ÚSTAV INFORMAČNÍCH STUDIÍ A KNIHOVNICTVÍ
FF UK V PRAZE

Vladimír Smetáček

Systém SEMAN

Rozpracovaná verze

Praha

Listopad 2007

UČEBNÍ TEXT K SYSTÉMU SEMAN

Vladimír Smetáček

(Stav k 10.11.2007)

Tento text je koncipován jako programovaný. V samotném textu jsou uvedeny pouze základní informace. Příklady, úlohy a kontroly úloh jsou dostupné po kliknutí na názvy doplňujících souborů. Ty jsou v textu vždy reprezentovány třímístným číslem v kulatých závorkách a v elektronické verzi jsou uvedeny v samostatné části, v tištěné verzi vůbec uváděny nejsou.
V samostatné části elektronické verze jsou uvedeny také názvy a obsahy činnosti programů. Samotné programy jsou přístupné pouze u pracovníků systému Seman, popisy však postačí k jejich pochopení. Popisy programů jsou dostupné kliknutím na číselné označení programu, které se nachází na vhodném místě tohoto textu. Popisy programů jsou shromážděny na konci textu. I když s nimi v této době nelze pracovat, k uvědomění si jejich funkce popisy postačí. Některé jsou tak jednoduché, že i nepříliš zkušený programátor si může podobné vytvořit, jiné jsou součástí jiných systému a textových editorů a student si je z těchto systémů a editorů může podle vlastního uvážení převzít.
Ukázky ze skutečných souborů Semanu a skutečných řešení zajímavých pracovních úkolů jsou někdy uvedeny v počítačové verzi. Pro tištěnou verzi nejsou vůbec vhodné, neboť by jimi příliš nabobtnala.
Text ze studenta neudělá odborníka na systém Seman, ale dostatečně ho s ním seznámí. Základním cílem je naučit studenta uvažovat způsobem, který spojuje lidské myšlení s pracovními postupy výpočetní techniky. Stroj není schopen uvažovat jako člověk, člověk však je schopen se naučit uvažovat v pojmech a postupech, které jsou vhodné pro práci výpočetní techniky při řešení těchto složitých problémů. I ti studenti, kteří nebudou pracovat se systémem Seman, se mohou díky zvládnutí tohoto textu ve značné míře naučit chápat vztahy mezi lidským myšlením a prací výpočetní techniky s plným vědomím toho, že výpočetní technika je nástroj člověka, nikoli že člověk je nástrojem výpočetní techniky.
Tento text neumožňuje přístup k základním souborům a k pracovním programům, příklady souborů a řešení však činnosti systému dostatečně odpovídají. Každý uživatel si může pro jakýkoli postup vytvořit vlastní program a ten použít. K základním souborům a programům se ovšem může dostat až po dohodě s tvůrci a správci systému.
Programy, s nimiž pracuje Seman, jsou uloženy stejně jako základní rozsáhlé soubory na chráněných počítačích. Ve vlastním textu nejsou uváděna označení programů, neboť student s programy nebude pracovat, ale má-li o ně zájem, může se na jejich popisy podívat v oddíle 1.11.1.
Text je vytvořen jako počítačový. Text neobsahuje příklady a doplňky stejným způsobem jako tištěný text, ale jako vyvolavatelné doplňky, které jsou uloženy

mimo vlastní text a označeny nezaměnitelnými číselnými kódy, které obsahují třímístné číslo v závorce. V počítačovém textu stačí klepnout na třímístné číslo v závorkách, a soubor se objeví na obrazovce na místě čísla. Po skončení práce s ním postačí klepnout na klávesu escape a soubor zmizí. Postupně bude tento text, především ve své přílohové, vyvolatelné části stále více doplňován příklady souborů, tvořících vstupy a výstupy skutečného nasazení programů a souborů Semanu. To samozřejmě znamená, že se rozsah tohoto počítačového souboru bude stále zvětšovat.

