OpenThesaurus-SK:
Otvorený slovenský synonymický slovník vytváraný pomocou komunity na webe

http://www.openthesaurus.tk/
2005-02-12

Abstrakt

Projekt OpenThesaurus je PHP/MySQL web rozhranie, ktoré umožňuje spoluprácu viacerých ľudí na tvorbe synonymického slovníka. Ktokoľvek sa môže prihlásiť, vyhľadávať synonymá v databáze, pridávať nové, prípadne opravovať a dopĺňať už vytvorené synonymické spojenia.

Zozbierané dáta sú voľne dostupné (licencia je v prílohe). Dáta sú exportované do formátov OpenOffice.org, KWord a ako prostý text. Zdrojové kódy, pomocou ktorých slovník pracuje, môžu byť použité na vytvorenie ďalších jazykových verzií.

Obsah

Všeobecne

Čo je to projekt OpenThesaurus

Synonymický slovník je zoznam slov, ktoré majú rovnaký alebo podobný obsahový význam. Najjednoduchším príkladom sú dve slová s rovnakým významom, napríklad *agresívny, útočný*.

Do doby než sa nám podarilo vytvoriť tento slovník, neexistoval slovenský synonymický slovník, ktorý by bol poskytovaný slobodne. Doteraz dostupné slovníky boli poskytované len komerčne. Slovenský synonymický slovník je jedným z jazykových mutácií projektu OpenThesaurus.

Projekt OpenThesaurus má za cieľ vytvoriť synonymický slovník, ktorý by bol poskytovaný pod otvorenou licenciou a tým pádom bol dostupný všetkým. Táto licencia zaručuje, že ani v budúcnosti nemôžu byť tieto dáta spoplatnené. Ďalšou významnou vecou je to, že na zdokonaľovaní a rozširovaní slovníka sa môže aktívne podieľať každý.

Stránky OpenThesaurus sú realizované pomocou programovacieho jazyka PHP a databázy MySQL, čo ho umožňuje relatívne jednoducho inštalovať na všetky web servery, ktoré túto technológiu podporujú. Pomerne jednoduchá štruktúra je ľahko zrozumiteľná aj pre používateľov ktorí nemajú dostatočné znalosti lingvistiky.

Aké môže byť využitie synonymického slovníka?

Pri tvorbe referátov, ročníkových prác, a iných textov, u ktorých chcete zvýšiť kvalitu. Ak sa často opakujú v texte tie isté slová, môžete ich nahradiť synonymami a text sa stáva živším. Taktiež pomôže nahradiť cudzie slová zrozumiteľnejšími alebo naopak jednoduché slová odbornými.

Pre prekladateľov a spisovateľov je slovník neoceniteľným pomocníkom. Hoci ich slovná zásoba je nanajvýš závideniahodná, vždy je čo zdokonalovať a nie vždy vám „slina prinesie na jazyk“ to správne slovo.

Ak vytvárate texty v cudzom jazyku, budú vám užitočné slovníky v iných jazykoch (pozri Projekt OpenThesaurus v iných jazykoch a ďalšie odkazy).

Vo vyhľadávacích službách, s cieľom nájsť dokumenty, ktoré hľadanú informáciu obsahujú, avšak používajú inú terminológiu.

Kde môžem získať dáta slovníka?

Linky na stiahnutie nájdete na úvodnej stránke projektu (obr.1G). Licenciu nájdete na http://rak.bb.euroweb.sk/~zdpo/thesaurus/licence.html. Zatiaľ je iba v angličtine. Prípadný preklad je vítaný.

Generovanie slovníkov prebieha automaticky každý piatok po 23.00. Podporované formáty sú:

• OpenOffice.org 1.1.4
• OpenOffice.org 2.x
• KThesaurus (súčasť projektu KOffice)
• textový formát

Ako začať?

Na úvodnej stránke je odkaz „Kontrola synonymických radov“ (obr.1C). Stránka slovenského Tezaura vám umožňuje náhodné vyhľadávanie synonymických radov. Toto je užitočné, keď chcete pomôcť projektu, avšak neviete kde začať.

Takto pomôžete skontrolovať databázu synoným. Databáza zaznamenáva koľkokrát boli synonymá zobrazené. Ak si dáte zobraziť ďalšie náhodné slová, vyberú sa prednostne také, ktoré ešte neboli zobrazené, alebo boli zobrazené najmenej často.

Keď prídete na to, že v synonymickom rade je chyba (napríklad nejaké slovo tam nepatrí) môžete naň kliknúť a chybu (ak ste registrovaný) opraviť.

Čo znamenajú pojmy synonymum, synonymický rad slov ?

Ak dve alebo viac slov majú v istom kontexte rovnaký význam, tak im hovoríme, že sú to synonymá. Napríklad:

agresívny, útočný
nemravný, neslušný, obscénny
žalobaba, donášač

Synonymá vytvárajú *synonymické rady* - skupiny slov s rovnakým významom. Preto sa slová s rôznym významom - ako napríklad agent - objavujú vo viacerých radoch synonym:

Synonymický rad 1: agent, sliedič, vyzvedač, špión
Synonymický rad 2: agent, sprostredkovateľ, zástupca

Poznámka: Angličtina pre synonymické rady používa označenie *synsets* (WordNet).

