The Theory behind Keyword Analysis

Václav Cvrček
Workshop on Quantitative Text Analysis for SSH
Brown University
April 8, 2016
Introduction
Introduction

Text interpretation

- at the core of the humanities' mission
Introduction

Text interpretation

▶ at the core of the humanities’ mission
▶ our interpretation + other people’s interpretation
## Introduction

### Text interpretation

- at the core of the humanities’ mission
- our interpretation + other people’s interpretation
- interpretation with minimum amount of extra-textual information and intuition
Introduction

Text interpretation

- at the core of the humanities’ mission
- our interpretation + other people’s interpretation
- interpretation with minimum amount of extra-textual information and intuition
- frame of reference, scheme, expectations, communicative norms…

there is no objective interpretation – depends on point of view (recipient)
Introduction

Text interpretation

- at the core of the humanities’ mission
- our interpretation + other people’s interpretation
- interpretation with minimum amount of extra-textual information and intuition
- frame of reference, scheme, expectations, communicative norms...
- there is no objective interpretation – depends on point of view (recipient)
Language corpus
What is a corpus?

- sample of naturally occurring written texts or transcribed speeches
- stored electronically (searchable)
- basis for linguistic analysis and description
CADS = corpus assisted discourse studies

“A Needle in a Haystack” (collaborative research project of Brown and Charles University)

► how language reflects the changing nature of the society?
CADS = corpus assisted discourse studies

“A Needle in a Haystack” (collaborative research project of Brown and Charles University)

- how language reflects the changing nature of the society?
- how different is the interpretation of the contemporary and historical reader?
CADS = corpus assisted discourse studies

“A Needle in a Haystack” (collaborative research project of Brown and Charles University)

- how language reflects the changing nature of the society?
- how different is the interpretation of the contemporary and historical reader?
- how can we test the limits of the corpus-based quantitative analysis of text?

http://brown.edu/research/projects/needle-in-haystack/
Prominent items
How do we start with interpretation?

- what is striking in a text?
- topics, motives, themes – expressed by words
- interaction between words, topics...
- function/meaning of words, topics...
- minimize researcher bias
Why don’t you simply count them?

Top 10 lemmas:

the  the  the
be   and  and
of   to   be
a    of   of
don't you simply count them? have
and  to  a
the  we  that
in   have  in
in   in  in
it   our  they
Why don’t you simply count them?

Top 10 lemmas:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>the</td>
<td>the</td>
</tr>
<tr>
<td>be</td>
<td>and</td>
<td>and</td>
</tr>
<tr>
<td>of</td>
<td>to</td>
<td>be</td>
</tr>
<tr>
<td>a</td>
<td>be</td>
<td>of</td>
</tr>
<tr>
<td>and</td>
<td>of</td>
<td>to</td>
</tr>
<tr>
<td>to</td>
<td>we</td>
<td>a</td>
</tr>
<tr>
<td>he</td>
<td>that</td>
<td>he</td>
</tr>
<tr>
<td>it</td>
<td>a</td>
<td>have</td>
</tr>
<tr>
<td>have</td>
<td>our</td>
<td>in</td>
</tr>
<tr>
<td>in</td>
<td>in</td>
<td>they</td>
</tr>
</tbody>
</table>
Thematic concentration

- content words with “abnormal” frequency
Thematic concentration

- content words with “abnormal” frequency
- Zipf’s word-frequency distribution
Thematic concentration

- content words with “abnormal” frequency
- Zipf’s word-frequency distribution
  - $h$-point: rank = frequency

© Popescu & Altmann 2006
Thematic concentration

- content words with “abnormal” frequency
- Zipf’s word-frequency distribution
  - $h$-point: rank = frequency
  - $h$-point – approximately separates autosemantic and synsemantic branch

© Popescu & Altmann 2006
Thematic concentration

- content words with “abnormal” frequency
- Zipf’s word-frequency distribution
  - $h$-point: rank = frequency
  - $h$-point – approximately separates autosemantic and synsemantic branch
  - content words above the $h$-point are TC words
Thematic concentration

- content words with “abnormal” frequency
- Zipf’s word-frequency distribution
  - $h$-point: rank $=$ frequency
  - $h$-point – approximately separates autosemantic and synsemantic branch
  - content words above the $h$-point are TC words

© Popescu & Altmann 2006
Thematic concentration

h - point definition: $f(h) = h$

frequency $f(r)$

synsemantic branch

$(1/2)h^2$

(1/2)$h^2$

autosemantic branch

$h$ - point

rank $r$
TC words

1984:
Winston, say, know, Party, face, word, O’Brien, seem, look, never, think, moment, always, hand, year, way, long, now, eye, day, possible, war...

SOTU:
year, job, work, America, new, people, american, know, need, help, country, business, time, world, economy, family, right, tax, Congress, nation...

