HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Terttu Nevalainen, Tanja Säily

VARIENG Research Unit, University of Helsinki

& Harri Siirtola

University of Tampere

Tools for comparing corpora: Text Variation Explorer (TVE)

ISLE 2, Boston, MA

17–21 June 2011

Outline

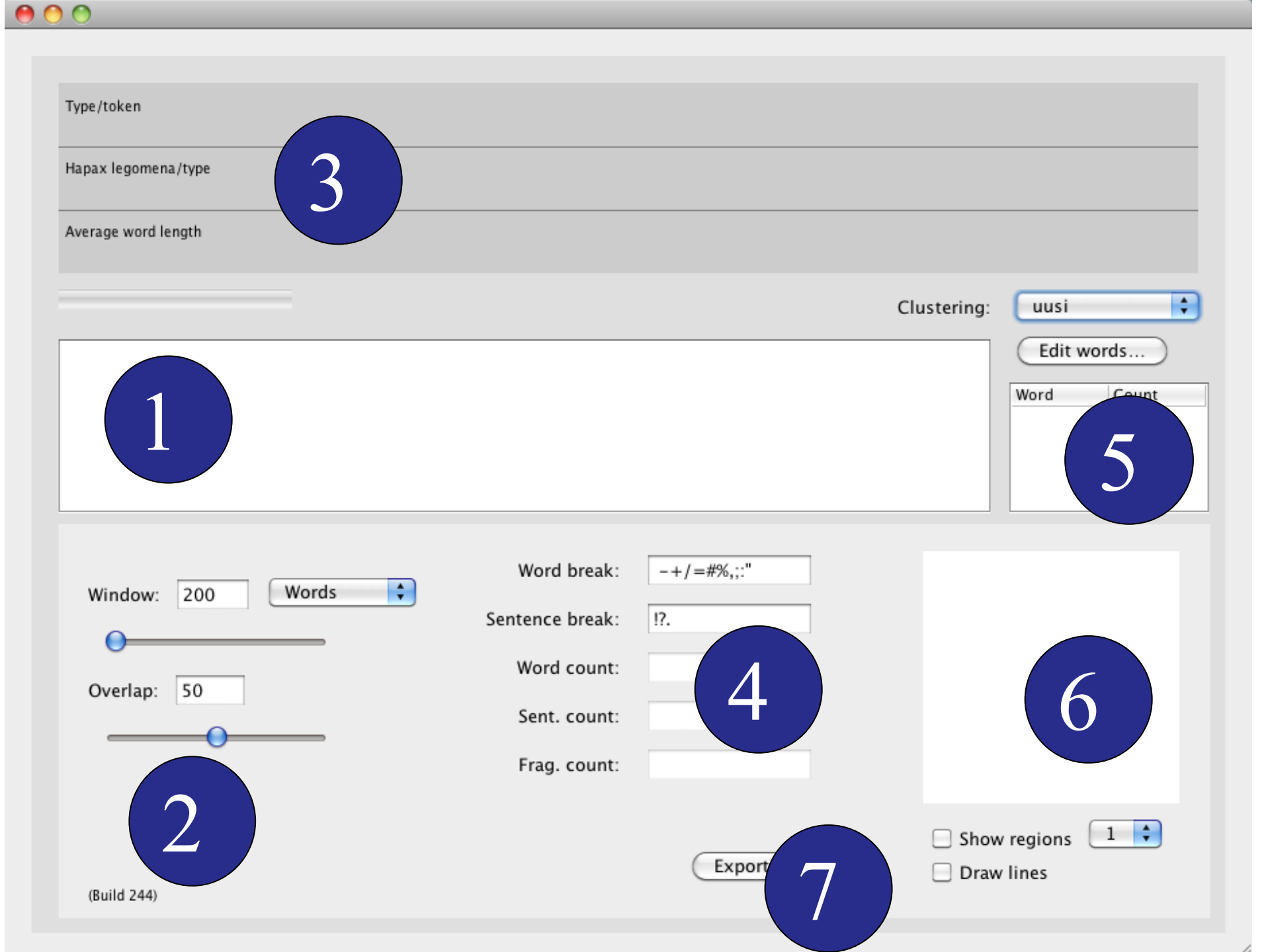
- Introduction
- Overview of TVE
- Exploring texts and text corpora
  - *Ulysses* by James Joyce
  - The Brown Family of Corpora
  - The Corpus of Early English Correspondence
- Conclusions

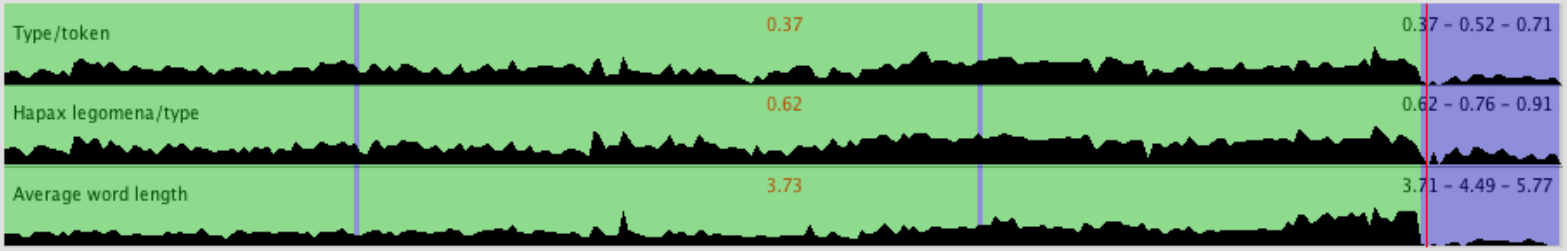
Text visualization: towards interactive tools

- tools needed for looking inside texts
- for finding out more about texts and corpora
  - before choosing data for analysis
  - for interpreting the results
- > “exploratory corpus linguistics”
- Text Variation Explorer (TVE) developed by Harri Siirtola & the DAMMOC project

Text Variation Explorer (TVE)

- can vary the size and overlap of the text fragments to be analyzed (**text view**)
- provides **line graphs** of three common text measures
  - type-token ratio (TTR)
  - proportion of hapax legomena
  - average word length
- clusters text fragments according to a given set of words (Principal Component Analysis)
  - the **PCA view** displays each text fragment as a point, and shows the values of first two PCs for it
- allows three-way **brushing**:
  - Click a point on any of the views (text, line graph, PCA) and the other two will update to show the relevant part


Type/token
Hapax legomena/type
Average word length
3
Clustering:
uusi
Edit words...
Word
Count
1
5
Window:
200
Words
Overlap:
50
2
(Build 244)
Word break:
-+/=#%,;:"
Sentence break:
!?.
Word count:
Sent. count:
Frag. count:
4
6
Export
7
Show regions
1
Draw lines



