CzechParl: Corpus of Stenographic Protocols from Czech Parliament

Miloš Jakubíček, Vojtěch Kovář

NLP Centre, Faculty of Informatics, Masaryk University, Brno
Botanická 68a, 602 00 Brno, Czech Republic

{xjakub,xkovar3}@fi.muni.cz

RASLAN 2010, 3. 12. 2010

Outline

1 Motivation

2 JSCDPL

3 CzechParl

4 Related Work

5 Future Development

Motivation

- language became the main weapon of politicians in modern democratic countries
- therefore political speech should be subject to linguistic analysis and introspection
- → Velký slovník floskulí (Just 2009)
  - for one-word entries it provides frequencies from Frequency dictionary of Czech
  - interesting examples from daily newspaper, but no corpus-evidence
- newspapers not freely available for crawling

Motivation

- language became the main weapon of politicians in modern democratic countries
- therefore political speech should be subject to linguistic analysis and introspection
- → Velký slovník floskulí (Just 2009)
  - for one-word entries it provides frequencies from Frequency dictionary of Czech
  - interesting examples from daily newspaper, but no corpus-evidence
- newspapers not freely available for crawling

Motivation

- language became the main weapon of politicians in modern democratic countries
- therefore political speech should be subject to linguistic analysis and introspection
- $\longrightarrow$ Velký slovník floskulí (Just 2009)
  - for one-word entries it provides frequencies from Frequency dictionary of Czech
  - interesting examples from daily newspaper, but no corpus-evidence
- newspapers not freely available for crawling

Motivation

- language became the main weapon of politicians in modern democratic countries
- therefore political speech should be subject to linguistic analysis and introspection
- $\longrightarrow$ Velký slovník floskulí (Just 2009)
  - for one-word entries it provides frequencies from Frequency dictionary of Czech
  - interesting examples from daily newspaper, but no corpus-evidence
- newspapers not freely available for crawling

Motivation

- language became the main weapon of politicians in modern democratic countries
- therefore political speech should be subject to linguistic analysis and introspection
- $\longrightarrow$ Velký slovník floskulí (Just 2009)
  - for one-word entries it provides frequencies from Frequency dictionary of Czech
  - interesting examples from daily newspaper, but no corpus-evidence
- newspapers not freely available for crawling

Motivation

- language became the main weapon of politicians in modern democratic countries
- therefore political speech should be subject to linguistic analysis and introspection
- $\longrightarrow$ Velký slovník floskulí (Just 2009)
  - for one-word entries it provides frequencies from Frequency dictionary of Czech
  - interesting examples from daily newspaper, but no corpus-evidence
- newspapers not freely available for crawling

JSCDPL I

- = Joint Czech and Slovak Digital Parliamentary Library
- free resource of all documents originating in Czech legislative institutions
- semi-annotated documents
- following types of documents are part of the JSCDPL:
  - Invitations for sessions
  - Debates
  - Bills
  - Resolutions
  - Materials of committees

JSCDPL I

- = Joint Czech and Slovak Digital Parliamentary Library
- free resource of all documents originating in Czech legislative institutions
- semi-annotated documents
- following types of documents are part of the JSCDPL:
  - Invitations for sessions
  - Debates
  - Bills
  - Resolutions
  - Materials of committees

JSCDPL I

- = Joint Czech and Slovak Digital Parliamentary Library
- free resource of all documents originating in Czech legislative institutions
- semi-annotated documents
- following types of documents are part of the JSCDPL:
  - Invitations for sessions
  - Debates
  - Bills
  - Resolutions
  - Materials of committees

JSCDPL I

- = Joint Czech and Slovak Digital Parliamentary Library
- free resource of all documents originating in Czech legislative institutions
- semi-annotated documents
- following types of documents are part of the JSCDPL:
  - Invitations for sessions
  - Debates
  - Bills
  - Resolutions
  - Materials of committees

JSCDPL I

- = Joint Czech and Slovak Digital Parliamentary Library
- free resource of all documents originating in Czech legislative institutions
- semi-annotated documents
- following types of documents are part of the JSCDPL:
  - Invitations for sessions
  - Debates
  - Bills
  - Resolutions
  - Materials of committees

