Comparing like with like?
Tools for exploring families of corpora

Terttu Nevalainen and Tanja Säily

with thanks to Jukka Suomela
Outline

• Representativeness vs. comparability of corpora?
  • How similar/different are the components of corpus families?
  • BROWN, CEEC, ICE etc.
• Methods of comparing corpora
  • Keyword analysis
  • Exploratory corpus linguistics: TVE
• Case study
  • Comparing ICE-Great Britain, ICE-New Zealand, ICE-India and ICE-Canada
Representativeness vs. comparability of corpora?

• Access to comparable materials is essential for the study of linguistic change and regional variation

• **International Corpus of English (ICE):** Each component corpus contains 500 texts of approximately 2,000 words each – a total of one million words
  • All texts date from 1990 or later
  • Authors and speakers of the texts are aged 18 or above
    – born in the country in whose corpus they are included, or moved there at an early age
    – educated through the medium of English
  • [http://ice-corpora.net/ice/design.htm](http://ice-corpora.net/ice/design.htm)
Issues with representativeness (1/2)

• ICE corpora contain samples of speech and writing by both males and females and different age groups

• General issues:
  • the proportions of speakers and writers are not representative of the proportions in the population as a whole because:
    – women are not equally represented in professions such as politics and law
    – various age groups are not equally represented among students or academic authors.
Issues with representativeness (2/2)

• Specific local issues:
  • “This task of creating a corpus that matches the general ICE framework, but which is still representative of the universe of the local and culture-specific production of standard English texts in Fiji, quickly started to look as difficult as squaring the circle.” (Biewer, Hundt & Zipp 2010: 5)
  • “While it makes sense to achieve success in both representativeness and comparability, there is a sense in which these two goals conflict: an attempt to achieve greater comparability may actually impede representativeness and vice versa.” (Leech 2007: 142, as cited by Biewer et al.)
How to deal with comparability: compilers’ view

- Biewer at al. (2010: 16): to provide an adequate database for the study of English worldwide, we need:
  - a consensus among all ICE teams on fundamental issues
  - a common agreement on changes in the general ICE framework, where necessary, to include the new generation of ESL/ESD components
  - a better coordination of all individual ICE projects
  - a comprehensive documentation of decisions past and present of all ICE teams, in manuals that go beyond a mere listing of filenames and markup conventions
How to deal with comparability: corpus users’ view

- **Exploration** of the basic similarities of/differences between corpora using statistical and corpus linguistic tools
- **Keyword analysis**: used to discover differences
  - E.g. Oakes and Farrow (2007) applied the chi-square test in their study of the Brown corpus family and the corpora of Australian (ACE), Indian (Kolhapur), New Zealand (WC) English, plus two East African English corpora (Kenya, Tanzania).
- Interactive visualization: **Text Variation Explorer**
  - Can be used to explore both similarities and differences
Interactive visualization: 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
TVE: basic functions

Line graphs for measures

Text pane

List of words for clustering

Sliders for window settings

PCA view
Sample data: Ulysses

Comparing like with like? Tools for exploring families of corpora / Terttu Nevalainen & Tanja Säily
Case study: Comparing four ICE corpora

- Comparing four varieties (converted to plain text)
  - ICE-Canada
  - ICE-Great Britain
  - ICE-India
  - ICE-New Zealand
- Contrasting
  - Medium: spoken vs. written
  - Genre: conversation, academic writing, press, fiction (e.g. Leech et al. 2009)
  - Sub-genre: e.g. academic writing
    - humanities, social sciences, natural sciences, technology
ICE: Canada, GB, India, NZ