POUŽÍVATEĽSKÝ MANUÁL

Registrácia a nastavenia používateľa

Registrovať sa nemusíte, pokiaľ chcete slovník používať len na vyhľadávanie synonym alebo si stiahnuť jeho súbory do externých aplikácií.

Naopak, ak máte v úmysle pomôcť zlepšiť slovník a zapisovať doň nové slová, registrácia je nutná. Odporúčam vám prečítať si celú túto dokumentáciu, aby ste sa zoznámili s pravidlami vytvárania slovníka.

Hneď na úvodnej stránke vpravo hore sú linky, kde sa môžete prihlásiť do projektu. Bude potrebné aby ste vyplnili e-mailovú adresu, na ktorú vám bude zaslané heslo potrebné na prihlásenie a vyjadrili súhlas s podmienkami projektu.

Toto heslo si môžete zmeniť a takisto nastaviť meno, ktoré bude zobrazované vo verejne prístupných štatistikách. Ak meno nevyplníte, zostanete anonymný.

V prípade, že sa vám nepodarí prihlásiť, je možné, že nemáte povolené „Cookies“. Skontrolujte si nastavenie vo svojom internetovom prehliadači. Spôsob nastavenia nájdete v jeho dokumentácií.

Čo by som mal brať do úvahy, keď chcem niečo zmeniť v slovníku?

Skôr ako vložíte alebo zmeníte nejakú položku musíte pochopiť, ako sú dáta v slovníku štruktúrované - podľa významu. Napríklad slovo *agent* môže byť význame *sprostredkovateľ* (poistenia), alebo vo význame *vyzvedač*. Pritom slová *sprostredkovateľ* a *špión* synonymami nie sú.

Preto pre agenta musia v slovníku existovať dva synonymické rady (významy): v jednom bude slovo *sprostredkovateľ* a v druhom bude *špión*. Takto:

agent, sprostredkovateľ
agent, špión

V našom úmysle je vytvoriť otvorený a slobodný slovník. Z tohto dôvodu nesmú byť do databázy vkladané synonymá spôsobom, ktorý by znemožňoval voľné šírenie. To znamená prepisovanie slov z existujúcich databáz, či už v elektronickej alebo inej forme, ktoré sú chránené proti neautorizovanému šíreniu.

Používajte iba slovenské slová. Cudzie slová používajte iba v prípade, ak sú bežne používané. Nemá zmysel vytvoriť synonymá zo slov, ktoré nie je možné použiť v bežnom texte.
Nevkladajte skratky.

Pred vložením dát skontrolujte pravopis. Slovník momentálne nemá kontrolu pravopisu pri vkladaní údajov do databázy.

Pokiaľ to je možné, tak nepoužívajte viacslovné spojenia. Nemá zmysel vkladať výraz *rozpočet na reklamu*, ak už slovník obsahuje synonymá pre slovo *rozpočet.*

Nevkladajte do databázy názvy miest (dedín, riek...), firiem :-) a podobne.

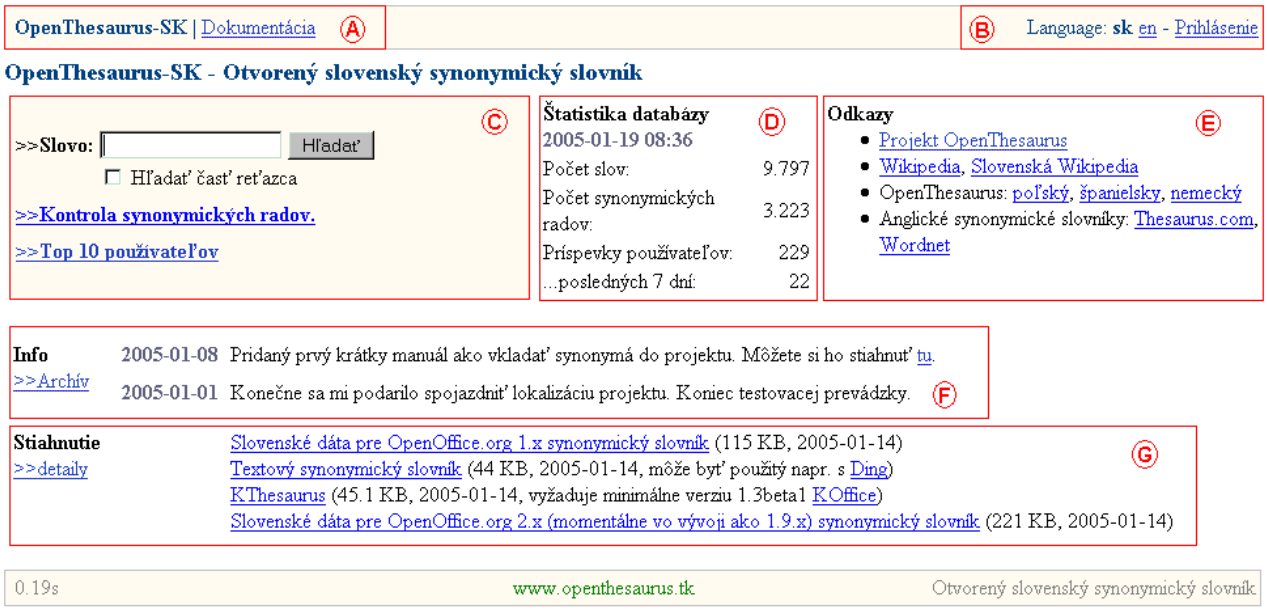
Čo to znamená vkladať slová „v základnom tvare“?