Clustering: Pronouns

Edit words...

eat at our table on Christmas day if you please O no thank you not in my house stealing my potatoes and the oysters 2/6 per doz going out to see her aunt if you please common robbery so it was but I was sure he had something on with that one it takes me to find out a thing like that he said you have no proof it was her proof O yes her aunt was very fond of oysters but I told her what I thought of her suggesting me to go out to be alone with her I wouldnt lower myself to spy on them the garters I found in her room the Friday she was out that was enough for me a little bit too much her face swelled up on her with temper when I gave her her weeks notice I saw to that better do without them altogether do out the rooms myself quicker only for the damn cooking and throwing out the dirt I gave it to him anyhow either she or me leaves the house I couldnt even touch him if I thought he was with a dirty barefaced liar and sloven like that one denying it up to my face and singing about the place in the W C too because she knew she was too well off yes because he couldnt possibly do without it that long so he must do it somewhere and the last

Word	Count
i	34
it	29
he	25
me	19
him	19
you	15
his	11
her	11
my	9
myself	4
your	4
she	4
its	3
them	3

2

Window: 996 Words

Overlap: 0

1

Word break: -+/=#%,;:"

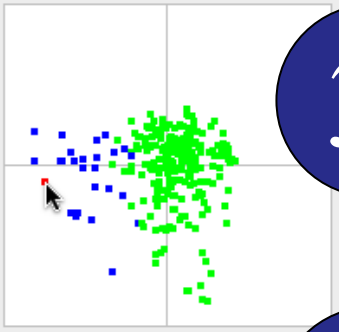
Sentence break: !?.

Word count: 266306

Sent. count: 24293

Frag. count: 268

Export...



3

Show regions 2

Draw lines

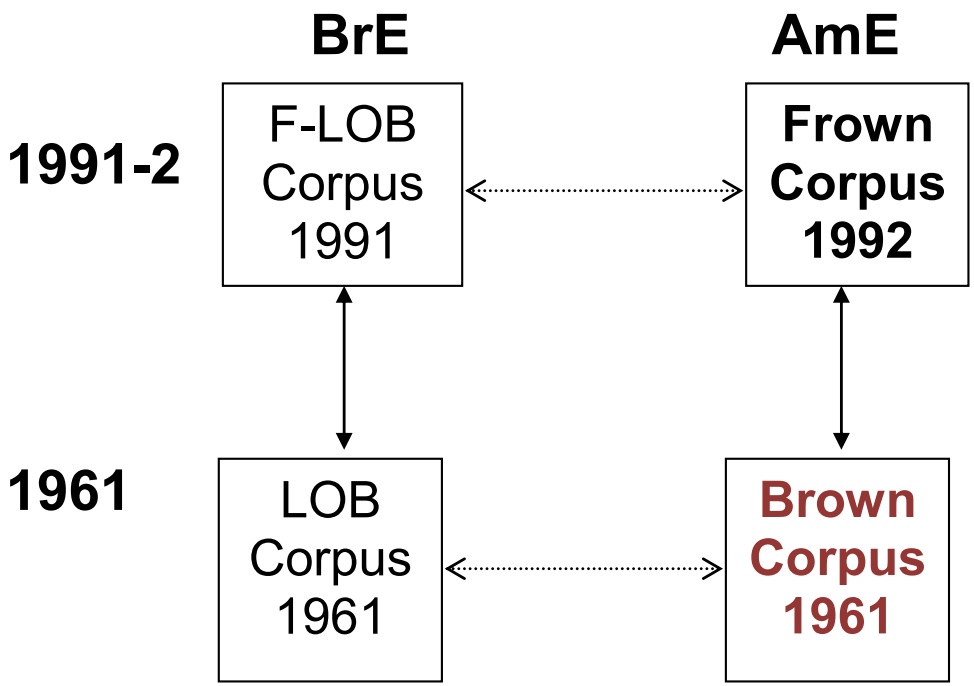
4

(Build 244)

Comparing and contrasting text corpora

- text corpora multiply
- e.g. the Brown family of matching corpora
  - structure: 15 text categories
  - size: 1 million words, 500 text samples, à 2,000 words
  - periods: 1961, 1991 (… 1930s, 1900s)
- how good a match?
- PCA: exploratory analysis using personal pronouns

The Brown Family of Corpora



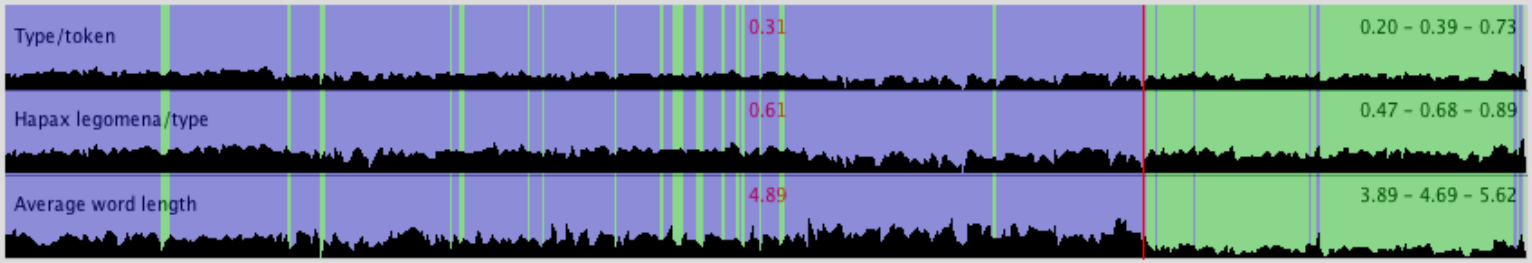
- The Frown (Freiburg-Brown) Corpus - American English, 1992
- The FLOB (Freiburg Lancaster-Oslo/Bergen) Corpus - British English, 1991

Text categories in the Brown Corpus Family

	Genre group	Category	Content of category
Informative prose	**Press**	A	Reportage
		B	Editorial
		C	Review
	General Prose	D	Religion
		E	Skills, trades and hobbies
		F	Popular lore
		G	Belles lettres, biographies, essays
		H	Miscellaneous
	Learned	J	Science
Imaginative prose	**Fiction**	K	General fiction
		L	Mystery and detective Fiction
		M	Science fiction
		N	Adventure and Western
		P	Romance and love story
		R	Humor

Non-fiction 75%

Fiction 25%



Clustering: Pronouns

Edit words...

