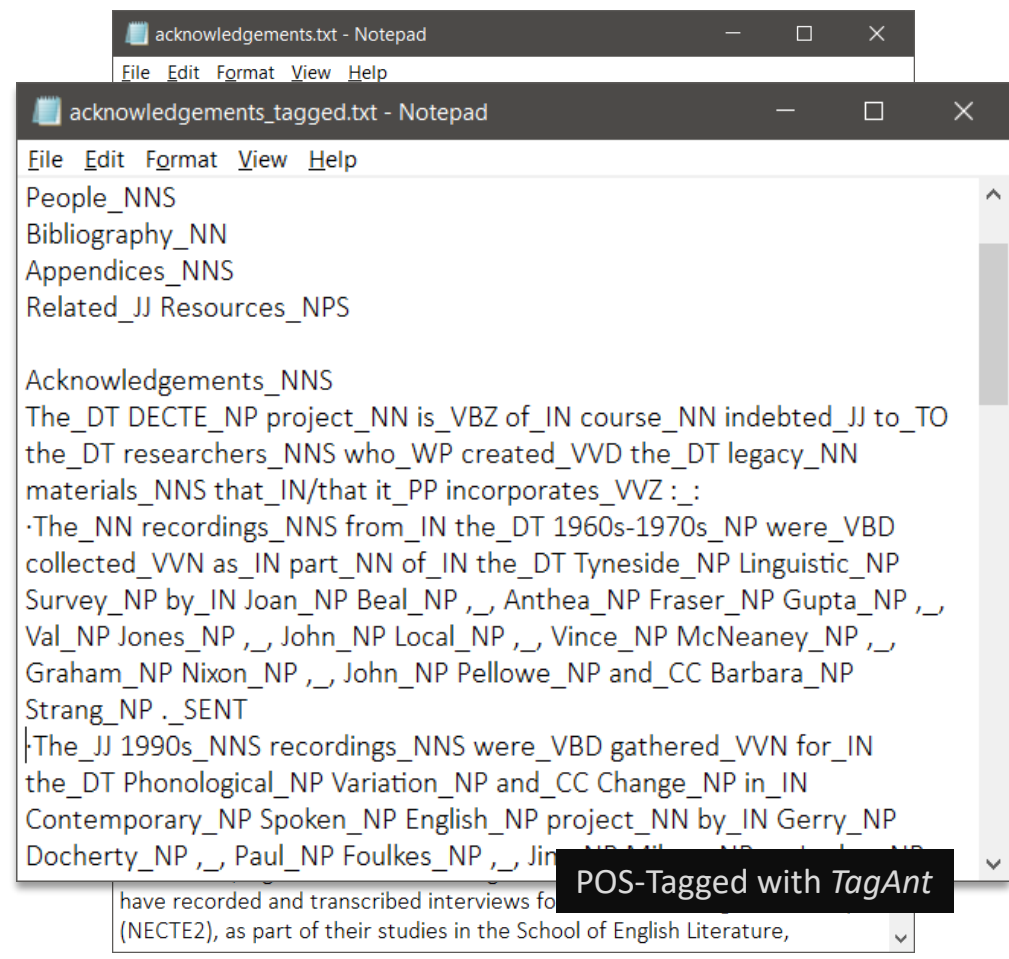
Windows, Mac, Linux; freeware

- **#LancsBox**  
<http://corpora.lancs.ac.uk/lancsbox>

also uses *TreeTagger*

English, Chinese; Arabic, Catalan, Czech, Danish, Dutch, Finnish, French, German, Italian, Korean, Latin, Mongolian, Portuguese, Romanian, Russian, Slovak, Slovenian, Spanish, Swahili

Windows, Mac, Linux; freeware [automatically POS-tags imported corpora]



## Text Editing

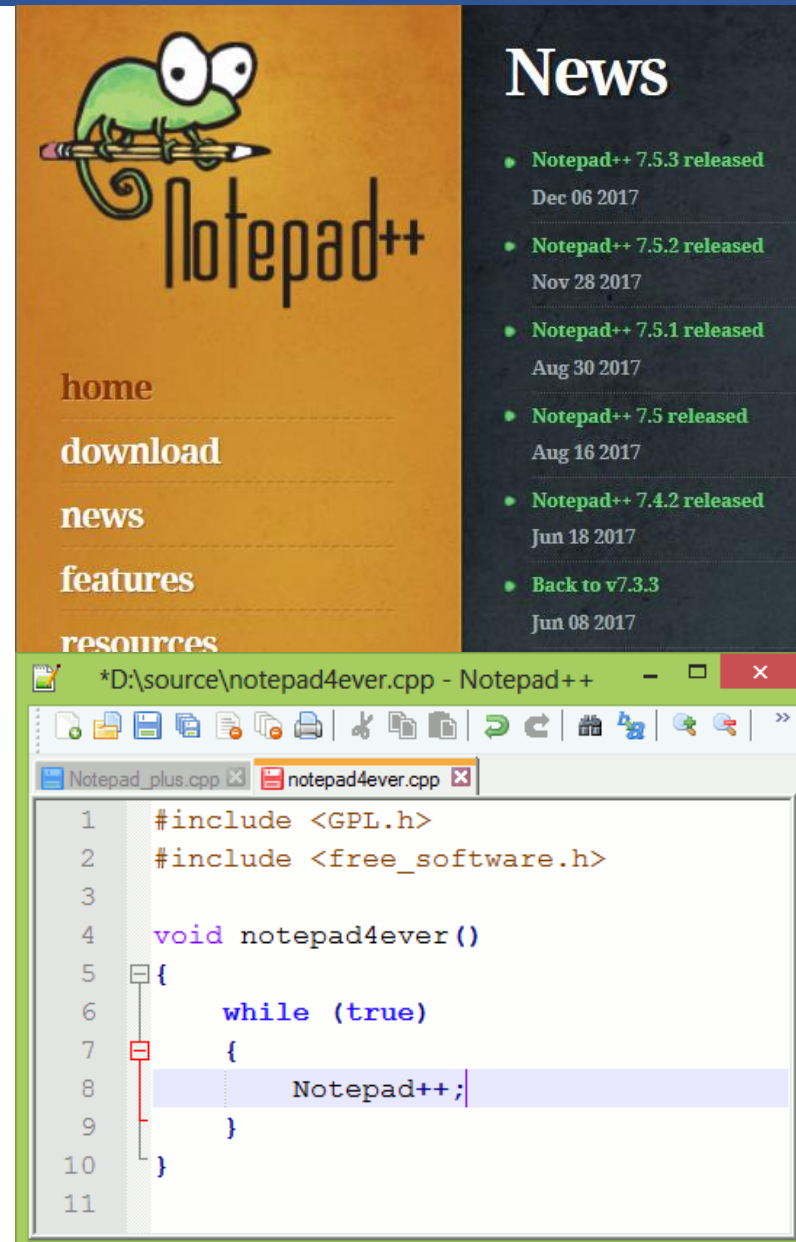
Batch find/replace operations:  
straightforward substitutions  
or using 'regular expressions'

- Text Editors – e.g. *Notepad++* (Don Ho)  
<https://notepad-plus-plus.org>

Windows; freeware

Portable version at *PortableApps.com*:  
[https://portableapps.com/apps/development/notepadpp\\_portable](https://portableapps.com/apps/development/notepadpp_portable)

?? Installing Notepad++ on a Mac:  
<https://mike.kronenberg.org/winebottler-how-to-install-notepad-on-a-mac>

The image shows two parts of the Notepad++ ecosystem. The top part is a screenshot of the Notepad++ website, which has an orange background. It features a green cartoon frog sitting on a pencil, with the text 'Notepad++' in a large, stylized font. Below the logo are navigation links: 'home', 'download', 'news', 'features', and 'resources'. To the right of the website is a 'News' section with a dark background, listing several release updates with dates, such as 'Notepad++ 7.5.3 released Dec 06 2017'. The bottom part of the image is a screenshot of the Notepad++ code editor. The title bar shows the file path '\*D:\source\notepad4ever.cpp - Notepad++'. The code is written in C++ and includes headers for 'GPL.h' and 'free\_software.h'. A function named 'notepad4ever()' is defined, which contains a 'while (true)' loop. Inside the loop, the text 'Notepad++;' is printed. The code is displayed in a light-colored font on a dark background, with line numbers 1 through 11 visible on the left side.

## Text Editing

Batch find/replace operations:  
straightforward substitutions  
or using ‘regular expressions’

- Regular Expressions
  - general/‘underspecified’ or coded expressions
  - intended to find character string patterns

