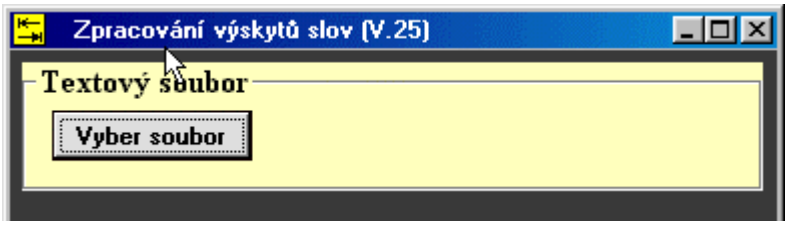
Popis programu

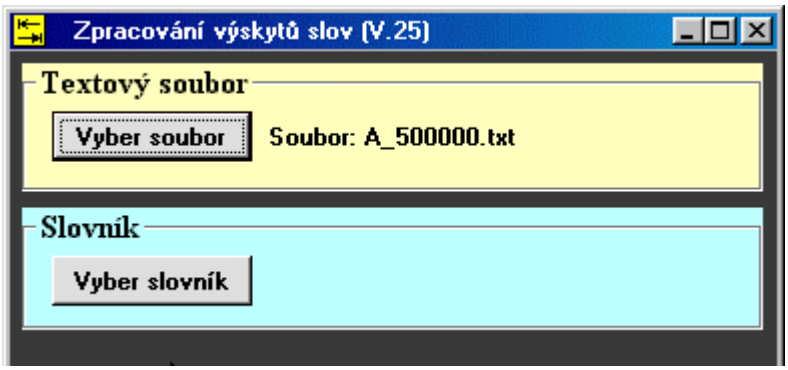
1. Výběr textového souboru ke zpracování.

Textový soubor musí mít příponu „txt“. V následně otevřeném okně se vybere příslušný datový soubor. Soubor se nemusí nacházet v adresáři odkud je program spouštěn. Soubor musí být čistý, neformátovaný (plain text) textový soubor.



2. Výběr slovníku s klíčovými slovy.

Slovník musí mít příponu „slv“. V následně otevřeném okně se vybere příslušný soubor slovníku. Soubor se nemusí nacházet v adresáři odkud je program spouštěn.



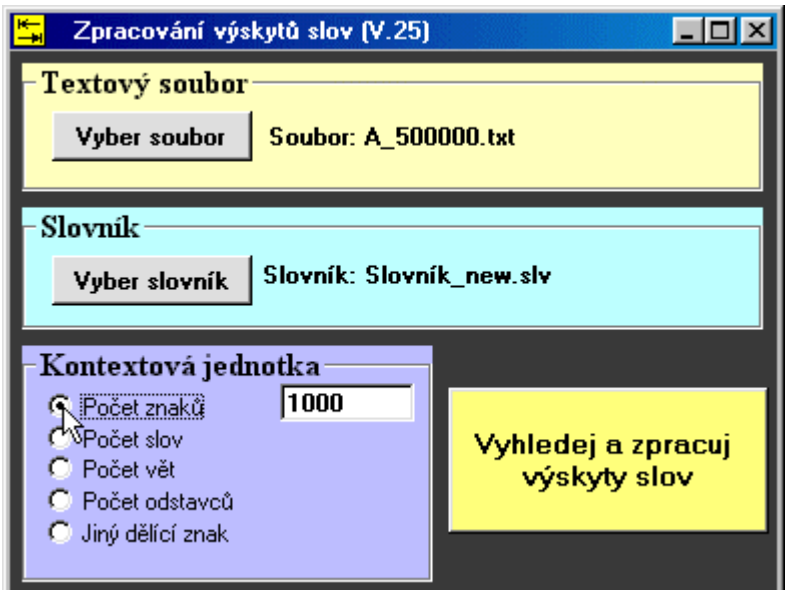
Slovník je obecný textový soubor sestavený v běžném textovém editoru (notepad).

