A Historical Sociolinguist's Digital Tools Starter Kit

Kelly E. Wright
University of Kentucky

Inaugural NARNiHS Conference
22 July 2017

http://www.uky.edu/~mrlaue2/narnihs2017/workshop.html

•••

Google Drive Folder

To Download

- A Text Editer
  - BBEdit: https://www.barebones.com/products/textwrangler/
  - PC↓ MAC↑
  - Notepad++: https://notepad-plus-plus.org
- AntConc: http://www.laurenceanthony.net/software/antconc/
- Gephi: https://gephi.org

PCEEC

http://ota.ox.ac.uk/desc/2510

- Parsed Corpus of Early English Correspondence
- Oxford Text Archive--one of the largest repositories for Digital Corpora
- 4970 personal letters
- 84 collections
- 666 writers
- 1410?-1681
- 2.2 million words

Metadata

- Author
- Recipient
- Letter

- Big 5
- Time Period
- Authenticity

Letter Formatting

../2510/2510/PCEEC/corpus_description/index.htm

<B_MARVELL> <Q_MAV_A_1653_T_AMARVELL>
<L_MARVELL_001>
<A_ANDREW_MARVELL_JR> <A-GENDER_MALE>
<A-REL_---> <A-DOB_1621>
<R_OLIVER_CROMWELL> <R-GENDER_MALE>
<R-REL_---> <R-DOB_1599>
<AREW_MARVELL_JR> <P_304> {ED:1.}

AUTHOR:ANDREW_MARVELL_JR:MALE:_:1621:32
RECIPIENT:OLIVER_CROMWELL:MALE:_:1599:54
LETTER:MARVELL_001:E3:1653:AUTOGRAPH:OTHER
{COM:ADDRESSED} For his Excellence , the Lord General Cromwell . these
with my most humble service : MARVELL,304.001.1

RegEx

- A special text string for describing a search pattern
- The most basic search is any string
  - You don't have to change your settings to do traditional searching
- RegEx will do exactly what you ask it to

RegEX

- You can use a hyphen inside a character class to specify a range of characters. [0-9] matches a *single* digit between 0 and 9. You can use more than one range, and you can combine ranges and single characters. [0-9a-fxA-FX] matches a hexadecimal digit or the letter X.

RegEx

Accuracy

- Recall
- Precision

RegEx

Accuracy

- Recall
  - Did I leave anything behind?
- Precision
  - How much noise is present?

RegEx

Standard Operating Procedures

- Consumption
- Negation

RegEx

Consumption

- \d{4}

RegEx

Negation

- A negated character class still must match a character. q[^u] does *not* mean: "a q not followed by a u". It means: "a q followed by a character that is not a u".
  - Does not match the q in the strin Iraq.
  - Does match the q and the space after the q in Iraq is a country.

RegEx Metacharacters

the asterisk or star * Zero (0) or more

the plus sign + One (1) or more

the question mark ? Zero (0) or one (1)

the parenthesis ( ) Grouping

the opening square bracket [Define a character class

and the opening curly brace { Introduce a quantifier

the backslash \ escape following character

the caret ^ marks the start of a string

the dollar sign $ marks the end of a string

the period or dot . matches any one character

the vertical bar or pipe symbol | or

RegEx Returns

- cat|dog food matches cat or dog food. To create a regex that matches cat food or dog food, you need to group the alternatives: (cat|dog) food.

Let’s try a basic search

Google Drive

- Open up BBEdit
- Load Marvell.txt from the workshop folder
- Search *her*

What do we notice in the results?

Let’s try a basic search

What do we notice in the results?

- RegEx does what you tell it.
- Now try, \sher\s

Once more, with AntConc

- Open up AntConc
- Load Marvell.txt
- Settings > Global Settings > Wildcards
- Repeat the *her* search

What is different about these results?

- Try the RegEx \sher\s

Do we get the same results?

Play!

With Cheat Sheets

- Dave Child's Basic Cheat Sheets

What did you come up with?