1 Úvod

1.1 Lidé a je obklopující skutečnost

Lidé se musí trvale vyrovnávat se svým prostředím. Toto vyrovnávání je stále více spojeno s praktickým uplatňováním poznatků. Naprostá většina poznatků, které je člověk schopen vytvořit, pochopit a předat jiným lidem, má formu jazykových sdělení. Vývoj prostředků výpočetní techniky zvyšuje naše možnosti při práci s jazykovými sděleními, grafickými zobrazeními i zvukovými soubory. Problém je ovšem v tom, že zpracování jazykových sdělení, grafických zobrazení i zvukových souborů je spojeno s ryze lidskými intelektuálními a citovými pochody a ty pravděpodobně nebude možno ani v budoucnosti plně implementovat do funkcí výpočetní techniky. V oblasti jazykových sdělení jde především o sémantické a překladové aspekty jazyka. Jazyk vhodný pro práci na prostředcích výpočetní techniky musí proto v dostatečné míře řešit sémantické aspekty lidmi používaných jazyků a to, že lidé používají mnoho různých přirozených i umělých jazyků a že rozdílnost jazyků je jednou z příčin vzájemného nepochopení v lidské společnosti.

1.2 Lidé a současný stupeň vývoje lidstva

Uvažovat o současném stupni vývoje lidstva je ošemetné. Tvůrci teorií mají často pocit, že jejich názory a navrhovaná řešení jsou součástí vrcholného stádia vývoje lidstva a přitom se většinou mýlí. Zatím ve všech obecně známých případech, kdy jednotlivec, skupina lidí, politicko-ekonomická ideologie nebo náboženství věřili, že dospěli k vrcholnému bodu, došlo ke zklamání. Nejvýznačnější světová náboženství, politické doktríny, vědecké a filosofické teorie se buď ukáží jako neúspěšné nebo, i když jsou zpočátku úspěšné, jsou postupně překonány dalším vývojem.
Etapy vývoje lidstva je možné definovat podle různých kritérií. Jedním z nich je cíl ovládnutí světa, v němž lidská společnost žije, k zajištění co nejlepšího života člověka a lidské společnosti. Žádná etapa a její způsob ovládání světa však není zcela samostatný a jednoznačně oddělen od předchozích a následujících.
V současné etapě lidstvo ví stále více o světě i o sobě, informace se staly jedním ze základních pojmů a to má řadu velmi kladných i dosti záporných důsledků.

Aby lidstvo dokázalo zvládnout složitost této existence, musí mimo jiné v dostatečné míře zvládnout své vlastní znalosti.

1.3 Lidstvo a vývoj jeho poznání

Neznáme bytost, která by byla schopna poznávat ji obklopující skutečnost i sebe sama tak dokonale, jako to dokáže člověk. Možné námitky věřících v boha, v různé spirituální pohledy, v dosud vědecky dostatečně neprozkoumané inteligence a podobně, nelze nikdy plně vyvrátit. Ani lidé v ně věřící jejich existenci nedokáží dostatečně prokázat. Pro člověka tedy je v současné době možno přijmout myšlenku, že vše co o světě i o sobě ví, je výsledkem lidského poznání. Ukáže-li se tato myšlenka jednou příliš úzkou, bude jistě změněna. Problém lidského poznání spočívá především ve dvou skupinách skutečností:

- Lidé jsou různě inteligentní, mají různé vzdělání, žijí v jiném prostředí. Proto jejich schopnost poznání se může podstatně odlišovat. Pro jednoho člověka může být pojem „černá díra“ zcela pochopitelný, pro jiného nikoli. Rozdíly ve vědomostech a inteligenci mezi lidmi se mohou trochu zmenšovat, ale úplně asi nezmizí nikdy.
- Člověk velmi často chápe poznávanou skutečnost s ohledem na své potřeby, a aniž si je toho někdy vědom, vnáší do ní rysy, které tam nejsou.

Vytváření systémů, které pomohou člověku poznávat a přitom se pokud možno vyhnout oněm omezením, by proto mohlo lidem pomoci překonat v budoucnosti řadu potíží. Je nepochybné, že to nebude lehké.

1.4. Jazyk a poznání

Protože téměř vše, co člověk ví o světě, v němž žijeme, je vyjadřováno jazykem (matematické a logické symboly jsou jen speciální druh jazyka), už mnoho staletí se odborníci snaží vytvořit jazyk, který by umožňoval zachytit co nejjednodušším a současně i nejúplnějším způsobem všechny poznatky o světě kolem nás. Dosud se takový jazyk nikomu nepodařilo vytvořit, a ani my nebudeme očekávat, že Seman tyto podmínky splňuje zcela, ale může to být další výrazný krok k budoucímu univerzálnímu jazyku.