### *RegEx Guides / Tutorials*

<http://www.regular-expressions.info>

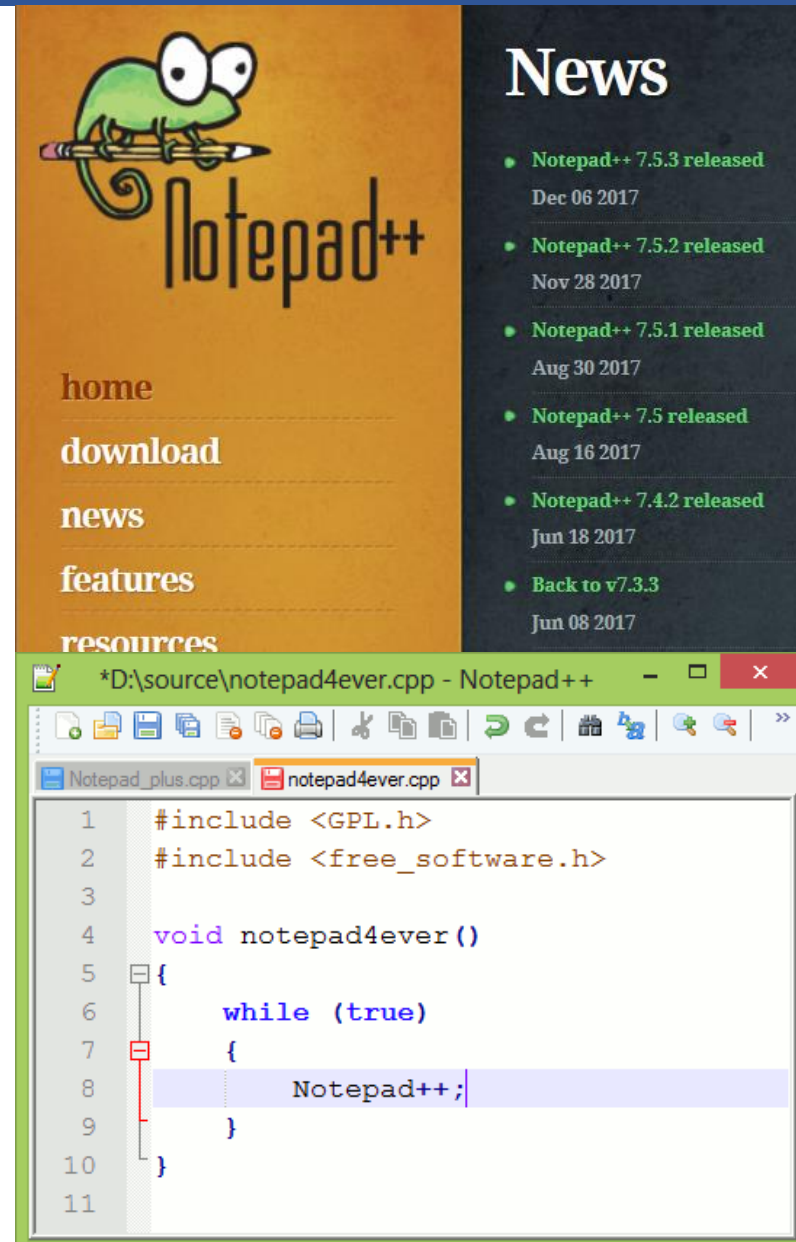
<http://www.rexegg.com>

<https://www.regexbuddy.com>

<https://regexone.com>

[http://docs.notepad-plus-plus.org/index.php/Regular\\_Expressions](http://docs.notepad-plus-plus.org/index.php/Regular_Expressions)

<http://www.zytrax.com/tech/web/regex.htm>

The image shows a composite of two screenshots. The top screenshot is the Notepad++ website homepage, featuring a green cartoon frog on a pencil and a navigation menu with links for 'home', 'download', 'news', 'features', and 'resources'. The 'news' link is highlighted. The right side of the website displays a 'News' section with a list of recent releases: 'Notepad++ 7.5.3 released' (Dec 06 2017), 'Notepad++ 7.5.2 released' (Nov 28 2017), 'Notepad++ 7.5.1 released' (Aug 30 2017), 'Notepad++ 7.5 released' (Aug 16 2017), 'Notepad++ 7.4.2 released' (Jun 18 2017), and 'Back to v7.3.3' (Jun 08 2017). The bottom screenshot shows the Notepad++ application window editing a C++ file named 'notepad4ever.cpp'. The code includes headers for 'GPL.h' and 'free\_software.h', and defines a 'notepad4ever()' function with a 'while (true)' loop. The current line of code is 'Notepad++;'.

home  
download  
news  
features  
resources

**News**

- Notepad++ 7.5.3 released  
Dec 06 2017
- Notepad++ 7.5.2 released  
Nov 28 2017
- Notepad++ 7.5.1 released  
Aug 30 2017
- Notepad++ 7.5 released  
Aug 16 2017
- Notepad++ 7.4.2 released  
Jun 18 2017
- Back to v7.3.3  
Jun 08 2017

```
1  #include <GPL.h>
2  #include <free_software.h>
3
4  void notepad4ever ()
5  {
6      while (true)
7      {
8          Notepad++;
9      }
10 }
11
```





# Text Editing – Regular Expressions: Examples

Find in Files

Find Replace Find in Files Mark

Find what: `(([a-z]+?)ing \\(\\(\\+R\\+\\1in\\)\\)`

Replace with: `\\1ing {*R=N*}`

Filters: `*.*`

Directory: `Q:\\DECTE_Files\\DECTE_Corpus_Files\\NECTE2\\Corpu`

Match whole word only

Match case

Search Mode

Normal

Extended (\\n, \\r, \\t, \\0, \\x...)

Regular expression  matches newline

Follow current doc.

In all sub-folders

In hidden folders

Transparency

On losing focus

Always

Buttons: Find All, Replace in Files, Close



ear} [/u]  
 ops were playing {\*R=N\*} a bigger part like you have l

Search "`(([a-z]+?)ing \\(\\(\\+R\\+\\1in\\)\\)`" (231 hits in 26 files)

`Q:\\DECTE_Files\\DECTE_Corpus_Files\\NECTE2\\Corpus_N2\\NECTE2_2007-2017\\02_Interview_Transcripts\\Interview_Trans`

`Q:\\DECTE_Files\\DECTE_Corpus_Files\\NECTE2\\Corpus_N2\\NECTE2_2007-2017\\02_Interview_Transcripts\\Interview_Trans`  
 Line 25: [2015/CM/1820] well it's a pain because you `((+R+yu))` leave all of your `((+R+u'yu))` friends at the time and obvious

`Q:\\DECTE_Files\\DECTE_Corpus_Files\\NECTE2\\Corpus_N2\\NECTE2_2007-2017\\02_Interview_Transcripts\\Interview_Trans`  
 Line 232: [2015/CM/1820] yeah like I think at school as well like social groups were `playing ((+R+playin))` a bigger part like

`Q:\\DECTE_Files\\DECTE_Corpus_Files\\NECTE2\\Corpus_N2\\NECTE2_2007-2017\\02_Interview_Transcripts\\Interview_Trans`

`Q:\\DECTE_Files\\DECTE_Corpus_Files\\NECTE2\\Corpus_N2\\NECTE2_2007-2017\\02_Interview_Transcripts\\Interview_Trans`  
 Line 230: [2015/IN/IP/8115] I was `trying ((+R+tryin))` to tell someone what a lemon top was the other day and they {uncle

`Q:\\DECTE_Files\\DECTE_Corpus_Files\\NECTE2\\Corpus_N2\\NECTE2_2007-2017\\02_Interview_Transcripts\\Interview_Trans`  
 Line 276: [2015/IN/IP/8115] My medic friends are `trying ((+R+tryin))` to pick hospitals now and one of them is `((+R+thems`

`Q:\\DECTE_Files\\DECTE_Corpus_Files\\NECTE2\\Corpus_N2\\NECTE2_2007-2017\\02_Interview_Transcripts\\Interview_Trans`

`Q:\\DECTE_Files\\DECTE_Corpus_Files\\NECTE2\\Corpus_N2\\NECTE2_2007-2017\\02_Interview_Transcripts\\Interview_Trans`  
 Line 28: [2015/ED/3580] and there's `((+ST+there are))` loads of people but no I wouldn't say I have like that level of friend



# Text Editing – Regular Expressions: Examples

FIND: `\\(\\(\\+comm\\+[ ]){0,1}([^\(]*?)\\)\\)`

Search "\\(\\(\\+comm\\+[ ]){0,1}([^\(]\*?)\\)\\)" (704 hits in 131 files)

Q:\DECTE\_Files\DECTE\_Corpus\_Files\NECTE2\Corpus\_N2\NECTE2\_2007-2017\02\_Interview\_Transcripts\Interview\_Transcripts--TXT\_(processed)\2012-2017\2013\_SEL2091\_027--Tra  
Line 49: [2013/CH/9281] The drinking {R=N\*} games the tequila shots everything {R=N\*} just get it out of ((realised outta)) our system 'cause we're meeting {R=N\*} some girls from  
Line 76: [2013/CH/9281] Aye mates mingin like she's not very nice ((+comm+ background noise)) [/u]  
Line 146: [2013/CH/9281] But if the gay persons a {INT} prick then it doesn't matter {INT} ((+comm+ phone signal interruption starts)) [/u]  
Line 153: [2013/OE/9281] Yeah {pause} I-well I have before but I don't think he m-knew I was being serious when I said it ((+comm+ phone interruption stops)) [/u]  
Line 173: [2013/OE/9281] Yeah but only 'cause your shagging {R=N\*} guys you {INT} fucking {R=N\*} ((+comm+ background noise)) [/u]  
Line 243: [2013/CH/9281] Aye B-Basshunter was mint like he was mint he was class {pause} but aye sweatiest night of my life {INT} got started on by the black fella in the er-toilets a

Q:\DECTE\_Files\DECTE\_Corpus\_Files\NECTE2\Corpus\_N2\NECTE2\_2007-2017\02\_Interview\_Transcripts\Interview\_Transcripts--TXT\_(processed)\2012-2017\2014\_SEL2091\_003--Tra  
Line 326: [2014/RC/4318] but obviously out there you get the real stuff bananas were about ((+comm+measuring length out with hands for illustration)) that big what ten centimetr  
Line 755: [2014/RC/4318] yeah I'm nothing {R=NG\*} like ((+comm+gesturing to interviewer)) [/u]

Q:\DECTE\_Files\DECTE\_Corpus\_Files\NECTE2\Corpus\_N2\NECTE2\_2007-2017\02\_Interview\_Transcripts\Interview\_Transcripts--TXT\_(processed)\2012-2017\2014\_SEL2091\_004--Tra  
Line 170: [2014/IN/KB/7215] {!!laugh!} {pause} what was your favourite subject at school ((+comm+interference from informant's foot on recording surface)) [/u]  
Line 225: [2014/IN/KB/7215] {!!laugh!} ((+comm+interference from informant's foot on recording surface)) did you ever have {R=Hdrop\*} any dodgy ones [/u]  
Line 254: [2014/RH/7215] and then in secondary school one one time I remember like walking {R=N\*} to school and thinking {R=N\*} this is ((+comm+interference from informant's  
Line 486: [2014/MJ/7215] ((+comm+informant makes hover craft noise)) ehm there's you see that it like going {R=N\*} along as if it's a massive model and then suddenly it'll go all a  
Line 486: [2014/MJ/7215] ((+comm+informant makes hover craft noise)) ehm there's you see that it like going {R=N\*} along as if it's a massive model and then suddenly it'll go all a