Hobbit:
come, go, Bilbo, see, dwarves, time, long, make, think, great, good, know, far, still, goblin, find, way, look, little, light...
<table>
<thead>
<tr>
<th>Pros and cons of TC words</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ objective – based on frequency distribution</td>
</tr>
</tbody>
</table>
Pros and cons of TC words

+ objective – based on frequency distribution
+ text analytical applications: comparing texts according to their thematic compactness
TC: discussion

Pros and cons of TC words

+ objective – based on frequency distribution
+ text analytical applications: comparing texts according to their thematic compactness
+ no reference corpus required
## Pros and cons of TC words

+ objective – based on frequency distribution
+ text analytical applications: comparing texts according to their thematic compactness
+ no reference corpus required
  - set of TC words is invariant
Pros and cons of TC words

+ objective – based on frequency distribution
+ text analytical applications: comparing texts according to their thematic compactness
+ no reference corpus required
  - set of TC words is invariant
  - “interpretation without interpreter” – interpretation always depends on the point of view
Keyword analysis
Keywords and KWA

Keywords

▶ homonymous term\(^1\)

\(^1\)For other meanings see e.g. Williams 1976 or Wierzbicka 1997.
Keywords and KWA

Keywords

- homonymous term\(^1\)
- words with higher relative frequency in a text

\(^1\)For other meanings see e.g. Williams 1976 or Wierzbicka 1997.
Keywords and KWA

### Keywords

- homonymous term\(^1\)
- words with higher relative frequency in a text
- based on comparison with reference corpus

---

\(^1\)For other meanings see e.g. Williams 1976 or Wierzbicka 1997.
Keywords and KWA

<table>
<thead>
<tr>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>➤ homonymous term(^1)</td>
</tr>
<tr>
<td>➤ words with higher relative frequency in a text</td>
</tr>
<tr>
<td>➤ based on comparison with reference corpus</td>
</tr>
<tr>
<td>➤ significance testing: (\chi^2) test, log-likelihood (G) test, Fisher test</td>
</tr>
</tbody>
</table>

\(^1\)For other meanings see e.g. Williams 1976 or Wierzbicka 1997.
Keywords and KWA

Keywords

- homonymous term\(^1\)
- words with higher relative frequency in a text
- based on comparison with reference corpus
- significance testing: \(\chi^2\) test, log-likelihood (G) test, Fisher test

A word-form which recurs within the text in question will be more likely to be key in it. (Scott–Tribble 2006)

\(^1\text{For other meanings see e.g. Williams 1976 or Wierzbicka 1997.}\)
Obtaining keywords – algorithm

Procedure

- count frequency of each word – most frequent words are *the*, *of*, *was*...
Obtaining keywords – algorithm

Procedure

▶ count frequency of each word – most frequent words are *the*, *of*, *was*…
▶ compare it with a frequency of the same word in a corpus

Keywords: Words which appear in a text or corpus that are statistically significantly more frequent than would be expected by chance when compared to a corpus which is larger or of equal size.
Obtaining keywords – algorithm

**Procedure**

- count frequency of each word – most frequent words are *the, of, was*...
- compare it with a frequency of the same word in a corpus
- use statistical tests: $\chi^2$, log-likelihood or Fisher to find out if the difference is significant
Obtaining keywords – algorithm

### Procedure

- count frequency of each word – most frequent words are *the, of, was*...
- compare it with a frequency of the same word in a corpus
- use statistical tests: $\chi^2$, log-likelihood or Fisher to find out if the difference is significant
- interpret *top* $X$ most prominent keywords

**Keywords**: Words which appear in a text or corpus that are statistically significantly more frequent than would be expected by chance when compared to a corpus which is larger or of equal size.
Obtaining keywords – algorithm

<table>
<thead>
<tr>
<th>Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>▶ count frequency of each word – most frequent words are <em>the</em>, <em>of</em>, <em>was</em>…</td>
</tr>
<tr>
<td>▶ compare it with a frequency of the same word in a corpus</td>
</tr>
<tr>
<td>▶ use statistical tests: $\chi^2$, log-likelihood or Fisher to find out if the difference is significant</td>
</tr>
<tr>
<td>▶ interpret <em>top X</em> most prominent keywords</td>
</tr>
</tbody>
</table>

Keywords: Words which appear in a text or corpus that are statistically significantly more frequent than would be expected by chance when compared to a corpus which is larger or of equal size.
Note on significance and effect size

Gabrielatos, C. & Marchi, A. (2012): there is a difference between (statistical) **significance** and (linguistic) **relevance** (effect size)
Note on significance and effect size

Gabrielatos, C. & Marchi, A. (2012): there is a difference between (statistical) **significance** and (linguistic) **relevance** (effect size)

 Metrics used to calculate keyness

  ▶ **significance** – level of certainty we have that the difference exists (N.B. $\chi^2$ test is “asymptotically true”)
Note on significance and effect size

Gabrielatos, C. & Marchi, A. (2012): there is a difference between (statistical) **significance** and (linguistic) **relevance** (effect size)

---

**Metrics used to calculate keyness**

- **significance** – level of certainty we have that the difference exists (N.B. $\chi^2$ test is “asymptotically true”)
- **relevance** – importance of the difference (for interpretation)
Note on significance and effect size

