CZECH NATIONAL CORPUS
Why do we read the same text differently?

Václav Cvrček
Sheridan Center, Brown University
March 16, 2015
Introduction
Introduction

Text interpretation

- interpreting a text is at the core of the humanities’ mission
# Introduction

## Text interpretation

- interpreting a text is at the core of the humanities’ mission
- our interpretation + other people’s interpretation
Introduction

Text interpretation

- interpreting a text is 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

- interpreting a text is 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...
## Introduction

### Text interpretation

- interpreting a text is 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?

A corpus is a sample of naturally occurring written texts or transcribed speeches stored electronically which can serve as a basis for linguistic analysis and description.
What is a corpus?

A corpus is a sample of naturally occurring written texts or transcribed speeches stored electronically which can serve as a basis for linguistic analysis and description.
## Brief history

### Brown corpus

- **Standard Corpus of Present-Day Edited American English**
- **Henry Kučera, W. Nelson Francis**
- 1 million words – 500 samples, each 2000 words
- different genres, works published in 1961

### LOB Corpus (Lancaster – Oslo/Bergen Corpus)

- British counterpart to the Brown Corpus
- 1970-1979, University of Lancaster, University of Oslo
- same size, also texts from 1961, same categories
- comparative studies (American and British English)
# Brief history

## Brown corpus

- Standard Corpus of Present-Day Edited American English
- Henry Kučera, W. Nelson Francis
- 1 million words – 500 samples, each 2000 words
- different genres, works published in 1961

## LOB Corpus (Lancaster – Oslo/Bergen Corpus)

- British counterpart to the Brown Corpus
- 1970-1979, University of Lancaster, University of Oslo
- same size, also texts from 1961, same categories
- comparative studies (American and British English)
Czech National Corpus
Czech National Corpus

SYN
2.2 bil.

ORAL
5 mil.

Diakorp
2.5 mil.

InterCorp
1.4 bil.
KORPUSOVÝ WORKSHOP
3. ÚNORA 2015
Ústav Českého národního korpusu srdce zve všechny zájemce na celodenní praktický workshop, který se bude konat v sobotu 25. 4. 2015 na Filozofické fakultě Univerzity Karlovy v Praze. 
Podrobnosti a registrace formulář.

INTERCORP VERZE 7
22. PROSINEC 2014
V prosinci 2014 byl zveřejněn korpus InterCorp verze 7, ve kterém oproti verzi 6 přibyl velký balík filmových titulků. Celkový rozsah cizojazyčných textů dosáhl 173 mil. slov v jadre a 1,2 mil. slov v kolekci, počet cizích jazyků se zvýšil na 38.

REFERENČNOST KORPUSŮ
22. PROSINEC 2014
Pocináme InterCorpem verze 7 budou všechny nově zveřejněné korpusy referenční, součástí jejich názvu bude také číslo verze. Starší verze korpusů zůstanou všem uživatelům přístupné.

Co je to korpus?
Jazykový korpus je elektronicky soubor autentických textů (psaných nebo mluvených), v němž je možné jednoduše vyhledávat jazykové jevy (zejména slova a slovní spojení) a zobrazovat je v jejich přirozeném kontextu.

Kdo jsme?
Český národní korpus je akademický projekt založený v roce 1994 při FF UK a spravován Ústavem Českého národního korpusu. Jeho cílem je systematicky mapovat češtinu a další jazyky ve snaze o přirozené jazyky k novému zájmu. Čerpe a zpracovává jazyky, k mova vědět nebo češtinou používá. 

Aplikace

Podpora a zdroje informací

víc...
...but what is it good for?

Use of language corpora

- compiling dictionaries
...but what is it good for?

<table>
<thead>
<tr>
<th>Use of language corpora</th>
</tr>
</thead>
<tbody>
<tr>
<td>▶ compiling dictionaries</td>
</tr>
<tr>
<td>▶ compiling grammars</td>
</tr>
</tbody>
</table>
...but what is it good for?

Use of language corpora

- compiling dictionaries
- compiling grammars
- source for NLP applications (machine translation, information retrieval...)

...but what is it good for?