JSCDPL I

- = Joint Czech and Slovak Digital Parliamentary Library
- free resource of all documents originating in Czech legislative institutions
- semi-annotated documents
- following types of documents are part of the JSCDPL:
  - Invitations for sessions
  - Debates
  - Bills
  - Resolutions
  - Materials of committees

JSCDPL I

- = Joint Czech and Slovak Digital Parliamentary Library
- free resource of all documents originating in Czech legislative institutions
- semi-annotated documents
- following types of documents are part of the JSCDPL:
  - Invitations for sessions
  - Debates
  - Bills
  - Resolutions
  - Materials of committees

JSCDPL I

- = Joint Czech and Slovak Digital Parliamentary Library
- free resource of all documents originating in Czech legislative institutions
- semi-annotated documents
- following types of documents are part of the JSCDPL:
  - Invitations for sessions
  - Debates
  - Bills
  - Resolutions
  - Materials of committees

JSCDPL I

- = Joint Czech and Slovak Digital Parliamentary Library
- free resource of all documents originating in Czech legislative institutions
- semi-annotated documents
- following types of documents are part of the JSCDPL:
  - Invitations for sessions
  - Debates
  - Bills
  - Resolutions
  - Materials of committees

JSCDPL II

- Austrian Constituent Imperial Diet 1848–1849 (Vienna, Kromeriz)
- Diet of the Czech Kingdom 1861–1913
- National Assembly of the Czechoslovak Republic and the Czechoslovak Socialist Republic 1918–1968
- Diet of the Slovak Republic 1939–1945
- Slovak National Council 1944–1960
- Czech National Council 1969–1992
- Resolutions of the presidium of the Slovak National Council 1970–1987
- Federal Assembly of the Czechoslovak Socialist Republic and the Czechoslovak Federal Republic (Chamber of the People and Chamber of Nations) 1969–1992
- Parliament of the Czech Republic (Chamber of Deputies and Senate) since 1993
- National Council of the Slovak Republic since 1993

JSCDPL II

- Austrian Constituent Imperial Diet 1848–1849 (Vienna, Kromeriz)
- Diet of the Czech Kingdom 1861–1913
- National Assembly of the Czechoslovak Republic and the Czechoslovak Socialist Republic 1918–1968
- Diet of the Slovak Republic 1939–1945
- Slovak National Council 1944–1960
- Czech National Council 1969–1992
- Resolutions of the presidium of the Slovak National Council 1970–1987
- Federal Assembly of the Czechoslovak Socialist Republic and the Czechoslovak Federal Republic (Chamber of the People and Chamber of Nations) 1969–1992
- Parliament of the Czech Republic (Chamber of Deputies and Senate) since 1993
- National Council of the Slovak Republic since 1993

JSCDPL II

- Austrian Constituent Imperial Diet 1848–1849 (Vienna, Kromeriz)
- Diet of the Czech Kingdom 1861–1913
- National Assembly of the Czechoslovak Republic and the Czechoslovak Socialist Republic 1918–1968
- Diet of the Slovak Republic 1939–1945
- Slovak National Council 1944–1960
- Czech National Council 1969–1992
- Resolutions of the presidium of the Slovak National Council 1970–1987
- Federal Assembly of the Czechoslovak Socialist Republic and the Czechoslovak Federal Republic (Chamber of the People and Chamber of Nations) 1969–1992
- Parliament of the Czech Republic (Chamber of Deputies and Senate) since 1993
- National Council of the Slovak Republic since 1993

JSCDPL II

- Austrian Constituent Imperial Diet 1848–1849 (Vienna, Kromeriz)
- Diet of the Czech Kingdom 1861–1913
- National Assembly of the Czechoslovak Republic and the Czechoslovak Socialist Republic 1918–1968
- Diet of the Slovak Republic 1939–1945
- Slovak National Council 1944–1960
- Czech National Council 1969–1992
- Resolutions of the presidium of the Slovak National Council 1970–1987
- Federal Assembly of the Czechoslovak Socialist Republic and the Czechoslovak Federal Republic (Chamber of the People and Chamber of Nations) 1969–1992
- Parliament of the Czech Republic (Chamber of Deputies and Senate) since 1993
- National Council of the Slovak Republic since 1993