Gabrielatos, C. & Marchi, A. (2012): there is a difference between (statistical) **significance** and (linguistic) **relevance** (effect size)

### Metrics used to calculate keyness

- **significance** – level of certainty we have that the difference exists (N.B. $\chi^2$ test is “asymptotically true”)
- **relevance** – importance of the difference (for interpretation)
- crucial for the *top X* approach:
Note on significance and effect size

Gabrielatos, C. & Marchi, A. (2012): there is a difference between (statistical) **significance** and (linguistic) **relevance** (effect size)

<table>
<thead>
<tr>
<th>Metrics used to calculate keyness</th>
</tr>
</thead>
<tbody>
<tr>
<td>▶ <strong>significance</strong> – level of certainty we have that the difference exists (N.B. $\chi^2$ test is “asymptotically true”)</td>
</tr>
<tr>
<td>▶ <strong>relevance</strong> – importance of the difference (for interpretation)</td>
</tr>
<tr>
<td>▶ crucial for the top $X$ approach:</td>
</tr>
<tr>
<td>1. <strong>identification</strong> of KWs – statistical tests</td>
</tr>
</tbody>
</table>
Note on significance and effect size

Gabrielatos, C. & Marchi, A. (2012): there is a difference between (statistical) **significance** and (linguistic) **relevance** (effect size)

<table>
<thead>
<tr>
<th><strong>Metrics used to calculate keyness</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>significance</strong> – level of certainty we have that the difference exists (N.B. $\chi^2$ test is “asymptotically true”)</td>
</tr>
<tr>
<td><strong>relevance</strong> – importance of the difference (for interpretation)</td>
</tr>
<tr>
<td>crucial for the <em>top X</em> approach:</td>
</tr>
<tr>
<td>1. <strong>identification</strong> of KWs – statistical tests</td>
</tr>
<tr>
<td>2. <strong>ranking</strong> of KWs – task for a different metric</td>
</tr>
</tbody>
</table>
DIN coefficient

Variation on the Sørensen–Dice’s coefficient:\(^2\):

\[
DIN = 100 \times \frac{\text{RelFq}(\text{Target}) - \text{RelFq}(\text{Reference})}{\text{RelFq}(\text{Target}) + \text{RelFq}(\text{Reference})}
\]

- values of DIN

DIN coefficient

Variation on the Sørensen–Dice’s coefficient\(^2\):

\[
DIN = 100 \times \frac{RelFq(Target) - RelFq(Reference)}{RelFq(Target) + RelFq(Reference)}
\]

- values of DIN
  - -100 (\(=\) when a word is present only in the RefC)

DIN coefficient

Variation on the Sørensen–Dice’s coefficient\(^2\):

\[
DIN = 100 \times \frac{\text{RelFq}(\text{Target}) - \text{RelFq}(\text{Reference})}{\text{RelFq}(\text{Target}) + \text{RelFq}(\text{Reference})}
\]

- values of DIN
  - -100 (= when a word is present only in the RefC)
  - 0 (= when a word occurs equally in target and RefC)

DIN coefficient

Variation on the Sørensen–Dice’s coefficient\(^2\):

\[
DIN = 100 \times \frac{\text{RelFq}(\text{Target}) - \text{RelFq}(\text{Reference})}{\text{RelFq}(\text{Target}) + \text{RelFq}(\text{Reference})}
\]

- values of DIN
  - -100 (= when a word is present only in the RefC)
  - 0 (= when a word occurs equally in target and RefC)
  - 100 (= when a word is present only in the target corpus)

**DIN coefficient**

Variation on the Sørensen–Dice’s coefficient\(^2\):

\[
DIN = 100 \times \frac{\text{RelFq}(\text{Target}) - \text{RelFq}(\text{Reference})}{\text{RelFq}(\text{Target}) + \text{RelFq}(\text{Reference})}
\]

- values of DIN
  - -100 (= when a word is present only in the RefC)
  - 0 (= when a word occurs equally in target and RefC)
  - 100 (= when a word is present only in the target corpus)
- represents the proportion of the difference of relative frequencies to their mean (\(\times 50\))

\(^2\text{cf. Hofland–Johansson (1982).}\)
DIN coefficient

Variation on the Sørensen–Dice’s coefficient\(^2\):

\[
DIN = 100 \times \frac{\text{RelFq}(\text{Target}) - \text{RelFq}(\text{Reference})}{\text{RelFq}(\text{Target}) + \text{RelFq}(\text{Reference})}
\]

- values of DIN
  - -100 (= when a word is present only in the RefC)
  - 0 (= when a word occurs equally in target and RefC)
  - 100 (= when a word is present only in the target corpus)
- represents the proportion of the difference of relative frequencies to their mean (\(\times 50\))
- identical value of DIN for words appearing in a target text only (!)