<table>
<thead>
<tr>
<th>Use of language corpora</th>
</tr>
</thead>
<tbody>
<tr>
<td>▶ compiling dictionaries</td>
</tr>
<tr>
<td>▶ compiling grammars</td>
</tr>
<tr>
<td>▶ source for NLP applications (machine translation, information retrieval…)</td>
</tr>
<tr>
<td>▶ as a mirror of societal changes</td>
</tr>
</tbody>
</table>
...but what is it good for?

Use of language corpora

- compiling dictionaries
- compiling grammars
- source for NLP applications (machine translation, information retrieval...)
- as a mirror of societal changes
- discourse analysis
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?
- can we approach the interpretation of a historical document as though we were readers from that historical period?
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?
- can we approach the interpretation of a historical document as though we were readers from that historical period?
- 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?
- can we approach the interpretation of a historical document as though we were readers from that historical period?
- 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?

⇒ compare similar historical documents against different frames of reference and observe the change in time
The Needle-in-a-Haystack Method (NHM)

The Needle-in-a-Haystack Method (NHM) is being developed as a collaborative project between members of the Department of Slavic Languages at Brown University and the Institute of the Czech National Corpus at Charles University in Prague, Czech Republic. Its aim is two-fold: to develop a quantitative method (software) for text analysis (KWords) and a set of Best Practice Guidelines for the software for the humanities students and scholars. We explore what and to what extent the application is capable of facilitating text analysis. The project studies the latter with a variety of texts from the Czech National Corpus including political speeches and literary texts. KWords now also allows analysis of texts in English.

NHM emerged in the fall of 2012. It links several disciplines: corpus linguistics, discourse analysis, Slavic linguistics, social sciences, and digital humanities.
Keyword analysis
Keyword analysis (KWA)

Romeo and Juliet vs. all Shakespeare plays (Scott et al. 2006)

<table>
<thead>
<tr>
<th>AH</th>
<th>DEATHLY</th>
<th>MARRIED</th>
<th>SLAIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>ART</td>
<td>EARLY</td>
<td>MERCUTIO</td>
<td>THEE</td>
</tr>
<tr>
<td>BACK</td>
<td>FRIAR</td>
<td>MONTAGUE</td>
<td>THOU</td>
</tr>
<tr>
<td>BANISHED</td>
<td>JULIET</td>
<td>MONUMENT</td>
<td>THURSDAY</td>
</tr>
<tr>
<td>BENVOLIO</td>
<td>JULIET’S</td>
<td>NIGHT</td>
<td>THY</td>
</tr>
<tr>
<td>CAPULET</td>
<td>KINSMAN</td>
<td>NURSE</td>
<td>TORCH</td>
</tr>
<tr>
<td>CAPULETS</td>
<td>LADY</td>
<td>O</td>
<td>TYBALT</td>
</tr>
<tr>
<td>CAPULET’S</td>
<td>LAWRENCE</td>
<td>PARIS</td>
<td>TYBALT’S</td>
</tr>
<tr>
<td>CELL</td>
<td>LIGHT</td>
<td>POISON</td>
<td>VAULT</td>
</tr>
<tr>
<td>CHURCHYARD</td>
<td>LIPS</td>
<td>ROMEO</td>
<td>VERONA</td>
</tr>
<tr>
<td>COUNTY</td>
<td>LOVE</td>
<td>ROMEO’S</td>
<td>WATCH</td>
</tr>
<tr>
<td>DEAD</td>
<td>MANTUA</td>
<td>SHE</td>
<td>WILT</td>
</tr>
</tbody>
</table>
Keywords

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

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

<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> was...</td>
</tr>
</tbody>
</table>
Keywords

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

Procedure

- count frequency of each word – most frequent words are the, of was...
- compare it with a frequency of the same word in general corpus
Keywords

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

Procedure

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

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

Procedure

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

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.
Diachronic Keyword Analysis

Suitable data for MD-CADS

- Presidential New Year’s addresses (NYA) of Gustáv Husák (1975–1989)
Diachronic Keyword Analysis

Suitable data for MD-CADS

- Presidential New Year’s addresses (NYA) of Gustáv Husák (1975–1989)
- presumed to be flat, ritualistic and monotonous, full of cliches
  (www.youtube.com/watch?v=QiBm4YX9y24)

