CzechParl: Corpus of Stenographic Protocols from Czech Parliament

Miloš Jakubíček and Vojtěch Kovář

Natural Language Processing Centre, Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
xjakub@fi.muni.cz, xkovar3@fi.muni.cz

Abstract. Within a single language, there is a large variety of styles used for written text and speech, which differ significantly and have their subtle specifics. Among all of those, the language of politicians represents an integral class that deserves detailed analysis. In this paper we present CzechParl, a corpus we built from stenographic protocols recorded during plenary meetings of the Czech parliament in its modern era from 1993 to 2010. We provide brief statistics of the corpus and discuss its intended future usage and further development.

1 Introduction

During the last century language became the main weapon of politicians in modern democratic countries, directed to the citizens (voters) which are exposed to it in everyday rush. Often it is political speech that changes their opinions and heavily influents elections results. Obviously this leads to situations when not only pure informative and communicative, but also demonstrative, manipulative and psychological language functions are exhibited, sometimes in a latent form that comes mostly unnoticed. It is therefore of great importance that also this kind of language becomes subject to linguistic analysis and introspection.

For this purpose one needs data that will be large enough to be representative. One straightforward source of political texts represent various news and newspapers that focus on public life. Unfortunately most of those are not available in electronic archives with free access, moreover all of them do not contain texts crafted by politicians directly, but rather comments and glosses written by journalists and also lots of non-political texts – and as such they are not suitable for the projected needs.

Therefore we turned our attention to the place where everyday politics is being performed, namely the Czech parliament, and where we can benefit from the fact that for legal reasons stenographic protocols of politicans' speeches are made during all plenary meetings and are freely available in the electronic form.

In further text we describe CzechParl, a corpus built from such stenographic protocols (known as Hansards in United Kingdom and other Commonwealth countries) recorded in both chambers of Czech parliament – the Chamber of Deputies (1993–2010) and Senate (1996–2010). We briefly refer on how the corpus

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2010, pp. 41–46, 2010. © Tribun EU 2010

was prepared, report on its structure and basic statistical characteristics and discuss further analysis that is going to be performed in the future.

2 Corpus Building

2.1 Joint Czech and Slovak Digital Parliamentary Library

In 1993 the Joint Czech and Slovak Digital Parliamentary Library [1] (further referred to as JCSDPL) has been announced, a shared initiative of Czech and Slovak parliaments which aimed at providing free access to both modern and historical parliamentary documents in electronic form. It contains documents since 1848 that were produced in several legislative institutions [2], besides these, the Czech (Bohemian) Assemblies Digital Library [3] was built on top of the JCSDPL, providing historical documents dated back to the 11th century. The institutions covered by JCSDPL are as follows:

- 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

Following types of documents are part of the JSCDPL:

- Invitations for sessions
- Debates
- Bills
- Resolutions
- Materials of committees

In the current work we have processed the protocols from the modern era of the Czech parliament (since 1993) that contain documents in Czech only and are of most interest with regard to the intended analyses: the debates that are stenographically recorded and as such represent a unique source of truly captured discourses.

While the JSCDPL definitely represents an invaluable language resource, it was not intended for any automatic processing or annotation. Historical documents are available in the form of scanned images, modern ones (including those very recent that have been processed) as HTML pages. Therefore extensive cleaning and post-processing was needed to obtain plain text accompanied with the desired annotation (as described in further text).

```
 <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>
```

Fig. 1. Random sample of corpus annotation.

2.2 Annotation Scheme

The source data have been converted from HTML into plain text, relieved of any boilerplate (e. g. HTML-related metadata like headers and footers) and afterwards tokenized, segmented into sentences and lemmatized as well as tagged using the `desamb` tagger [4] which works on top of the morphological analyser of Czech called `majka` [5,6].

Furthermore, following structures have been annotated using an XML-like markup (as given in parentheses):

- **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

A random sample of corpus source text is provided in Figure 1. Next, the corpus data in such form have been encoded using the Manatee/Bonito corpus management system [7,8], components empowering the Sketch Engine [9], enabling fast and effective search and analysis including lookup of individual speeches by speaker name or advanced querying using the Corpus Query Language [10,11].