Q:\DECTE\_Files\DECTE\_Corpus\_Files\NECTE2\Corpus\_N2\NECTE2\_2007-2017\02\_Interview\_Transcripts\Interview\_Transcripts--TXT\_(processed)\2012-2017\2014\_SEL2091\_005--Tra  
Line 259: [2014/OP/RE/6264] ((+comm+this is a voice on a video)) go [/u]  
Line 260: [2014/OP/AF/6264] ((+comm+this is a voice on a video)) hi {#}Bec thank you for supporting {R=N\*} me love {unclear} [/u]

Q:\DECTE\_Files\DECTE\_Corpus\_Files\NECTE2\Corpus\_N2\NECTE2\_2007-2017\02\_Interview\_Transcripts\Interview\_Transcripts--TXT\_(processed)\2012-2017\2014\_SEL2091\_007--Tra  
Line 397: [2014/SK/4583] seventeenth {unclear} ((+comm+ informant's phone vibrate goes off)) [/u]  
Line 425: [2014/SK/4583] {unclear} ((+comm+informants overlapping)) [/u]

Q:\DECTE\_Files\DECTE\_Corpus\_Files\NECTE2\Corpus\_N2\NECTE2\_2007-2017\02\_Interview\_Transcripts\Interview\_Transcripts--TXT\_(processed)\2012-2017\2014\_SEL2091\_012--Tra  
Line 172: [2014/AS/3956] oh I got a skiing {R=N\*} one and sewing {R=N\*} not that I can sew and I think it was like cooking {R=NG\*} ((+comm+informant's mobile goes off)) and stu  
Line 275: [2014/AS/3956] I used to go and visit one of my best friends she did ehm physiotherapy at Leeds University and whenever I used to go down there it was very much like ki  
Line 292: [2014/AI/3956] honestly I just I do so many dumb ass things when I'm drunk I don't do anything {R=N\*} massively regrettable I don't think unless like I've said something

Q:\DECTE\_Files\DECTE\_Corpus\_Files\NECTE2\Corpus\_N2\NECTE2\_2007-2017\02\_Interview\_Transcripts\Interview\_Transcripts--TXT\_(processed)\2012-2017\2014\_SEL2091\_013--Tra  
Line 35: [2014/GD/4936] oh my God I love it you got to {R=gotta\*} sleepover ((+comm+informant sings this)) I love it {INT} it's still on now [/u]

Q:\DECTE\_Files\DECTE\_Corpus\_Files\NECTE2\Corpus\_N2\NECTE2\_2007-2017\02\_Interview\_Transcripts\Interview\_Transcripts--TXT\_(processed)\2012-2017\2014\_SEL2091\_017--Tra

((+comm+ background noise)) ((+comm+informant makes hover craft noise))



# Text Editing – Regular Expressions: Examples

FIND:

`(([a-z]*?)(o[uw]))([a-z]*?) \\(\\([\\+]{0,1}R[\\+]{0,1}[ ]{0,1}\\1oo\\3\\)\\)`

```
Search "[a-z]*?(o[uw])[a-z]*? \\(\\([\\+]{0,1}R[\\+]{0,1}[ ]{0,1}\\1oo\\3\\)\\)" (83 hits in 11 files)
Q:\DECTE_Files\DECTE_Corpus_Files\NECTE2\Corpus_N2\NECTE2_2007-2017\02_Interview_Transcripts\Interview_Transcripts--TXT_(processed)\2012-2017\2012
Line 326: [2014/DT/4870] right out ((+R+oot)) {INT} legend [/u]
Q:\DECTE_Files\DECTE_Corpus_Files\NECTE2\Corpus_N2\NECTE2_2007-2017\02_Interview_Transcripts\Interview_Transcripts--TXT_(processed)\2012-2017\2012
Line 280: [2014/MR/2981] {!laugh!} I've played him a thousand ((+R+thoosand)) times {!/laugh!} [/u]
Line 286: [2014/MR/2981] yeah we did and one day I was at th- at the baths and we were stood by this horrible woman who used to take us for swimming {*R=
Q:\DECTE_Files\DECTE_Corpus_Files\NECTE2\Corpus_N2\NECTE2_2007-2017\02_Interview_Transcripts\Interview_Transcripts--TXT_(processed)\2012-2017\2012
Line 278: [2014/AC/5982] nah I'll do my {*R=me*} get my {*R=me*} degree out ((+R+oot)) the way and then er and then see what it's like but nor it's good I like
Line 508: [2014/AC/5982] I'll ((+R+al)) be about ((+R+about)) twenty-four {INT} when I finish [/u]
Line 611: [2014/AC/5982] he {*R=Hdrop*} was only about ((+R+about)) like my height though wasn't ((+R+wan)) he {*R=Hdrop*} [/u]
Q:\DECTE_Files\DECTE_Corpus_Files\NECTE2\Corpus_N2\NECTE2_2007-2017\02_Interview_Transcripts\Interview_Transcripts--TXT_(processed)\2012-2017\2012
Line 22: [2014/RD/2728] but they would put bits in instead ((+R+instead)) of ((+R+uh)) you didn't sit and if there was a bit missing {*R=N*} yee didn't sit and kni
Line 58: [2014/RD/2728] well it's not ((+R+no)) it's not ((+R+no)) tourists it's the ones that's ((+ST+that have)) come here to brought all the houses ((+R+hooses))
Line 60: [2014/RD/2728] the locals cannot {*R=cannit*} get a house ((+R+hoose)) now [/u]
Line 69: [2014/RW/2728] and there's ((+ST+there are)) no jobs nat {INT} not unless they're cleaning {*R=N*} holiday houses ((+R+hooses)) {!/laugh!} [/u]
Line 136: [2014/RD/2728] it's died out ((+R+oot)) and aall now {!/laugh!} [/u]
Line 173: [2014/RD/2728] that's if he happened to get ower and get drowned {*R=drooned*} and they found ((+R+foond)) his body {!/laugh!} [/u]
Line 201: [2014/RD/2728] aye used to cut the feet cut the feet off the bairns socks when when they were worn out ((+R+oot)) and and keep {!/laugh!} their you k
Line 217: [2014/RW/2728] when the when we first got the oil skin trousers ((+R+troosers)) you know wi the {#}Trot got them and {!/laugh!} he wasn't impressed
Line 237: [2014/RW/2728] we used to go to ((+R+tu)) the North East Bank which was about ((+R+about)) five hours away [/u]
Line 252: [2014/RD/2728] well it was a plastic cup {!/cough!} a little toy camel they were gutting {*R=N*} the fish and the fish had swallowed the plastic toy the c
Line 252: [2014/RD/2728] well it was a plastic cup {!/cough!} a little toy camel they were gutting {*R=N*} the fish and the fish had swallowed the plastic toy the c
Line 252: [2014/RD/2728] well it was a plastic cup {!/cough!} a little toy camel they were gutting {*R=N*} the fish and the fish had swallowed the plastic toy the c
Line 258: [2014/RW/2728] aye {#}Roma will ((+R+Roma'll)) still have it in the house ((+R+hoose)) Aa bet {!/laugh!} [/u]
Line 280: [2014/RD/2728] well yes Aa watched it a bit but er Aa mean the ones that's coming {*R=N*} on now most of them have ((+R+thems)) been on but er t
Line 282: [2014/RW/2728] oh that's er that's a bit of acting {*R=NG*} that and when the when the you see the water {INT} on the wheelhouse ((+R+wheelhoose))
Line 310: [2014/RD/2728] in Newfoundland for of course that's into Canada farther north again they did an awful ((+R+aawful)) lot of {*R=lotta*} cod and herring
```

house ((+R+hoose))    out ((+R+oot))    found ((+R+foond))

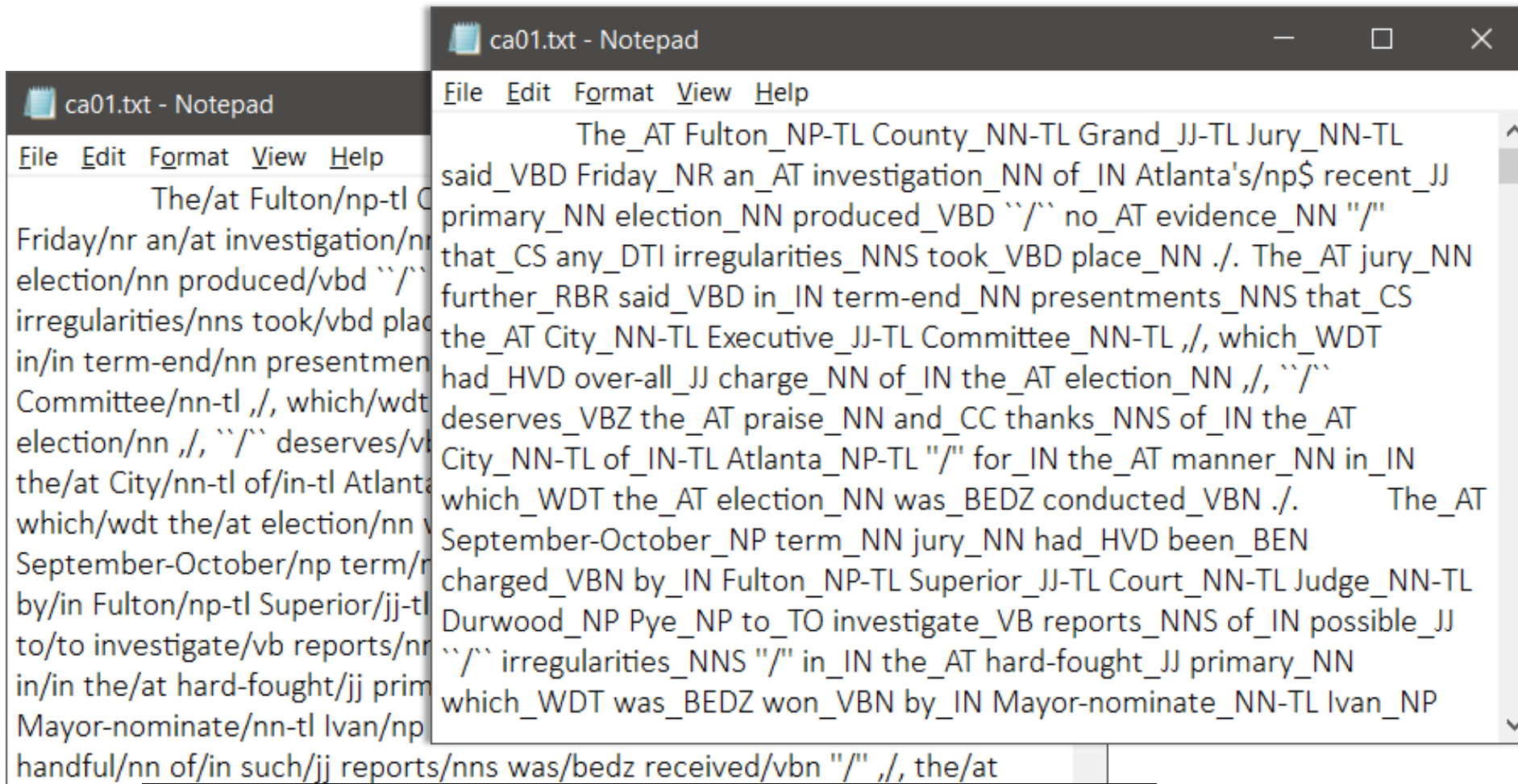




# Text Editing – Regular Expressions: Examples

FIND: `/([a-z\ -]+?)`

REPLACE: `_\U\1`



Brown Corpus (Francis and Kucera 1964; <https://archive.org/details/BrownCorpus>)