Slovník sestává z několika základních (klíčových slov, ke kterým je možné připojit synonyma. Klíčová slova s jejich synonymy jsou vždy na jednom řádku a jsou oddělena čárkami. Kolik klíčových slov, tolik je ve slovníku řádků. Při zadávání slov a synonym je možné použít hvězdičkovou “*“ konvenci. Pokud slovo nebo synonymo začíná nebo končí znakem “*“, pak hvězdička nahrazuje jakékoliv znaky (zákon* = zákonný, zákonodárný atd; *zákon* = nezákonný,zákonný, zákonodárný atd.). Pokud slovo hvězdička není použita, je v textu vyhledána jen přímá shoda.

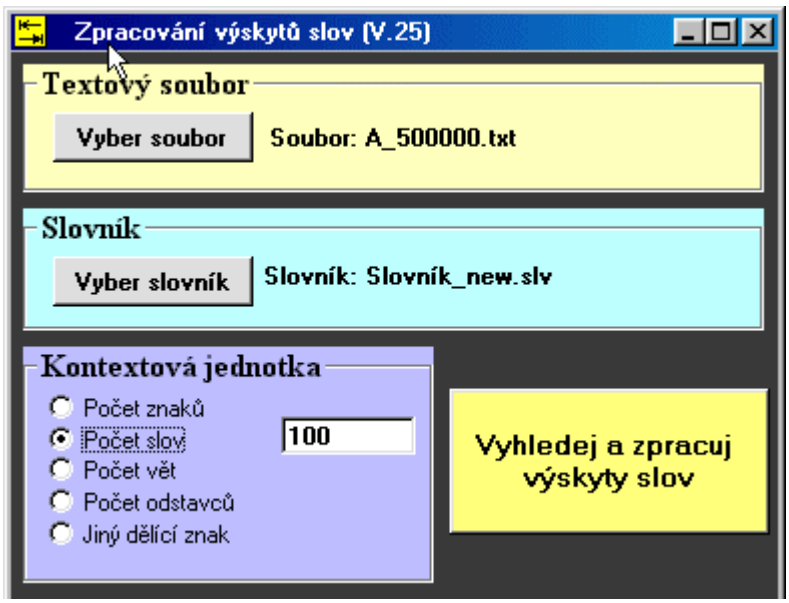
3. Volba kontextové jednotky.

Celý zpracovávaný textový soubor je rozdělen do malých částí - kontextových jednotek, ve kterých jsou pak sledovány výskyty zadaných slov. Způsob a velikost kontextové jednotky je možné volit několika způsoby:

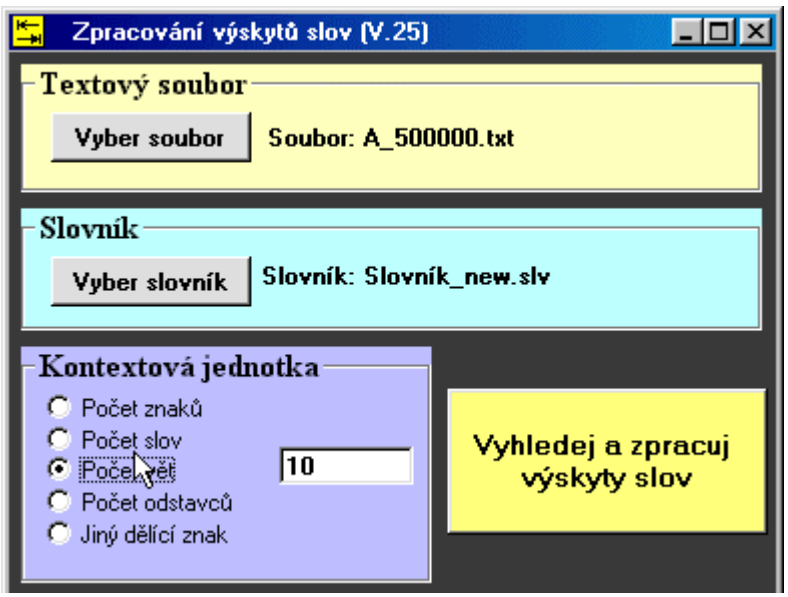
a) V počtu znaků. Celý textový soubor je rozdělen na části se zadaným počtem znaků.



b) V počtu slov. Celý textový soubor je rozdělen na části se zadaným počtem slov. Jako slovo je považován sled znaků ohraničený mezerou.