DIN coefficient

Variation on the Sørensen–Dice’s coefficient\(^2\):

\[
DIN = 100 \times \frac{RelFq(Target) - RelFq(Reference)}{RelFq(Target) + RelFq(Reference)}
\]

- values of DIN
  - -100 (= when a word is present only in the RefC)
  - 0 (= when a word occurs equally in target and RefC)
  - 100 (= when a word is present only in the target corpus)

- represents the proportion of the difference of relative frequencies to their mean (× 50)
- identical value of DIN for words appearing in a target text only (!)
- useful for ranking of KWs (not for their identification!)

Example 1: Grammatical words

Keywords from all Husák’s New Year’s Addresses
Tools for KWA
Welcome to KWords!

How can we read a text from the viewpoint of the reader from the past? What did she expect in the text, and what did she find surprising? What was the author trying to do accentuate while revising his/her text multiple times? How do we account for different receptions of the same text by audiences? Can we detect subtle societal and cultural changes by following texts over time?

These are just samples questions that the principles used in this application can help answer.

KWords is an application to identify word forms that are often closely connected to the overarching themes and genre of a text. This application (currently implemented in Czech) is an important part of the on-going research entitled "A Needle in a Haystack", a project led by faculty from Charles University in Prague and Brown University (cf. our powerpoint presentations [here](http://www.korpus.cz) and [here](http://www.korpus.cz)).

Keywords, used in this sense, are not pre-select. On the contrary, keywords are outputs that this tool yields as possibly prominent word forms. Keywords thus serve as a starting point for its interpretation. In this regard, our approach is data and corpus-driven and minimizes the role played by intuition and subjectivity.

The application identifies keywords by comparing their frequency in analyzed text to frequencies of these words in reference corpus (cf. [SYN2010](http://syn2010.cz) - a balanced and reference corpus of contemporary written Czech). If the difference is statistically significant (according to the x2 or log-likelihood tests) the effect size of the difference is counted (see the list of keywords in the second tab).

Keyword analysis not only identifies keywords in a text, but also reveals relations between them. You can examine the dispersion plot (to see where the keywords appear in the text) or check the keyword links (i.e. frequent co-occurrences of keywords in a certain span). The application also comes with simple collocation analysis, which lets you look at the context in which KWs occur and the word forms (collocates) that tend to occur in nearby contexts.

This is a first version of this software. We are working to apply this approach to other languages very soon (our next goal is to apply it to English). We will also set up the best-practice guidelines for keyword analysis. Meanwhile, please try out the application and send your comments and suggestions to help improve this tool.

Masaiko Fider, Brown University, USA
masaiko.fider@brown.edu

Václav Čvsek, Charles University in Prague, Czech Republic
vaclav.cvosek@ff.cuni.cz

© 2012 Václav Čvsek, Pavel Vondrůka (ÚČNK)
Vážené soudružky, vážení soudruzi, dřež spoluobčané!

Vstoupili jsme do nového roku 1975. Doufáme, abych vás všechny na jeho prahu jménem ústředního výboru Komunistické strany Československa, ústředního výboru Národní fronty, jménem československé vlády i námocního prezidenta republiky a jménem svým srdečně pozdraví a popřál vám v novém roce hodně zdraví, štěstí, osobní spokojenosti i mnoho nových pracovních úspěchů.

V životě člověka i v životě společnosti a státu je to příležitost k zamýšlení, co se nám v minulém roce zdařilo a co ne, co nás očekává a co si silujeme od příštího roku.

Můžeme říci, že uplynulý rok 1974 byl ve všech oblastech pro Československo dobrým rokem. Důstojně se připadlo k několika předešlého rokům v rozvoji našeho hospodářství, životního úrovně a celé společnosti. Dobře pracovala naš průmysl, stavebnictví, zemědělství, přínosy jsou výsledky v ostatních oblastech. Úspěšně jsme zakončili čtvrtý rok pátého pětiletého plánu.

Období, kterým jsme prošli od XIV. sjezdu Komunistické strany Československa, je charakterizováno příznivým vnitropolitickým vývojem, upevňováním našeho socialistického státu, rozvojem sociálně demokratické politiky a angažování pracovní aktivity dělnické třídy, rolnictva, inteligence, žen, mládeže i prohlušováním morálně-politické jednoty pracujícího lidu i obou bratrských národů a všech národů.