## Text Analysis Software

- **AntConc** (Anthony; <http://www.laurenceanthony.net/software/antconc>)  
KWIC Concordances; Wordlists; Collocates; Clusters/N-grams; Keywords  
Windows, Mac, Linux; freeware
- **#LancsBox** (Brezina, McEnery, Wattam; <http://corpora.lancs.ac.uk/lancsbox>)  
KWIC Concordances; 'Whelk' and 'Words' (freq, dispersion, keywords);  
GraphColl (listing and visualization of collocations and colligations)  
Windows, Mac, Linux; freeware
- **Wordsmith Tools** (Scott; <http://www.lexically.net/wordsmith>)  
Concord (KWIC Concordances); Wordlist; Keywords; *utility programs*  
Windows; single user licence £50





# Corpus Analysis: a few examples

## Wordsmith Tools

ed) — □ ×

Compute Settings Windows Help

N

	Overall	D-N0=	D-N1
text file			
file size	97,576,184	76,840	71,100
tokens (running words) in text	4,136,576	4,268	3,670
tokens used for word list	4,129,082	4,260	3,670
sum of entries			
types (distinct words)	46,283	803	66
type/token ratio (TTR)	1.12	18.85	17.9
standardised TTR	32.87	31.15	28.6
STTR std.dev.	67.34	55.43	54.3
STTR basis	1,000	1,000	1,000
mean word length (in characters)	3.91	3.85	3.8
word length std.dev.	2.00	1.91	1.8
sentences	87,832	1	
mean (in words)	47.01	4,260	3,670
std.dev.	259.56		
paragraphs	1,106	1	
mean (in words)	3,733.35	4,260	3,670
std.dev.	1,618.59		
headings			

statistics filenames notes

0% T S

Word list (unsaved) — □ ×

File Edit View Compute Settings Windows Help

N	Word	Freq.	%	Texts	%	Dispersion
1	I	169,474	4.10	1,106	100.00	0.96
2	AND	128,024	3.09	1,106	100.00	0.97
3	THE	120,896	2.92	1,106	100.00	0.96
4	LIKE	102,247	2.47	1,104	99.82	0.93
5	TO	94,620	2.29	1,106	100.00	0.97
6	A	85,994	2.08	1,106	100.00	0.98
7	YOU	77,645	1.88	1,105	99.91	0.92
8	IT	76,768	1.86	1,106	100.00	0.98
9	WAS	74,228	1.79	1,105	99.91	0.96
10	THAT	53,065	1.28	1,106	100.00	0.97
11	OF	51,348	1.24	1,106	100.00	0.97
12	IN	49,638	1.20	1,106	100.00	0.98
13	YEAH	48,478	1.17	1,050	94.94	0.90
14	BUT	43,920	1.06	1,106	100.00	0.97
15	AS	42,305	1.02	1,103	99.73	0.72
16	JUST	40,151	0.97	1,104	99.82	0.96
17	WE	35,343	0.85	1,101	99.55	0.95
18	IT'S	34,216	0.83	1,099	99.37	0.98
19	SO	33,819	0.82	1,105	99.91	0.94
20	KNOW	33,435	0.81	1,101	99.55	0.86
21	MY	29,951	0.72	1,103	99.73	0.97
22	ON	28,342	0.69	1,104	99.82	0.98
23	REALIZED	27,425	0.66	557	50.36	0.55

frequency alphabetical statistics filenames notes

46,283 entries Row 4 0% T S

# Corpus Analysis: a few examples



## Wordsmith Tools

Concord

File Edit View Compute Settings Windows Help

Concordance

N	Concordance	Set Tag	Wor
1	is er really the real broad isn't it Geordie hinny that's the real Tyneside		4,3
2	some of them are still Tyneside broad Geordie but it'll go on for years and		4,8
3	and then you could let yourself go into Geordie your voice might sound nicer		5,0
4	give us any aye yes no howay man Geordie never takes us to the races		5,4
5	I do at times I forget myself eh the Geordie language I got wrong for that		3,9
6	if I get myself annoyed I just let the Geordie language fly and that's it I		4,0
7	like sh- shortening sort of style in your Geordie language you know eh to talk		4,7
8	I mean round here eh this is a real Geordie accent along this particular		4,2
9	eh yes Gateshead's a lot more broad Geordie than Newcastle is I think so if		4,3
10	that kind of language n- not the strong Geordie but eh now th- I mean down		3,6
11	where they speak the real strong Geordie you know but I do change my		3,7
12	she she has oh oh well course I'm a Geordie aren't I oh yes she does I		4,1
13	came from you don't have to be pure Geordie for people to say where you		4,2
14	from mm no I don't disapprove of Geordie accents as as you're speaking		4,3
15	never heard of that one don't think Geordie mm squify or one over the		1,4
16	pay so much tax I don't like the true Geordie no this "ganen yem" and things		3,6
17	time I think uh-huh I never go broad Geordie and I never go oh so posh		4,0
18	I mean language is going out I think Geordie is completely different to what		4,6
19	go on you know doesn't talk eh Geordie but a person working in a		4,1
20	a lapse it's eh it's just accepted isn't it Geordie eh you like it or you don't and		4,7
21	a there's a book on Larn Yersel Geordie uh-huh I'll tell you what I'm		23
22	I hadn't been allowed to speak Geordie all the time myself when I was		3,7
23	Geordie and I wouldn't let them speak Geordie all the time I don't know why		3,7
24	light please today ehm well eh it's all Geordie all the questions you were		5,5

concordance collocates plot patterns clusters timeline filenames source text notes

1,461 entries Row 1 0% T S



# Corpus Analysis: a few examples

## Wordsmith Tools

Concord

File Edit View Compute Settings Windows Help

N	Word	Set	Texts	Total	Total Left	Total Right	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	GEORDIE		418	1,56	51	51	12	18	8	6	7	1,465	7	6	8	18	12
2	A		277	707	597	110	37	44	64	105	347		3	25	24	26	32
3	I		234	583	212	371	73	69	59	11			42	121	82	65	61
4	LIKE		213	494	279	215	49	55	53	80	42		27	51	43	48	46
5	THE		200	405	280	125	30	27	28	37	158		12	25	34	22	32
6	ACCENT		181	337	28	309	7	4	14	2	1		285	1	6	8	9
7	AND		171	312	107	205	32	29	18	15	13		68	62	20	36	19
8	YOU		166	295	141	154	50	41	28	21	1		26	35	35	29	29
9	BUT		139	228	53	175	21	18	7	6	1		73	51	20	16	15
10	TO		142	212	121	91	29	29	37	20	6		3	9	28	24	27
11	OF		133	194	135	59	21	21	18	37	38		1	10	17	16	15
12	IT'S		92	144	61	83	19	14	11	16	1		15	20	14	15	19
13	THINK		99	144	64	80	21	14	17	10	2			12	21	22	25
14	DON'T		92	143	70	73	18	24	21	7			2	10	31	18	12
15	THAT		99	139	75	64	17	14	22	14	8		3	25	10	10	16
16	NOT		95	135	87	48	7	14	28	33	5		6	10	12	12	8
17	AS		82	132	67	65	8	13	11	21	14		12	11	14	17	11
18	IT		94	130	49	81	8	14	11	10	6		7	15	16	22	21
19	REALLY		86	130	90	40	5	11	18	32	24		3	10	9	10	8
20	I'M		82	128	73	55	13	11	17	27	5		7	13	12	11	12
21	IS		85	122	62	60	10	11	20	15	6		16	21	10	3	10
22	KNOW		87	121	55	66	11	17	13	11	3			16	18	24	8

concordance collocates plot patterns clusters timeline filenames source text notes

694 entries Row 1 0% T S < > GEORDIE



# Corpus Analysis: a few examples

## Wordsmith Tools

Concordance Cluster List (unsaved)

File Edit View Compute Settings Windows Help

N	Cluster	Freq.	Set	Length	Related
1	A GEORDIE ACCENT	94		3	
2	THE GEORDIE ACCENT	87		3	
3	YOU'RE A GEORDIE	37		3	
4	GEORDIE ACCENT BUT	30		3	
5	GEORDIE ACCENT AND	25		3	
6	GEORDIE ACCENT I	23		3	
7	LIKE A GEORDIE	23		3	
8	LIKE THE GEORDIE	23		3	
9	I'M A GEORDIE	21		3	
10	GEORDIE BUT I	19		3	
11	YES REALIZED AS	19		3	
12	WITH A GEORDIE	19		3	
13	STRONG GEORDIE ACCENT	18		3	
14	A GEORDIE AND	18		3	
15	I DON'T KNOW	18		3	
16	NOT A GEORDIE	17		3	
17	GEORDIE ACCENT LIKE	16		3	
18	BUT I DON'T	16		3	
19	GOT A GEORDIE	15		3	
20	OF A GEORDIE	15		3	
21	BROAD GEORDIE ACCENT	15		3	
22	GEORDIE YOU KNOW	14		3	
23	I DON'T THINK	14		3	
24	A GEORDIE BUT	14		3	
25	I DON'T LIKE	13		3	

concordance collocates plot patterns clusters **timeline** filenames source text notes