Subcorpora

With RegEx

- Separate by salient metadata
- Put each letter onto a single line

Subcorpora

Unique and Universal Delimiters

- Separate by salient metadata
- Each letter is preceded by the *text identifier*, labelled **Q**
- <Q_BAC_A_1569_FN_N2BACON>

Contains five codes separated by underscores:

- Text_from the Bacon collection_written by a single author_date_to a member of their nuclear family_writer code

Metadata Encoding

((CODE <B_BACON>))
((CODE
<Q_BAC_A_1569_FN_N2BACON>))
((CODE <L_BACON_001>))
((CODE <A_NICHOLAS_BACON_II>))
((CODE <A-GENDER_MALE>))
((CODE <A-REL_BROTHER>))
((CODE <A-DOB_1543>))
((CODE <R_NATHANIEL_BACON_I>))
((CODE <R-GENDER_MALE>))
((CODE <R-REL_BROTHER>))
((CODE <R-DOB_1546?>))

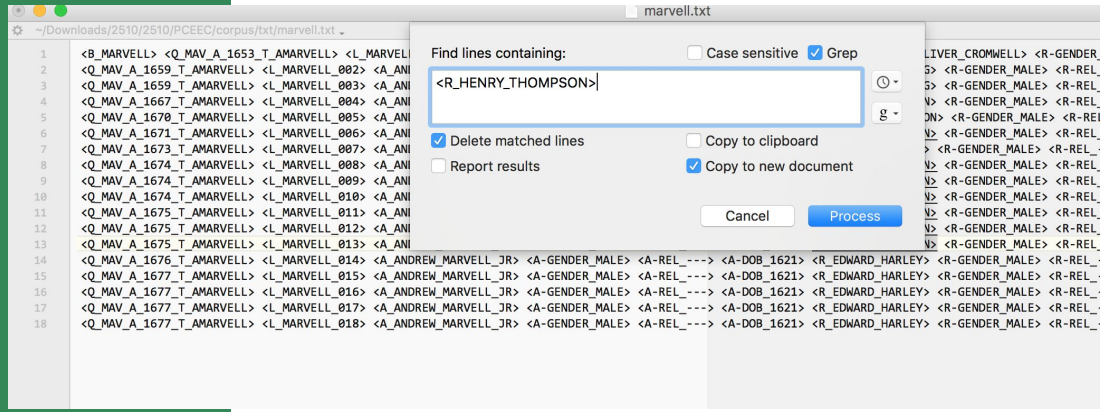
Subcorpora

Unique and Universal Delimiters

- Open BBedit
- Functions by using *Find/Replace*
  - Find: TextWrangler = \r(?!<Q)
    Notepad++ = \n(?!<Q)
  - Replace: with a “space”
- Carriage return (negative lookahead text identifier)

Play!

- Choose something to separate by
- In BBedt: Text > Process Lines Containing



Addressing Predictable Spelling Errors

With Character Classes

- Character classes are one of the most commonly used RegEx features.
- You can find a word, even if it is misspelled, such as sep[ae]r[ae]te or li[cs]en[cs]e.

Vard2

Because Orthography is a lie, and our minds aren't algorithms

The software assists with manual normalisation by suggesting candidate normalisations for detected spelling variants. As decisions are made by the user, VARD learns how to best normalise the spelling variation in your corpus to the point where it can successfully automatically normalise the entire corpus after training.

VARD2

- VARD2 has to be opened in the command line
- Navigate to your copy of the folder
- Select run.command shell script

VARD2

- Open Harvey.txt in BBedit
- Find *my*

How many results?

VARD2

- Open Vard2
- Load Harvey.txt
- Normalize *mai*
- Save With XML Tags
- Load the varded file into BBEdit

VARD2 Output

```
AUTHOR:GABRIEL_HARVEY:MALE:_:1545?:28?
RECIPIENT:JOHN_YOUNG:MALE:_:_:_
LETTER:HARVEY_001:E2:1573:AUTOGRAPH:OTHER
But it <normalised orig="mai" auto="false">my</normalised> pleas your wurship to remember
famus poet : Quis tulerit Gracchos de seditione querentes ?
HARVEY,5.001.71

AUTHOR:GABRIEL_HARVEY:MALE:_:1545?:28?
RECIPIENT:JOHN_YOUNG:MALE:_:_:_
LETTER:HARVEY_001:E2:1573:AUTOGRAPH:OTHER
for it were nedles for me to go about to point out his pride and <P_6>
lustines , whereas his oun gai gallant gaskins , his kut dublets , his
staring hare , with sum other gudli and gentlemanlike ornaments , do
and wil discri it sufficiently . HARVEY,6.001.72
```

VARD2

Output

How many results when we search for *my* now??