Na krajine českého národa jsou přítomnosti
KWords: list of keywords

Two types of prominent units: **keywords** and **thematic concentration**

<table>
<thead>
<tr>
<th>Pořadí</th>
<th>Tvar</th>
<th>FQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>lidu</td>
<td>15</td>
</tr>
<tr>
<td>10</td>
<td>roce</td>
<td>11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tvar</th>
<th>LL</th>
<th>DIN</th>
<th>Fq(text)</th>
<th>Fq(ref)</th>
<th>TC</th>
</tr>
</thead>
<tbody>
<tr>
<td>draží</td>
<td>49.773</td>
<td>99.8553</td>
<td>4</td>
<td>205</td>
<td>-</td>
</tr>
<tr>
<td>pracujícího</td>
<td>36.426</td>
<td>99.8315</td>
<td>3</td>
<td>179</td>
<td>-</td>
</tr>
<tr>
<td>svazem</td>
<td>42.521</td>
<td>99.6397</td>
<td>4</td>
<td>511</td>
<td>-</td>
</tr>
<tr>
<td>socialistického</td>
<td>53.132</td>
<td>99.6390</td>
<td>5</td>
<td>640</td>
<td>-</td>
</tr>
<tr>
<td>sovětským</td>
<td>42.475</td>
<td>99.6375</td>
<td>4</td>
<td>514</td>
<td>-</td>
</tr>
<tr>
<td>národné</td>
<td>31.787</td>
<td>99.6333</td>
<td>3</td>
<td>390</td>
<td>-</td>
</tr>
<tr>
<td>lidu</td>
<td>149.003</td>
<td>99.4889</td>
<td>15</td>
<td>2720</td>
<td>+</td>
</tr>
<tr>
<td>pracujících</td>
<td>28.847</td>
<td>99.3803</td>
<td>3</td>
<td>660</td>
<td>-</td>
</tr>
<tr>
<td>československého</td>
<td>27.194</td>
<td>99.2100</td>
<td>3</td>
<td>842</td>
<td>-</td>
</tr>
<tr>
<td>hospodářství</td>
<td>87.706</td>
<td>99.0847</td>
<td>10</td>
<td>3254</td>
<td>-</td>
</tr>
<tr>
<td>socialistické</td>
<td>25.765</td>
<td>98.9972</td>
<td>3</td>
<td>1070</td>
<td>-</td>
</tr>
<tr>
<td>zeměmi</td>
<td>25.743</td>
<td>98.9935</td>
<td>3</td>
<td>1074</td>
<td>-</td>
</tr>
<tr>
<td>surovín</td>
<td>25.222</td>
<td>98.9021</td>
<td>3</td>
<td>1172</td>
<td>-</td>
</tr>
<tr>
<td>komunistické</td>
<td>41.057</td>
<td>98.7889</td>
<td>5</td>
<td>2156</td>
<td>-</td>
</tr>
<tr>
<td>československa</td>
<td>32.726</td>
<td>98.7706</td>
<td>4</td>
<td>1751</td>
<td>-</td>
</tr>
</tbody>
</table>
Dispersion plot

SOTU 2016: terrorism × economy
Keyword links

<table>
<thead>
<tr>
<th>Keyword links according to the size of the window</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Distant KW Links</strong> = KWs appearing in distant context (-15;-5) and (5;15); these KW links indicate that the themes represented by these KWs may form a discourse-semantic network</td>
</tr>
</tbody>
</table>

- **Custom** = arbitrary size of the span
Keyword links

Keyword links according to the size of the window

**Distant KW Links** = KWs appearing in distant context (-15; -5) and (5; 15); these KW links indicate that the themes represented by these KWs may form a discourse-semantic network.

**Immediate KW Links (multi-word KWs)** = co-occurrence of two or more KWs within immediate or near context (-2; 2); these adjacent KW may signal multi-word KW unit (e.g. *American people, better politics*).
Keyword links

Keyword links according to the size of the window

**Distant KW Links** = KWs appearing in distant context (-15;-5) and (5;15); these KW links indicate that the themes represented by these KWs may form a discourse-semantic network

**Immediate KW Links (multi-word KWs)** = co-occurrence of two or more KWs within immediate or near context (-2;2); these adjacent KW may signal multi-word KW unit (e.g. *American people, better politics*)

**custom** = arbitrary size of the span
Comparison

For comparing texts – time series (SOTU)