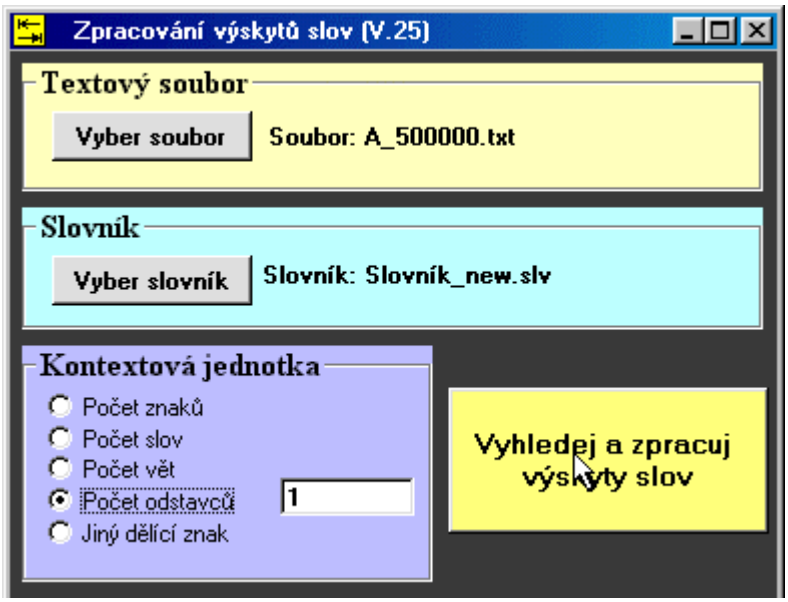


c) V počtu vět. Celý textový soubor je rozdělen na části se zadaným počtem vět. Věta je část textového souboru mezi znaky: "." nebo "!" nebo "?". Na začátku a na konci může být kterýkoliv z nich.

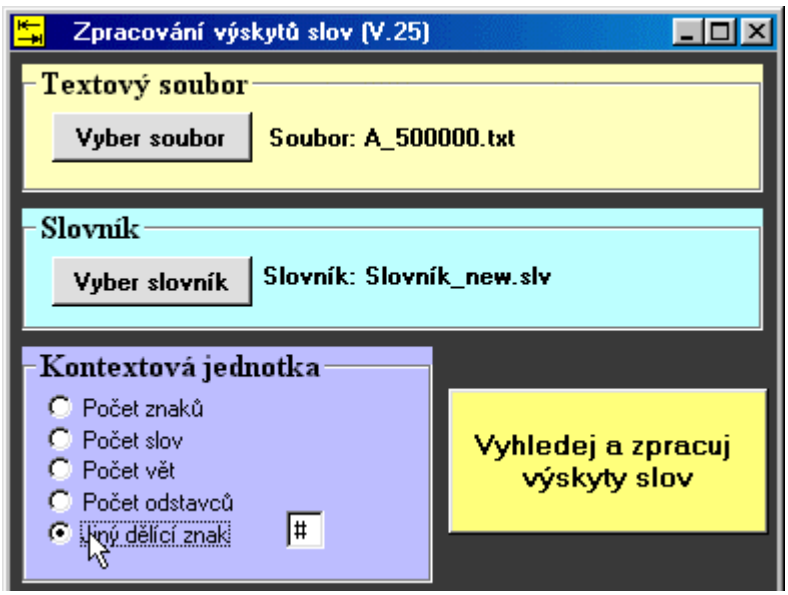


Poznámka: Při dělení na věty nelze vždy rozpoznat jednotlivé věty. Jako konce vět jsou například vzaty všechny tečky, které se v textu nacházejí, tedy i tečky za řadovými číslovkami nebo za tituly.

d) V počtu odstavců. Celý textový soubor je rozdělen na části se zadaným počtem odstavců. Odstavec je část textového souboru mezi znaky konců odstavců (CR, chr(13))



e) Dělení speciálním znakem. Do textového souboru je do určitých míst vložen speciální znak, který je dělící značkou při dělení do kontextových jednotek.