1.4.0. Historické návaznosti

Již mnoho set let se filozofové, jazykovědci a logici pokoušejí vytvořit jednotný univerzální jazyk, který by sloužil k jednotnému popisu všech jevů světa a umožnil by jednotnou komunikaci všech lidí. Vytvořit takový jazyk se dosud nikomu nepodařilo, mnoho pokusů však bylo velice inspirativních. Seman využívá všech dosažitelných historických zkušeností.

1.4.1. Komenský a pansofie

Komenský ve snaze umožnit směřování všech lidí k pochopení božích záměrů a dosažení jejich určité vzdělanostní úrovně, kromě pedagogických postupů také promýšlel jednotný jazyk. Ten dostatečně nerozpracoval, ale v českém národě mu patří neopominutelné místo v těchto snahách. Kdyby nedošlo k požáru Lešna a životní osudy nezmítaly Komenským z místa na místo, vnímali bychom ho

pravděpodobně nejen jako velkého pedagoga, ale také jako velkého encyklopedistu a velkého tvůrce předpokladů k univerzálnímu jazyku.

1.4.2. Leibnitz a logika jazyka

Leibnitz kromě práce na matematických objevech stále pracoval s jazykem a dospěl k představě, že všechny pojmy lze chápat jako jednoduchá sémantická vyjádření s případnou matematickou hodnotou. V moderní době je znám jako tvůrce jednoho z nejzajímavějších početních strojků, jehož jméno se vyskytuje v názvech několika programovacích principů a jazyků i názvech databází. Kdyby v jeho době již existovala výpočetní technika umožňující práci s většími objemy dat, asi by patřil k lidem, kteří by hledali dosti úspěšně cesty k vytvoření strojově zpracovatelného univerzálního jazyka. Vynásobením nebo jiným matematickým usouvztažněním číselných hodnot přisouzených jednoduchým, základním slovům lze podle něho dospět k hodnotám odpovídajím významu složitějších pojmů. Např. bude-li hodnota pojmu býti ženského rodu 2, hodnota pojmu býti mužského rodu 3, hodnota pojmu býti člověkem 4 a hodnota pojmu býti hovězím dobytkem 5, pak hodnota pojmu žena bude 8, hodnota pojmu muž 12, hodnota pojmu býk bude 15, hodnota pojmu kráva 10. Tento projekt se nikdy nepodařilo uskutečnit, mimo jiné proto, že jazyk je pro prostý převod na číselného vyjádření příliš složitý, ale snaha chápat pojmy jako kombinovatelné veličiny se i v pozdějších dobách stále znovu v jiných podobách vracela.

1.4.3. Carnap a budování vědeckého jazyka

Carnap a tzv. Vídeňský kruh vycházeli z nepříjemné situace vzájemného vzdalování vědců v chápání jazyka vědy a tím i chápání vědecké práce a jejích výsledků vědci z jiných oborů. Sjednocení jazyka vědy a pravidel práce s ním mělo vést ke zlepšení vědecké činnosti. Vznik specializovaných časopisů snažících se o jednotnou terminologii vědy (např. Scientific American) je spojen (alespoň volně) s těmito snahami. Prudké vzdalování se věd je v rovině jazyka uvědomovaným problémem a především producenti některých vědeckých časopisů přispívají k jeho zmírňování

1.4.4. Jazyky typu lingua franca

V každém historickém období některá národní území a jazyky hrály klíčovou úlohu, proto příslušníci jiných národů museli tohoto jazyka používat, aby se dostali nejen k poznatkům, ale také si usnadnili cestu k úspěchu. Pojem lingua franca koresponduje s dobou významu franské říše. Linguou francou byly v Evropě postupně latina, francouzština, dnes je to angličtina. Neznalost v té které době tohoto jazyka s velkou pravděpodobností snižovala úspěšnost odborníků a veřejných činitelů.

1.4.6. Prekoordinované selekční jazyky

Mnohé informační prekoordinované jazyky se pokoušejí vidět skutečnost jako hierarchicky utříděnou strukturu, v níž je každý obsahově specifičtější pojem podřazen pojmu obsahově obecnějšímu. Tyto vazby nutně tříští mnohostranně se prostupující pojmy a způsobují jistou rigiditu prekoordinovaných systémů. Hierarchické vazby mezi pojmy a snaha popsat svět jako jednotnou strukturu vyjadřovanou jazykem však dělá z prekoordinovaných jazyků významného předchůdce a inspirátora a pravděpodobně ještě na dlouhou dobu souputníka každého budoucího univerzálního jazyka.