JSCDPL II

- Austrian Constituent Imperial Diet 1848–1849 (Vienna, Kromeriz)
- Diet of the Czech Kingdom 1861–1913
- National Assembly of the Czechoslovak Republic and the Czechoslovak Socialist Republic 1918–1968
- Diet of the Slovak Republic 1939–1945
- Slovak National Council 1944–1960
- Czech National Council 1969–1992
- Resolutions of the presidium of the Slovak National Council 1970–1987
- Federal Assembly of the Czechoslovak Socialist Republic and the Czechoslovak Federal Republic (Chamber of the People and Chamber of Nations) 1969–1992
- Parliament of the Czech Republic (Chamber of Deputies and Senate) since 1993
- National Council of the Slovak Republic since 1993

JSCDPL II

- Austrian Constituent Imperial Diet 1848–1849 (Vienna, Kromeriz)
- Diet of the Czech Kingdom 1861–1913
- National Assembly of the Czechoslovak Republic and the Czechoslovak Socialist Republic 1918–1968
- Diet of the Slovak Republic 1939–1945
- Slovak National Council 1944–1960
- Czech National Council 1969–1992
- Resolutions of the presidium of the Slovak National Council 1970–1987
- Federal Assembly of the Czechoslovak Socialist Republic and the Czechoslovak Federal Republic (Chamber of the People and Chamber of Nations) 1969–1992
- Parliament of the Czech Republic (Chamber of Deputies and Senate) since 1993
- National Council of the Slovak Republic since 1993

JSCDPL II

- Austrian Constituent Imperial Diet 1848–1849 (Vienna, Kromeriz)
- Diet of the Czech Kingdom 1861–1913
- National Assembly of the Czechoslovak Republic and the Czechoslovak Socialist Republic 1918–1968
- Diet of the Slovak Republic 1939–1945
- Slovak National Council 1944–1960
- Czech National Council 1969–1992
- Resolutions of the presidium of the Slovak National Council 1970–1987
- Federal Assembly of the Czechoslovak Socialist Republic and the Czechoslovak Federal Republic (Chamber of the People and Chamber of Nations) 1969–1992
- Parliament of the Czech Republic (Chamber of Deputies and Senate) since 1993
- National Council of the Slovak Republic since 1993

JSCDPL II

- Austrian Constituent Imperial Diet 1848–1849 (Vienna, Kromeriz)
- Diet of the Czech Kingdom 1861–1913
- National Assembly of the Czechoslovak Republic and the Czechoslovak Socialist Republic 1918–1968
- Diet of the Slovak Republic 1939–1945
- Slovak National Council 1944–1960
- Czech National Council 1969–1992
- Resolutions of the presidium of the Slovak National Council 1970–1987
- Federal Assembly of the Czechoslovak Socialist Republic and the Czechoslovak Federal Republic (Chamber of the People and Chamber of Nations) 1969–1992
- Parliament of the Czech Republic (Chamber of Deputies and Senate) since 1993
- National Council of the Slovak Republic since 1993

JSCDPL II

- Austrian Constituent Imperial Diet 1848–1849 (Vienna, Kromeriz)
- Diet of the Czech Kingdom 1861–1913
- National Assembly of the Czechoslovak Republic and the Czechoslovak Socialist Republic 1918–1968
- Diet of the Slovak Republic 1939–1945
- Slovak National Council 1944–1960
- Czech National Council 1969–1992
- Resolutions of the presidium of the Slovak National Council 1970–1987
- Federal Assembly of the Czechoslovak Socialist Republic and the Czechoslovak Federal Republic (Chamber of the People and Chamber of Nations) 1969–1992
- Parliament of the Czech Republic (Chamber of Deputies and Senate) since 1993
- National Council of the Slovak Republic since 1993

JSCDPL II

- Austrian Constituent Imperial Diet 1848–1849 (Vienna, Kromeriz)
- Diet of the Czech Kingdom 1861–1913
- National Assembly of the Czechoslovak Republic and the Czechoslovak Socialist Republic 1918–1968
- Diet of the Slovak Republic 1939–1945
- Slovak National Council 1944–1960
- Czech National Council 1969–1992
- Resolutions of the presidium of the Slovak National Council 1970–1987
- Federal Assembly of the Czechoslovak Socialist Republic and the Czechoslovak Federal Republic (Chamber of the People and Chamber of Nations) 1969–1992
- Parliament of the Czech Republic (Chamber of Deputies and Senate) since 1993
- National Council of the Slovak Republic since 1993

