Tanja Säily, Terttu Nevalainen
VARIENG Research Unit, University of Helsinki
& Harri Siirtola
University of Tampere

Nothing beats a picture except an interface:
Interactive visualisation for corpus linguistics

AVML 2012, York
5–7 September 2012
Outline

- Introduction
- Overview of Text Variation Explorer
- Exploring corpora
  - The Brown Family of Corpora
  - The Corpus of Early English Correspondence
- Conclusions
Corpus linguists need…

- Exploratory visualisation and analytic tools
  - ‘Quick and dirty’, multiple visualisations
  - Interactive: continuous, direct manipulation

- Tools for statistical, confirmatory analysis
  - Usually passive, stepped mode of interaction

- Explanatory tools for presenting the results
  - Polished, limited number of visualisations
Exploratory corpus linguistics

- Tools needed for looking inside texts
- For finding out more about texts and corpora
  - Before choosing data for analysis
  - For interpreting the results

- Key features
  - Interaction → pattern discovery, insight generation
    (Pike et al. 2009)
  - Connection between text and visualisation
  - Usability
Text Variation Explorer (TVE)

Video clip introducing TVE:
http://www.cs.helsinki.fi/u/tsaily/presentations/tve.mp4
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
<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 Religion</td>
<td></td>
</tr>
<tr>
<td></td>
<td>E Skills, trades and hobbies</td>
<td></td>
</tr>
<tr>
<td></td>
<td>F Popular lore</td>
<td></td>
</tr>
<tr>
<td></td>
<td>G Belles lettres, biographies, essays</td>
<td></td>
</tr>
<tr>
<td>Learned</td>
<td>H Miscellaneous</td>
<td></td>
</tr>
<tr>
<td>Imaginative prose</td>
<td>Fiction</td>
<td></td>
</tr>
<tr>
<td></td>
<td>K General fiction</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L Mystery and detective Fiction</td>
<td></td>
</tr>
<tr>
<td></td>
<td>M Science fiction</td>
<td></td>
</tr>
<tr>
<td></td>
<td>N Adventure and Western</td>
<td></td>
</tr>
<tr>
<td></td>
<td>P Romance and love story</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R Humor</td>
<td></td>
</tr>
</tbody>
</table>

Non-fiction: 75%

Fiction: 25%
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, ~\le is a rotation about the ~\langle Z\rangle axis of the platform
**vf. Since the earth is rotating and the unlevelled 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
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
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,
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 25 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

Word break: -+/=#%,;":
Sentence break: !?.
Word count: 1020558
Sent. count: 60232
Frag. count: 524
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.
## Corpora of Early English Correspondence

http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>words</td>
<td>450,085</td>
<td>2,159,132</td>
</tr>
<tr>
<td>collections</td>
<td>23</td>
<td>84</td>
</tr>
<tr>
<td>letters</td>
<td>1,123</td>
<td>4,970</td>
</tr>
<tr>
<td>writers</td>
<td>194</td>
<td>666</td>
</tr>
<tr>
<td>time span</td>
<td>1418–1681</td>
<td>1410–1681</td>
</tr>
</tbody>
</table>

* Historical Sociolinguistics Team’s collaborative project with the University of York, Prof. Anthony Warner, Dr. Susan Pinzuk and Dr. Ann Taylor. Tagging by Arja Nurmi and parsing by Ann Taylor.
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?
therin more a targe. And as touching your demeaning in marriage, that for special causes greatly resteth in our mynde and pleasure, we have in likewise shewed unto hym the same by our said instructions, to whom in declaring therof, and of every other thing concerning the premisses, we desire you to yeve unto hym ful faith and credence, and with al effect apple and endeavor you to the execution and performing of the same, as our great trust is in you. Yeven, &c. the xxix. day of Septembre.

To Gerald FitzGerald
A.D. 1484. September.
To The er of Kildare,
Right trust, &c. Certif ieng you that as touching the

The Corpus of Early English Correspondence Sampler (1410–1680)
good. My sister gies you thankes for seending him to her. I pray you remember that I reckon the days you are away; and I hope you are nowe well at Heartford, where 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 newes I can send you is, that my Lo. Brooke is nowe at Beaethams Court. My hope is to see you heare this day senet, or to-morrow senet, and I pray God glue us a happy meeting, and preserfe you safe; which will be the great comfort of
Your most true affectionat wife, Briliana Harley.
("Reply the 30 of Sep. 1625")

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

- Interaction is crucial
  - Preserves connection between text and visualisation
  - Enhances data exploration

- TVE is a simple but flexible tool
  - Multiple measures are useful

- TVE is freely available:

- More tools like this needed in corpus linguistics


How different were women writers?

Dorothy Osborne (1671)

Lady Arabella Stuart (1605)
Dorothy Osborne vs. Arabella Stuart

Eventually, with great reluctance, I went back to my side and sat at the receipt in the next chamber and never treated them to give or take blows for my sake, was content to send 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.
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 jelly, 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
agew, but now she is well, and merry. Thus with my humble duty
unto your Ladyship and humble thanks for the token, you sent
<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, 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 Geliey, which my Servant made; I pray God you find it good. My Aunt Cavendish was here on Monday last, she certify 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 agew, but now she is well, and merry. Thus with my humble duty unto your Ladyship and humble thanks for the token, you sent