180 entries Row 1 0% T S < >



# Corpus Analysis: a few examples

#LancsBox

#LancsBox v 3.0.2

KWIC    Whelk    GraphColl    Words    Text

Corpora **KWL...**

**Search Term** Geordie **Occurrences** 177 (2.75) **Texts** 53/161 **Corpus** Corpus 1 **Context** 7 **Display Text**

Index	File	Left	Node	Right
1	decten1pvc14	Tongues everybody knows about it that's a	Geordie	and went to s live in Spital
2	decten1t1sg0'	yes with with the bits of real	Geordie	comes out now and again yes yes
3	decten1t1sg0f	something to somebody they still pick the	Geordie	a lot of people do if you
4	decten1t1sg0f	way or something they say you're a	Geordie	the Newcastle Gateshead area ah no but
5	decten1t1sg1:	really because eh you never notice a	Geordie	accent with living up here ehm I'd
6	decten1t1sg1:	like that get good money naturally their	Geordie	accent's starting to go a bit because
7	decten1t1sg1:	no to me that's more Scottish than	Geordie	no not often kep to catch yeah
8	decten1t1sg1:	aye aye couldn't say well I'm a	Geordie	me like aye aye aye mm oh
9	decten1t1sg1:	Tees for the news and they're talking	Geordie	and it sounds funny you know and
10	decten1t1sg1:	over with a lot of eh like	Geordie	a bit you know but it just
11	decten1t1sg1f	she was Welsh or something and she's	Geordie	find mind fly bill well men head
12	decten1t1sg1f	do ehm I don't really talk very	Geordie	suppose I say it myself you know
13	decten1t1sg1f	you know suppose I thought it was	Geordie	the way I talked but they said
14	decten1t1sg1f	makes me cringe yes I I like	Geordie	songs mind I like to hear them
15	decten1t1sg1f	eh songs but I don't like the	Geordie	voice very few and far between I
16	decten1t1sg1f	my mother we don't talk eh real	Geordie	you never hear we talk you know
17	decten1t1sg1f	now you see her husband talks is	Geordie	and this lady always talks like that
18	decten1t1sg1:	I've been off over twelve month yes	Geordie	no I can't say I do well
19	decten1t1sg1:	he's been well sozzled see it's all	Geordie	slang up here you see fourteen well
20	decten1t1sg1f	understand them but if they put a	Geordie	on the Londoners wouldn't understand him see
21	decten1t1sg2f	up am I giving all the standard	Geordie	answers I don't mind well I wouldn't
22	decten1t1sg2f	that I was worried about having a	Geordie	accent I like accents I think they're
23	decten1t1sg2'	but I wouldn't say you spoke broad	Geordie	yes I think so because people from
24	decten1t1sg2:	Thompson years ago on here on the	Geordie	programme he's never on the telly though
25	decten1t1sg2:	he was what you called a real	Geordie	well I didn't think I was a
26	decten1t1sg2:	I didn't think I was a real	Geordie	until I went on one of these

Filtering complete





# Corpus Analysis: a few examples

#LancsBox

#LancsBox v 3.0.2

KWIC
Whelk
GraphColl
Words
Text

Corpora KWL... X

**Search Term** Geordie **Occurrences** 177 (2.75) **Texts** 53/161 **Corpus** Corpus 1 **Context** 7 **Display Text**

Index	File	Left	Node	Right
1	decten1pvc14	Tongues everybody knows about it that's a	Geordie	and went to s live in Spital
2	decten1t1sg0'	yes with with the bits of real	Geordie	comes out now and again yes yes
3	decten1t1sg0f	something to somebody they still pick the	Geordie	a lot of people do if you
4	decten1t1sg0f	way or something they say you're a	Geordie	the Newcastle Gateshead area ah no but
5	decten1t1sg1;	really because eh you never notice a	Geordie	accent with living up here ehm I'd
6	decten1t1sg1;	like that get good money naturally their	Geordie	accent's starting to go a bit because
7	decten1t1sg1;	no to me that's more Scottish than	Geordie	no not often kep to catch yeah
8	decten1t1sg1;	aye aye couldn't say well I'm a	Geordie	me like aye aye aye mm oh
9	decten1t1sg1;	Tees for the news and they're talking	Geordie	and it sounds funny you know and
10	decten1t1sg1;	over with a lot of eh like	Geordie	a bit you know but it just
11	decten1t1sg1f	she was Welsh or something and she's	Geordie	find mind fly bill well men head
12	decten1t1sg1f	do ehm I don't really talk very	Geordie	suppose I say it myself you know
13	decten1t1sg1f	you know suppose I thought it was	Geordie	the way I talked but they said
14	decten1t1sg1f	makes me cringe yes I I like	Geordie	songs mind I like to hear them
15	decten1t1sg1f	eh songs but I don't like the	Geordie	voice very few and far between I
16	decten1t1sg1f	my mother we don't talk eh real	Geordie	you never hear we talk you know
17	decten1t1sg1f	now you see her husband talks is	Geordie	and this lady always talks like that
18	decten1t1sg1;	I've been off over twelve month yes	Geordie	no I can't say I do well
19	decten1t1sg1;	he's been well sozzled see it's all	Geordie	slang up here you see fourteen well
20	decten1t1sg1f	understand them but if they put a	Geordie	on the Londoners wouldn't understand him see
21	decten1t1sg2f	up am I giving all the standard	Geordie	answers I don't mind well I wouldn't
22	decten1t1sg2f	that I was worried about having a	Geordie	accent I like accents I think they're
23	decten1t1sg2'	but I wouldn't say you spoke broad	Geordie	yes I think so because people from
24	decten1t1sg2;	Thompson years ago on here on the	Geordie	programme he's never on the telly though
25	decten1t1sg2;	he was what you called a real	Geordie	well I didn't think I was a
26	decten1t1sg2'	I didn't think I was a real	Geordie	until I went on one of these

Filtering complete
⌂ | ? | ! | !

# Corpus Analysis: a few examples

#LancsBox

Search Term **Geordie** Occurrences **177** (2.75) Texts 53/161 Corpus Corpus 1 Context 7 Display Text

Index	File	Left	Node	Right
1	decten1pvc14	Tongues everybody knows about it that's a	Geordie	and went to s live in Spital
2	decten1t1sg0r	yes with with the bits of real	Geordie	comes out now and again yes yes
3	decten1t1sg0f	something to somebody they still pick the	Geordie	a lot of people do if you
4	decten1t1sg0f	way or something they say you're a	Geordie	the Newcastle Gateshead area ah no but
5	decten1t1sg1:	really because eh you never notice a	Geordie	accent with living up here ehm I'd
6	decten1t1sg1:	like that get good money naturally their	Geordie	accent's starting to go a bit because
7	decten1t1sg1:	no to me that's more Scottish than	Geordie	no not often kep to catch yeah
8	decten1t1sg1:	aye aye couldn't say well I'm a	Geordie	me like aye aye aye mm oh
9	decten1t1sg1:	Tees for the news and they're talking	Geordie	and it sounds funny you know and
10	decten1t1sg1:	over with a lot of eh like	Geordie	a bit you know but it just

File	Tokens	Frequency	Relative frequency per 10k
decten2y10i005b-SPW.txt	2426	8	32.976093
decten1t1sg24-SPW.txt	2909	6	20.625645
decten1t1sg27-SPW.txt	6324	13	20.55661
decten2y07i008b-SPW.txt	3959	8	20.207123
decten1t1sg12a-SPW.txt	1613	3	18.598885
decten2y08i003a-SPW.txt	4879	9	18.446402
decten2y10i009b-SPW.txt	2869	5	17.427675
decten1t1sg26-SPW.txt	4931	7	14.195903
decten1t1sg16a-SPW.txt	4321	6	13.885675
decten2y07i012a-SPW.txt	4760	6	12.605042
decten2y08i003b-SPW.txt	4846	6	12.381346
decten2y07i008a-SPW.txt	7012	8	11.409014
decten1t1sg36-SPW.txt	3538	4	11.305823



# Corpus Analysis: a few examples

#LancsBox

▼ Span 5<=>5

▼ Statistics 11 - LogRatio

▼ Threshold

▼ Corpus Corpus 1

▼ Unit

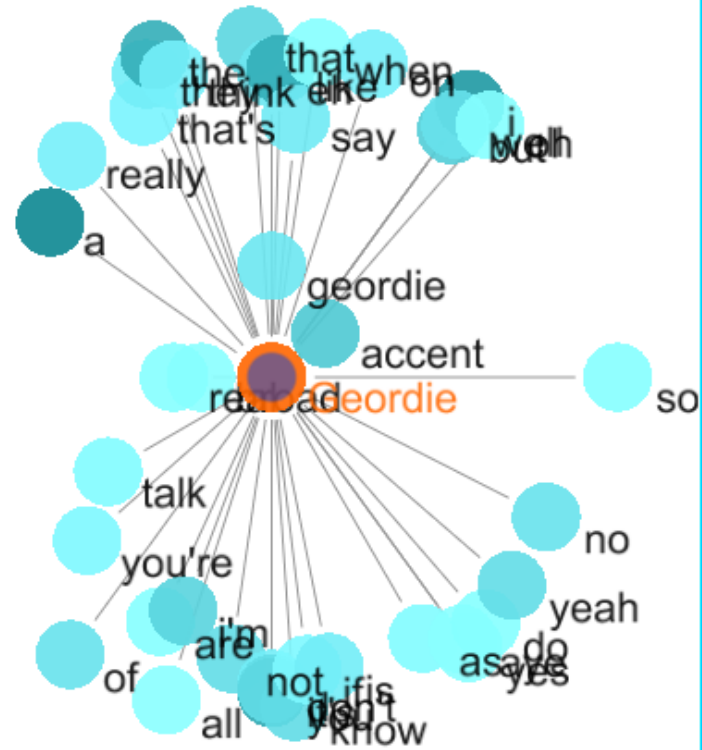
Clear

## Geordie

Spread out

Freq: 177 Collocates: 40

Status	Position	Collocate	▼ Stat	Freq (coll.)	Freq (corpus)
○	R	accent	9.667426261...	32	175
○	L	broad	9.625664695...	10	56
○	L	real	9.001327647...	11	89
○	M	geordie	8.684339665...	18	177
○	L	talk	6.861513563...	12	387
○	L	you're	5.513717060...	13	1047
○	L	i'm	5.440274720...	26	2202
○	R	say	5.203954249...	17	1693
○	L	are	4.992108708...	11	1267
○	L	a	4.950509278...	113	13393
○	L	not	4.628517993...	22	3254
○	R	if	4.386907774...	13	2271
○	L	that's	4.362606229...	14	2487
○	R	is	4.350045506...	16	2867
○	L	really	4.302690829...	15	2777
○	R	as	4.218734976...	12	2354
○	M	eh	4.140797318...	18	3726
○	R	but	4.120168827...	30	6299
○	M	don't	4.094706403...	16	3419
○	R	like	4.070953807...	55	11947
○	R	no	4.064085472...	20	4365
○	M	it's	3.971863028...	22	5117
○	R	well	3.956008621...	20	4703
○	M	you	3.887522756...	60	14792
○	R	yeah	3.840938623...	22	5601
○	L	think	3.831059764...	15	3845
○	R	ave	3.694926629...	10	2816














































