1.4.7. Postkoordinované selekční jazyky

Všechny kvalitní tezaury ve svých deskriptorových odstavcích zachovávají vztahy: desktripor – nedeskriptor, nadřazený desktiptor – podřazený deskriptor, deskriptor – asociovaný deskriptor. Mnohé tezaury navíc přiřazují k některým těmto vztahům různé role a váhy a tím se přibližují sémantickým řešením univerzálního sémantického jazyka.

1.4.8. Unitermy

Unitermové systémy pracují s jednoznačně formulovanými pojmy, a všechny ostatní pojmy zakazují. To vedlo často k nedokonalému popisu a vyhledávání. Jasně však prokázaly, že pouze jazyk, který svým výrazům dává jednoznačný význam, se může stát nástrojem exaktního popisu a zkoumání čehokoli, co se ve světě, v němž žijeme, děje.

1.4.9. Sémantické primitivy

Sémantický primitiv je pojem, slovo, výraz, který už nelze dále rozložit na kombinaci obsahově specifičtějších jednotek. Sémantický primitiv je vlastně přímým předchůdcem termínu sem (viz dále). Problém sémantické primitivu spočívá v nejistotě, zda jsme ho již dosáhli a zda jakýkoli výraz nelze převést na kombinaci ještě specifičtějších sémantických primitivů. Sém sice směřuje k tomu být sémantickým primitivem, ale tohoto stavu nedosahuje vždy a to především ze tří důvodů:

0. nemůžeme si být jisti, že výraz nelze rozložit na kombinaci jiných sémanticky obecnějších výrazů
1. lidské myšlení je navyklé na některé pojmy a významy (být mužem, být ženou, být velký, být červený). Mohlo by být velmi obtížné zvyknout si na pojmy a významy hlubší. I v případě dnes formulovaných sémů už to může být pro řadu lidí dosti obtížné.
2. čím větší je množina sémů na úrovni kombinace sémantických primitivů, tím sice přesnější, ale také výpočetně složitější může být práce se semy. Teorie mlhavých množin a podobné teorie ukazují, jaké problémy mohou vznikat při kombinačně-množinovém řešení jazykových problémů.

2. Informační a komunikační orgány a systémy.

Člověk a všichni živočichové svět kolem sebe poznávají díky smyslovým (informačním) orgánům a mozku, společnost poznává svět především díky informačním systémům (odborníci, výzkumné ústavy, knihovny, archivy atd.) a mění ho existenčně působící činností. Změnit svět jen na základě poznatků nelze. Sebelepší informační systém pouze poskytne podklady o možném způsobu zásahu do lidskou společnost obklopujícího a na ni působícího světa, ale prováděné změny závisí na kombinaci vhodných poznatků a použití vhodných prostředků fyzického a organizačního působení.
Sebelepší výsledky činnosti Semanu mohou vést pouze k nalezení některých vhodných a nutných řešení, ne však všech a jejich uskutečnění vždy bude záviset na organizačních, politických, stavebních, vědeckých a dalších podobných postupech. Dokonce i kdyby jednou bylo s využitím Semanu, nebo jiného budoucího systému vybudováno něco, co by snílek označoval jako centrální mozek lidstva, nebude možno nikdy zaručit, že

a) tento mozek se v některých výpočtech a odhadech nezmýlí
b) tento mozek nebo jeho části se nedostanou do rukou lidí, kterým nejde o nejsprávnější poznání a z něho vyplývající řešení reality, ale především o získání výhod pro vybranou skupinu lidí.

Proto musí každý, kdo bude na Semanu pracovat, s těmito nebezpečími počítat a přemýšlet o tom, jak by bylo možno se s nimi v co největším stupni vyrovnat. Zastavit vývoj Semanu, abychom se vyhnuli možným nebezpečím, nemá smysl. Neboť principy jsou velmi jednoduché a je velmi nepravděpodobné, že dalším lidem nenapadne něco podobného a vývoj technických podmínek lidstva dává možnost tyto nápady rozvíjet, třeba i s jiným způsobem kódování a logicko-matematického řešení.