Do databázy majú byť vkladané slová v základnom tvare, tzn. slovesá v neurčitku, podstatné mená v jednotnom čísle a v prvom páde, prídavné mená v prvom stupni:

správne: *bežať*, nesprávne: *bežal*
správne: *dom*, nesprávne: *domy*
správne: *dlhý*, nesprávne: *dlhší*

Úvodná stránka

Úvodná stránka projektu je vytvorená tak, aby zobrazila údaje, ktoré používateľ najviac využíva:



Obrázok 1: Úvodná stránka

Dokumentácia (obr.1A)- prístup na stránku, kde sú stručne zodpovedané niektoré otázky. Na túto stránku sa môžete dostať aj tak, že kliknete na malý otáznik na stránke a zobrazí sa príslušná časť stránky s vysvetlením (obr.3C,F).

Language sk en (obr.1B) - prepnutie jazyka sk/en (slovensky/anglicky) webové rozhranie (interface) bude vo zvolenom jazyku.

Prihlásenie (obr.1B) - musíte byť registrovaný používateľ, ak chcete meniť dáta v slovníku. Ak chcete v slovníku len vyhľadávať, nepotrebujete sa registrovať, ale nebudete môcť ani pridávať, či opravovať slovník. Po tom, ako sa zaregistrujete vám bude zaslaný kód na e-mailovú adresu, ktorú ste zadali pri registrácii.

Vyhľadanie synonyma umožňuje pole „>>*Slovo“,* kde napíšete slovo ku ktorému hľadáte synonymum a kliknete na *Enter* alebo tlačítko *Hľadať*. Ak slovo nie je nádejné, je vám ponúknuté, aby ste toto slovo vložili do slovníka „*Pridať 'slovo' a synonymá do synonymického slovníka*“.

Taktiež sa slovník pokúsi nájsť podobné slová, ktoré by mohli vám mohli pomôcť. Ak chcete, aby slovník zobrazil podobné slová, aj keby vaše slovo bolo v slovníku, zaškrtnite políčko „*Hľadať časť reťazca*“ Potom slovník nebude hľadať len striktne napísané slovo, ale aj slová jemu podobné (s príponou, predponou)(obr.1C), respektíve hľadané slovo bude brané ako podreťazec. Napríklad ak dáte vyhľadať slovo „*sám*“ a zaškrtnete políčko „*Hľadať časť reťazca*“, ponúkne vám okrem slova „*sám*“ aj tieto ďalšie nájdené slová:

*o**sam**elosť, o**sam**elý, os**am**otenosť, o**sam**otený, **sam**a, **sam**á ruka **sam**á noha, **sam**o sebou, **sam**očinný, **sam**o lepiaca páska, **sam**ostatne, **sam**ostatný, **sam**ota, **sam**otársky, **sam**otný, **sam**ovláda, **sam**ovoľný, **sam**ozrejme, **sam**ozrejmosť*

Na hociktoré slovo môžete kliknúť a pozrieť si k nemu synonymá.

Kontrola synonymických radov - Niektoré slová majú viacero významov a preto sa môžu nachádzať vo viacerých synonymických radoch (1 synonymický rad=1 význam), pričom jednotlivé slová z rôznych radov, nemusia byt navzájom synonymami.

Synonymický slovník v projekte OpenOffice.org 1.1.4 umožňuje definovať len jeden synonymický rad pre jedno slovo. V OpenOffice.org 2.x toto obmedzenie už nie je. Na toto obmedzenie bol braný ohľad aj pri vytváraní vstupných dát. To malo za následok, že jednom riadku (synonymickom rade) sa nachádzajú slová s rôznym významom (t.j. nie všetky slová v takýchto synonymických radoch sú si vzájomne synonymami). Postupne budú tieto synonymické rady identifikované a rozdeľované do samostatných radov podľa významu slov. Aj vy sa môžete zapojiť do ich odhaľovania.(obr.1C)

Štatistika databázy – zobrazuje aktuálny stav databázy.(obr.1D) Viac o štatistike nájdete v časti Štatistika.