VARD2

Training

- Return to Vard
- Load your new version of Harvey.txt into the Trainer

The AIF File

https://drive.google.com/open?id=0BzlGStEoNAf0dlViU3Y1bU9XODg

- Associated Personal Information

	A	B	C	D	E	F	G	H	I	J	K
1	ALLEN_001	WILLIAM_ALLEN	CARDINAL_OF_ENGLAND	MALE	1532		RICHARD_HOPKINS	SCHOLAR	MALE	1546?	
2	ALLEN_002	WILLIAM_ALLEN	CARDINAL_OF_ENGLAND	MALE	1532		OWEN_LEWIS	DR/BISHOP_OF_CASSANO	MALE		
3	ALLEN_003	WILLIAM_ALLEN	CARDINAL_OF_ENGLAND	MALE	1532		JOHN_ARDEN		MALE		
4	ALLEN_004	WILLIAM_ALLEN	CARDINAL_OF_ENGLAND	MALE	1532		JOHN_ARDEN		MALE		
5	ARUNDEL_001	ANNE_HOWARD[N.DACRE]	COUNTESS_OF_ARUNDEL	FEMALE	1553?		WILLIAM_CECIL	1ST_LORD_BURGHLEY/ROYAL_MINISTER(DNB)	MALE	1520	
6	ARUNDEL_002	ANNE_HOWARD[N.DACRE]	COUNTESS_OF_ARUNDEL	FEMALE	1553?	KIN	GILBERT_TALBOT_1	7TH_EARL_OF_SHREWSBURY	MALE	1553	KIN
7	ARUNDEL_003	ANNE_HOWARD[N.DACRE]	COUNTESS_OF_ARUNDEL	FEMALE	1553?	KIN	GILBERT_TALBOT_1	7TH_EARL_OF_SHREWSBURY	MALE	1553	KIN
8	ARUNDEL_004	THOMAS_HOWARD_III	2ND_EARL_OF_ARUNDEL_	MALE	1585	SON	ANNE_HOWARD[N.DACRE]	COUNTESS_OF_ARUNDEL	FEMALE	1553?	MOTHER
9	ARUNDEL_005	THOMAS_HOWARD_III	2ND_EARL_OF_ARUNDEL_	MALE	1585	SON	ANNE_HOWARD[N.DACRE]	COUNTESS_OF_ARUNDEL	FEMALE	1553?	MOTHER
10	ARUNDEL_006	ANNE_HOWARD[N.DACRE]	COUNTESS_OF_ARUNDEL	FEMALE	1553?	KIN	MARY_TALBOT[N.CAVENDISH]	COUNTESS_OF_SHREWSBURY	FEMALE		KIN
11	ARUNDEL_007	THOMAS_HOWARD_III	2ND_EARL_OF_ARUNDEL_	MALE	1585	SON	ANNE_HOWARD[N.DACRE]	COUNTESS_OF_ARUNDEL	FEMALE	1553?	MOTHER
12	ARUNDEL_008	ALETHEIA_HOWARD[N.TALI	COUNTESS_OF_ARUNDEL	FEMALE	1585?	DAUGHTER	GILBERT_TALBOT_1	7TH_EARL_OF_SHREWSBURY	MALE	1553	FATHER
13	ARUNDEL_009	ANNE_HOWARD[N.DACRE]	COUNTESS_OF_ARUNDEL	FEMALE	1553?	KIN	MARY_TALBOT[N.CAVENDISH]	COUNTESS_OF_SHREWSBURY	FEMALE		KIN
14	ARUNDEL_010	ANNE_HOWARD[N.DACRE]	COUNTESS_OF_ARUNDEL	FEMALE	1553?	MOTHER	THOMAS_HOWARD_III	2ND_EARL_OF_ARUNDEL_AND_SURREY/POLITIC	MALE	1585	SON
15	ARUNDEL_011	THOMAS_HOWARD_III	2ND_EARL_OF_ARUNDEL_	MALE	1585		DUDLEY_CARLETON	VISCOUNT_DORCHESTER/DIPLOMAT(DNB)	MALE	1573	
16	ARUNDEL_012	THOMAS_HOWARD_III	2ND_EARL_OF_ARUNDEL_	MALE	1585	SON	ANNE_HOWARD[N.DACRE]	COUNTESS_OF_ARUNDEL	FEMALE	1553?	MOTHER
