OBSAH

ÚVOD **9**

1 Vytěžování korpusových dat a dotazovací jazyk CQL **17**
1.1 Kvantitativní metody vyhodnocování korpusových dat 28
1.1.1 Vytvoření frekvenčního slovníku (v *Manatee/Bonitu*) 30
1.1.2 Disperze výrazů 32
1.1.3 Kolokace a koligace 33
1.1.4 Testy kolokační významnosti – asociační míry 39
1.1.4.1 *MI-score* – vzájemná informace 40
1.1.4.2 *T-test, t-score* – míra kontrastu 42
1.2 Základní regulární výrazy CQL (Corpus Query Language) 48

2 Anotace a technické aspekty tvorby korpusů **51**
2.1 Formát dat a kódování znaků 56
2.1.1 Konce řádků ve zdrojovém souboru – typy a jejich nastavení 61
2.1.2 Segmentace a tokenizace textu 62
2.1.3 Korpusové formáty vertikály a sloupcová (lingvistická) anotace 68
2.2 Anotace (metadata) a formát XML 71
2.2.1 Element nebo atribut? – strukturování XML dokumentu 83
2.2.2 Strukturace s elementy 84
2.2.3 Strukturace s atributy a prázdným elementem 86
2.2.4 Strukturace kombinací elementů a atributů 87
2.3 Možnosti XML pro lingvistické zpracování dat 89
2.3.1 Varianty korpusové vertikály 97
2.4 Pokročilejší práce s regulárními výrazy a jazykem CQL 100
2.4.1 Přegenerovávání komplexních strukturovaných vzorů a víceznačnost tagsetu 109
2.4.2 Rozšířené sady regulárních metaznaků 117
2.4.3 Souvislosti XML a CQL: způsob anotace a formát vyhledávacích masek 214

3 Nástroje pro vytěžování korpusových dat **139**
3.1 *WordList* 140
3.2 *ConcApp* 140
3.3 SCP – *Simple Concordance Program* 142
3.4 *Corsis-TenkaText* 142
3.5 *WConcord* 146

3.6 *TextSTAT* 149
3.7 *AntConc* 152
3.7.1 Regulární výrazy 154
3.7.2 Disperze výrazů a detekce jejich výskytu v definovaných úsecích 156
3.7.3 Slovníky, frekvenční seznamy a statistické testy 158
3.7.4 Klastry/n-gramy 159
3.7.5 Kolokace/koligace (*MI-score, t-score*) 159
3.7.6 Klíčová/tematická slova (*logLikelihood, Chi-squared*) 160
3.7.7 Indexace slovníku (lemma list) 162
3.8 *Xaira* 166
3.8.1 Kompilace korpusu 166
3.8.2 Typy dotazů a možnosti vytěžování dat 168
3.8.2.1 *Phrase Query* (fráze) 168
3.8.2.2 *Word Query* (slovní tvar) 169
3.8.2.3 *Addkey Query* (atributivní dotaz) 171
3.8.2.4 *Pattern Query* (řetězec) 172
3.8.2.5 *XML Query* (XML tagy) 173
3.8.2.6 *Query Builder Visual Interface* (grafická tvorba dotazu) 174
3.8.2.7 *CQL/XQL Query* (CQL/XQL dotaz) 176
3.8.2.8 *Collocation* (kolokace) 178

4 Možnosti automatického zpracování textu 183
4.1 Softwarové nástroje 184
4.1.1 *jTokenizer* 184
4.1.2 MLTC – *Multilingual Corpus Toolkit* 190
4.2 Počítačové skripty pro automatické zpracování textu 192
4.2.1 Příkazový řádek 192
4.2.1.1 Základy práce s příkazovým řádkem 193
4.2.1.2 Tipy pro práci s příkazovým řádkem 198
4.2.2 Počítačové skripty a jejich využití v korpusové lingvistice 198
4.2.2.1 *fsmTokenize* 199
4.2.2.2 *sentence-boundary.pl* 199
4.2.2.3 *join-files.py* 200
4.2.2.4 *concordance01.pl* 202
4.2.2.5 *concordance02.pl* 204
4.2.2.6 *simple_vocab.pl* 204
4.2.2.7 *freq_list.pl* 205
4.2.2.8 *bigramcount.pl* 206
4.2.3 Unixové/linuxové nástroje pro zpracování textu 206

4.2.3.1 Číslování řádků 208
4.2.3.2 Konverze konců řádků 210
4.2.3.3 Konverze kódování znaků 211
4.2.4 Linuxové nástroje a vytěžování korpusových dat 212
4.2.4.1 *awk*: manipulace s vertikálou a sloupcovou anotací 218
4.2.4.2 *sed*: textové konverze a frekvenční slovníky 221
4.2.4.3 Tvorba slovníku z příkazového řádku 223
4.2.5 Nástroje pro automatickou lingvistickou anotaci: lemmatizace a taggování 226
4.2.5.1 *MorphoDiTa* – tvorba automaticky anotované databáze 232
4.2.5.1.1 *Tokenizer* 233
4.2.5.1.2 *Morphological Generation* 235
4.2.5.1.3 *Morphological Analysis* 235
4.2.5.1.4 *Morphological Tagger* 236
4.2.6 *MorphCon* (konvertor morfologických tagsetů) 239
4.2.6.1 *SimpleTag-Conversion* 240
4.2.6.2 *KWIC/Tag-Format* 240
4.2.6.3 *WPL-VerticalMode* 241

5 Kompilace korpusu v systému *Manatee/Bonito* 243
5.1 Základní úprava zdrojového souboru: nastavení kódování znaků a typů konců řádků a souborového formátu 244
5.2 Konverze zdrojového souboru do vertikály 244
5.3 Anotace zdrojové vertikály 245
5.4 Příprava a konfigurace serveru *Manatee* 246
5.5 Konfigurace (deklarace) zdrojové vertikály 247
5.6 Kompilace: konverze zdrojové vertikály do korpusové databáze 250
5.7 Formát a tvorba paralelního korpusu 251

Závěr 257
Literatura a prameny 261
Přílohy 273
Rejstřík 282
Summary 287