

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 <i>Manatee/Bonitu</i>)	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 <i>MI-score</i> – vzájemná informace	40
1.1.4.2 <i>T-test</i> , <i>t-score</i> – 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 Přegenerování 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 <i>WordList</i>	140
3.2 <i>ConcApp</i>	140
3.3 SCP – <i>Simple Concordance Program</i>	142
3.4 <i>Corsis-TenkaText</i>	142
3.5 <i>WConcord</i>	146
3.6 <i>TextSTAT</i>	149
3.7 <i>AntConc</i>	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 (<i>MI-score</i> , <i>t-score</i>)	159
3.7.6	Klíčová/tematická slova (<i>logLikelihood</i> , <i>Chi-squared</i>)	160
3.7.7	Indexace slovníku (lemma list)	162
3.8	<i>Xaira</i>	166
3.8.1	Kompilace korpusu	166
3.8.2	Typy dotazů a možnosti vytěžování dat	168
3.8.2.1	<i>Phrase Query</i> (fráze)	168
3.8.2.2	<i>Word Query</i> (slovní tvar)	169
3.8.2.3	<i>Addkey Query</i> (atributivní dotaz)	171
3.8.2.4	<i>Pattern Query</i> (řetězec)	172
3.8.2.5	<i>XML Query</i> (XML tagy)	173
3.8.2.6	<i>Query Builder Visual Interface</i> (grafická tvorba dotazu)	174
3.8.2.7	<i>CQL Query</i> (CQL dotaz)	176
3.8.2.8	<i>Collocation</i> (kolokace)	178
4	Možnosti automatického zpracování textu	183
4.1	Softwarové nástroje	184
4.1.1	<i>jTokenizer</i>	184
4.1.2	MLTC	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	<i>fsmTokenize</i>	199
4.2.2.2	<i>sentence-boundary.pl</i>	199
4.2.2.3	<i>join-files.py</i>	200
4.2.2.4	<i>concordance01.pl</i>	202
4.2.2.5	<i>concordance02.pl</i>	204
4.2.2.6	<i>simple_vocab.pl</i>	204
4.2.2.7	<i>freq_list.pl</i>	205
4.2.2.8	<i>bigramcount.pl</i>	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 <i>awk</i> : manipulace s vertikálou a sloupcovou anotací	218
4.2.4.2 <i>sed</i> : 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 <i>MorphoDiTa</i> – tvorba automaticky anotované databáze	232
4.2.5.1.1 <i>Tokenizer</i>	233
4.2.5.1.2 <i>Morphological Generation</i>	235
4.2.5.1.3 <i>Morphological Analysis</i>	235
4.2.5.1.4 <i>Morphological Tagger</i>	236
4.2.6 <i>MorphCon</i> (konvertor morfologických tagsetů)	239
4.2.6.1 <i>SimpleTag-Conversion</i>	240
4.2.6.2 <i>KWIC/Tag-Format</i>	240
4.2.6.3 <i>WPL-VerticalMode</i>	241
5 Kompilace korpusu v systému <i>Manatee/Bonito</i>	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 <i>Manatee</i>	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