Poznámka: Pokud je zpracovávaný textový soubor příliš objemný a současně je zadáno kritérium na jeho rozdělení do kontextových jednotek nevhodně tak, že jejich počet je příliš vysoký, může dojít z příčiny nedostatku paměťového prostoru PC k ukončení programu. Kritéria je pak třeba vhodně upravit.

4. Zpracování souboru podle zadaných kritérií.

Spuštění zpracování se provede stiskem tlačítka „Vyhledej a zpracuj výskyty slov“. Program před započetím vyhledávání slov a zjišťováním vztahů (vzdáleností) mezi nimi musí zdrojový textový soubor upravit. Úpravu se provádějí v několika krocích a jsou závislé na zadaném dělícím prvku kontextové jednotky. Postup úprav je zobrazen textem ve spouštěcím tlačítku, které má v tomto případě rudou barvu. Úpravy ve stručnosti spočívají v odstranění nadbytečných a pro dané zpracování nevýznamných znaků (vícenásobný sled teček, mezer, konců odstavců atd.).

Výsledek zpracování je po ukončení výpočtů zobrazen na obrazovce jako textový soubor jehož příklad je uveden v Příloze 1. Výčet obsahuje základní data o zvoleném textovém souboru a vybraném slovníku. Následuje tabulka se souhrnem nalezených výskytů slov v kontextových jednotkách.

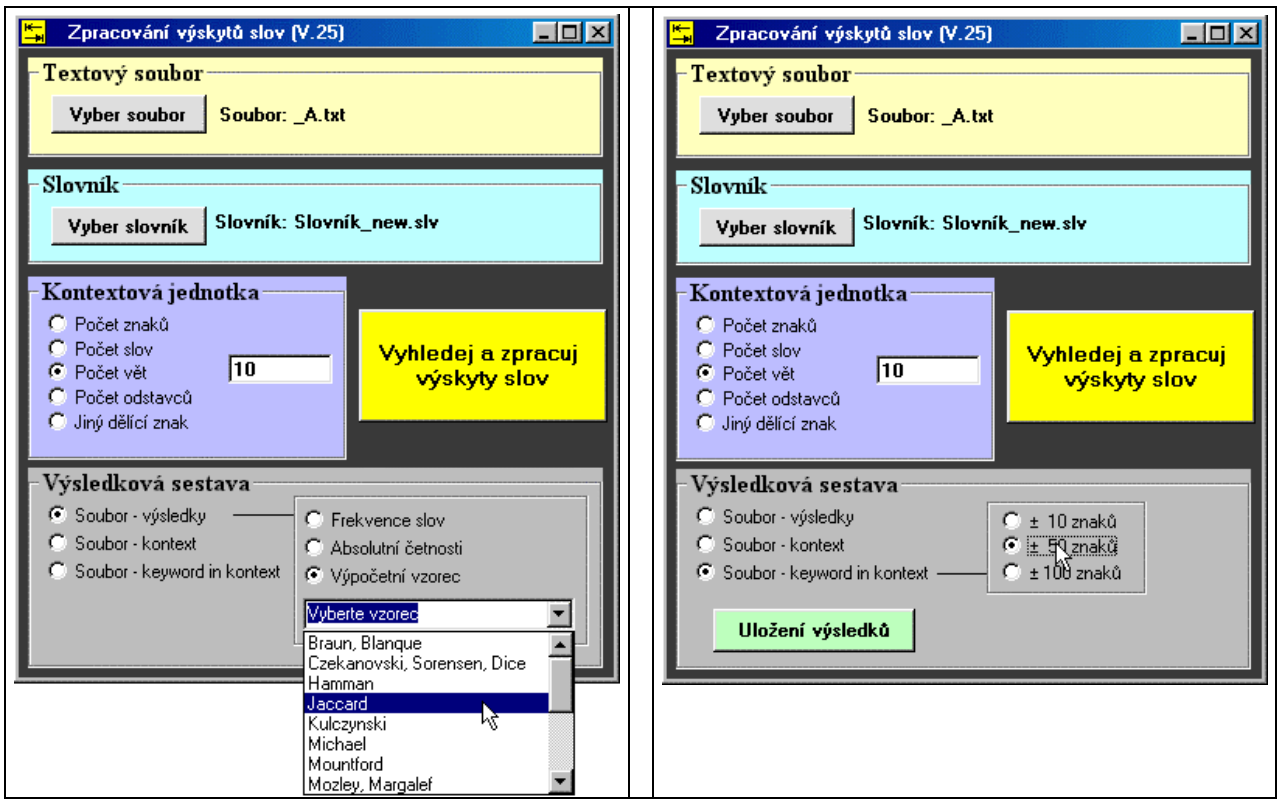
Pak jsou uvedeny výsledky výpočtů absolutních a relativních koeficientů podle všech použitých výpočetních vztahů. Jejich podrobnější popis je v Příloze 2.