Odkazy – tu sú linky na ďalšie slovníky v slovenčine a iných jazykoch.(obr.1E)

Info – zobrazuje najčerstvejšie správy o zmenách v projekte. Viac správ sa zobrazí po kliknutí na odkaz *Archív*.(obr.1F)

Stiahnutie (obr.1F) – obsahuje odkazy na súbory ktoré si môžete stiahnuť do externých aplikácií. Link „*detaily*“ zobrazí podrobnejšie informácie o tom ako tieto súbory vznikajú (Pozri časť: Export dát).

Ako robiť zmeny v slovníku

Pole s hľadaním synonyma sa nachádza hneď na úvodnej stránke(obr.1C) a aj na každej stránke projektu v záhlaví (obr.3A).

Skôr, ako by ste začali vkladať synonymá do databázy, musíte sa presvedčiť, či už dané slovo nie je vložené resp. v akom význame je vložené. Napríklad slovo „*letieť*“ mi zobrazilo tieto výsledky (obr.2):

Obrázok 2: Výsledok hľadania slova "letieť"

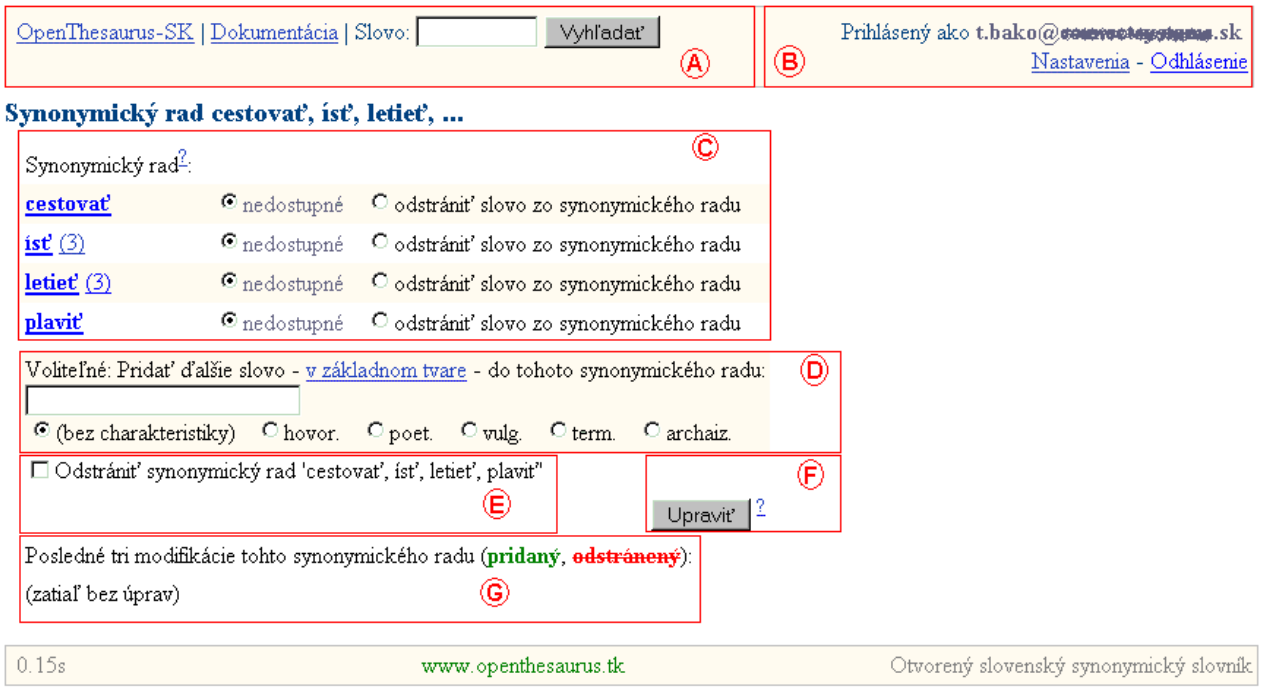
Niekedy sa vám stane, že hľadané slovo už je zaradené do synonymického radu, ale v inom význame, s akým vy chcete pracovať. V tomto prípade môžete pridať ďalší význam slova a tým pádom vytvoriť nový synonymický rad (obr.2B). Pri vytváraní sa vám otvorí stránka „*Pridať ďalší synonymický rad*“ na ktorej môžete synonymickému radu priradiť kategóriu na ktorú bude odkazovať:

Anatómia, Astronómia, Automobilizmus, Biochémia, Biológia, Botanika, Chémia, Dejepis, Ekonómia, Elektro, Fyzika, Gastronómia, Geológia, Hudba, Matematika, Medicína, Náboženstvo, Námorníctvo, Politika, Právo, Technika, Vojsko, Výpočtová technika, Zoológia, Šport

Pokiaľ vám chýba nejaká kategória, napíšte administrátorovi slovníka.

O hľadanom slove môžete zistiť viac na internete alebo Wikipédii (obr.2C).