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Match</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>2015</th>
<th>2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>america</td>
<td>8</td>
<td>90.057</td>
<td>91.791</td>
<td>94.173</td>
<td>94.931</td>
<td>93.484</td>
<td>95.561</td>
<td>95.270</td>
<td>95.770</td>
</tr>
<tr>
<td>american</td>
<td>8</td>
<td>92.402</td>
<td>87.344</td>
<td>87.956</td>
<td>93.687</td>
<td>89.404</td>
<td>89.801</td>
<td>88.153</td>
<td>91.302</td>
</tr>
<tr>
<td>americans</td>
<td>8</td>
<td>96.949</td>
<td>98.504</td>
<td>96.352</td>
<td>97.922</td>
<td>96.051</td>
<td>98.544</td>
<td>97.992</td>
<td>97.852</td>
</tr>
<tr>
<td>businesses</td>
<td>8</td>
<td>94.285</td>
<td>96.588</td>
<td>93.184</td>
<td>95.898</td>
<td>94.784</td>
<td>96.803</td>
<td>95.261</td>
<td>89.733</td>
</tr>
<tr>
<td>congress</td>
<td>8</td>
<td>93.430</td>
<td>91.764</td>
<td>92.926</td>
<td>95.147</td>
<td>95.626</td>
<td>96.447</td>
<td>94.607</td>
<td>85.127</td>
</tr>
<tr>
<td>country</td>
<td>8</td>
<td>72.281</td>
<td>63.427</td>
<td>61.832</td>
<td>73.273</td>
<td>65.681</td>
<td>73.975</td>
<td>80.216</td>
<td>68.627</td>
</tr>
<tr>
<td>economy</td>
<td>8</td>
<td>94.282</td>
<td>82.977</td>
<td>81.678</td>
<td>86.599</td>
<td>87.110</td>
<td>86.977</td>
<td>91.689</td>
<td>91.535</td>
</tr>
<tr>
<td>jobs</td>
<td>8</td>
<td>91.377</td>
<td>93.390</td>
<td>94.009</td>
<td>94.858</td>
<td>94.862</td>
<td>93.136</td>
<td>91.850</td>
<td>86.734</td>
</tr>
<tr>
<td>nation</td>
<td>8</td>
<td>95.239</td>
<td>94.017</td>
<td>94.820</td>
<td>91.845</td>
<td>90.918</td>
<td>90.822</td>
<td>92.112</td>
<td>87.954</td>
</tr>
<tr>
<td>our</td>
<td>8</td>
<td>91.249</td>
<td>89.494</td>
<td>90.163</td>
<td>85.799</td>
<td>92.062</td>
<td>89.556</td>
<td>89.595</td>
<td>88.885</td>
</tr>
<tr>
<td>tonight</td>
<td>8</td>
<td>92.853</td>
<td>90.103</td>
<td>94.594</td>
<td>92.738</td>
<td>95.025</td>
<td>92.333</td>
<td>93.499</td>
<td>85.544</td>
</tr>
<tr>
<td>we</td>
<td>8</td>
<td>73.582</td>
<td>74.344</td>
<td>77.690</td>
<td>65.049</td>
<td>73.408</td>
<td>68.662</td>
<td>77.826</td>
<td>75.927</td>
</tr>
<tr>
<td>energy</td>
<td>7</td>
<td>88.320</td>
<td>86.587</td>
<td>81.443</td>
<td>90.664</td>
<td>90.494</td>
<td>79.321</td>
<td>-</td>
<td>81.414</td>
</tr>
<tr>
<td>every</td>
<td>7</td>
<td>80.140</td>
<td>-</td>
<td>63.884</td>
<td>79.055</td>
<td>70.643</td>
<td>80.389</td>
<td>80.460</td>
<td>73.231</td>
</tr>
<tr>
<td>let</td>
<td>7</td>
<td>-</td>
<td>84.176</td>
<td>76.733</td>
<td>82.360</td>
<td>86.461</td>
<td>86.322</td>
<td>86.372</td>
<td>71.427</td>
</tr>
<tr>
<td>more</td>
<td>7</td>
<td>40.651</td>
<td>46.588</td>
<td>38.618</td>
<td>47.972</td>
<td>48.623</td>
<td>64.598</td>
<td>48.356</td>
<td>-</td>
</tr>
<tr>
<td>new</td>
<td>7</td>
<td>48.135</td>
<td>-</td>
<td>62.168</td>
<td>51.924</td>
<td>49.090</td>
<td>55.555</td>
<td>65.794</td>
<td>44.152</td>
</tr>
<tr>
<td>people</td>
<td>7</td>
<td>44.789</td>
<td>46.885</td>
<td>54.560</td>
<td>42.980</td>
<td>-</td>
<td>44.223</td>
<td>42.404</td>
<td>48.855</td>
</tr>
<tr>
<td>us</td>
<td>7</td>
<td>61.388</td>
<td>53.550</td>
<td>65.755</td>
<td>-</td>
<td>52.091</td>
<td>57.705</td>
<td>67.335</td>
<td>73.777</td>
</tr>
<tr>
<td>democrats</td>
<td>6</td>
<td>92.852</td>
<td>95.421</td>
<td>96.721</td>
<td>-</td>
<td>91.685</td>
<td>93.631</td>
<td>96.325</td>
<td>-</td>
</tr>
<tr>
<td>families</td>
<td>6</td>
<td>83.069</td>
<td>91.392</td>
<td>-</td>
<td>-</td>
<td>93.035</td>
<td>87.196</td>
<td>92.988</td>
<td>82.279</td>
</tr>
</tbody>
</table>
State of the Union
Obama’s State of the Union Address

Eight addresses (2009–2016)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokens</td>
<td>6346</td>
<td>8024</td>
<td>7611</td>
<td>7743</td>
<td>7427</td>
<td>7506</td>
<td>7479</td>
<td>6671</td>
</tr>
<tr>
<td>Types</td>
<td>1347</td>
<td>1531</td>
<td>1505</td>
<td>1555</td>
<td>1561</td>
<td>1598</td>
<td>1521</td>
<td>1422</td>
</tr>
</tbody>
</table>

source: http://www.whitehouse.gov

Average length $N = 7351$
Average vocabulary $V(N) = 1505$

Reference corpus: British National Corpus (BNC)
Permanent KWs (Key KWs)