17	ARUNDEL_013	THOMAS_HOWARD_III	2ND_EARL_OF_ARUNDEL_	MALE	1585	SON	ANNE_HOWARD[N.DACRE]	COUNTESS_OF_ARUNDEL	FEMALE	1553?	MOTHER
18	ARUNDEL_014	ANNE_HOWARD[N.DACRE]	COUNTESS_OF_ARUNDEL	FEMALE	1553?		THOMAS_EDMONDES	SIR	MALE	1564?	
19	ARUNDEL_015	THOMAS_HOWARD_III	2ND_EARL_OF_ARUNDEL_	MALE	1585	SON	ANNE_HOWARD[N.DACRE]	COUNTESS_OF_ARUNDEL	FEMALE	1553?	MOTHER
20	ARUNDEL_016	THOMAS_HOWARD_III	2ND_EARL_OF_ARUNDEL_	MALE	1585	SON	ANNE_HOWARD[N.DACRE]	COUNTESS_OF_ARUNDEL	FEMALE	1553?	MOTHER
21	ARUNDEL_017	THOMAS_HOWARD_III	2ND_EARL_OF_ARUNDEL_	MALE	1585	SON	ANNE_HOWARD[N.DACRE]	COUNTESS_OF_ARUNDEL	FEMALE	1553?	MOTHER
22	ARUNDEL_018	THOMAS_HOWARD_III	2ND_EARL_OF_ARUNDEL_	MALE	1585	SON	ANNE_HOWARD[N.DACRE]	COUNTESS_OF_ARUNDEL	FEMALE	1553?	MOTHER
23	ARUNDEL_019	THOMAS_HOWARD_III	2ND_EARL_OF_ARUNDEL_	MALE	1585	SON	ANNE_HOWARD[N.DACRE]	COUNTESS_OF_ARUNDEL	FEMALE	1553?	MOTHER
24	ARUNDEL_020	THOMAS_HOWARD_III	2ND_EARL_OF_ARUNDEL_	MALE	1585	SON	ANNE_HOWARD[N.DACRE]	COUNTESS_OF_ARUNDEL	FEMALE	1553?	MOTHER
25	ARUNDEL_021	THOMAS_HOWARD_III	2ND_EARL_OF_ARUNDEL_	MALE	1585	SON	ANNE_HOWARD[N.DACRE]	COUNTESS_OF_ARUNDEL	FEMALE	1553?	MOTHER
26	ARUNDEL_022	DUDLEY_CARLETON	VISCOUNT_DORCHESTER/	MALE	1573		HORACE_VERE	BARON/ARMY_OFFICER(DNB)	MALE	1565	
27	ARUNDEL_023	THOMAS_HOWARD_III	2ND_EARL_OF_ARUNDEL_	MALE	1585	SON	ANNE_HOWARD[N.DACRE]	COUNTESS_OF_ARUNDEL	FEMALE	1553?	MOTHER
28	ARUNDEL_024	THOMAS_HOWARD_III	2ND_EARL_OF_ARUNDEL_	MALE	1585	SON	ANNE_HOWARD[N.DACRE]	COUNTESS_OF_ARUNDEL	FEMALE	1553?	MOTHER
29	ARUNDEL_025	THOMAS_HOWARD_III	2ND_EARL_OF_ARUNDEL_	MALE	1585	SON	ANNE_HOWARD[N.DACRE]	COUNTESS_OF_ARUNDEL	FEMALE	1553?	MOTHER
30	ARUNDEL_026	THOMAS_HOWARD_III	2ND_EARL_OF_ARUNDEL_	MALE	1585	SON	ANNE_HOWARD[N.DACRE]	COUNTESS_OF_ARUNDEL	FEMALE	1553?	MOTHER
31	ARUNDEL_027	THOMAS_HOWARD_III	2ND_EARL_OF_ARUNDEL_	MALE	1585	SON	ANNE_HOWARD[N.DACRE]	COUNTESS_OF_ARUNDEL	FEMALE	1553?	MOTHER
32	ARUNDEL_028	THOMAS_HOWARD_III	2ND_EARL_OF_ARUNDEL_	MALE	1585	SON	ANNE_HOWARD[N.DACRE]	COUNTESS_OF_ARUNDEL	FEMALE	1553?	MOTHER
33	ARUNDEL_029	INIGO_JONES	ARCHITECT	MALE	1573		THOMAS_HOWARD_III	2ND_EARL_OF_ARUNDEL_AND_SURREY/POLITIC	MALE	1585	