an error in heading east. The earth is spinning at an angular velocity
~\q equal to one revolution per 24 hr&. When the platform is
level, ~|e is a rotation about the ~<Z> axis of the platform
**f. Since the earth is rotating and the unleveled gyro-stabilized platform
is fixed with respect to a reference in space, an observer on
the earth will see the platform rotating (with respect to the earth).
#THIRTY-THREE# SCOTTY did not go back to school. His parents
talked seriously and lengthily to their own doctor and to a specialist
at the University Hospital- Mr& McKinley was entitled to a
discount for members of his family- and it was decided it would be
best for him to take the remainder of the term off, spend a lot of time
in bed and, for the rest, do pretty much as he chose- provided, of
course, he chose to do nothing too exciting or too debilitating. His
teacher and his school principal were conferred with and everyone agreed

Word	Count
it	18
his	13
he	8
their	3
him	2
its	2
i	1
she	1
her	1
they	1
i'm	1

Window: 2000 Words

Overlap: 50

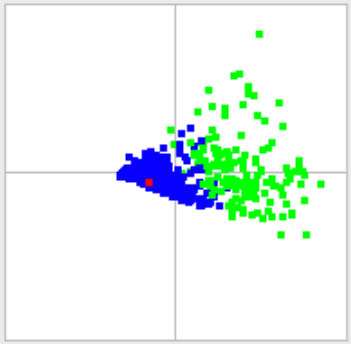
Word break: -+/=#%,;:"

Sentence break: !?.

Word count: 1025842

Sent. count: 48023

Frag. count: 527



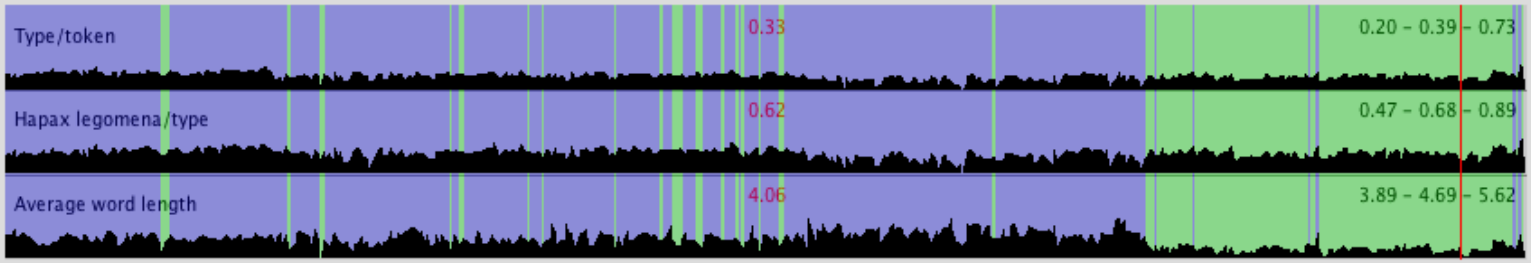
Show regions 2

Draw lines

Brown Corpus

Export...

(Build 244)



Clustering: Pronouns

Edit words...

some months later to move to Funk Furnaces. The job at Funk wasn't particularly better, but it got him away from being subordinate to John and assured him steady advancement, since Funk was owned to a large degree by various branches of Linda's family. Poor John's rise continued to be meteoric. When he was made a vice president only a year after the new sales job, a leading business magazine ran his photograph with a brief biography in a series on NATIONAL BUSINESS LEADERS OF THE FUTURE.
She called then to say she had a baby-sitter for that night. "Shirley appreciated the chance to make some money. Such a nice little thing- lives right in the building". "That's swell", I said sweetly. I could get along without that three dollars. In some ways it was worth being out the money- just knowing I was no longer obligated to Nadine! It was past midnight and we were in

Word	Count
i	81
she	26
it	24
her	22
you	20
him	18
his	13
we	13
he	12
they	9
me	8

Window: 2000 Words

Overlap: 50

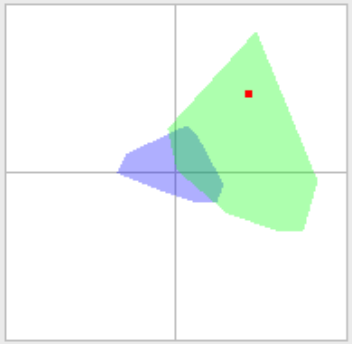
Word break: -+/=#%,;:"

Sentence break: !?.

Word count: 1025842

Sent. count: 48023

Frag. count: 527

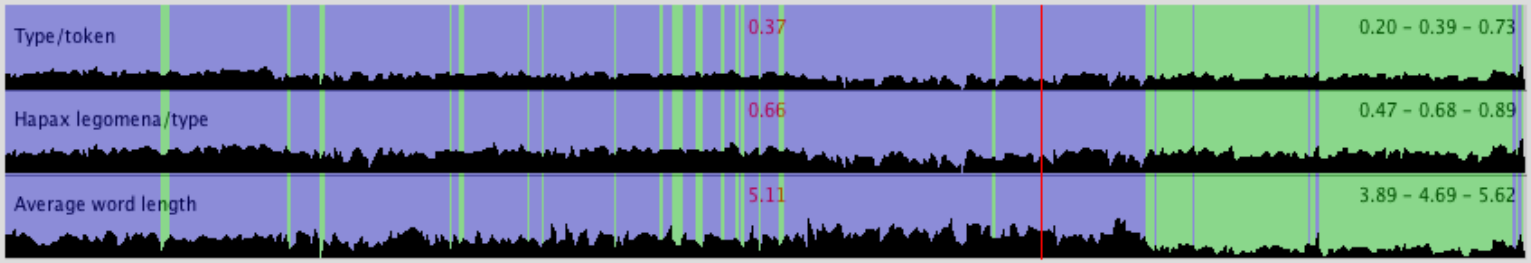


Show regions 2

Draw lines

Export...

(Build 244)



Clustering: Pronouns

change and broaden. _EMOTIONAL CHARACTERISTICS_ How a child
feels about himself, about other people, and about the tasks confronting
him in school may have as much influence on his success in school
as his physical and intellectual characteristics. A considerable amount
of evidence exists to show that an unhappy and insecure child is not
likely to do well in school subjects. Emotional maturity is the result
of many factors, the principal ones being the experiences of the
first few years of the child's life. However, the teacher who understands
the influence of emotions on behavior may be highly influential
in helping pupils gain confidence, security, and satisfaction.
Concerning this responsibility of the teacher, suggestions for helping
children gain better control of the emotions are presented in Chapter
11. The following generalizations about the emotional characteristics
of elementary-school children may be helpful. _1._ Typically,

Edit words...

