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
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 wouldn't 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 week's notice I saw to it 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 couldn't even touch him if I thought he was with a dirty bare faced liar and slyman 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 couldn't possibly do without it that long so he must do it somewhere and the last time he came on my button when yes it the right Roder gave my hand.
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

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

**Non-fiction**
- 75%

**Fiction**
- 25%
an error in heading east. The earth is spinning at an angular velocity
\( \omega \) equal to one revolution per 24 hr\&. When the platform is
level, \( \omega \) is a rotation about the \( \omega < z \) axis of the platform
**f. Since the earth is rotating and the unlevied 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

Brown Corpus

(Build 244)
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 obliged to Nadine! It was past midnight and we were in...
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,
Frown Corpus

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.
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
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

Window: 2000  Words
Overlap: 50

Word break: - / = % , ; : " 
Sentence break: ?,.

Word count: 1020558
Sent. count: 60232
Frag. count: 524

Export...
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
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, \~ie 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
course several orders stronger than the echo signal received back from 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 2 S derivative, the U-boat became an ineffective weapon and the coiffin of its crew.

Ceno 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.
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?
The Corpus of Early English
Correspondence Sampler (1410–1680)
good. My sister gives you thanks for sending him to her. I pray you remember that I reckon the days you are away, and I hope you are nowe well at Heariford, where it may be, this letter will put you in minde of me, and let you knowe, all your friends heare are well; and all the newes I can send you is, that my Lo. Brocke is nowe at Beaethams Court. My hope is to see you heare this day senet or to-morrow senet, and I pray God give us a happy meeting, and preserve you safe, which will be the great comfort of

Your most true affectionate wife, Brilliana Harley.

[March the 19 of Sept. 1625.]

<Q HArr 1625 BHarley>

<Q A LADY BRILLIANA HARLEY>
How different were women writers?

Dorothy Osborne (1671)

Lady Arabella Stuart (1605)
Dorothy Osborne vs. Arabella Stuart
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 geli; which my Servant made; I
pray God you find ill good. My Aunt Cavendish was here on
Monday last, she certified me, of your Ladyship's good health,
and disposition, which I pray God long to continue. I am in
good health; my Cousin Mary hath had three little fits of an
age, but now she is well, and merry. Thus with my humble duty
unto your Ladyship and humble thanks for the token, you sent.
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 Cavendish was here on Monday last, she certified me, of your Ladyship's good health, and disposition, which I pray God long to continue. I am in good health; my Cousin Mary hath had three little fits of an age, but now she is well, and merry. Thus with my humble duty unto your Ladyship and humble thanks for the token, you sent.
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
References


Measures used in TVE (1)

- **Type-token ratio (TTR)**
  - varies according to sample size
  - useful in comparative analysis if sample size is fixed

- **Hapax legomena**
  - vary according to sample size
  - e.g. in the Brown Corpus of American English, about 50% of the 50,000 word types are *hapax legomena* (Leech 2004: 87)

- **Average word length**
  - less affected by sample size
  - but texts with the same average word-length might have different spectra
Measures used in TVE (2)

- Principal Component Analysis (PCA)
  - the first component accounts for the largest variance, the second for the second largest

- Word lists
  - fifty common function words (Binongo 2003)
  - personal pronouns
  - modal auxiliaries
  - [or any other list]