3 CzechParl Statistics

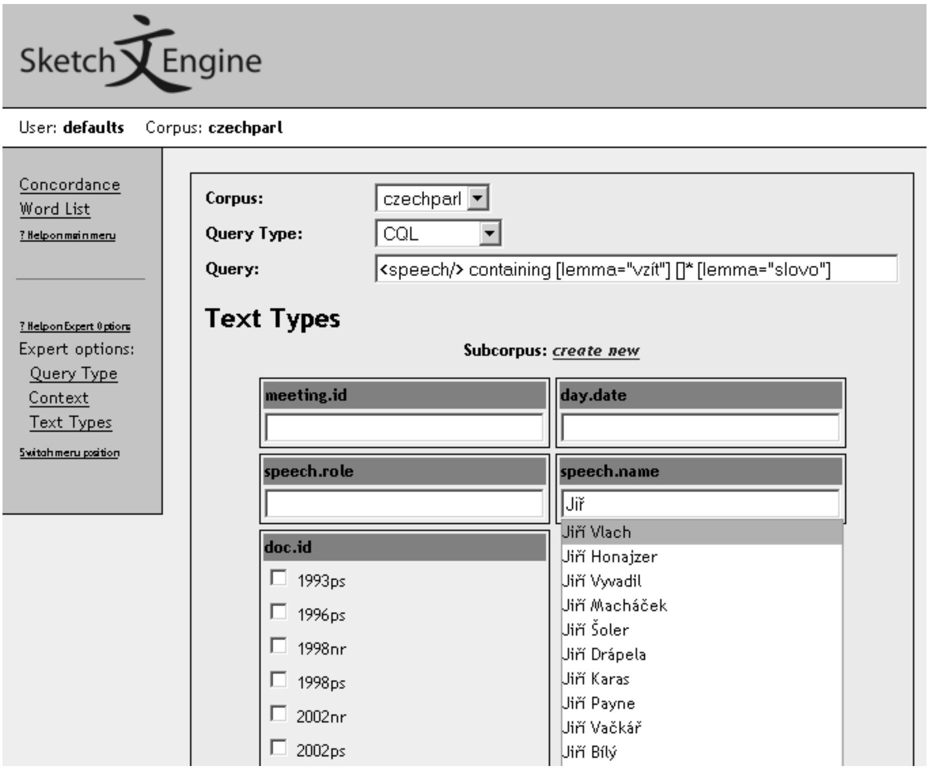
The corpus is available for view and search upon registration on `http://corpora.fi.muni.cz`. Bonito, the web interface built on top of Manatee, enables submitting of powerful queries and creating sophisticated statistical reports. A screenshot of the web interface is provided in Figure 2.

In Table 1, a summary of basic statistical properties of CzechParl is provided. Notable is the significant difference in the size between the part originating in

Table 1. Statistical summary of attributes and structures present in CzechParl.

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**

the Chamber of Deputies and the part recorded in Senate: over 90% of the corpus comes from the Chamber of Deputies. This corresponds to the hypothesis that most political debates occur in Chamber of Deputies, which also convenes more often than the Senate.



User: **defaults** Corpus: **czechparl**

Corpus: czechparl

Query Type: CQL

Query: <speech/> containing [lemma="vzít"] []* [lemma="slovo"]

Text Types

Subcorpus: *create new*

meeting.id

day.date

speech.role

speech.name Jiř

Jiří Vlach
Jiří Honajzer
Jiří Vyvadil
Jiří Macháček
Jiří Šoler
Jiří Drápela
Jiří Karas
Jiří Payne
Jiří Vačkář
Jiří Bílý

doc.id

- □ 1993ps
- □ 1996ps
- □ 1998nr
- □ 1998ps
- □ 2002nr
- □ 2002ps

Fig. 2. Screenshot of the Bonito query interface.

4 Related Work

Similar attempts to build corpus of parliamentary documents have been performed for Dutch [12] and Spanish [13]. An online demo for searching small part (1994–1997) of German parliamentary documents is available as well as part of the Corpus Workbench project [14]. An important resource in this domain is also the EuroParl [15], a parallel corpus of documents originating in the European parliament, which however focuses on a different goal, namely statistical machine translation.

Even though parliamentary documents in most countries are publicly available (including stenographic protocols), the prevailing majority still waits for being processed into the form of an annotated and searchable text corpus that will encourage researchers to provide corpus-based evidence for their theories on political discourse, following the notable example of [16] where corpus-motivated studies in this domain are presented for 11 European parliaments.

5 Conclusions and Future Development

In this paper we presented CzechParl, a corpus of parliamentary documents from both chambers of the modern Czech parliament – the Chamber of Deputies and Senate. Source texts have been obtained from the Joint Czech and Slovak Digital Parliamentary Library. The corpus contains annotation of individuals speeches and as such it is suitable for further linguistic analysis and introspection focused on political discourse.