CzechParl

- built from stenographic protocols available in JSCDPL
- from modern era of Czech Parliament, both Chamber of Deputies (since 1993) and Senate (since 1996)
- only the debates included
- annotation:
  - **sentences** (<s>)
  - **paragraphs** (<p>), as given in the stenographic protocols
  - **discourses** (<speech>), extracted from the stenographic protocols and containing the speaker name and role
  - **meeting days** (<day>), containing the date of the meeting
  - **documents** (<doc>), where each document represents an electoral term of either the Chamber of Deputies or Senate
- (will be:) available on http://corpora.fi.muni.cz

CzechParl

- built from stenographic protocols available in JSCDPL
- from modern era of Czech Parliament, both Chamber of Deputies (since 1993) and Senate (since 1996)
- only the debates included
- annotation:
  - **sentences** (<s>)
  - **paragraphs** (<p>), as given in the stenographic protocols
  - **discourses** (<speech>), extracted from the stenographic protocols and containing the speaker name and role
  - **meeting days** (<day>), containing the date of the meeting
  - **documents** (<doc>), where each document represents an electoral term of either the Chamber of Deputies or Senate
- (will be:) available on http://corpora.fi.muni.cz

CzechParl

- built from stenographic protocols available in JSCDPL
- from modern era of Czech Parliament, both Chamber of Deputies (since 1993) and Senate (since 1996)
- only the debates included
- annotation:
  - **sentences** (<s>)
  - **paragraphs** (<p>), as given in the stenographic protocols
  - **discourses** (<speech>), extracted from the stenographic protocols and containing the speaker name and role
  - **meeting days** (<day>), containing the date of the meeting
  - **documents** (<doc>), where each document represents an electoral term of either the Chamber of Deputies or Senate
- (will be:) available on http://corpora.fi.muni.cz

CzechParl

- built from stenographic protocols available in JSCDPL
- from modern era of Czech Parliament, both Chamber of Deputies (since 1993) and Senate (since 1996)
- only the debates included
- annotation:
  - **sentences** (<s>)
  - **paragraphs** (<p>), as given in the stenographic protocols
  - **discourses** (<speech>), extracted from the stenographic protocols and containing the speaker name and role
  - **meeting days** (<day>), containing the date of the meeting
  - **documents** (<doc>), where each document represents an electoral term of either the Chamber of Deputies or Senate
- (will be:) available on http://corpora.fi.muni.cz

CzechParl

- built from stenographic protocols available in JSCDPL
- from modern era of Czech Parliament, both Chamber of Deputies (since 1993) and Senate (since 1996)
- only the debates included
- annotation:
  - **sentences** (<s>)
  - **paragraphs** (<p>), as given in the stenographic protocols
  - **discourses** (<speech>), extracted from the stenographic protocols and containing the speaker name and role
  - **meeting days** (<day>), containing the date of the meeting
  - **documents** (<doc>), where each document represents an electoral term of either the Chamber of Deputies or Senate
- (will be:) available on http://corpora.fi.muni.cz

CzechParl

- built from stenographic protocols available in JSCDPL
- from modern era of Czech Parliament, both Chamber of Deputies (since 1993) and Senate (since 1996)
- only the debates included
- annotation:
  - **sentences** (<s>)
  - **paragraphs** (<p>), as given in the stenographic protocols
  - **discourses** (<speech>), extracted from the stenographic protocols and containing the speaker name and role
  - **meeting days** (<day>), containing the date of the meeting
  - **documents** (<doc>), where each document represents an electoral term of either the Chamber of Deputies or Senate
- (will be:) available on http://corpora.fi.muni.cz

CzechParl

- built from stenographic protocols available in JSCDPL
- from modern era of Czech Parliament, both Chamber of Deputies (since 1993) and Senate (since 1996)
- only the debates included
- annotation:
  - **sentences** (<s>)
  - **paragraphs** (<p>), as given in the stenographic protocols
  - **discourses** (<speech>), extracted from the stenographic protocols and containing the speaker name and role
  - **meeting days** (<day>), containing the date of the meeting
  - **documents** (<doc>), where each document represents an electoral term of either the Chamber of Deputies or Senate
- (will be:) available on http://corpora.fi.muni.cz