Network Analysis

https://www.youtube.com/watch?v=3bBkZbqzyY4

- The Uniformitarian Principle and Data-Driven Research
- Nodes, Edges, Density, Multiplexity
- Centralities

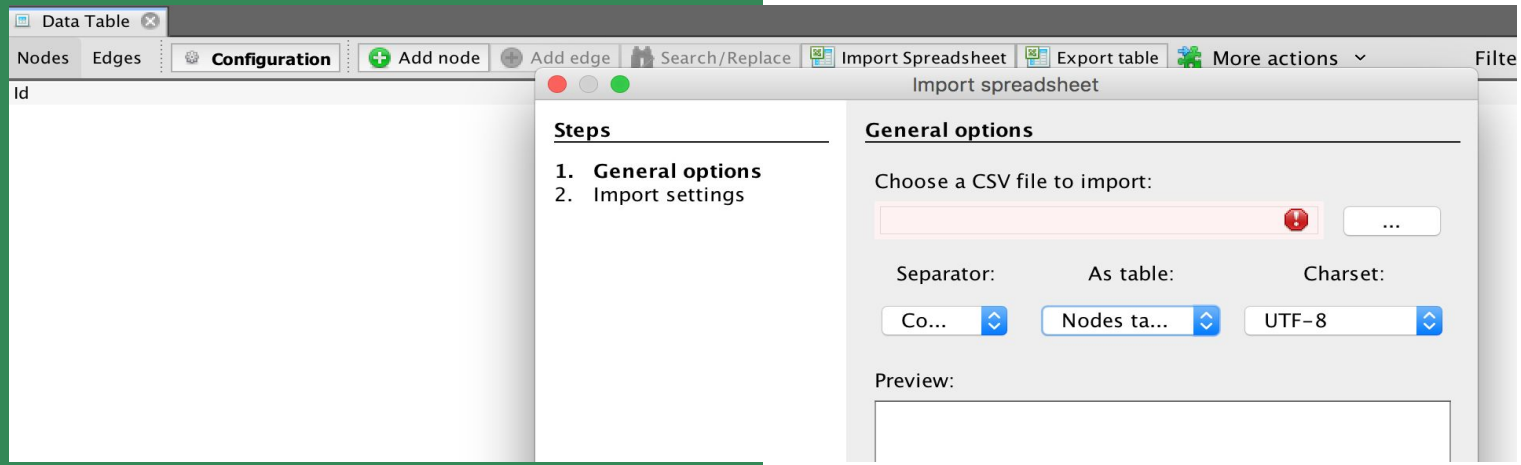
Gephi

Visualizing Centralities

- Betweenness
  - The shortest path
- Degree
  - Total connections
- Closeness
  - Sum of the shortest distances between each node and every other node in the network

Gephi

- In Data Laboratory, load Tremendous Node List and 00Edge from the Google Drive Folder.
- Make sure when you load Nodes, the Nodes Tab and Nodes Table selections are marked. So too with Edges.



Let's Visualize!

Gephi Play

- Filters
  - Typology > Degree Range > (drag down)
- Statistics (centrality)
  - Network diameter > Run

Let's Visualize!

Gephi Play

- Allow us to think critically about the multifarious connections in All Our Data
- Navigate to the Layout panel and run the Yifan Hu Projection
- Play with Appearance options

I <3 AIF

Best Practices in Documentation

- Translates Easily
- Potential for industry standard
- 500 schmunks

NetLogo

Because sometimes a day is better when you tip the scales in favor of grass.

- Agent-based modeling
- Get at the untenable experiments

http://www.netlogoweb.org/launch#http://www.netlogoweb.org/assets/modelslib/Sample%20Models/Biology/Wolf%20Sheep%20Predation.nlogo

THANKS Y'ALL!

•••

Kelly E. Wright
University of Kentucky
kellywright5.wixsite.com/raciolinguistics