2.4. Co je seman

Naprostá většina lidských znalostí má formu jazykových výroků a to výroků zachycených na nosičích. Na výroky lze převést téměř vše, s čím se v životě setkáváme. Informační pracovníci pracují především s odbornými texty, které jsou nejfrekventovanější a nejdostupnější formou informací.
Jazyk má dva zásadní problémy:
- různé přirozené jazyky. Řešením je vytváření překladových slovníků, převodových systémů gramatických pravidel, zajišťování překladů. Dlouhodobě jistě vzniknou počítačové překladové systémy. Již dnes se na nich intenzivně pracuje.
- složité sémantické vazby v rámci každého jazyka bez vztahu ke znění slova a výrazu (to platí i pro jazyky používající obrázkové písmo, nebo pro matematiku a logiku). Řešením je vytváření složitých sémantických slovníků typu tezaurů, MDT a dalších typů informačních jazyků.

Seman je nástroj umožňující stále složitější práci s výroky, slovy a texty s využitím prostředků výpočetní techniky. Seman pracuje s přirozenými a

umělými jazyky ve tvaru sémových ekvivalentů, přičemž sémové ekvivalenty zastupují slova a výrazy přirozených a umělých jazyků v míře dostačující pro práci na prostředcích výpočetní techniky.
Nelze očekávat, že by Seman hrál někdy úlohu jazyka zajišťujícího přímou komunikaci a porozumění mezi lidmi. Lze však počítat s tím, že usnadní vytvoření prostředků, které přiblíží výpočetní techniku k možnostem lidského uvažování, neboť převádí lidský jazyk na jazyk formalizovaný a současně zachovávající sémantické obsahy a vazby.

2.4.6. Seman a jeho vazba na výpočetní techniku

Člověk a lidská společnost stále své znalosti rozšiřuje a prohlubuje, a většina společenských akcí probíhá s využitím jazyka. Jazyk a jeho využití jsou možné jen díky nesmírně složitým vazbám mezi lidským mozkem a znakovým systémem jazyka. Schopnost lidské inteligence pracovat s jazykem, je postupně převáděna na možnosti práce výpočetní techniky. V současné době je nejvýraznějším krokem v tomto směru Internet a můžeme vycházet z předpokladu, že se budou neustále zdokonalovat tři parametry výpočetní techniky.

a) neustále poroste rychlost operací
b) neustále poroste velikost zpracovatelných dat
c) ještě dlouho se bude zmenšovat fyzická velikost zařízení výpočetní techniky

2.4.7. Práce s textem jako textem

Naprostá většina textů dnes existuje jako texty, do nich žádné programy spojené se zpracováním sémů nebo s překladem z jednoho jazyka do jiného, nezasahují. Systém Seman v současné době využívá „jednoduchých“ textových informací, předpokládá se, že texty zpracovávané v systému Seman budou číst, chápat a využívat lidé, a že výpočetní technika zde pouze poslouží jen jako pomocný prostředek, umožňující například uchování, přenos, komprimaci těchto textů. Proto musí systém Seman být schopen pracovat s textem v té podobě, v jaké existuje. Zde se pracuje se slovy, větami a odstavci. Textů je obrovské množství. Pro mnoho uživatelů je důležité, tvořit seznamy slov z textu, frekvenční seznamy slov z jednoho nebo skupiny textů, výběr jistých částí textu podle subnázvu, podle kombinací slov, dokonce i podle typu písma. Tyto úkoly je možné provádět s využitím sémů nebo jen s využitím programů týkajících se slovní podoby textu.
Podstatným problémem jsou obrázky, grafy a tabulky, které je třeba držet v paměti počítače, a poskytovat je uživateli na správném místě a ve vhodné chvíli, případně dokonce ve vhodném zpracování. V této verzi systém Seman obrázky, grafy a tabulky zachovává, ale nezpracovává. Je to však jen otázka času, kdy to bude umět a míry, v níž to bude umět.