<table>
<thead>
<tr>
<th>KWs appearing in all eight addresses:</th>
</tr>
</thead>
<tbody>
<tr>
<td>america, american, americans, businesses, congress, country, economy, jobs, nation, tonight</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>KWs appearing in six or seven addresses:</th>
</tr>
</thead>
<tbody>
<tr>
<td>energy, every, let, more, new, people, democrats, families, make, millions, republicans, tax, why</td>
</tr>
</tbody>
</table>
SOTU: Topics – economy/politics
Reference corpus in KWA
## Reference corpus in KWA

<table>
<thead>
<tr>
<th>What does reference corpus affect?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>size:</strong> bigger reference corpus $\Rightarrow$ more KWs</td>
</tr>
</tbody>
</table>
Reference corpus in KWA

What does reference corpus affect?

size: bigger reference corpus $\Rightarrow$ more KWs

composition: different reference corpora represent different readers (conceptualized reader)
# Reference corpus in KWA

## What does reference corpus affect?

**size:** bigger reference corpus $\Rightarrow$ more KWs

**composition:** different reference corpora represent different readers  
(conceptualized reader)

- balanced corpus $\sim$ general reader
# Reference corpus in KWA

## What does reference corpus affect?

<table>
<thead>
<tr>
<th>Size</th>
<th>Bigger reference corpus $\Rightarrow$ more KWs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Composition</td>
<td>Different reference corpora represent different readers (conceptualized reader)</td>
</tr>
<tr>
<td></td>
<td>▶ Balanced corpus $\sim$ general reader</td>
</tr>
<tr>
<td></td>
<td>▶ Specialized corpus $\sim$ specific reader (e.g. from the past, with specific background...)</td>
</tr>
</tbody>
</table>
Different readers = different interpretations

Contrastive KWA analysis

▶ different RefCs: interferences – time, style, topic differences
Different readers = different interpretations

<table>
<thead>
<tr>
<th>Contrastive KWA analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>▶ different RefCs: interferences – time, style, topic differences</td>
</tr>
<tr>
<td>▶ New Year’s Addresses of the last communist president of the Czechoslovakia Gustáv Husák (1975–1989)</td>
</tr>
</tbody>
</table>
Different readers = different interpretations

<table>
<thead>
<tr>
<th>Contrastive KWA analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>▶ different RefCs: interferences – time, style, topic differences</td>
</tr>
<tr>
<td>▶ New Year’s Addresses of the last communist president of the Czechoslovakia Gustáv Husák (1975–1989)</td>
</tr>
<tr>
<td>▶ contemporary reader (SYN2010) × reader from the past (Totalita)</td>
</tr>
</tbody>
</table>
Different readers = different interpretations

### Contrastive KWA analysis

- different RefCs: interferences – time, style, topic differences
- New Year’s Addresses of the last communist president of the Czechoslovakia Gustáv Husák (1975–1989)
  - contemporary reader (SYN2010) × reader from the past (Totalita)
- State of the Union addresses of Barack Obama (2009–2016)
**Different readers = different interpretations**

<table>
<thead>
<tr>
<th>Contrastive KWA analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>▶ different RefCs: interferences – time, style, topic differences</td>
</tr>
<tr>
<td>▶ New Year’s Addresses of the last communist president of the Czechoslovakia Gustáv Husák (1975–1989)</td>
</tr>
<tr>
<td>▶ contemporary reader (SYN2010) × reader from the past (Totalita)</td>
</tr>
<tr>
<td>▶ State of the Union addresses of Barack Obama (2009–2016)</td>
</tr>
<tr>
<td>▶ general reader (BNC) × politician/expert (rest of Obama’s speeches)</td>
</tr>
</tbody>
</table>
Husák: Influence of the reference corpora

What happens if we compare texts to different RefCs?

- the inventory of KWs does not differ substantially

- ranking (prominence of KWs – DIN)

- genre differences

- Modal verbs: want, can

- Verbs: 1. sg./pl.

- Contemporary reader (SYN2010)

- connected with historical events

- ideology

- archaisms, historism
Husák: Influence of the reference corpora

What happens if we compare texts to different RefCs?

- the inventory of KWs does not differ substantially
- the difference is in ranking (prominence of KWs – DIN)
Husák: Influence of the reference corpora

What happens if we compare texts to different RefCs?

- the inventory of KWs does not differ substantially
- the difference is in ranking (prominence of KWs – DIN)

Historical reader (Totalita)

→ genre differences
  - Modal verbs: want, can
  - Verbs: 1. sg./pl.
Husák: Influence of the reference corpora

What happens if we compare texts to different RefCs?

- the inventory of KWs does not differ substantially
- the difference is in ranking (prominence of KWs – DIN)

Historical reader (Totalita)
- genre differences
  - Modal verbs: want, can
  - Verbs: 1. sg./pl.