CzechParl

- built from stenographic protocols available in JSCDPL
- from modern era of Czech Parliament, both Chamber of Deputies (since 1993) and Senate (since 1996)
- only the debates included
- annotation:
  - **sentences** (<s>)
  - **paragraphs** (<p>), as given in the stenographic protocols
  - **discourses** (<speech>), extracted from the stenographic protocols and containing the speaker name and role
  - **meeting days** (<day>), containing the date of the meeting
  - **documents** (<doc>), where each document represents an electoral term of either the Chamber of Deputies or Senate
- (will be:) available on http://corpora.fi.muni.cz

CzechParl

- built from stenographic protocols available in JSCDPL
- from modern era of Czech Parliament, both Chamber of Deputies (since 1993) and Senate (since 1996)
- only the debates included
- annotation:
  - **sentences** (<s>)
  - **paragraphs** (<p>), as given in the stenographic protocols
  - **discourses** (<speech>), extracted from the stenographic protocols and containing the speaker name and role
  - **meeting days** (<day>), containing the date of the meeting
  - **documents** (<doc>), where each document represents an electoral term of either the Chamber of Deputies or Senate
- (will be:) available on http://corpora.fi.muni.cz

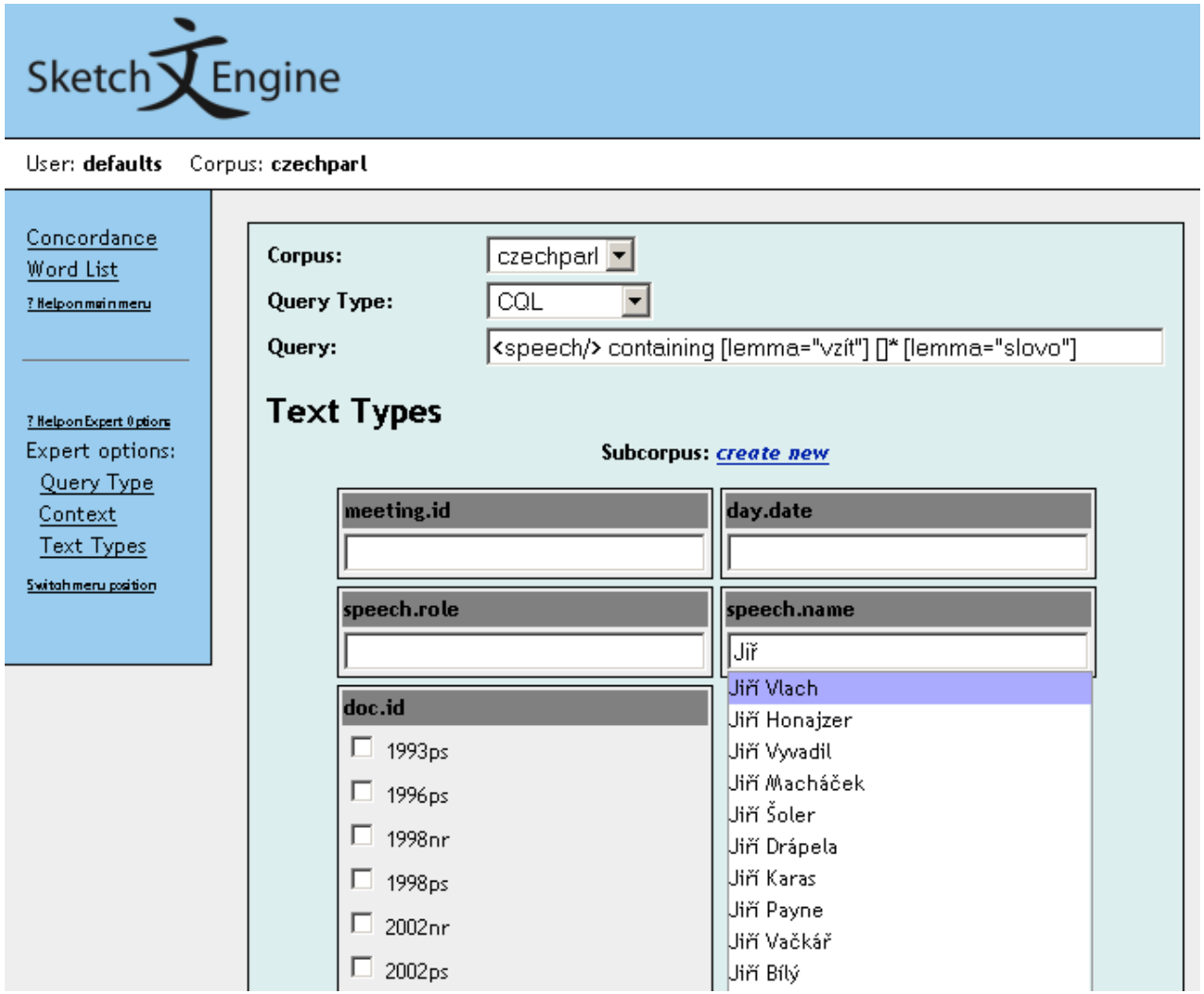
CzechParl

- built from stenographic protocols available in JSCDPL
- from modern era of Czech Parliament, both Chamber of Deputies (since 1993) and Senate (since 1996)
- only the debates included
- annotation:
  - **sentences** (<s>)
  - **paragraphs** (<p>), as given in the stenographic protocols
  - **discourses** (<speech>), extracted from the stenographic protocols and containing the speaker name and role
  - **meeting days** (<day>), containing the date of the meeting
  - **documents** (<doc>), where each document represents an electoral term of either the Chamber of Deputies or Senate
- (will be:) available on http://corpora.fi.muni.cz

CzechParl III

```
<p><s><speech name="Miroslav Kalousek" role="Poslanec"> :
Nebojte se , já nechci reagovat na pana poslance Ratha .
</s><s> Jsou příspěvky , na které se dá reagovat pouze
nonverbálně a to mi Schwarzenberg zakázal .  </s><s> (
Ohlas .  )  </s></p><p><s> Rád bych ale zareagoval na pana