- Totalita – 15 mil. words (1952–1977) of written Czech; communist newspaper reader from the past
# Diachronic Keyword Analysis

## Suitable data for MD-CADS

- Presidential New Year’s addresses (NYA) of Gustáv Husák (1975–1989)
  - presumed to be flat, ritualistic and monotonous, full of cliches ([www.youtube.com/watch?v=QiBm4YX9y24](https://www.youtube.com/watch?v=QiBm4YX9y24))
  - same author, same genre/situation × time (and topic)

<table>
<thead>
<tr>
<th>Reference Corpora</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Czech National Corpus</strong> (<a href="http://www.korpus.cz">www.korpus.cz</a>):</td>
</tr>
<tr>
<td>Totalita – 15 mil. words (1952–1977) of written Czech; communist newspaper reader from the past</td>
</tr>
</tbody>
</table>
Diachronic Keyword Analysis

Suitable data for MD-CADS

- Presidential New Year’s addresses (NYA) of Gustáv Husák (1975–1989)
- presumed to be flat, ritualistic and monotonous, full of cliches ([www.youtube.com/watch?v=QiBm4YX9y24](http://www.youtube.com/watch?v=QiBm4YX9y24))
- same author, same genre/situation × time (and topic)
Diachronic Keyword Analysis

Suitable data for MD-CADS

- Presidential New Year’s addresses (NYA) of Gustáv Husák (1975–1989)
- presumed to be flat, ritualistic and monotonous, full of cliches (www.youtube.com/watch?v=QiBm4YX9y24)
- same author, same genre/situation × time (and topic)
## Diachronic Keyword Analysis

### Suitable data for MD-CADS

- Presidential New Year’s addresses (NYA) of Gustáv Husák (1975–1989)
- presumed to be flat, ritualistic and monotonous, full of cliches ([www.youtube.com/watch?v=QiBm4YX9y24](https://www.youtube.com/watch?v=QiBm4YX9y24))
- same author, same genre/situation × time (and topic)
  - Totalita – 15 mil. words (1952–1977) of written Czech; communist newspaper ~ reader from the past
Tools for KWA
KWords

http://kwords.korpus.cz

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 of 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 ongoing 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 and here).

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 roles played by intuition and subjectivity.

The application identifies keywords by comparing their frequency in analyzed text to frequencies of these words in reference corpus (e.g., SYN2010 - a balanced and reference corpus of contemporary written Czech). If the difference is statistically significant (according to the z2 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.

Masatko Fiddler, Brown University, USA
masatko_fiddler@brown.edu

Václav Cvrček, Charles University in Prague, Czech Republic
vacekr.cvrc@ff.cuni.cz

© 2012 Václav Cvrček, Pavel Vondráčka (ÚČNK)
KWords: analysed text

Vážené soudruži, vážení soudruzi, děkuji spoluobčané!

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

V životě člověka i v životě společnosti a státu je to příležitost k zamyšlení, 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 dobrym rokem. Důsledně se přiřadila ke několika předcházejícím rokům v rozvoji našeho hospodářství, životní úroveň a celé společnosti. Dobře pracovala naš právna, stavebnictví, zemědělství, přiznivě jsou výsledky v ostatních oblastech. Úspěšné jsme zakončili čtvrtý rok pátého pětiletého plánu.

Období, které se začalo v roce 1974, je charakterizováno příznivým vnitropolitickým vývojem, upevňováním našeho socialistického státu, rozmachem socialistické demokracie, politickou angažovaností i pracovní aktivitou občanského světa, inteligence, žen, mládeže i prohlubováním morálně-politického jednoty pracujícího lidu i obou bratských národů a všech národů.