Potom, keď si kliknutím vyberiete jeden z troch nájdených synonymických radov, zobrazí sa vám vybraný synonymický rad a možnosti manipulácie s ním (obr.3).



Obrázok 3: Synonymický rad cestovať, ísť, letieť, plaviť, ...

V prípade, že hľadané slovo má viac významov (vyskytuje sa vo viacerých synonymických radoch - významoch), tak sa za ním zobrazí číslo, ktoré vyjadruje počet významov(obr.3C – slová *ísť* a *letieť*).

Kliknutím na slovo sa zobrazí stránka „*Detaily pre slovo*...“, kde môžete slovu priradiť charakteristiku t.j. či patrí medzi hovorové slová, poetické, vulgarizmy, termíny alebo archaizmy.

Kliknutím na číslo za ním (3) sa vám zobrazia všetky významy (synonymické rady) daného slova.(obr.2) Takže ak kliknete na číslo (3) za slovom „*letieť*“ vrátite sa vlastne späť na zoznam synonymických radov v ktorých sa toto slovo vyskytuje. Ak kliknete na číslo za slovom „*ísť*“, zobrazia sa vám synonymické rady, v ktorých sa toto slovo nachádza:

*dariť sa, excelovať, **ísť**, mať sa, mať úspech, vynikať, žiariť*
*chodiť, **ísť**, kráčať, prechádzať*
*cestovať, **ísť**, letieť, plaviť*

Hociktoré slovo môže byť odstránené zo synonymického radu slov (obr.3C). Označíte toto slovo kliknutím na položku „*odstrániť slovo zo synonymického radu*“ a potvrďte zmenu kliknutím na tlačítko „*Upraviť*“(obr.3F).

Ak chcem slovo do tohto radu pridať, napíšem ho *v základnom tvare* (pozri: Čo to znamená vkladať slová „*v základnom tvare*“?) do políčka „*Pridať ďalšie slovo*“(obr.3D).

Ku pridanému slovu môže byť pripojená charakteristika hneď (hovorové slovo, vulgarizmus, atď.), čiže nemusíte potom dodatočne na slovo v synonymickom rade klikať, aby ste mu mohli priradiť charakteristiku (obr.3D).

Celý synonymický rad slov môžete odstrániť zaškrtnutím políčka „*odstrániť synonymický rad*“(obr.3F) a potvrďte zmenu kliknutím na tlačítko „*Upraviť*“(obr.3F). Slovník sa vás spýta na dôvod, prečo chcete celý rad vymazať.

V dolnej časti stránky sa zobrazuje zoznam posledných zmien používateľov tejto synonymického radu - pridané a odstránené slová (obr.3G).

Štatistika

Slovník zaznamenáva všetky aktivity ktoré sa zobrazujú v štatistikách. Štatistika databázy umiestnená na úvodnej stránke (obr.1D) zobrazuje aktuálny stav slovníka.

„Top 10 používateľov“ je odkaz na hlavnej stránke, ktorý zobrazuje aktivitu za posledných 7 dní.

Osobná štatistika zobrazuje počet pridaných a odstránených slov. Je na stránke nastavenia používateľa po kliknutí na „Nastavenia“(obr.3B).Tu si môžete tiež zmeniť heslo a zobrazované meno v štatistikách.

Pri zvolení synonymického radu sú v dolnej časti stránky zobrazené posledné zmeny – pridané, odstránené slová s polu s dátumom a časom zmeny (obr.3G).

Na stránkach administrátora sú ďalšie štatistiky (pozri Stránky administrátora).

Toto je štatistika slovníka pri tvorbe tohoto dokumentu:

Počet rôznych slov: 9 838.
Počet synonymických spojení: 3 238.

ĎALŠIE INFORMÁCIE

Stránky administrátora

Všetky zmeny používateľov v databáze sú zaznamenávané, takže administrátor má plnú kontrolu nad slovníkom. Úlohou administrátora je kontrolovať správnosť dát v slovníku a kontrolovať aj zmeny používateľov, ktorí databázu upravovali. Má prístup ku špeciálnej stránke, ktorá mu v tejto úlohe výrazne pomáha.

Tabuľky v strede stránky zobrazujúce:
- 5 posledných registrovaných používateľov a možnosť zobrazenia všetkých používateľov (obr.4A)
- 8 posledných hľadaných slov a možnosť zobrazenia ďalších (obr.4B)

Akcie v ľavom stĺpci:
Linky ktoré okamžite spúšťajú okamžité vytvorenie súborov pre export (OpenOffice.org thesaurus, Kword, text,...)(obr.4D)

Linky zobrazujúce štatistiky a pomôcky pre správu databázy: (obr4C,E)

Calculate average size of synsets – kalkulácia štatistiky:
synonymických radov: 3 238
slov v synonymických radoch: 10 390
slov/synonymických radov: 3,2087708462

Random synsets – synonymické rady, ktoré boli náhodne zobrazené

Import new words from text – obsahuje pole, kde môže administrátor hromadne nakopírovať/importovať slová