<table>
<thead>
<tr>
<th>CAN</th>
<th>GB</th>
<th>IND</th>
<th>NZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type/token</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypothesis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average word length</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mum was fine.
Henry complained of the 'old Jap problem.'
But why, I wondered, had they sent the postcard of a marlin?
I thought about this a good deal.
Did they suspect on my part an interest in marlin?
Or had the card simply been wiped off a revolving stand in an act of duty?
I placed it in a cupboard with fifty other such cards: Australia Square, the Opera House, Cold Coast meter maids, koala bears, one-legged abo's standing in the desert; on the other side my mother's concerned words to 'take care...': wave riders; Canberra spread out like the blank digits on a Seiko watch; resorts, river cruises, Bondi; lifesavers; Ayers Rock; and the one time I thought ill of these cards I reminded myself of the reporter's question all those years ago - when she asked if I had drunk all those cans?
Or for that matter had I a lick of all those candy striped iceblocks the sticks of which spiralled down from the motorway.
Likewise the rest of the debris of consumption I stuffed in the ramparts of the house.
ICE: Canada, GB, India, NZ

Mum was fine.
Henry complained of the 'old jib problem'.
But why, I wondered, had they sent the postcard of a marlin?
I thought about this a good deal.
Did they suspect on my part an interest in marlin?
Or had the card simply been wiped off a revolving stand in an act of duty?
I placed it in a cupboard with fifty other such cards: Australia Square, the Opera House, Gold Coast meter maids, koala bears, one-legged abo's standing in the desert; on the other side my mother's concerned words to 'take care...'; wave riders; Canberra spread out like the blank digits on a Seiko watch; resorts, river cruises, Bondi, lifesavers; Ayers Rock; and the one time I thought ill of these cards I reminded myself of the reporter's question all those years ago - when she asked if I had drunk all those cans? Or for that matter had I a lick of all those candy striped iceblocks the sticks of which spiralled down from the motorway.
Likewise the rest of the debris of consumption I stuffed in the ramparts of the house.
ICE: Canada, GB, India, NZ

Mum was fine.
Henry complained of the 'old jap problem'.
But why, I wondered, had they sent the postcard of a marlin?
I thought about this a good deal.
Did they suspect on my part an interest in marlin?
Or had the card simply been swept off a revolving stand in an act of duty?
I placed it in a cupboard with fifty other such cards: Australia Square, the Opera House, Gold Coast meter maids, koala bears, one-legged abo's standing in the desert; on the other side my mother's concerned words to 'take care ...'; wave riders; Canberra spread out like the blank digits on a Seiko watch; resorts, river cruises, Bondi; lifesavers; Ayers Rock; and the one time I thought ill of these cards I reminded myself of the reporter's question all those years ago — when she asked if I had drunk all those cans? Or for that matter had I a lick of all those candy striped iceblocks the sticks of which spiralled down from the motorway.
Likewise the rest of the debris of consumption I stuffed in the ramparts of the house.

Window: 20000 words
Word break: -+/=#%":",,?

Word count: 4265439
Frag. count: 214
Conversation

Comparing like with like? Tools for exploring families of corpora / Terttu Nevalainen & Tanja Säily
Conversation

Comparing like with like? Tools for exploring families of corpora / Terttu Nevalainen & Tanja Säily
Academic writing

However the costs of such equipment needs to be balanced against the potential reuse factors in New Zealand.

Alternative screen and barrier protection systems can be put in place which will provide an upgrading in safety sought both on behalf of the worker and public.

So whether the gay nineties?

There is no doubt the industry fortunes of concrete construction are linked to the general building activity.

Depressed activity as at present perhaps suppresses the drive towards pursuing different techniques and materials.

However there is no doubt the industry is now well aware of the opportunities and responsibilities it must take for the future to face competition.

The developments during the past decade show the industry's capability to respond to the changing scenes, it just has to keep it up.

<table>
<thead>
<tr>
<th>Window: 2000 words</th>
<th>Word break: &quot;+/=/#%&quot;:&quot;?;&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overlap: 0</td>
<td>Word count: 362885</td>
</tr>
<tr>
<td></td>
<td>Frag. count: 182</td>
</tr>
</tbody>
</table>

Export...
However the costs of such equipment needs to be balanced against the potential reuse factors in New Zealand. Alternative screen and barrier protection systems can be put in place which will provide an upgrading in safety sought both on behalf of the worker and public. So whether the gay nineties?

