

# Oxford WordSmith Tools 4.0

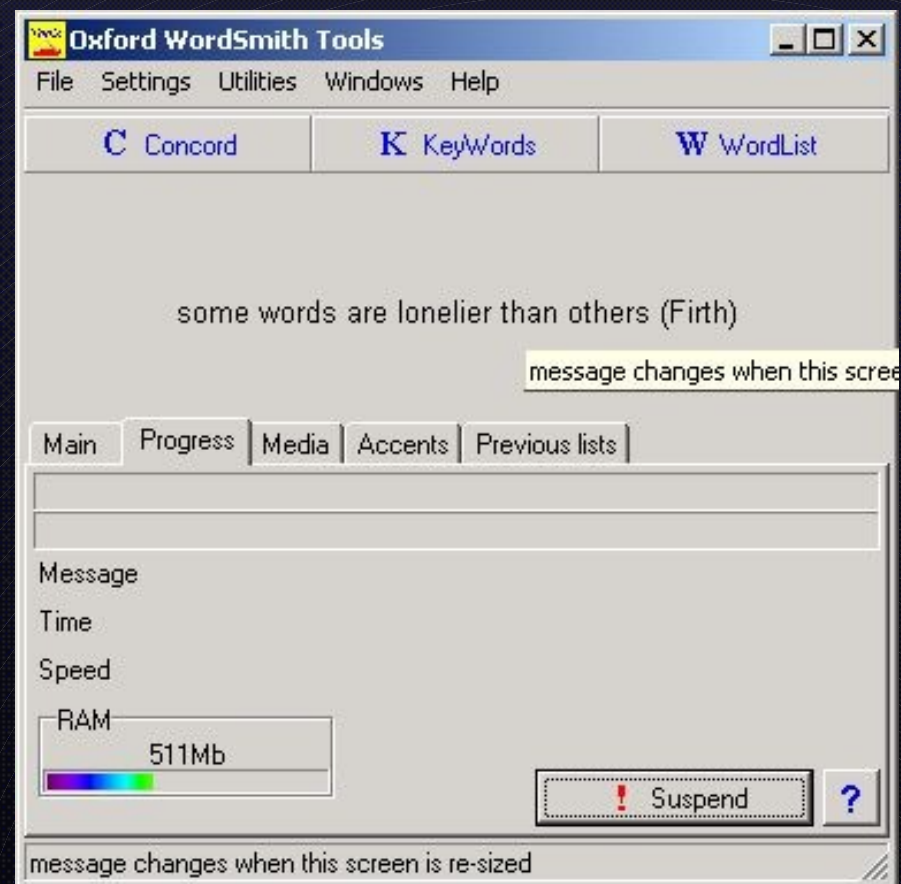
Bc. Jiří Mikulášek

PA154 – Nástroje pro korpusy  
FI MUNI

duben 2006

# WordSmith 4.0

- Základní informace
- WordList
- Concordancer
- Keywords
- Další nástroje



# Základní informace

- Autor: Mike Scott, Oxford University Press od roku 1996, aktuálně ve verzi 4.0.
- Komerční, demo verze pro práci nepoužitelná.
- Soubor nástrojů pro lexikální analýzu.
- Tři hlavní nástroje – WordList, Concordancer, Keywords.
- Množství dalších užitečných utilit – WebGetter, Language chooser, File

# WordList

- Analyzuje vstupní text a generuje frekvenční seznamy slov.
- Vstupem může být několik textových souborů (značkováného) textu.
- Výstupem je frekvenční tabulka výskytů.
- Výstup je možné řadit dle různých parametrů.
- Umožňuje lematizaci (po přidání seznamu lemat).

# WordList

	Word	Freq.	%	Texts	%	Lemmas	Set
1	THE	6 055 105	6,09	4 050	99,90		
2	OF	3 049 564	3,07	4 040	99,65		
3	AND	2 624 341	2,64	4 050	99,90		
4	TO	2 599 505	2,61	4 049	99,88		
5	A	2 181 592	2,19	4 045	99,78		
6	IN	1 946 021	1,96	4 047	99,83		
7	#	1 604 421	1,61	3 167	78,12		
8	THAT	1 052 259	1,06	4 026	99,31		
9	IS	974 293	0,98	4 027	99,33		
10	IT	922 687	0,93	4 022	99,21		
11	FOR	880 848	0,89	4 036	99,56		
12	WAS	863 917	0,87	3 931	96,97		
13	I	732 523	0,74	3 830	94,47		
14	ON	731 319	0,74	4 027	99,33		
15	WITH	659 997	0,66	4 012	98,96		
16	AS	655 259	0,66	3 992	98,47		
17	BE	651 535	0,66	4 011	98,94		
18	HE	593 609	0,60	3 739	92,23		
19	YOU	588 503	0,59	3 619	89,27		
20	AT	534 875	0,53	4 000	99,87		

frequency | alphabetical | statistics | filenames | notes

512 588 Type-in demo limit = 100

# WordList – oblasti použití

- Vytvoření podkladů pro hledání klíčových slov a vytváření konkordancí.
- Analýza konzistence – porovnávání různých verzí téhož textu apod.
- Lematizace – soubor lemat, nebo vytváření lemat ručně.
- Analýza souvislosti slov.