Na konci roka 1974 očekáváme různé tlumočené.
KWords: list of keywords

Two types of prominent units: **keywords** and **thematic concentration**
## 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
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. *národní fronta* “national front”)
Keyword links

<table>
<thead>
<tr>
<th>Description</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Distant KW Links</strong></td>
<td>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>
<tr>
<td><strong>Immediate KW Links (multi-word KWs)</strong></td>
<td>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. <em>národní fronta</em> “national front”)</td>
</tr>
<tr>
<td><strong>custom</strong></td>
<td>arbitrary size of the span</td>
</tr>
</tbody>
</table>
KWords: multi-analysis – comparison

⇒ Are NYAs “flat”?
<table>
<thead>
<tr>
<th>Year</th>
<th>Word-form</th>
<th>Fq</th>
<th>Left context*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1975</td>
<td>komunistické</td>
<td>5</td>
<td>greeting, party’s organs, “correct” policy</td>
</tr>
<tr>
<td>1976</td>
<td>komunistické</td>
<td>6</td>
<td>greeting, party’s organs, successful policy</td>
</tr>
<tr>
<td>1977</td>
<td>komunistické</td>
<td>5</td>
<td>greeting, party’s organs, support of policy</td>
</tr>
<tr>
<td>1978</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>1979</td>
<td>komunistické</td>
<td>5</td>
<td>greeting, party’s organs, support of policy</td>
</tr>
<tr>
<td>1980</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>1981</td>
<td>komunistické</td>
<td>4</td>
<td>greeting, party’s organs, criticism</td>
</tr>
<tr>
<td>1982</td>
<td>komunistické</td>
<td>5</td>
<td>greeting, party’s organs, support of principles</td>
</tr>
<tr>
<td>1983</td>
<td>KSČ</td>
<td>5</td>
<td>greeting, party’s organs, “correct” policy</td>
</tr>
<tr>
<td>1984</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>1985</td>
<td>komunistické</td>
<td>4</td>
<td>greeting, party’s organs, trust in policy</td>
</tr>
<tr>
<td>1986</td>
<td>komunistické</td>
<td>3</td>
<td>greeting, party’s organs</td>
</tr>
<tr>
<td>1987</td>
<td>komunistické</td>
<td>3</td>
<td>greeting, party’s organs, KSSS</td>
</tr>
<tr>
<td>1988</td>
<td>KSČ</td>
<td>3</td>
<td>party’s organs</td>
</tr>
<tr>
<td>1989</td>
<td>KSČ</td>
<td>4</td>
<td>party’s organs</td>
</tr>
</tbody>
</table>

*Right context – strany (party) – remains the same.
State of the Union
Obama’s State of the Union Address

Seven addresses (2009–2015)

<table>
<thead>
<tr>
<th></th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>2015</th>
<th>Mean</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>7448</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>1517</td>
</tr>
</tbody>
</table>

source: http://www.whitehouse.gov

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

**KWs appearing in all seven addresses:**

america, american, americans, businesses, congress, country, economy, jobs, nation, tonight

**KWs appearing in five or six addresses:**

energy, every, let, new, people, tax, why, college, deficit, families, kids, make, millions, reform
SOTU: Topics – economy/politics
SOTU: Topics – health/education
Two readers
Different readers = different interpretations

Contrastive KWA analysis

- different readers are represented by different reference corpora
Different readers = different interpretations

Contrastive KWA analysis

- different readers are represented by different reference corpora
- interferences: time, style, topic differences
Different readers = different interpretations

Contrastive KWA analysis

- different readers are represented by different reference corpora
- interferences: time, style, topic differences
- Husák: contemporary reader (SYN2010) × reader from the past (Totalita)
Different readers = different interpretations

Contrastive KWA analysis

- different readers are represented by different reference corpora
- interferences: time, style, topic differences
- Husák: contemporary reader (SYN2010) × reader from the past (Totalita)
- Obama: general reader (BNC) × politician/expert (rest of Obama’s speeches)
What happens if we compare texts to different RefCs?

- the larger the reference corpus is, the more KWs we obtain
What happens if we compare texts to different RefCs?

- the larger the reference corpus is, the more KWs we obtain
- the inventory of KWs does not differ substantially
**Husák: Influence of the reference corpora**

What happens if we compare texts to different RefCs?

- the larger the reference corpus is, the more KWs we obtain
- 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 larger the reference corpus is, the more KWs we obtain
- 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
Husák: Influence of the reference corpora

What happens if we compare texts to different RefCs?