# Corpus Analysis: a few examples

## Aside

- ‘transparent’ filenames are usually a good idea (information/labels going from general to specific)

-  D-N2=decten2y10i005b=F=15-20=19.txt
-  D-N2=decten2y10i006a=M=21-30=22.txt
-  D-N2=decten2y10i006b=M=21-30=22.txt
-  D-N2=decten2y10i007a=M=15-20=19.txt
-  D-N2=decten2y10i007b=M=15-20=19.txt
-  D-N2=decten2y10i008a=F=15-20=19.txt
-  D-N2=decten2y10i008b=F=15-20=19.txt
-  D-N2=decten2y10i009a=M=41-50=43.txt
-  D-N2=decten2y10i009b=M=51-60=53.txt
-  D-N2=decten2y10i010a=F=51-60=52.txt
-  D-N2=decten2y10i010b=M=51-60=53.txt
-  D-N2=decten2y10i011a=M=15-20=19.txt
-  D-N2=decten2y10i011b=F=15-20=20.txt

- |   |   |   |
|---|---|---|
|  2012_SEL2091_060_Transcript.txt   |  2012_SEL8163_004_Transcript.txt   |  2013_SEL2091_017_Transcript.txt   |
|  2012_SEL2091_061_Transcript.txt   |  2012_SEL8163_005_Transcript.txt   |  2013_SEL2091_018_Transcript.txt   |
|  2012_SEL2091_062_Transcript.txt   |  2012_SEL8163_006_Transcript.txt   |  2013_SEL2091_019_Transcript.txt   |
|  2012_SEL2091_063_Transcript.txt   |  2012_SEL8163_007_Transcript.txt   |  2013_SEL2091_020_Transcript.txt   |
|  2012_SEL2091_064_Transcript.txt   |  2013_SEL2091_001_Transcript.txt   |  2013_SEL2091_021_Transcript.txt   |
|  2012_SEL2091_065_Transcript.txt   |  2013_SEL2091_002_Transcript.txt   |  2013_SEL2091_022_Transcript.txt   |
|  2012_SEL2091_066_Transcript.txt   |  2013_SEL2091_003_Transcript.txt   |  2013_SEL2091_023_Transcript.txt   |
|  2012_SEL2091_067_Transcript.txt   |  2013_SEL2091_004_Transcript.txt   |  2013_SEL2091_024_Transcript.txt   |
|  2012_SEL2091_068_Transcript.txt   |  2013_SEL2091_005_Transcript.txt   |  2013_SEL2091_025_Transcript.txt   |
|  2012_SEL2091_069_Transcript.txt |  2013_SEL2091_006_Transcript.txt |  2013_SEL2091_026_Transcript.txt |
|  2012_SEL2091_070_Transcript.txt |  2013_SEL2091_007_Transcript.txt |  2013_SEL2091_027_Transcript.txt |
|  2012_SEL2091_071_Transcript.txt |  2013_SEL2091_008_Transcript.txt |  2013_SEL2091_028_Transcript.txt |
|  2012_SEL2091_072_Transcript.txt |  2013_SEL2091_009_Transcript.txt |  2013_SEL2091_029_Transcript.txt |
|  2012_SEL2091_073_Transcript.txt |  2013_SEL2091_010_Transcript.txt |  2013_SEL2091_030_Transcript.txt |
|  2012_SEL2091_074_Transcript.txt |  2013_SEL2091_011_Transcript.txt |  2013_SEL2091_031_Transcript.txt |
|  2012_SEL2091_075_Transcript.txt |  2013_SEL2091_012_Transcript.txt |  2013_SEL2091_032_Transcript.txt |
|  2012_SEL2091_076_Transcript.txt |  2013_SEL2091_013_Transcript.txt |  2013_SEL2091_033_Transcript.txt |
|  2012_SEL8163_001_Transcript.txt |  2013_SEL2091_014_Transcript.txt |  2013_SEL2091_034_Transcript.txt |
|  2012_SEL8163_002_Transcript.txt |  2013_SEL2091_015_Transcript.txt |  2013_SEL2091_035_Transcript.txt |
|  2012_SEL8163_003_Transcript.txt |  2013_SEL2091_016_Transcript.txt |  2013_SEL2091_036_Transcript.txt |



# Corpus Analysis: a few examples

YES / YEAH / AYE / YEP

AntConc

AntConc 3.5.0 (Dev) (Windows) 2017

File Global Settings Tool Preferences Help

Corpus Files

Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List

Concordance Hits 98176

Hit	KWIC	File
24885	e all left on the beach <b>aye</b> I suppose <b>aye</b> that's how I think to	N2=2013_SEL2091_014b=F=61-70=69.txt
24886	? No, my point is that. <b>aye</b> , I suppose. <b>aye</b> . did they not think	N2P=2007_SEL2091_043a=M=21-30=21.txt
24887	hat bad Right there <b>Aye</b> I suppose <b>aye</b> if you're -- if you're at	N2P=2007_SEL2091_045b=M=21-30=21.txt
24888	e Yeah you would <b>Aye</b> , I suppose <b>Aye</b> Should we?	N2P=2009_SEL2091_041a=M=15-20=19.txt
24889	ge that's true yeah <b>yeah</b> I suppose <b>Bruge</b> you need a six figu	N2=2012_SEL2091_021b=M=21-30=21.txt
24890	he'd be brilliant why <b>yeah</b> I suppose <b>but</b> no I know there's ple	D-N1b=decten1pvc10a=M=16-20=18.txt
24891	have both can you it's <b>Yeah</b> I suppose <b>but</b> . Some things you ca	D-N2=decten2y07i004b=F=15-20=19.txt
24892	And work somewhere? <b>Aye</b> I suppose. <b>But</b> I can't be arsed to go t	N2=2007_SEL2091_010b=M=21-30=22.txt
24893	people from the North <b>Yeah</b> I suppose <b>But</b> theirs is more lool	N2=2008_SEL2091_005a=M=15-20=19.txt
24894	ant to see" Yeah <b>yeah</b> I suppose <b>but</b> I'm I'm sad that I can't	N2=2008_SEL2091_015b=F=15-20=20.txt
24895	could've but yeah <b>yeah</b> I suppose <b>but</b> err it's not so not so gi	N2=2012_SEL2091_036a=F=15-20=19.txt
24896	ilvia'z nursery as well? <b>Yeah</b> I suppose <b>but</b> we were really really lit	N2P=2007_SEL2091_020b=F=15-20=19.txt
24897	crash into something or <b>yeah</b> . I suppose <b>but</b> at least you'd be a	N2P=2007_SEL2091_020b=F=15-20=19.txt
24898	asking why you do that. <b>yeah</b> I suppose, <b>But</b> boys are a lot more c	N2P=2007_SEL2091_041b=F=21-30=21.txt
24899	people from the North <b>Yeah</b> I suppose <b>But</b> theirs is more lo	N2P=2008_SEL2091_005a=M=15-20=19.txt