5. Volba výstupních tiskových sestav.

K volbě souboru výstupní tiskové sestavy se dostaneme po zavření okna s náhledem výsledků zpracování.

Podle níže uvedených možností jsou vybrány příslušné výsledky rozborů a uloženy do prostého textového souboru. Soubor je uložen ve stejném adresáři, ve kterém byl spuštěn výpočetní program.

Možnosti pro volbu výstupní sestavy jsou v podstatě tři:
- výsledková sestava - výsledky,
- výsledková sestava - kontextové jednotky,
- výsledková sestava - výpis nalezených klíčových slov s okolím (keyword in kontext).



Název textového souboru výsledkové sestavy je vytvořen z vybraných kritérií pro zpracování. Má následující strukturu.

- Jednotlivé položky názvy jsou odděleny znakem podtržítka „_“.
- Název zpracovaného souboru bez přípony. Název je zkrácen na maximální délku 12 znaků.
- Název vybraného slovníku bez přípony. Název je zkrácen na maximální délku 12 znaků.
- Vybraný typ kontextové jednotky:
  - Z pro znaky, následovaný počtem znaků v kontextové jednotce
  - S pro slova, následovaný počtem slov v kontextové jednotce
  - V pro věty, následovaný počtem vět v kontextové jednotce
  - O pro odstavce, následovaný počtem odstavců v kontextové jednotce
  - K# pro jiný oddělovací znak
- Datum a čas ve tvaru _rok-měsíc-den_hodina-minuta-sekunda
- Typ výstupu:
  - RES pro sestavu výsledků
  - KON pro soubor výpisu kontextových jednotek
  - KWIC pro soubor výpisu okolí kolem nalezených klíčových slov

a) Výsledková sestava - výsledky.

Výsledková sestava vždy začíná základními údaji o vybraném textovém souboru a slovníku. Pak následuje výběr:

a1) Frekvence slov - tabulka se souhrnem nalezených výskytů slov v kontextových jednotkách. Popis tabulky je podle příkladu z Přílohy 1.

```
Výskyty slov v kontextových jednotkách
                        Výskyty      Výskyty v     % v       % v
                      slova celkem  jednotkách  jednotkách   sumě
 ------------------------------------------------------------------------
 1.  SPRAVEDL:           [  57]       [  55]     [ 96.49%] [  4.38%]
 2.  ZÁKON:              [1129]       [ 670]     [ 59.34%] [ 86.71%]
 3.  _PRÁVO_:            [  83]       [  75]     [ 90.36%] [  6.37%]
 4.  ČEST:               [  33]       [  32]     [ 96.97%] [  2.53%]
 ------------------------------------------------------------------------
     SUMA VÝSKYTŮ:        1302                                 100%
```