Unknown words from search log - 12, 20, 50, 100 hľadaných slov, ktoré používatelia hľadali, ale neboli v slovníku nájdené

Most active users – 5, 20, 50, 100 najaktívnejších používateľov

Top 20 searches – dvadsať najčastejšie hľadaných slov

Free text search – voľné vyhľadanie textu, zadané pomocou SQL syntaxe

Duplicates - funkcia ktorá umožní vyhľadať duplikáty synonymických radov. Napríklad, keď sa viac slov v dvoch radoch opakuje:

defraudovať, oklamať, ošudiť, podviesť, spreneveriť
oklamať, ošudiť, podvádzať, podviesť, spreneveriť

Uses – zoznam slov, pre ktoré bola zadefinovaná charakteristika

Subjects – zoznam tém, ku ktorým je možné každé slovo priradiť. Napr. *Fyzika, Medicína, Zoológia, Astronómia*

Phrases – úslovia, slovné zvraty napr: *bez ladu a skladu, samá ruka samá noha, dať sa do poriadku*

Ellipsis - výrazy obsahujúce tri bodky „...“

Large synsets - synonymické rady tvorené 5 a viac slovami

Small synsets - synonymické rady tvorené len jedným slovom. Môže sa stať že používateľ vymaže zo spojenia napr. *cval – útek*, slovo *útek*, keďže toto spojenie nie je synonymom, ale slovo *cval* zabudne vymazať a zostáva v databáze osamotené

Prefix/suffix - slová so zátvorkami v ktorých sú vyjadrené možné predpony, prípony slova

Short forms - skratky, skrátené slová končiace bodkou

Multi occurences – slová ktoré sa vyskytujú v troch a viac synonymických radoch:

menštruácia,perióda

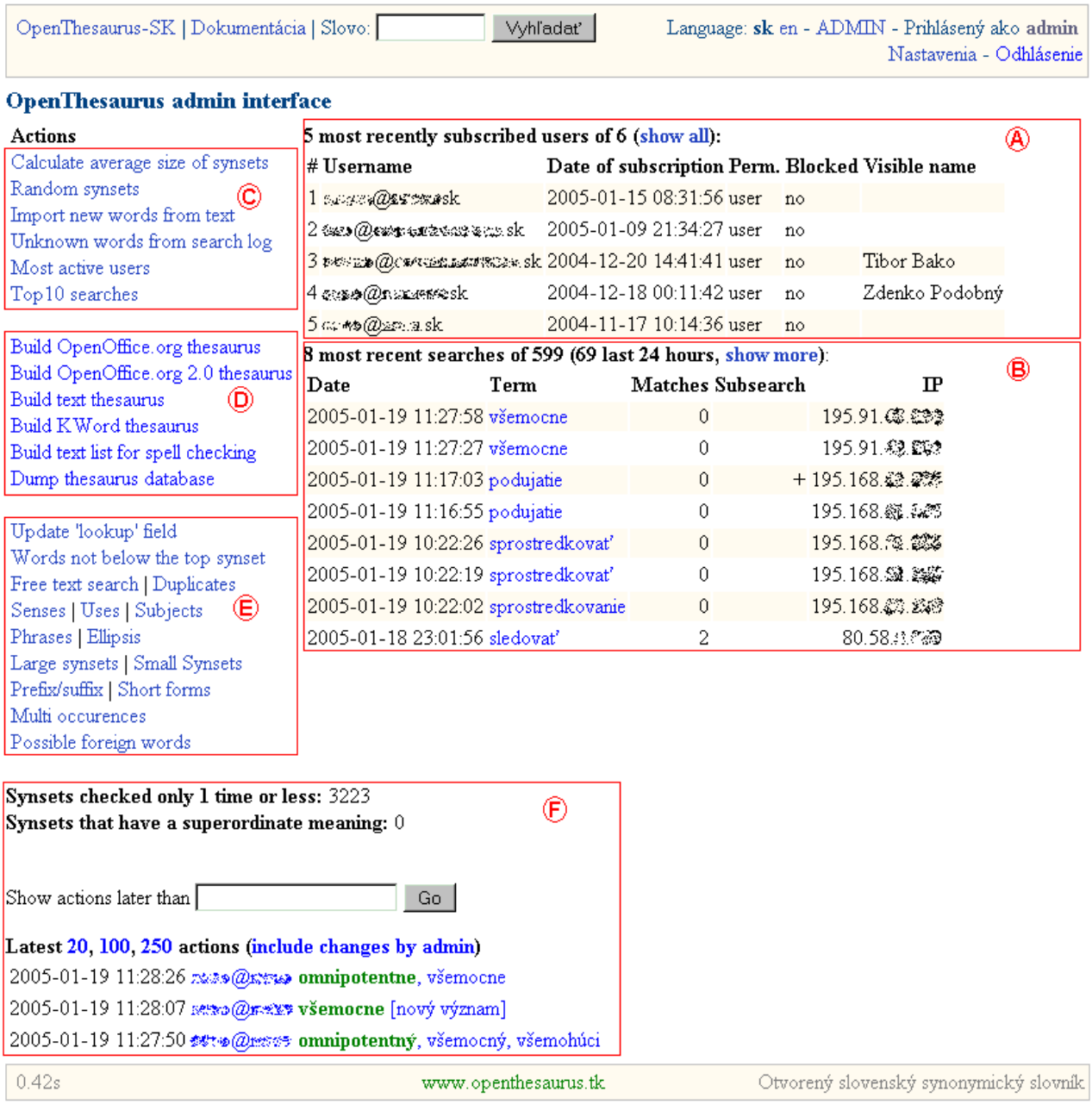
kmit,perióda

obdobie,perióda,fáza,štádium,stupeň

Possible foreign words – zobrazí slová u ktorých predpokladá že sú cudzie. Zisťuje to podľa predpony alebo prípony slova. Napr: končiace -us: *alkoholizmus, nacizmus, egoizmus*

V dolnej časti: (obr.4F)

Počet synonymických radov, ktoré boli skontrolované len raz alebo vôbec.
Počet synonymických radov s nadradeným významom.
Zoznam posledných zmien v slovníku (pridané a vymazané slová, kedy a kým) a ich možnosť zobrazenia od zadaného dátumu.



Obrázok 4: Stránka administrátora

Štruktúra databázy

Web stránka OpenThesaurus je vytvorená pomocou skriptovacieho jazyka PHP (www.php.net) a MySQL (www.mysql.com), jednou z najznámejších relačných databáz. Oboje, skriptovací jazyk aj databáza, sú zdarma. Najpodstatnejšie tabuľky sú:

words – zoznam všetkých slov (každé slovo tu môže byť len raz)
meanings – zoznam všetkých významov
word_meanings – relácia (prepojenie) *n:m* medzi tabuľkami *words* a *meanings.*

To znamená, že zoznam významov môže obsahovať viacej ako jedno slovo a slovo sa môže objaviť vo viac ako jednom význame.

Export dát

Dáta je možné šíriť v zmysle tejto licencie.

Tabuľky, ktoré obsahujú aktuálne slová obsiahnuté v slovníku, významy a ich vzťahy (prepojenia) sa automaticky exportujú každý piatok pred polnocou pomocou príkazu MySQL – *mysqldump*.

Súbory sa potom skomprimujú a je možné ich stiahnuť z webstránky slovníka. Používatelia ich následne môžu vložiť do vlastnej (lokálnej) MySQL databázy a používať vlastné databázové požiadavky (query), ktoré im exportujú dáta. Avšak upravovať stiahnuté dáta lokálne nie je celkom dobrý nápad, nakoľko je potom dosť obtiažne tieto upravené dáta vložiť späť do databázy OpenThezaurus.

Tabuľky ktoré obsahujú detailné informácie o používateľových zmenách v databáze (napr. ktorý používateľ robil aké zmeny a kedy) sa neexportujú vzhľadom k zachovaniu anonymity používateľov.

Na PHP stránke prístupnej administrátorovi sa vytvára zoznam slov a ďalší zoznam obsahuje synonymá ku každému slovu. Oba zoznamy sa zapíšu do dvoch súborov a následne sú použité ako vstupné dáta pre skript *Parse_Thes.awk* Pavla Janíka (to je jeden zo šikovných ľudí podieľajúcich sa na projekte OpenOffice.org). Tento skript vytvára súbory *th_sk_SK.dat* a *th_sk_SK.idx* ktoré používa OpenOffice.org ako zdroj svojho tezaura. Podrobnejšie informácie na inštaláciu nájdete v exportovanom ZIP súbore README.

Slovenské dáta pre synonymický slovník programu OpenOffice.org 1.x. Poznámka: Tieto súbory správne fungujú iba v slovenskej verzii programu OpenOffice 1.1.4 a vyššej. Nižšie verzie nedokážu správne pracovať s niektorými písmenami (nesprávne zobrazujú diakritiku), a preto veľa synoným nie je nájdených, prípadne sú zle zobrazené. Inštalácia je popísaná v README súbore.

Slovenské dáta pre synonymický slovník programu OpenOffice.org 2.x
Táto verzia OpenOffice.org je momentálne (12.2.2005) vo vývoji.

Synonymický slovník pre KWord min. verzia 1.3beta1 Inštalácia: rozbaľte súbor do priečinku $KDEDIR/share/apps/thesaurus/ a potom v KThesaurus vyberte *zmeňte jazyk*.

Synonymický slovník v textovom formáte komprimovaný, použiteľný napr. s Ding.

Pre vývojárov: MySQL-dump.

Ako vytvoriť inú jazykovú verziu slovníka?