Word	Count
his	16
their	11
he	10
they	10
it	7
him	3
we	1
them	1

Window: 2000 Words

Overlap: 50

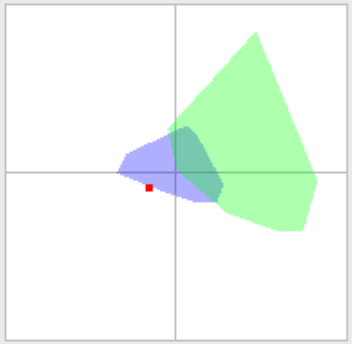
Word break: -+/=#%,;:"

Sentence break: !?.

Word count: 1025842

Sent. count: 48023

Frag. count: 527

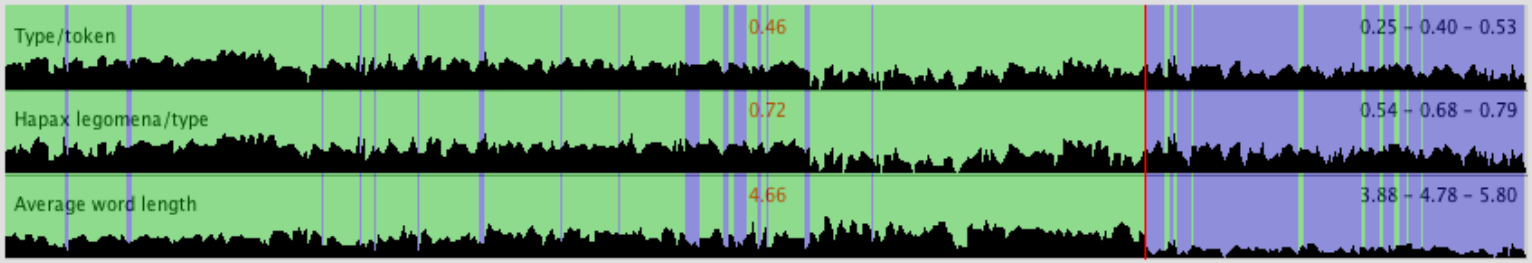


Show regions 2

Draw lines

Export...

(Build 244)



Clustering: Pronouns

Edit words...

radar itself. German equipment to warn submarines of ASV and night bombers of AI became a technology in its own right.
After the introduction in March 1943 of microwave ASV Mk III, an H 2S derivative, the U-boat became an ineffective weapon and the coffin of its crew.

Geno called B&B taxi, which took him to the Barrington campus along Route 9 in a rusty blue '84 Chevy. This road had been nothing but a slice of macadam through a cornfield when Geno arrived in Barrington seventeen years ago, but for him it was ruined now. Neon signs had arrived a decade ago, advertising a pizza shop and a bowling alley. Gas stations went up quickly,

Word	Count
i	15
he	14
it	14
his	13
you	11
its	11
him	7
us	5
she	3
me	2
your	2

Window: 2000 Words

Overlap: 50

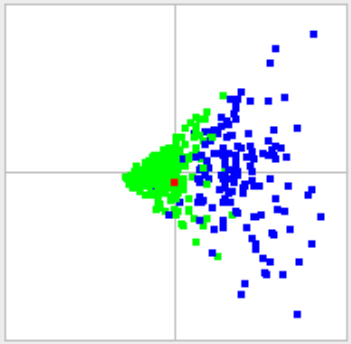
Word break: -+/=#%,;:"

Sentence break: !?.

Word count: 1020558

Sent. count: 60232

Frag. count: 524



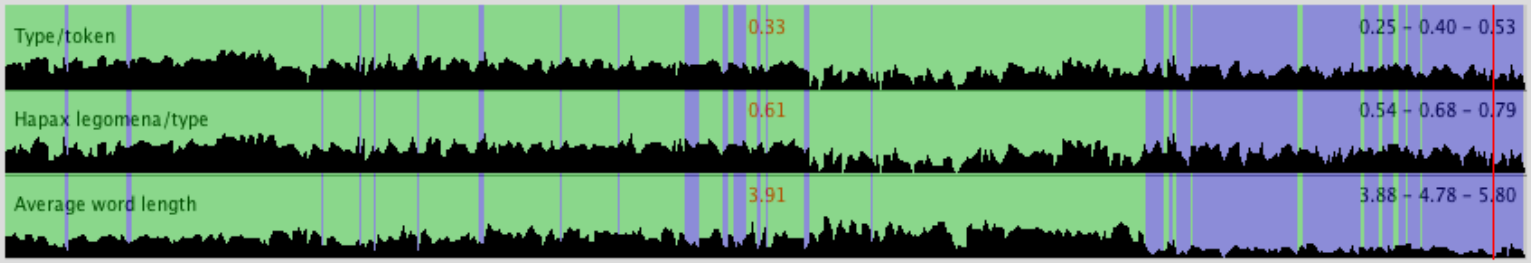
Show regions 2

Draw lines

Frown Corpus

Export...

(Build 244)



Clustering: Pronouns

Edit words...

feet to keep them dry. The night was still; the air began to smell sticky and old. They both kept their doors open, Tammy Wynette blending with the slow song of the crickets.
"You must think Espy's right special, you've stayed with him so long."

Separating
My mom is standing in the rain talking to the guy whose pickup she just rear-ended. It's getting dark. We've pulled off the road and the two of them are under a tree next to his truck. He's younger than Mom, wearing jeans, a T-shirt, and cowboy boots. When they laugh, Mom looks like she does with her dates, these guys that shake my hand and call me Sport. It's Michael, I tell them, but they don't listen to a fourteen-year-old. He gets a notebook and

Word	Count
i	63
she	44
he	41
her	33
you	24
me	23
it	22
we	15
my	13
his	12
our	11

Window: 2000 Words

Overlap: 50

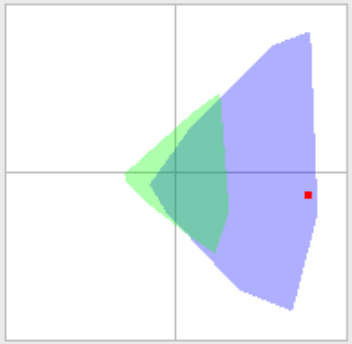
Word break: -+/=#%,;:"

Sentence break: !?.