V celém textovém souboru se klíčové slovo „spravedl“ (bez ohledu na velikosti písmen a i slova, která klíčové slovo obsahovala [hvězdičková konvence - zde hvězdičky neuvedeny]) vyskytlo celkem 57 krát. Slovo bylo nalezeno v 55ti kontextových jednotkách. Znamená to, že v několika kontextových jednotkách se slovo vyskytlo vícekrát než jednou. Klíčové slovo „_právo_“ (podtržítka znamenají, že ve slovníku bylo slovo zadáno přesně - vyhledáván je jen tento sled znaků ohraničený z obou stran mezerami) se celkem vyskytlo 83 krát a v 75ti kontextových jednotkách.

Třetí sloupec je procentní poměr počtu kontextových jednotek s výskytem slova k celkovému počtu výskytů:100*55/57 = 96,49 % pro slovo „spravedl“.

Poslední sloupec je procentní poměr počtu výskytu slova k celkovému počtu výskytů všech slov: 100*1129/1302 = 4,38 % pro slovo „zákon“.

a2) Absolutní četnosti - tabulka absolutních výskytů.

Velikost (rozsah) „trojúhelníkové“ tabulky závisí na počtu klíčových slov ve slovníku. Počet řádků je o 1 menší než počet klíčových slov. Tabulka má následující strukturu (pro uvedený příklad):

2:1
3:1 3:2
4:1 4:2 4:3

kde

2:1 je počet současných výskytů klíčového slova 2 (*zákon*) a klíčového slova 1 (*spravedl*) ve stejných kontextových jednotkách.

4:3 je počet současných výskytů klíčového slova 4 (*čest*) a klíčového slova 3 (právo) ve stejných kontextových jednotkách.

a3) Výpočetní vzorec - tabulka relativních výskytů

Výsledkem je opět trojúhelníková tabulka s relativními vzájemnými výskyty podle výše uvedeného pořadí. Výskyty jsou vypočteny dle zvoleného výpočetního vzorce. Způsob výpočtu je podrobněji uveden v Příloze 2.

b) Výpis kontextových jednotek.

Výsledkový prostý textový soubor obsahuje číslovaný výpis všech kontextových jednotek vytvořených podle zvoleného způsobu. Jako příklad následuje část textového souboru a výpis pěti kontextových jednotek při zadání slova v počtu 10.

```
Zájmy strojvůdců  Zájmy strojvůdce č. 9  z roku 2003
      Vydáno 15. května  2003  Alespoň „někteří˝ lidé byli  asi
doposud na dráze nenahraditelní  Z diskuze generálního ředitele
Českých drah ing. Petra Kousala se  členy prezídia Federace
strojvůdců ČR  Jak jsme již uvedli  v minulém vydání Zájmů
strojvůdce, dne 3. dubna se části jednání  prezídia Federace
strojvůdců ČR zúčastnil i nový generální ředitel ČD  ing. Petr
Kousal.
```

```
1:
zájmy strojvůdců zájmy strojvůdce č. 9 z roku 2003 vydáno
2:
 15. května 2003 alespoň „někteří˝ lidé byli asi doposud na
3:
 dráze nenahraditelní z diskuze generálního ředitele českých drah ing. petra
4:
 kousala se členy prezídia federace strojvůdců čr jak jsme již
5:
 uvedli v minulém vydání zájmů strojvůdce, dne 3. dubna se
```

c) Výpis nalezených klíčových slov s okolím (keyword in kontext).