Search Term  Words  Case  Regex

Search Window Size 250

Advanced

Start Stop Sort Show Every Nth Row 1

Kwic Sort

Level 1 1R  Level 2 2R  Level 3 3R

Total No. 1446

Files Processed

Clone Results





# Corpus Analysis: a few examples

**YES / YEAH / AYE / YEP**

1970s: 6146 tokens  
 1990s: 3413 tokens  
 2000s: 55157 tokens

	A	B	C	D	E	F	G	H	I	J
1	FILE/INFORMANT	PRE	TOKEN	POST	SET	FILE/INFORMANT	SPEAKER SEX	AGE GROUP	AGE	FILE TOKEN
251	decten0t1sg41	<informantTLSG41>	YES	</u>	D-NO	decten0t1sg41	F	41-50		76
252	decten0t1sg41	<informantTLSG41>	YES	</u>	D-NO	decten0t1sg41	F	41-50		77
253	decten0t1sg41	<informantTLSG41>	YES	</u>	D-NO	decten0t1sg41	F	41-50		78
254	decten0t1sg41	<informantTLSG41>	YES	I suppose I could </u>	D-NO	decten0t1sg41	F	41-50		79
255	decten0t1sg41	<informantTLSG41>	YEAH	I would say that </u>	D-NO	decten0t1sg41	F	41-50		80
256	decten0t1sg41	<informantTLSG41>	YES	I might say that </u>	D-NO	decten0t1sg41	F	41-50		81
257	decten0t1sg41	<informantTLSG41>	YES	</u>	D-NO	decten0t1sg41	F	41-50		82
258	decten0t1sg41	<informantTLSG41>	YES	</u>	D-NO	decten0t1sg41	F	41-50		83
259	decten0t1sg41	<informantTLSG41>	YES	I could </u>	D-NO	decten0t1sg41	F	41-50		84
260	decten0t1sg42a	<informantTLSG42a> eh	YES	</u>	D-NO	decten0t1sg42a	M	21-30		1
261	decten0t1sg42a	<informantTLSG42a> oh	YES	uh-huh yes </u>	D-NO	decten0t1sg42a	M	21-30		2
262	decten0t1sg42a	<informantTLSG42a> oh	YES	</u>	D-NO	decten0t1sg42a	M	21-30		3
263	decten0t1sg42a	<informantTLSG42a> oh	YES	</u>	D-NO	decten0t1sg42a	M	21-30		4
264	decten0t1sg42a	<informantTLSG42a>	YEAH	night shift was a eh eight	D-NO	decten0t1sg42a	M	21-30		5
265	decten0t1sg42a	<informantTLSG42a>	YES	that's a good question De	D-NO	decten0t1sg42a	M	21-30		6
266	decten0t1sg42a	<informantTLSG42a> <la	YEAH	well <pause/> no I canno	D-NO	decten0t1sg42a	M	21-30		7
267	decten0t1sg42a	<informantTLSG42a>	AYE	it's not what I would usua	D-NO	decten0t1sg42a	M	21-30		8
268	decten0t1sg42a	<informantTLSG42a> oh	YES	yes yes </u>	D-NO	decten0t1sg42a	M	21-30		9
269	decten0t1sg42a	<informantTLSG42a> oh	YES	yes </u>	D-NO	decten0t1sg42a	M	21-30		10
270	decten0t1sg42a	<informantTLSG42a> oh	YES	</u>	D-NO	decten0t1sg42a	M	21-30		11
271	decten0t1sg42a	<informantTLSG42a> oh	YES	I liked my school days </u	D-NO	decten0t1sg42a	M	21-30		12
272	decten0t1sg42a	<informantTLSG42a> oh	YES	I know if I had the chance	D-NO	decten0t1sg42a	M	21-30		13
273	decten0t1sg42a	<informantTLSG42a>	YEAH	the South Street you see	D-NO	decten0t1sg42a	M	21-30		14
274	decten0t1sg42a	<informantTLSG42a> eh	YEAH	</u>	D-NO	decten0t1sg42a	M	21-30		15

Tokens

Stats





# Corpus Analysis: a few examples

## YES / YEAH / AYE / YEP

DECTE\_YES\_antconc\_results.xlsx - Excel

Home Insert Page Layout Formulas Data Review View Developer Help PDF-XChange V6 Tell me Share

Clipboard Font Alignment Number Styles Cells Editing

AutoSave Off

B5 =COUNTIFS(Tokens!E2:E70820,"D-N0",Tokens!H2:H70820,"<>",Tokens!C2:C70820,"AYE")+COUNTIFS(Tokens!E2:E70820,"D-N1a",Tokens!H2:H70820,"<>",Tokens!C2:C70820,"AYE")

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	<b>Tokens</b>	TLS	PVC	N2		<b>Tokens</b>	N2-2007	N2-2008	N2-2009	N2-2010	N2-2011	N2-2012	N2-2013
2	N	6146	3413	55157		N	8107	6782	9858	3705	10878	9663	6164
3	YES	4167	985	4873		YES	1225	537	976	364	845	648	278
4	YEAH	455	1218	43502		YEAH	4791	5561	7861	3050	8470	8597	5172
5	AYE	1514	1208	6350		AYE	2040	621	914	264	1480	374	657
6	YEP/YUP	10	2	432		YEP/YUP	51	63	107	27	83	44	57
7													
8	<b>%</b>	1970s	1990s	2000s		<b>%</b>	2007	2008	2009	2010	2011	2012	2013
9	YES	67.80	28.86	8.83		YES	15.11	7.92	9.90	9.82	7.77	6.71	4.51
10	YEAH	7.40	35.69	78.87		YEAH	59.10	82.00	79.74	82.32	77.86	88.97	83.91
11	AYE	24.63	35.39	11.51		AYE	25.16	9.16	9.27	7.13	13.61	3.87	10.66
12	YEP/YUP	0.16	0.06	0.78		YEP/YUP	0.63	0.93	1.09	0.73	0.76	0.46	0.92
13													
14	<b>Sp.Sex 2000s</b>	FEMALE	MALE										
15	N	29464	25693										
16	YES	3283	1590										
17	YEAH	24554	18948										
18	AYE	1379	4971										
19	YEP/YUP	248	184										
20													
21	<b>%</b>	FEMALE	MALE	BOTH									
22	YES	11.14	6.19	8.83									
23	YEAH	83.34	73.75	78.87									

### YES in the 2000s (speaker sex)

Speaker Sex	Female %	Male %	Both %
FEMALE	83.34	16.66	
MALE	19.35	80.65	
BOTH	8.83	91.17	