Word count: 1020558

Sent. count: 60232

Frag. count: 524

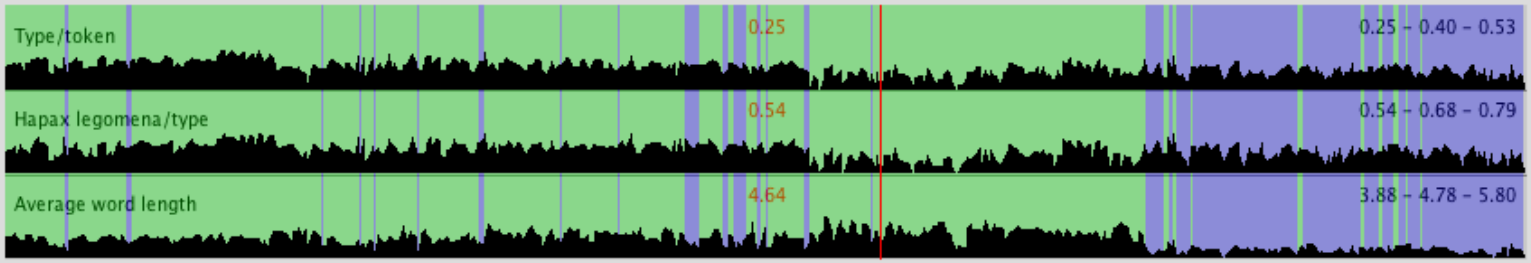


Show regions 2

Draw lines

Export...

(Build 244)



Clustering: Pronouns

Edit words...

expressed concern about the prospects for passage and the fact that "it mixes and matches entitlements and discretionary programs" (Perlman 1992, 23).
In light of the recession and the upcoming presidential election, Bush's overall $1.52 trillion fiscal 1993 budget proposal drew mixed reviews.

Fuels Used in Farming
You may be eligible to claim a credit or refund of excise taxes included in the price of fuel used on a farm for farming purposes, if you are the owner, tenant, or operator of a farm. You may claim only a credit for gasoline and special motor fuel used on a farm for farming purposes. You may claim either a credit or refund for diesel fuel and aviation fuel used on a farm for farming

Word	Count
you	27
your	17
it	15
its	3
his	1
her	1

Window: 2000 Words

Overlap: 50

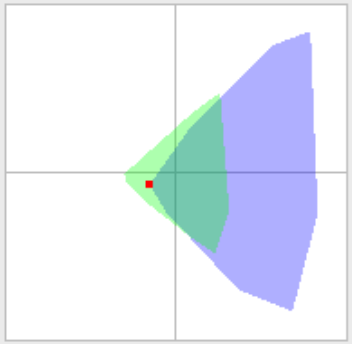
Word break: -+/=#%,;:"

Sentence break: !?.

Word count: 1020558

Sent. count: 60232

Frag. count: 524



Show regions 2

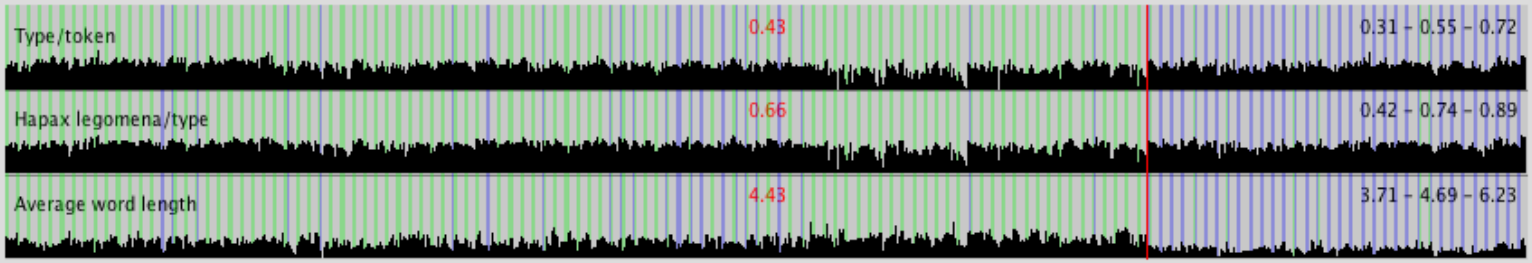
Draw lines

Export...

(Build 244)

Window size

- three **text measures**: TTR, proportion of hapax legomena, average word length
  - describe vocabulary richness, style of text
  - at least the first two **dependent on window size**
- Biber uses 400-word samples for TTR
- hapax-based measures said to stabilize around 1,300 words (Keim & Oelke)
- experiment: change window size from 2,000 to 400 words



Clustering: Pronouns

Edit words...

an error in heading east. The earth is spinning at an angular velocity
~\q equal to one revolution per 24 hr&. When the platform is
level, ~|e is a rotation about the ~<Z> axis of the platform
**f. Since the earth is rotating and the unleveled gyro-stabilized platform
is fixed with respect to a reference in space, an observer on
the earth will see the platform rotating (with respect to the earth).
#THIRTY-THREE# SCOTTY did not go back to school. His parents
talked seriously and lengthily to their own doctor and to a specialist
at the University Hospital- Mr& McKinley was entitled to a
discount for members of his family- and it was decided it would be
best for him to take the remainder of the term off, spend a lot of time
in bed and, for the rest, do pretty much as he chose- provided, of
course, he chose to do nothing too exciting or too debilitating. His
teacher and his school principal were conferred with and everyone agreed

Word	Count
his	4
he	3
it	2
him	1
their	1

Window: 400 Words

Overlap: 50

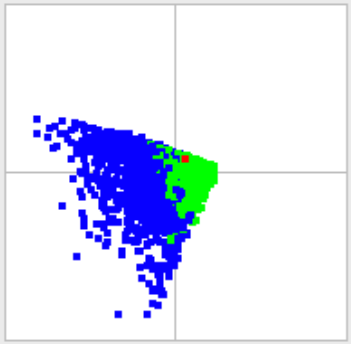
Word break: -+/=#%,;:"

Sentence break: !?.

Word count: 1025842

Sent. count: 48023

Frag. count: 2931



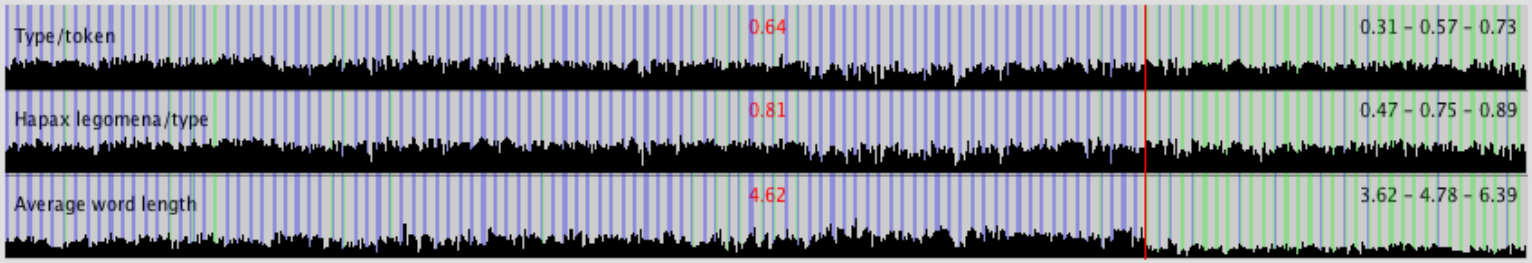
Show regions 2

Draw lines

Brown Corpus

Export...