Výsledkový prostý textový soubor v tomto případě obsahuje okolí kolem všech nalezených klíčových slov (včetně synonym). Velikost okolí je volitelná v počtu znaků 10, 50 a 100 na obě strany. Ukázka výpisu je pro okolí 10 znaků. Vybrány jsou jen některé řádky výpisu pro ilustraci.

```
SPRAVEDL:          ezávisle, SPRAVEDLivě a průh
SPRAVEDL:          ucí získá SPRAVEDLivý a nedi
SPRAVEDL:           měla být SPRAVEDLivější pro
SPRAVEDL:          espektuje SPRAVEDLivé požada
SPRAVEDL:           vede k neSPRAVEDLnosti. k t
SPRAVEDL:          stanovení SPRAVEDLivého rozd
SPRAVEDL:          ve až po „SPRAVEDLivém” zása
ZÁKON:              dohodnut ZÁKONem garanto
ZÁKON:              zmíněným ZÁKONem je přím
ZÁKON:             ní novely ZÁKONa č. 120/1
ZÁKON:             h je protiZÁKONný a proti
ZÁKON:             y nějakým ZÁKONným opatře
ZÁKON:             ívá všech ZÁKONných i nez
ZÁKON:              pro část ZÁKONodárců to
ZÁKON:             od návrh „ZÁKONa“ podepsa
 PRÁVO :           mají totiž PRÁVO nahlížet d
 PRÁVO :            pro deník PRÁVO uvedl, že
 PRÁVO :           5. 3. 2005 PRÁVO jiří novot
 PRÁVO :           rá by měla PRÁVO rozhodovat
 PRÁVO :           k základní PRÁVO jednat za
ČEST:              váta, kus ČESTného a spo
ČEST:              u možnost ČESTného odsto
```

Příloha 1: Celkový výpis výsledků zpracování

```
Zpracováno dne:            05.01.2006    13:08:20

Zpracovaný soubor:         _A.txt
Počet znaků v souboru:     5415207

Volba zpracování:
Typ kontextové jednotky:                Kontextová jednotka v počtu vět
Počet vět v kontextové jednotce:        10
Počet kontextových jednotek v souboru:  5833
Počet vět v souboru celkem:             58332
Průměrná délka věty:                    13,83 slov

Slovník:                   Slovník_new.slv
Počet základních slov:     4
 1: *spravedl*
 2: *zákon*
 3: právo
 4: *čest*

Výskyty slov v kontextových jednotkách
                       Výskyty      Výskyty v     % v       % v
                     slova celkem  jednotkách  jednotkách   sumě
-----------------------------------------------------------------------
1.  SPRAVEDL:          [  57]       [  55]     [ 96,49%] [  4,38%]
2.  ZÁKON:             [1129]       [ 670]     [ 59,34%] [ 86,71%]
3.  _PRÁVO_:           [  83]       [  75]     [ 90,36%] [  6,37%]
4.  ČEST:              [  33]       [  32]     [ 96,97%] [  2,53%]
-----------------------------------------------------------------------
    SUMA VÝSKYTŮ:       1302                                 100%

Absolutní četnosti ..............................................
 8
 3           37
 1           8             0

Výchozí hodnoty pro výpočty koeficientu podobnosti .................
              a             b             c             d
             ---------------------------------------------
   2:1        8             662           47            5115

   3:1        3             72            52            5705
   3:2        37            38            633           5124

   4:1        1             31            54            5746
   4:2        8             24            662           5138
   4:3        0             32            75            5725

Koeficienty podobnosti pro různé výpočetní vzorce .................
 Braun, Blanque:
 0,01194
 0,04000  0,05522
 0,01818  0,01194     1,00000

 Czekanovski, Sorensen, Dice:
 0,02207
 0,04615  0,09933
 0,02299  0,02279     1,00000

 Hamman:
 0,75686
 0,95748  0,76989
 0,97085  0,76475     0,96331

 Jaccard:
 0,01116
 0,02362  0,05226
 0,01163  0,01153     1,00000

 Kulczynski:
 0,07870
 0,04727  0,27428
 0,02472  0,13097     1,00000
```