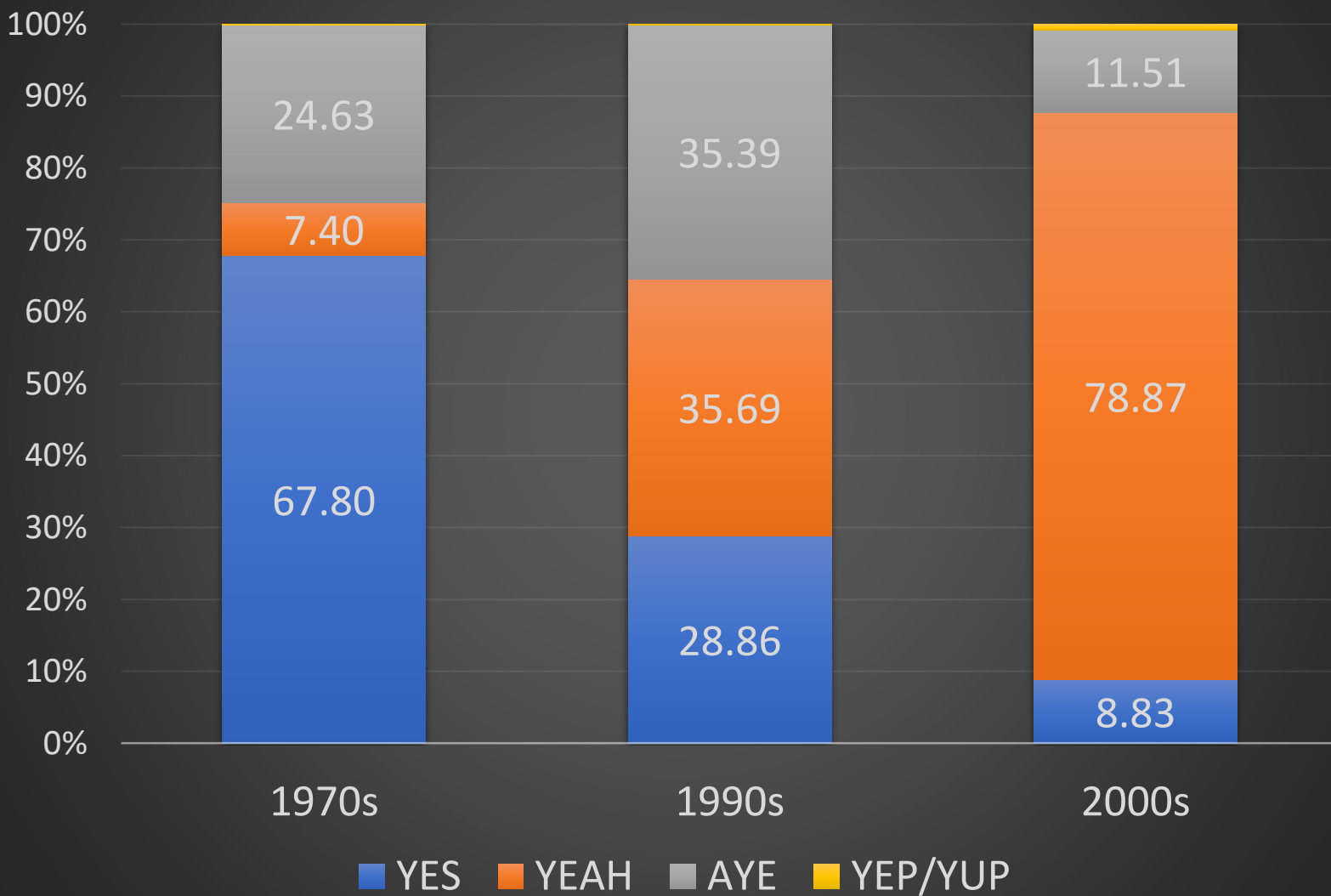
Ready



# Corpus Analysis: a few examples

YES / YEAH / AYE / YEP

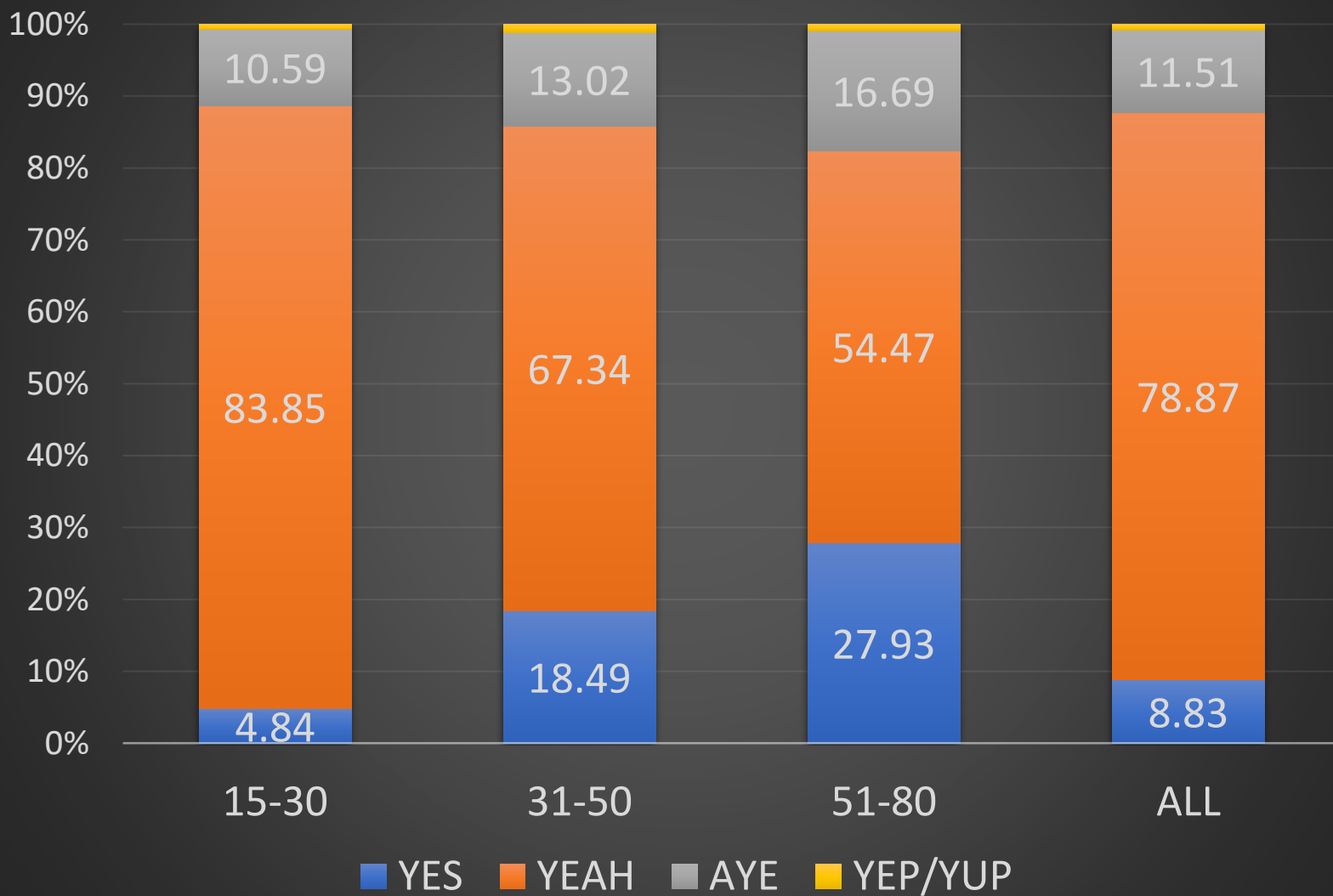
YES, 1970s-2000s





YES / YEAH / AYE / YEP

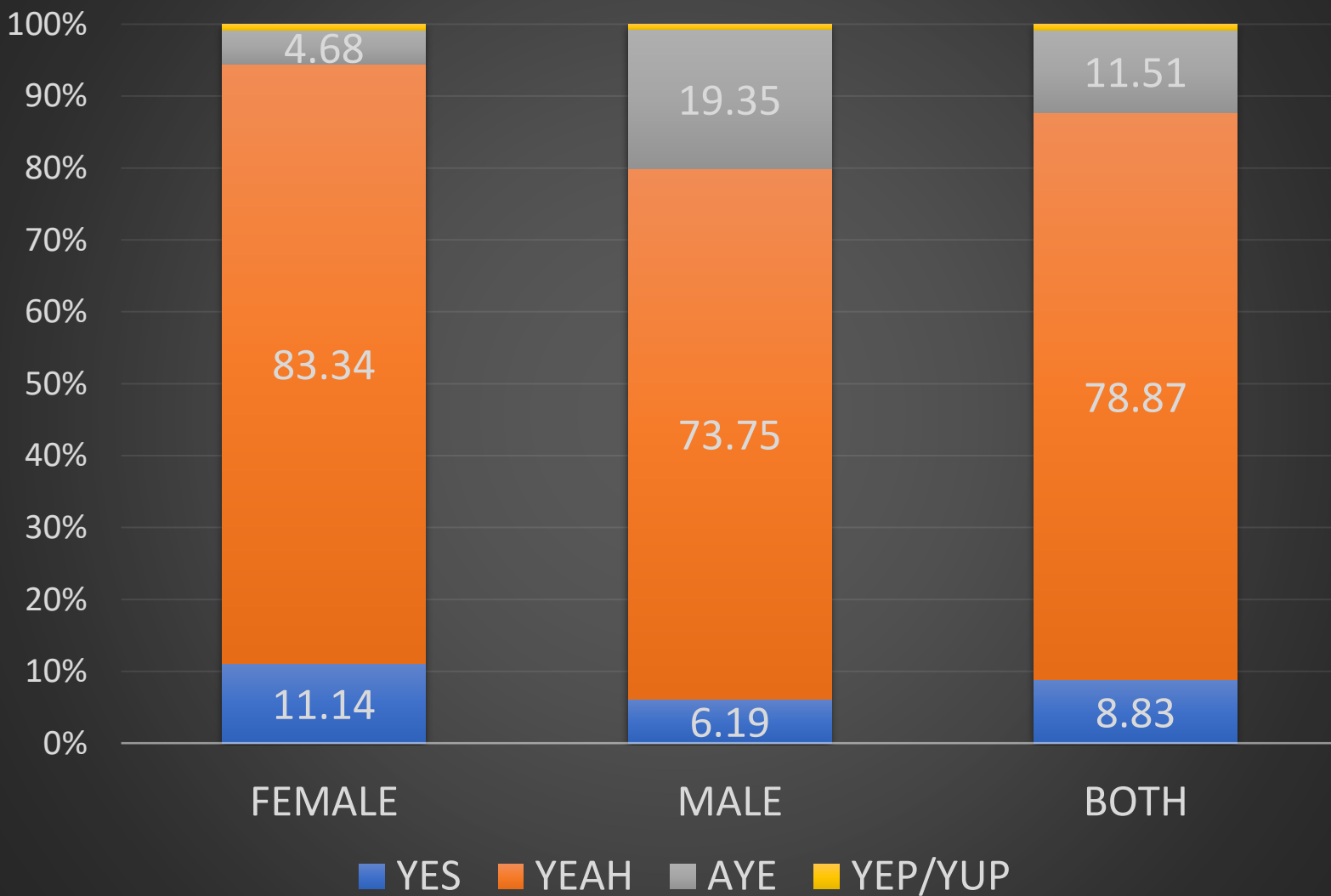
YES in the 2000s (speaker age)





YES / YEAH / AYE / YEP

### YES in the 2000s (speaker sex)







- Centre for Corpus Research. *Introduction to Corpus investigative techniques*. University of Birmingham  
<http://www.birmingham.ac.uk/research/activity/corpus/publications/introduction-corpus-investigative-techniques.aspx>
  - inc. Unit 2 – compiling a corpus, Unit 3 – available corpora / software
- CoRD: *Corpus Resource Database*. Varieng, University of Helsinki  
<http://www.helsinki.fi/varieng/CoRD/index.html>
  - inc. Corpus Finder (search by characteristics: spoken v. written, etc)
- Dillon, George. *Corpus Resources*. University of Washington.  
<http://courses.washington.edu/englhtml/engl560/corplingresources.htm>
- Linguist List, *Web Resource Listings: Texts and Corpora*  
<http://linguistlist.org/sp/GetWRListings.cfm?wrtypeid=1>
- Linguist List, *Software Related to Text/Corpus Linguistics*  
<http://linguistlist.org/sp/SearchWRListing-action.cfm?subclassid=7223&SearchType=LF&WRTypeID=2>



## Useful Links

- List of Corpora and Databases. Dept. of Language and Linguistic Science. University of York. <https://www.york.ac.uk/language/current/resources/corpora>
- OLAC: Open Language Archives Community. <http://www.language-archives.org/archives>
- Smith, Jen. Lexical Databases and Corpora. UNC Chapel Hill. <https://www.unc.edu/~jlsmith/lex-corp.html>
- UCREL: University Centre for Computer Corpus Research on Language. Lancaster University. <http://ucrel.lancs.ac.uk>
- Wynne, Martin (ed.). 2004. *Developing Linguistic Corpora: a Guide to Good Practice*. AHDS / OTA (Oxford Text Archive). <https://ota.ox.ac.uk/documents/creating/dlc/index.htm>
- Xiao, Richard (<http://www.lancaster.ac.uk/staff/xiaoz>). *Well-known and influential corpora: A survey*. <http://www.lancaster.ac.uk/staff/xiaoz/papers/corpus%20survey.htm>



### **AntConc**

- Laurence Anthony's YouTube Channel  
<https://www.youtube.com/user/AntlabJPN>
- Monika Bednarek's YouTube Channel  
<https://www.youtube.com/channel/UC3HWqldtJZpisxSiUZL0oKQ>
- Heather Froehlich – Getting Started with AntConc  
<https://hfroehli.ch/workshops/getting-started-with-antconc>
- Heather Froehlich – Corpus Analysis with AntConc  
<https://programminghistorian.org/lessons/corpus-analysis-with-antconc>

### **AntConc and Wordsmith Tools**

- Dawn Knight – Corpus Linguistics Exercises  
<https://sites.google.com/site/clexercises/antconc>

### **LancsBox**

- PDF guide and videos: <http://corpora.lancs.ac.uk/lancsbox/help.php>



## Select Bibliography

- Adolphs, Svenja and Ronald Carter. 2013. *Spoken Corpus Linguistics: From Monomodal to Multimodal*. London: Routledge.
- Baker, Paul. 2010. *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Baker, Paul, Andrew Hardie and Tony McEnery. 2006. *A Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Beal, Joan C., Karen P. and Hermann Moisl (eds). 2007. *Creating and Digitizing Language Corpora, Volume 1: Synchronic Databases*. Houndmills: Palgrave.
- Beal, Joan C., Karen P. and Hermann Moisl (eds). 2007. *Creating and Digitizing Language Corpora, Volume 2: Diachronic Databases*. Houndmills: Palgrave.
- Corrigan, Karen P. and Adam Mearns (eds). 2016. *Creating and Digitizing Language Corpora, Volume 3: Databases for Public Engagement*. Houndmills: Palgrave.
- McEnery, Tony and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.
- McEnery, Tony, Richard Xiao and Yukio Tono. 2006. *Corpus-Based Language Studies: An advanced resource book*. London: Routledge.  
companion website: <http://cw.routledge.com/textbooks/0415286239>
- O’Keeffe, Anne and Michael McCarthy (eds). 2010. *The Routledge Handbook of Corpus Linguistics*. London: Routledge.