(Build 244)



Clustering: Pronouns

course several orders stronger than the echo signal received back at the radar transmitter. The range of receivers designed to give warning of radar surveillance may therefore exceed the range of the radar itself. German equipment to warn submarines of ASV and night bombers of AI became a technology in its own right. After the introduction in March 1943 of microwave ASV Mk III, an H 2S derivative, the U-boat became an ineffective weapon and the coffin of its crew.

Geno called B&B taxi, which took him to the Barrington campus along Route 9 in a rusty blue '84 Chevy. This road had been nothing but a slice of macadam through a cornfield

Edit words...

Word	Count
his	3
its	3
he	2
him	2
it	2
itself	1
they	1
their	1

Window: 400 Words

Overlap: 50

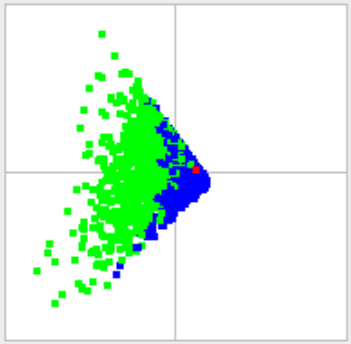
Word break: -+/=#%,;:"

Sentence break: !?.

Word count: 1020558

Sent. count: 60232

Frag. count: 2916



Show regions 2

Draw lines

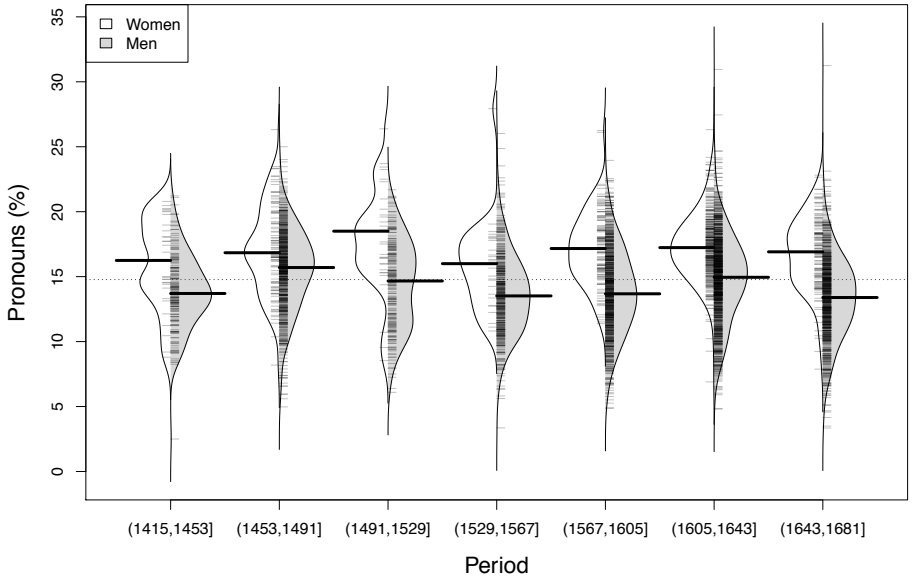
Frown Corpus

Export...

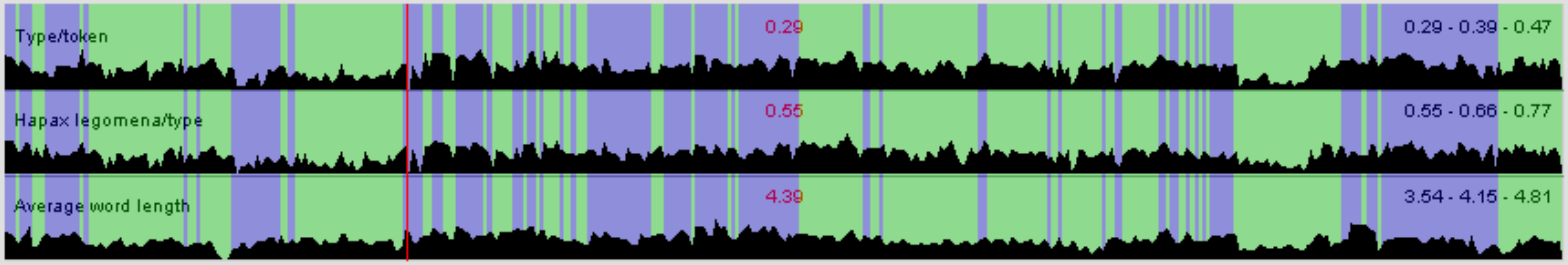
(Build 244)

Gender differences in a historical sociolinguistic corpus?

- women use more pronouns than men do in the Corpus of Early English Correspondence (CEEC; Säily, Nevalainen & Siirtola 2011)



- does gender difference show in the PCA view?



Clustering: Pronouns

therin more at large. And as touching your demeanyng in mariage, that for special causes greatly resteth in our mynde and pleasure, we have in likewise shewed unto hym the same by our said instruccions, to whom in declaring therof, and of everi othre thing concernyng the premisses, we desire you to yeve unto hym ful feith and credence, and with al effect applie and endevoir you to thexecucion and performyng of the same, as our great trust is in you. Yeven, &c. the xxix. day of Septembre.

<Q RER 1484 RICHARD3>
<A RICHARD III>
<P I,74>

[\A.D. 1484. September.\]
To Therle of Kildare.
Right trusti, &c. Certifieng you that as touching the

Edit words...

Word	Count
you	46
our	37
we	26
your	24
i	9
he	2
they	2
them	2
his	1

Window: 1300 Words

Overlap: 50

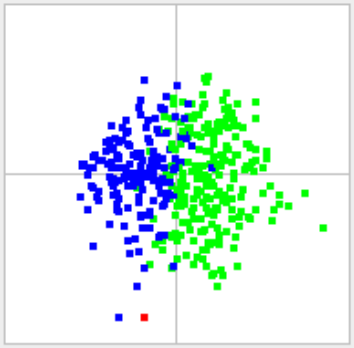
Word break: -+/=#%,;:"

Sentence break: !?.