- the larger the reference corpus is, the more KWs we obtain
- the inventory of KWs does not differ substantially
- the difference is in **ranking** (prominence of KWs – DIN)

<table>
<thead>
<tr>
<th>Historical reader (Totalita)</th>
<th>Contemporary reader (SYN2010)</th>
</tr>
</thead>
<tbody>
<tr>
<td>→ genre differences</td>
<td>→ connected with historical events</td>
</tr>
<tr>
<td>▶ Modal verbs: want, can</td>
<td>▶ ideology</td>
</tr>
<tr>
<td>▶ Verbs: 1. sg./pl.</td>
<td>▶ archaisms, historism</td>
</tr>
</tbody>
</table>
Cold war topics

Cold War KWs in SYN−KWA and TOT−KWA

Fidler & Cvrček, forthcoming
Collective possession

KWs "our" in SYN–KWA and TOT–KWA

Fidler & Cvrček, forthcoming
Ideological markers

Ideological markers KWs in SYN−KWA and TOT−KWA

Year

DIN

SYN−KWATOT−KWA


Fidler & Cvrček, forthcoming
Difference is in the sensitivity

- lines for both readers have similar tendencies
Difference is in the sensitivity

- lines for both readers have similar tendencies
- contemporary reader has higher overall level of prominence (DIN)
Difference is in the 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 is in the 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 is in the 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>
<th>Rank</th>
<th>Keyword</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rebekah</td>
<td>1</td>
<td>ben</td>
</tr>
<tr>
<td>2</td>
<td>bipartisan</td>
<td>2</td>
<td>economics</td>
</tr>
<tr>
<td>3</td>
<td>hardworking</td>
<td>3</td>
<td>childcare</td>
</tr>
<tr>
<td>4</td>
<td>loopholes</td>
<td>4</td>
<td>rebekah</td>
</tr>
<tr>
<td>5</td>
<td>childcare</td>
<td>5</td>
<td>west</td>
</tr>
<tr>
<td>6</td>
<td>republicans</td>
<td>6</td>
<td>believed</td>
</tr>
<tr>
<td>7</td>
<td>folks</td>
<td>7</td>
<td>striving</td>
</tr>
<tr>
<td>8</td>
<td>veterans</td>
<td>8</td>
<td>tools</td>
</tr>
<tr>
<td>9</td>
<td>diplomacy</td>
<td>9</td>
<td>sick</td>
</tr>
<tr>
<td>10</td>
<td>striving</td>
<td>10</td>
<td>spread</td>
</tr>
<tr>
<td>11</td>
<td>terrorists</td>
<td>11</td>
<td>write</td>
</tr>
<tr>
<td>12</td>
<td>americans</td>
<td>12</td>
<td>paid</td>
</tr>
<tr>
<td>13</td>
<td>infrastructure</td>
<td>13</td>
<td>surely</td>
</tr>
<tr>
<td>14</td>
<td>fastest</td>
<td>14</td>
<td>issues</td>
</tr>
<tr>
<td>15</td>
<td>democrats</td>
<td>15</td>
<td>scientists</td>
</tr>
</tbody>
</table>
2015 SOTU: Differences

### Specialist’s view:

<table>
<thead>
<tr>
<th>Rhetorics/style:</th>
<th>believed, write, thing, tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Important topics:</td>
<td>economics, leave, paid, childcare</td>
</tr>
</tbody>
</table>
2015 SOTU: Differences

Specialist’s view:

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

General view:

Specifics of political speech: bipartisan, republicans, folks, veterans, diplomacy, terrorists, americans, infrastructure, democrats, combat, sanctions
Topical words: hardworking, loopholes, childcare, striving, fastest
CADS and keyword analysis

- data-driven method facilitating text interpretation
<table>
<thead>
<tr>
<th>Summary</th>
</tr>
</thead>
</table>

CADS and keyword analysis

- data-driven method facilitating text interpretation
- minimum amount of extra-textual information
## Summary

### CADS and keyword analysis

- data-driven method facilitating text interpretation
- minimum amount of extra-textual information
- keywords = starting point of analysis
Summary

CADS and keyword analysis

- data-driven method facilitating text interpretation
- minimum amount of extra-textual information
- keywords = starting point of analysis
- accounting for addressee
Thank you for your attention!