# Concordancer

- Vytváření konkordancí – příkladů slov nebo frází s jejich kontexty.
- Vyhledá ve vstupních text(ech) konkordance podle vyhledávacích kritérií.
- Vyhledávání podle kontextu.
- Zobrazení kolokací, možnost nastavení kolokačních vzdáleností.
- Zpracování značkováného textu.
- Možnost editace vlastních kategorií





# Concordancer - použití

- Studium jazyka, hledání slovních spojení apod.
- Pomůcka pro učitele.
- Zkoumání chování slov vzhledem k jejich kontextu.

# Keywords

- Pokusí se identifikovat klíčová slova v textu na základě porovnávání frekvencí výskytu.
- Pracuje s výstupy WordListu – jeden wordlist pro vyhledání klíčových slov a druhý (větší) pro porovnání frekvencí.
- Možnost přímého nalezení konkordancí k nalezeným klíčovým slovům.
- Asociace klíčových slov.
- ...

# Keywords

The screenshot shows the KeyWords software interface. The main window displays a table with columns: Key word, Freq., %, . Freq., RC. %, Keyness, P, emmas, and Set. The first 10 rows list common words like #, THE, I, YOU, YEAH, A, AND, THAT, TO, and WE. Rows 11-19 contain a repetitive phrase 'past demo limit' with various colored highlights. The bottom of the window features a menu bar (File, Edit, View, Compute, Settings, Windows, Help), a toolbar with buttons for 'KW's', 'plot', 'links', 'clusters', 'filenames', 'notes', and 'source text', and a status bar showing '201' and 'Type-in demo limit = 10'.

	Key word	Freq.	%	. Freq.	RC. %	Keyness	P	emmas	Set
1	#	3 116	10,40	04 421	1,61	6 584,60	000000		
2	THE	1 249	4,17	55 105	6,09	-216,16	000000		
3	I	615	2,05	32 523	0,74	476,77	000000		
4	YOU	545	1,82	88 503	0,59	492,72	000000		
5	YEAH	538	1,79	83 012	0,08	2 281,22	000000		
6	A	534	1,78	81 592	2,19	-25,30	004888		
7	AND	531	1,77	24 341	2,64	-98,95	000000		
8	THAT	529	1,76	52 259	1,06	119,11	000000		
9	TO	522	1,74	99 505	2,61	-101,26	000000		
10	WE	432	1,44	00 833	0,30	669,82	000000		
11	past demo limit	o limit	o limit	o limit	o limit	ast demo limit	o limit	o limit	imit
12	past demo limit	o limit	o limit	o limit	o limit	ast demo limit	o limit	o limit	imit
13	past demo limit	o limit	o limit	o limit	o limit	ast demo limit	o limit	o limit	imit
14	past demo limit	o limit	o limit	o limit	o limit	ast demo limit	o limit	o limit	imit
15	past demo limit	o limit	o limit	o limit	o limit	ast demo limit	o limit	o limit	imit
16	past demo limit	o limit	o limit	o limit	o limit	ast demo limit	o limit	o limit	imit
17	past demo limit	o limit	o limit	o limit	o limit	ast demo limit	o limit	o limit	imit
18	past demo limit	o limit	o limit	o limit	o limit	ast demo limit	o limit	o limit	imit
19	past demo limit	o limit	o limit	o limit	o limit	ast demo limit	o limit	o limit	imit

# Keywords - využití

- Dobrá pomůcka pro charakterizaci textu – styl, žánr, ...
- Analýza obsahu.
- Vyhledávání informací.
- Klasifikace textu.

# Další nástroje

- Slušné množství užitečných nástrojů.
- WebGetter – vytváření vlastního korpusu pomocí vyhledávání textů na webu.
- Minimal Pairs – hledá páry co nejpodobnějších slov.
- File utilities – porovnávání, rozdělování, ...
- Text Converter – nahrazování částí textu

# Reference

- <http://www.lexically.net/wordsmith/>