Contemporary reader (SYN2010)
- connected with historical events
  - ideology
  - archaisms, historism
Detailed comparison – 3 thematic groups

**Cold war:** mír, míru, mírová, mírové, mírového, mírovému, mírovou, mírový, mírových, mírovými, mírumilovné, mírumilovných, mírumilovným; napětí; odzbrojení, výzbroje, zbrojení, zbrojením, ozbrojených

**Collective possession:** náš, naše, našeho, našem, našemu, naši, naší, našich, naším, naším, našimi

**Ideo markers:** socialismu, socialismus, socialistická, socialistické, socialistického, socialistickém, socialistickému, socialistickou, socialistický, socialistických, socialistickým, socialistickými; komunismu, komunisté, komunistů, ksč, komunistům, komunisty komunistická, komunistické, komunistickým
Cold war

Cold War KWs in SYN−KWA and TOT−KWA

Year


DIN

SYN−KWA

TOT−KWA

Fidler–Cvrček (2015)
Collective possession

KWs "our" in SYN−KWA and TOT−KWA

Fidler–Cvrček (2015)
Ideological markers

Ideological markers KWs in SYN−KWA and TOT−KWA

Fidler–Cvrček (2015)
**Reader from the past × contemporary reader**

<table>
<thead>
<tr>
<th>Totalita</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. style &amp; genre: <em>fellow citizens, friends</em></td>
</tr>
<tr>
<td>2. propaganda: <em>blossom, succeed</em></td>
</tr>
<tr>
<td>3. 1st pers. sg. (<em>greet, wish</em>)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SYN2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. ideology-related: <em>comrade(s), socialist, five-year plan</em></td>
</tr>
<tr>
<td>2. period-specific: <em>brotherly, liberating, feverish, dutiful, imperialistic</em></td>
</tr>
</tbody>
</table>
Difference in sensitivity

- lines for both readers have similar tendencies
Difference in sensitivity

- lines for both readers have similar tendencies
- contemporary reader has higher overall level of prominence (DIN)
Difference in sensitivity

- lines for both readers have similar tendencies
- contemporary reader has higher overall level of prominence (DIN)
- tendencies are more visible for reader from the past
Difference in sensitivity

- lines for both readers have similar tendencies
- contemporary reader has higher overall level of prominence (DIN)
- tendencies are more visible for reader from the past
- contemporary reader cannot distinguish subtle changes (overwhelmed by unusual lexicon)
Difference in sensitivity

- lines for both readers have similar tendencies
- contemporary reader has higher overall level of prominence (DIN)
- tendencies are more visible for reader from the past
- contemporary reader cannot distinguish subtle changes (overwhelmed by unusual lexicon)
- astute reader from the past might notice slight shifts in the discourse
### 2015 SOTU: General vs. specialist’s view

<table>
<thead>
<tr>
<th>Rank</th>
<th>Keyword</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rebekah</td>
</tr>
<tr>
<td>2</td>
<td>bipartisan</td>
</tr>
<tr>
<td>3</td>
<td>hardworking</td>
</tr>
<tr>
<td>4</td>
<td>loopholes</td>
</tr>
<tr>
<td>5</td>
<td>childcare</td>
</tr>
<tr>
<td>6</td>
<td>republicans</td>
</tr>
<tr>
<td>7</td>
<td>folks</td>
</tr>
<tr>
<td>8</td>
<td>veterans</td>
</tr>
<tr>
<td>9</td>
<td>diplomacy</td>
</tr>
<tr>
<td>10</td>
<td>striving</td>
</tr>
<tr>
<td>11</td>
<td>terrorists</td>
</tr>
<tr>
<td>12</td>
<td>americans</td>
</tr>
<tr>
<td>13</td>
<td>infrastructure</td>
</tr>
<tr>
<td>14</td>
<td>fastest</td>
</tr>
<tr>
<td>15</td>
<td>democrats</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Keyword</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ben</td>
</tr>
<tr>
<td>2</td>
<td>economics</td>
</tr>
<tr>
<td>3</td>
<td>childcare</td>
</tr>
<tr>
<td>4</td>
<td>rebekah</td>
</tr>
<tr>
<td>5</td>
<td>west</td>
</tr>
<tr>
<td>6</td>
<td>believed</td>
</tr>
<tr>
<td>7</td>
<td>striving</td>
</tr>
<tr>
<td>8</td>
<td>tools</td>
</tr>
<tr>
<td>9</td>
<td>sick</td>
</tr>
<tr>
<td>10</td>
<td>spread</td>
</tr>
<tr>
<td>11</td>
<td>write</td>
</tr>
<tr>
<td>12</td>
<td>paid</td>
</tr>
<tr>
<td>13</td>
<td>surely</td>
</tr>
<tr>
<td>14</td>
<td>issues</td>
</tr>
<tr>
<td>15</td>
<td>scientists</td>
</tr>
</tbody>
</table>
2015 SOTU: Differences

**Specialist’s view:**

Rhetorics/style: *believed, write, thing, tools*

Important topics: *economics, leave, paid, childcare*
# 2015 SOTU: Differences

## Specialist’s view:

| Rhetorics/style: believed, write, thing, tools |
| Important topics: economics, leave, paid, childcare |

## General view:

| Political speeches: bipartisan, republicans, folks, veterans, diplomacy, terrorists, americans, infrastructure, democrats, combat, sanctions |
| Topical words: hardworking, loopholes, childcare, striving, fastest |
References

Thank you for your attention!