Tu je zopár rád, ktorými by ste sa mali riadiť, ak chcete vytvoriť synonymický slovník v inom jazyku:

PHP súbory sú prístupné cez CVS - OpenThesaurus Project (http://sourceforge.net/projects/openthesaurus). Tie potom nainštalujte do webroot adresára na serveri a vytvoríte štruktúru databázy. Používateľské prostredie (interface) by malo byť preložené pomocou technológie GNU i18n (gettext). Podrobné inštrukcie nájdete v README súbore. Bude potrebné mať isté skúsenosti s PHP a MySQL na nastavenie servera.

Počiatočné dáta musia byť pod licenciou Open Source. Nový slovník by mal na počiatku obsahovať prinajmenšom 5000 slov. Je dôležité investovať dostatok času nájsť rozsiahly zdroj počiatočných dát, napríklad prehľadať web stránky univerzitných projektov alebo projektov Open Source s podobným cieľom. Začínať s menším počtom by mohlo odradiť používateľov tým, že veľa bežných slov by neobsahoval. Dobrým začiatkom by mohol byť viacjazyčný slovník, ale ak nemáte nič iné, pomôže aj obyčajný zoznam slov.

Dáta by mali mať štruktúru podobnú štruktúre súborov v OpenThesaure. Príklad ako by mohli takéto dáta vyzerať:

výplata,gáža,mzda,plat,príjem
beznádej,zúfalstvo
ošúchaný,obnosený,použitý

Je dôležité, aby jazyk, v ktorom je vytváraný projekt bol rodným jazykom administrátora. Jeho úlohou je totiž kontrolovať správnosť dát v slovníku a kontrolovať aj zmeny používateľov, ktorí databázu upravovali.

ZÁVER

Projekt OpenThesaurus v iných jazykoch a ďalšie odkazy

Nemecký Tezaurus http://www.openthesaurus.de/
Poľský Tezaurus http://synonimy.sourceforge.net/
Španielsky Tezaurus http://openoffice-es.sourceforge.net/thesaurus/
Anglický Tezaurus http://thesaurus.reference.com/,
http://www.cogsci.princeton.edu/%7Ewn/
Wikipedia http://www.wikipedia.org/
Slovenská Wikipedia http://sk.wikipedia.org/wiki/Main_Page

História

2005-01-21 - vytvorená tretia, zatiaľ posledná verzia. Začínam chápať čo to znamená vytvárať dokumentáciu (a prečo túto činnosť programátori z duše nenávidia :))
2005-01-17 - ďalšia, zdokonalená verzia s narýchlo urobenými obrázkami
2005-01-11 - vytvoril som prvú verziu tohoto návodu
2005-01-08 - existencia Slovenského OpenThesaura bola oznámená v konferenciách OpenOffice.org Announce a Users
2005-01-07 - Zdenko vytvoril nultú verziu návodu aby mohol projekt publikovať
2004-12-20 - testovacia prevádzka stránky www.openthesaurus.tk
2004-11-19 - úprava vstupných dát pre staršie verzie OpenOffice.org
2004-11-03 - riešenie problému s nesprávnym zobrazovaním diakritiky v OpenOffice.org
2004-11-02 - dokončená kontrola duplicít v slovníku – finálna verzia vstupných dát
2004-10-15 - prvé pokusy implementovať vytvorený slovník do OpenOffice.org
2004-10-12 - prvá verzia 100 % prekladu slov, ešte neskontrolovaná
2004-10-04 - je preložených asi 37 % slov
2004-10-01 - začal som pracovať na preklade a úprave českej verzie vstupných dát, ktorú mi poskytli na konferencii
2004-09-27 - člen konferencie users@sk.openoffice.org „MeX“ sa pýta, či existuje slovenský synonymický slovník pre OpenOffice.org, žiaľ odpoveď znie „nie“.

Tvorcovia slovenského projektu

- Daniel Naber - vytvoril pôvodný projekt OpenThesaurus ktorého engine používame.
- Laboratórium spracovania prirodzeného jazyka Fakulty informatiky Masakykovej univerzity v Brne vytvorilo český tezaurus (http://nlp.fi.muni.cz/projekty/czthes/), z ktorého sme vychádzali (konkrétne sú to Pavel Rychlý a Pavel Smrž).
- Administrácia a rozbehnutie web stránky a pomoc s manuálom - Zdenko Podobný
- Juraj Bednár - poskytnutie priestoru na serveri a pomoc pri spustení projektu.
- Tibor Bako - preklad, vytvorenie vstupných dát (t.j. takmer všetkých dát, ktoré sú momentálne dostupné) a tvorba tohto manuálu.
- ...a všetci registrovaní používatelia, ktorí prispeli do projektu.

Aktualizované 12. február 2005

Príloha 1: Licencia

Slovenské dáta sú vytvorené na základe prekladu a následnej úpravy českých dát vytvorených Pavlom Rychlým a Pavlom Smržom z Laboratória spracovania prirodzeného jazyka Fakulty informatiky Masakykovej univerzity v Brne. Z tohto dôvodu majú slovenská dáta rovnakú licenciu ako české.

Slovak thesaurus database licence