In particular we plan in the future the corpus to be subject to analysis with regard to what Just calls "floscula", a sort of thought-terminating clichés that lost their original meaning and suite to fool their readers/hearers: "Floscula is – be it deliberate, intentional and malae fidei (propaganda, ideology, advertisement, kitch) or subconscious, automatic and mechanic (style, trend, snobbish slang) – hiding and decorating of emptiness by words". [1]A dictionary of flosculae compiled by Just [17] represents an excellent basis for such analyses.

Since the collected corpus data contains over 100 millions of tokens, the content of CzechParl is also going to become a part of czes [18], a big Czech web corpus that is currently under development.

Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536 and in the National Research Programme II project 2C06009 and by the Czech Science Foundation under the project P401/10/0792.

[1] In Czech original: „Floskule je – ať už záměrné, účelové a obmyslné (propaganda, ideologie, reklama, kýč), nebo podvědomé, automatické a mechanické (móda, trendovost, snobský slang) – zakrývání, zdobení prázdna slovy." [17]

References

1. Parliamentary Library, Information Technology Department of the Office of the Chamber of Deputies, Information Technology Department of the Senate Chancellery of the Czech Parliament, Parliamentary Library, parliamentary archive and Information Technology Department of the National Council of the Slovak Republic: Joint Czech and Slovak Digital Parliamentary Library. [online] `http://www.psp.cz/cgi-bin/eng/sqw/hp.sqw?k=82` (2003) [cit. 16. 11. 2010].
2. Parliamentary Library, Information Technology Department of the Office of the Chamber of Deputies, Information Technology Department of the Senate Chancellery of the Czech Parliament, Parliamentary Library, parliamentary archive and Information Technology Department of the National Council of the Slovak Republic: Joint Czech and Slovak Digital Parliamentary Library. [online] `http://www.psp.cz/cgi-bin/eng/eknih/info.htm` (2003) [cit. 16. 11. 2010].
3. Parliamentary Library of the Czech Republic: Czech (Bohemian) Assemblies Digital Library. [online] `http://www.psp.cz/cgi-bin/eng/sqw/hp.sqw?k=82` (2003) [cit. 16. 11. 2010].
4. Šmerk, P.: Unsupervised Learning of Rules for Morphological Disambiguation. In: Lecture Notes in Computer Science, Springer Berlin / Heidelberg (2004).
5. Šmerk, P.: Towards Computational Morphological Analysis of Czech. Ph.D. thesis, Faculty of Informatics, Masaryk University, Brno (2010).
6. Šmerk, P.: Fast Morphological Analysis of Czech. In: Proceedings of the RASLAN Workshop 2009, Brno (2009).
7. Rychlý, P.: Korpusové manažery a jejich efektivní implementace. Ph.D. thesis, Fakulta informatiky, Masarykova univerzita, Brno (2000).
8. Rychlý, P.: Manatee/Bonito – A Modular Corpus Manager. In: 1st Workshop on Recent Advances in Slavonic Natural Language Processing, Brno (2007) 65.
9. Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D.: The Sketch Engine. In: Proceedings of EURALEX. (2004) 105–116.
10. Christ, O., Schulze, B.M.: The IMS Corpus Workbench: Corpus Query Processor (CQP) User's Manual. University of Stuttgart, Germany, `http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench` (1994).
11. Jakubíček, M., Rychlý, P., Kilgarriff, A., McCarthy, D.: Fast syntactic searching in very large corpora for many languages. In: PACLIC 24 Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation, Tokyo (2010) 741–747.
12. Marx, M., Schuth, A.: DutchParl. A Corpus of Parliamentary Documents in Dutch. In: Proceedings of LREC 2010. (2010) `http://politicalmashup.nl/dutchparl`.
13. Martin, C., Marx, M.: Parliamentary documents from Spain. In: Proceedings of LREC 2010. (2010) `http://politicalmashup.nl/SpanishParliament`.
14. CWB open-source community: `BUNDESTAG` corpus. [online] `http://cogsci.uni-osnabrueck.de/~korpora/ws/CQPdemo/Bundestag/` (2010) [cit. 16. 11. 2010].
15. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: MT Summit. (2005).
16. Ilie, C., ed.: European Parliaments Under Scrutiny: Discourse Strategies and Interaction Practices. John Benjamins Publishing, Amsterdam, Netherlands (2010).
17. Just, V.: Velký slovník floskulí. LEDA, Praha (2009).
18. Horák, A., Rychlý, P., Kilgarriff, A.: Czech word sketch relations with full syntax parser. In: After Half a Century of Slavonic Natural Language Processing, Brno, Czech Republic, Masaryk University (2009) 101–112.