There is no doubt the industry fortunes of concrete construction are linked to the general building activity. Depressed activity as at present perhaps suppresses the drive towards pursuing different techniques and materials. However there is no doubt the industry is now well aware of the opportunities and responsibilities it must take for the future to face competition. The developments during the past decade show the industry's capability to respond to the changing scenes, it just has to keep it up.
Introduction rests its political destiny, and that of the nation.
Prolonged industrial counteraction to the legislation can only aggravate New Zealand’s financial woes.

A new PM in Paris
French politics traditionally have a sense of the dramatic, Gallic fervour and passion.

Though the days have gone when the French changed their governments with unusual frequency life as a politician there is certainly no sinecure and there is no security of office as our Parliamentarians know it.

Take the case of the prime minister, Michel Rocard who visited New Zealand two weeks ago.

He resigned this week and while the ostensible reason for that decision is that he was failing to grapple with rising unemployment and economic problems, his strained relations with the French president, Francois Mitterrand, is believed to be another dominant factor.

However, an apology he offered to New Zealand over the sinking of the Greenpeace ship Rainbow Warrior could have been instrumental in his fall from power.

This apology, it is said, was too much for Gallic honour, though some would suggest arrogance may be a better word than honour.
Introduction rests its political destiny, and that of the nation.
Prolonged industrial counteraction to the legislation can only aggravate New Zealand's financial woes.

A new PM in Paris
French politics traditionally have a sense of the dramatic, Gallic fervour and passion.
Though the days have gone when the French changed their governments with unusual frequency life as a politician there is certainly no sinecure and there is no security of office as our Parliamentarians know it.
Take the case of the prime minister, Michel Rocard who visited New Zealand two weeks ago.
He resigned this week and while the ostensible reason for that decision is that he was failing to grapple with rising unemployment and economic problems, his strained relations with the French president, Francois Mitterrand, is believed to be another dominant factor.
However, an apology he offered to New Zealand over the sinking of the Greenpeace ship Rainbow Warrior could have been instrumental in his fall from power.
This apology, it is said, was too much for Gallic honour, though some would suggest arrogance may be a better word than honour.
Mum was fine.
Henry complained of the 'old Jap problem.'
But why, I wondered, had they sent the postcard of a marlin?
I thought about this a good deal.
Did they suspect on my part an interest in marlin?
Or had the card simply been swiped off a revolving stand in an act of duty?
I placed it in a cupboard with fifty other such cards: Australia Square, the Opera House, Gold Coast meter maids, koala bears, one-legged abo's standing in the desert; on the other side my mother's concerned words to 'take care...'; wave riders; Canberra spread out like the blank digits on a Seiko watch; resorts, river cruises, Bondi; lifesavers; Ayers Rock; and the one time I thought ill of these cards I reminded myself of the reporter's question all those years ago — when she asked if I had drunk all those cans?
Or for that matter had I a lick of all those candy striped iceblocks the sticks of which spiralled down from the motorway.
Likewise the rest of the debris of consumption I stuffed in the ramparts of the house.
Results

• **PCA** (pronouns & function words): differences and similarities emerge at the level of
  • Corpus: GB+India vs. Canada+NZ?
  • Genre: India singled out in conversation, GB (+India) in press
  • Subgenre: humanities vs. science across corpora (dividing line varies)

• **Line graphs** show overall trends, e.g.
  • conversation & fiction: NZ has lower avg word length
Conclusions

• TVE can provide a quick, interactive overview of similarities and differences across corpora
  • Preserves the connection between text and visualisation
• Multiple measures are useful
  • Besides TTR, hapax legomena & avg word length, user-defined word lists (e.g. pronouns) for PCA
  • To test the robustness of similarities/differences, vary the window size and export the results e.g. to R for statistical testing
• Developed by Harri Siirtola as part of our joint DAMMOC project, TVE is freely available from:
  • http://www.uta.fi/sis/tauchi/virg/projects/dammoc/tve.html
References