Michael:
0,00147
0,00164 0,02445
0,00049 0,00374 -0,00029

Mountford:
0,00024
0,00076 0,00101
0,00058 0,00043 1,00000

Mozley, Margalef:
1,26611
4,24145 4,29421
3,31364 2,17612 1,00000

Ochiai:
0,04167
0,04671 0,16506
0,02384 0,05464 1,00000

Phi:
0,00935
0,03610 0,13546
0,01677 0,03147 -0,00848

Rorers, Tanimoto:
0,78321
0,95836 0,79363
0,97127 0,78951 0,96397

Russell, Rao:
0,00137
0,00051 0,00634
0,00017 0,00137 1,00000

Simple matching:
0,87843
0,97874 0,88495
0,98543 0,88237 0,98165

Simpson:
0,14545
0,05455 0,49333
0,03125 0,25000 1,00000

Sokal, Sneath, Anderberg:
0,00561
0,01195 0,02683
0,00585 0,00580 1,00000

Yule:
0,13613
0,64102 0,77482
0,54879 0,44245 -0,00000

Příloha 2: Popis prováděných výpočtů

Velikost (rozsah) „trojúhelníkové“ matice relativních hodnot (koeficientů podobnosti) závisí na počtu klíčových slov ve slovníku. Počet řádků je o 1 menší než počet klíčových slov. Matice má následující strukturu (pro uvedený příklad):

2:1

3:1 3:2

4:1 4:2 4:3

kde:

2:1 je výpočtem zjištěná hodnota koeficientu odlišnosti, podle vybraného vzorce, z matice binárních dat zjištěných z výskytů klíčového slova 2 (*zákon*) a klíčového slova 1 (*spravedl*) ve stejných kontextových jednotkách.

4:3 je výpočtem zjištěná hodnota koeficientu odlišnosti, podle vybraného vzorce, z matice binárních dat zjištěných z výskytů klíčového slova 2 (*čest*) a klíčového slova 1 (právo) ve stejných kontextových jednotkách.

Číslování slov odpovídá řádkům ve slovníku. V uváděném příkladu je:

```
1: *spravedl*
2: *zákon*
3: právo
4: *čest*
```

Matice binárních koeficientů má tvar:

a	b	a +b
c	d	c +d
a+c	b+d	a +b +c +d

kde:

a je počet současných výskytů „prvního“ a „druhého“ slova v téže kontextové jednotce. První je odpovídající slovo z poměru první : druhé z trojúhelníkové matice relativních hodnot.

b je počet výskytu „prvního“ slova při nevýskytu „druhého“ slova v téže kontextové jednotce.

c je počet nevýskytu „prvního“ slova při výskytu „druhého“ slova v téže kontextové jednotce.

c je počet současných nevýskytů „prvního“ a „druhého“ slova v téže kontextové jednotce.

Binární koeficienty a, b, c, d jsou pro úplnost zobrazeny na obrazovce po proběhnutí výpočtů. Z nich je pak vypočten koeficient podobnosti Srs. Pro výpočet koeficientu podobnosti Srs existuje mnoho metod. Program koeficient vypočítává podle níže uvedených 16ti metod.

Braun, Blanque: $$Srs = \frac{a}{\max\{(a+b),(a+c\}}$$

Czekanovski, Sorensen, Dice: $$Srs = \frac{2a}{2a+b+c}$$

Hamman: $$Srs = \frac{a-(b+c)+d}{a+b+c+d}$$

Jaccard: $$Srs = \frac{a}{a+b+c}$$

Kulczynski: $$Srs = \frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$$

Michael: $$Srs = \frac{4(ad-bc)}{\left[(a+d)^2+(b+c)^2\right]}$$

Mountford: $$Srs = \frac{2a}{2(b+c)+2bc}$$

Mozley, Margalef: $$Srs = \frac{a(a+b+c+d)}{(a+b)(a+c)}$$

Ochiai: $$Srs = \frac{a}{\sqrt{(a+b)(a+c)}}$$

Phi: $$Srs = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

Rorers, Tanimoto: $$Srs = \frac{a+d}{a+2b+2c+d}$$

Russell, Rao: $$Srs = \frac{a}{a+b+c+d}$$

Simple matching: $$Srs = \frac{a+d}{a+b+c+d}$$

Simpson: $$Srs = \frac{a}{\min\{(a+b),(a+c)\}}$$

Sokal, Sneath, Anderberg: $$Srs = \frac{a}{a+2(b+c)}$$

Yule: $$Srs = \frac{ad - bc}{ad + bc}$$