poslance Sobotku a odmítl jeho tvrzení , že z větší části
mé vystoupení nesouviselo s projednávaným návrhem .
</s><s> Pokud jste nepochopil přímou souvislost mnou
prezentovaných indikátorů s vaším návrhem , pak se nedivím
, že ten návrh předkládáte , pane poslanče .  </s><s>
Příště prosím nechte své svědomí plout po vlnách mých vět
a otevřete srdce mým slovům a poznáte pravdu .  </s><s> (
Pobavení v pravé části sálu .  )  </s></speech></p>
```

CzechParl III



CzechParl IV

Parliament chamber	Chamber of Deputies	Senate	Total
Tokens	75,050,917	6,823,205	**81,874,122**
Sentences	3,987,910	198,816	**4,186,726**
Paragraphs	1,549,717	70,655	**1,620,372**
Documents	9	7	**16**
Days	1,985	140	**2,125**
Discourses	85,983	5,964	**91,947**

Related Work

- DutchParl (Marx and Schuth, 2010)
- SpanishParl (Marx and Martin, 2010)
- Samples of German Parliament protocols part of Corpus Workbench (CWB open-source community, 2010)
- EuroParl (Koehn, 2005)

Related Work

- DutchParl (Marx and Schuth, 2010)
- SpanishParl (Marx and Martin, 2010)
- Samples of German Parliament protocols part of Corpus Workbench (CWB open-source community, 2010)
- EuroParl (Koehn, 2005)

Related Work

- DutchParl (Marx and Schuth, 2010)
- SpanishParl (Marx and Martin, 2010)
- Samples of German Parliament protocols part of Corpus Workbench (CWB open-source community, 2010)
- EuroParl (Koehn, 2005)

Related Work

- DutchParl (Marx and Schuth, 2010)
- SpanishParl (Marx and Martin, 2010)
- Samples of German Parliament protocols part of Corpus Workbench (CWB open-source community, 2010)
- EuroParl (Koehn, 2005)

- provide frequencies and corpus-based examples for all entries in Velký slovník floskulí
- start cooperation with *Parl developers and put them together?
- CzechParl as basis for linguistic analysis similar to (Ilie, 2010)

- provide frequencies and corpus-based examples for all entries in Velký slovník floskulí
- start cooperation with *Parl developers and put them together?
- CzechParl as basis for linguistic analysis similar to (Ilie, 2010)

- provide frequencies and corpus-based examples for all entries in Velký slovník floskulí
- start cooperation with *Parl developers and put them together?
- CzechParl as basis for linguistic analysis similar to (Ilie, 2010)