Word count: 488309

Sent. count: 19508

Frag. count: 391



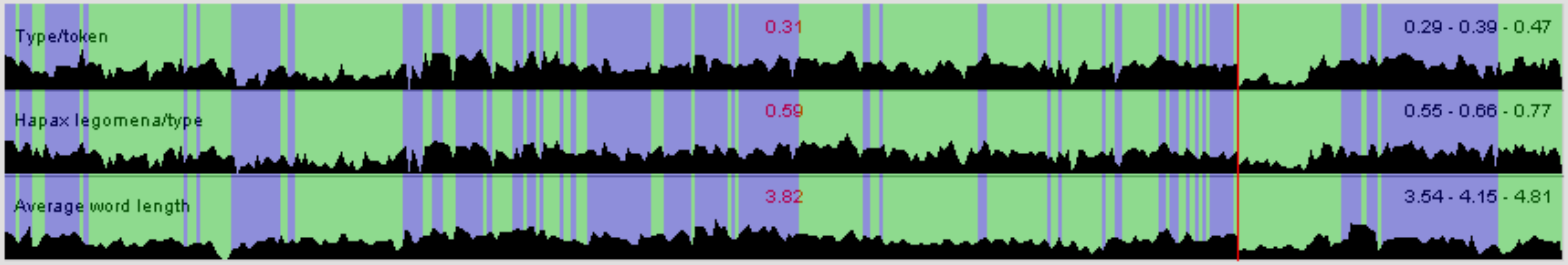
Show regions 2

Draw lines

The Corpus of Early English Correspondence Sampler (1410–1680)

(Build 244)

Export...



Clustering: Pronouns

good. My sister giues you thankes for seending him to her. I
pray you remember that I recken the days you are away; and I
hope you are nowe well at Heariford, wheare it may be, this
letter will put you in minde of me, and let you knowe, all your
frinds heare are well; and all the nwes I can seend you is, that
my Lo. Brooke is nowe at Beaethams Court. My hope is to see you
heare this day senet, or to-morrowe senet, and I pray God giue
vs a happy meeting, and presarfe you safe; which will be the
great comfort of
Your most true affectionat wife, Brilliana Harley.
(^Ragly: the 30 of Sep. 1625.^)

<Q HAR 1625 BHARLEY>
<A LADY BRILLIANA HARLEY>

[} [\II.\] }]

Edit words...

Word	Count
i	70
you	66
your	29
my	27
it	20
me	14
he	5
his	5
they	4
her	2
our	2
them	2
myself	1

Window: 1300 Words

Overlap: 50

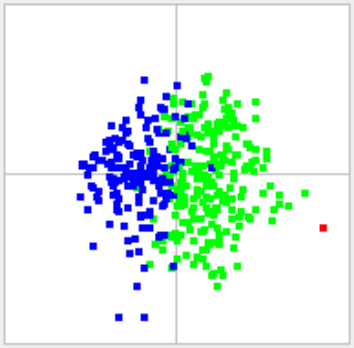
Word break: -+/=#%,;:"

Sentence break: !?.

Word count: 488309

Sent. count: 19508

Frag. count: 391



Show regions 2

Draw lines

The Corpus of Early English Correspondence Sampler (1410–1680)

(Build 244)

Export...

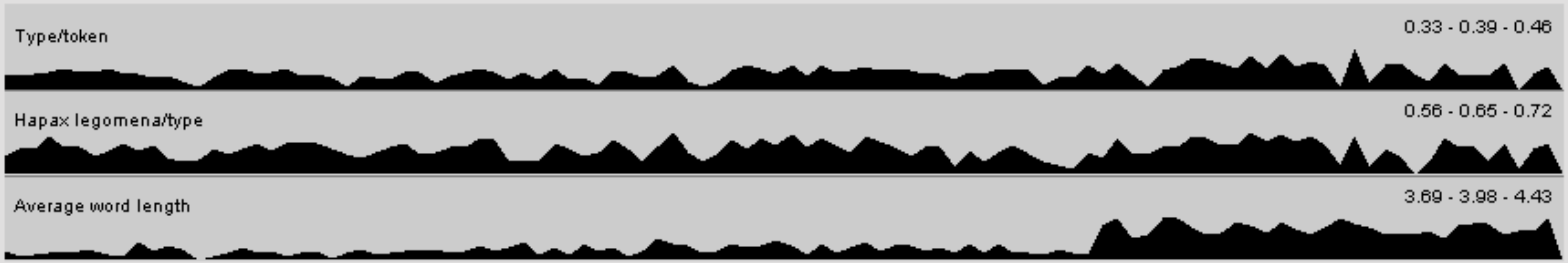
How different were women writers?

Dorothy Osborne (1671)

http://en.wikipedia.org/wiki/File:Dorothy,_Lady_Temple_by_Gaspar_Netscher.jpg

Lady Arabella Stuart (1605)

http://en.wikipedia.org/wiki/File:Lady_Arabella_Stuart.jpg



Clustering: Minim

of Euridices relapse) and vowing I would never answear to those
names by which I was called, and recalled, and cried out upon
(for if I should my love might be ashamed of me as now he may
well be of him self) I took my way down with a heavy heart
and being followed by them whom it might better have become us
both I should have followed, I was fain to set a good face on
bad fortune and there we had another skirmish where you and I
sat scribbling till .12. of the clock at night. but I finding my
self scarce able to stand on my feet what for my side and what
for my head, yet with a commanding voice called a troupe of
such viragoes as Virgilles Camilla that stood at the receipt in
the next chamber and never intreat[\ing\] them to give nor take
blows for my sake, was content to send
<P 153>
you the first news of this conflict; but though he were my
own man I sent for yet he being not so forward as certain

Edit words...

Word	Count

Window: 1300 Words

Overlap: 50

Word break: -+/=#%,;:"

Sentence break: !?.

Word count: 105377

Sent. count: 2592

Frag. count: 106

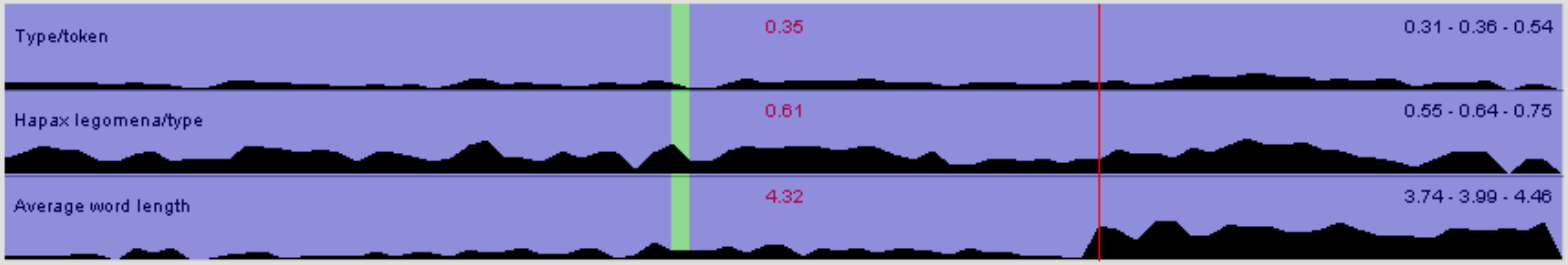
Dorothy Osborne vs. Arabella Stuart

Show regions 1

Draw lines

(Build 244)

Export...



Clustering: Pronouns

<Q A 1588 FO ASTUART>
<X ARABELLA STUART>
<P 119>
[} [\1 TO ELIZABETH TALBOT, COUNTESS OF SHREWSBURY, 8 FEBRUARY 1587/8\] }]
[\Addressed:\] To the right honourable my very good Lady and Grandmother the Countess of Shrewsbury.
Good Lady Grandmother, I have sent your Ladyship, the ends of my hair which were cut the sixth day of the moon, on saturday last; and with them, a pot of Gelly, which my Servant made; I pray God you find it good. My Aunt Cavendishe was here on Monday last, she certified me, of your Ladyship's good health, and dispositione, which I pray God long to continue. I am in good health, my Cousin Mary hath had three little fits of an agew, but now she is well, and merry. Thus with my humble duty unto your Ladyship and humble thanks for the token, you sent

Edit words...

Word	Count
i	48
your	34
my	31
her	19
me	13
it	13
you	7
they	5
his	3
them	1
their	1
themselves	1

Window: 1300 Words

Overlap: 50

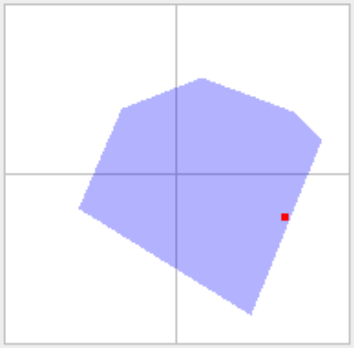
Word break: -+/=#%,;:"

Sentence break: !?.

Word count: 105377

Sent. count: 2592

Frag. count: 85

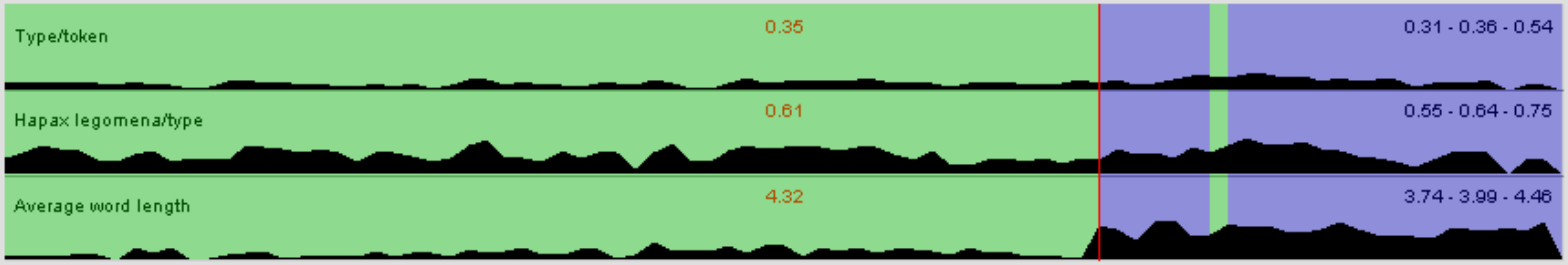


Show regions 2

Draw lines

Export...

(Build 244)



Clustering: Binongo's list

<Q A 1588 FO ASTUART>
<X ARABELLA STUART>
<P 119>
[} [\1 TO ELIZABETH TALBOT, COUNTESS OF SHREWSBURY, 8 FEBRUARY 1587/8\] }]
[\Addressed:\] To the right honourable my very good Lady and Grandmother the Countess of Shrewsbury.
Good Lady Grandmother, I have sent your Ladyship, the ends of my hair which were cut the sixth day of the moon, on saturday last; and with them, a pot of Gelly, which my Servant made; I pray God you find it good. My Aunt Cavendishe was here on Monday last, she certified me, of your Ladyship's good health, and dispositione, which I pray God long to continue. I am in good health, my Cousin Mary hath had three little fits of an agew, but now she is well, and merry. Thus with my humble duty unto your Ladyship and humble thanks for the token, you sent

Edit words...

Word	Count
and	60
to	45
of	40
the	36
all	18
that	16
it	13
so	13
not	12
with	12
for	11
in	10
which	10

Window: 1300 Words

Overlap: 50

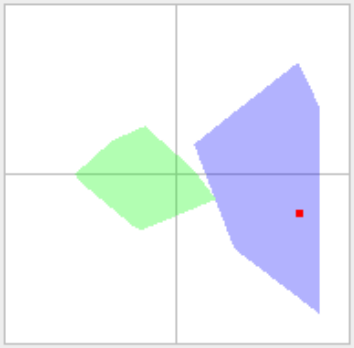
Word break: -+/=#%,;:"

Sentence break: !?.

Word count: 105377

Sent. count: 2592

Frag. count: 85

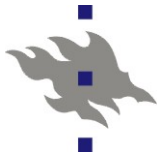


Show regions 2

Draw lines

Export...

(Build 244)



Conclusion

- TVE is a simple but flexible tool
- multiple measures are useful
- window size matters
  - 400 words is small for TTR and hapaxes
- TVE will be freely available by the end of 2011 via the DAMMOC home page
  - http://tauchi.cs.uta.fi/virg/projects.html

References

- Biber, Douglas (1988). *Variation across Speech and Writing*. Cambridge: CUP.
- Binongo, José Nilo G. (2003). Who wrote the 15th Book of Oz? An application of multivariate analysis to authorship attribution. *Chance* 16(2): 9–17.
- CEEC = *The Corpus of Early English Correspondence* (1998), comp. by Terttu Nevalainen, Helena Raumolin-Brunberg, Jukka Keränen, Minna Nevala, Arja Nurmi & Minna Palander-Collin. Helsinki: University of Helsinki.
- Keim, Daniel A. & Daniela Oelke (2007). Literature fingerprinting: A new method for visual literary analysis. *IEEE Symposium on Visual Analytics Science and Technology 2007*. October 30 – November 1, Sacramento, CA, USA.
- Säily, Tanja, Terttu Nevalainen & Harri Siirtola (2011). Variation in noun and pronoun frequencies in a sociohistorical corpus of English. *Literary and Linguistic Computing* 26(2): 167–188.