2.4.8. Porovnání doslovného znění dvou a více textů

Základním cílem je porovnat dva texty a zjistit, v čem se liší. To je úloha, kterou např. řeší tvůrci norem, tvůrci a příjemci opakovaných textů jako jsou dopisy, připomínky k úředním materiálům i ti kdo kontrolují původnost prací. Velmi často existuje potřeba porovnat dva texty u studentských prací. Existují různé systémy porovnávající stupeň podobnosti textu. Existují dokonce systémy zjišťující totožnost či její možnost v hudebních dílech. My navrhujeme v Semanu postupovat takto:

Každý porovnávaný text označíme dvoumístným nezaměnitelným označením. Toto označení připojíme ke každému slovu a interpunkčnímu znaménku. Těchto možných označení je 1296 (od 00 do zz). Systém Seman si zapamatuje označení textů a místa uložení textů.
Za znak „ : “ ke každému slovu a samostatnému interpunkčnímu znaménku připojíme pořadové číslo.
Bude-li to vhodné, k pořadovému číslu textu a slova připojíme kód, který určuje formální postavení slova či interpunkčního znaménka v textu. Významu tohoto kódu říkáme status slova.

Pořadová čísla jsou od 000001 do 999999. Text tedy může mít až milion slov. (bez jednoho slova). Zde jsou zapsány pouze nejčastější statusy. Kód statusu slova se zaznamenává za pořadové číslo slova a znak *.
Přiděluje se automaticky:
0 slovo zapsané běžným písmem - minuskulí
1 slovo má na počátku velké písmeno
2 slovo je zapsáno proloženě
3 celé slovo je zapsáno velkými písmeny
4 slovo je zapsáno zmenšenými písmeny
5 slovo je zapsáno červeně (další barvy mohou být zapsány písmeny)

U každého slova může být za znakem * zapsáno číslicemi a znaky více statusů.

Přidělíme pořadové číslo textu, který chceme porovnávat, přidělíme každému slovu pořadové číslo. Porovnáme slova s pořadovými čísly v obou textech. Doporučuje se porovnávat druhý text s textem prvním.

Je možno pracovat v několika stupních:

-Srovnání vždy jednoho slova s jedním. Za slovo z prvního textu se za znak = zaznamenají čísla totožných slov.

-srovnání dvojice slov v prvním a druhém textu. Porovnává se vždy slovo a slovo za ním v prvním textu se stejnými dvěmi slovy ve druhém textu.

-srovnání více slov (jejich počet je snadno zjistitelný podle pořadových čísel slov v textech). Odborníci z Českého normalizačního ústavu doporučují, aby se řady slov pokládaly za totožné při nejmenší délce 7 slov. Podobně usuzují i odborníci z jiných oblastí.

Výstupem je seznam totožných skupin slov ve dvou textech. Každá tato skupina je zapsána s pořadovým číslem prvního slova v druhém textu / a pořadovým číslem posledního slova skupiny v druhém textu. Před první skupinou té které délky je zapsána jejich délka v počtu slov a interpunkčních znamének.

Ukázka velmi krátkého textu je uvedena v souboru (001)

Porovnávání obsahu více textů

Program, který s využitím zahraničních podkladů a zkušeností vyvinul Mgr. Roman Chýla, umožňuje porovnávat více textů a zjišťovat stupeň jejich překrývání. Prvním krokem je vytváření otisků slovních skupin a jejich následné porovnávání, vyhledávání stejných otisků v databázi již zpracovaných textů. Viz ukázku přidělení pořadových čísel k soubor uvedenému v 001, v souboru (002). Potřeba porovnávání textů vedla k vytvoření programů, které umožňují zjišťování rozdílů ve znění dvou textů. Je hotov také program, který umožňuje porovnábání dvou nebo i více textů, přičemž červeně jsou zapsány části textu, které jsou v textu 2 a nebyly v textu 1. Ukázku zde neuvádíme.

Když chceme srovnat text (001) s dalším textem, zvolíme si ho vyvoláním souboru (003).

Pak provedeme opět přidělení pořadových čísel k tomuto textu (004).

Z obou textů stroj snadno pozná, že se texty liší jen ve dvou slovech a jsou si tedy obsahově velice blízké.

2.4.9. Náhrada slov a prvků jazyka krátkými kódy

Jazyková vyjádření již dnes představují nepřehledné množství textu. Lidský mozek má však značně omezenou kapacitu. Proto je třeba, i když bereme v úvahu, že se mozek a inteligence mohou značně vyvíjet, hledat cesty, jak záznam znalostí co nejvíce zkomprimovat, aniž bychom ztráceli jejich obsah. Přitom je zřejmé, že lidské myšlení a inteligence (umělou inteligenci nutně pokládáme a ještě dlouho budeme pokládat za rozšíření přirozené lidské inteligence) pracuje se slovy a slovům podobnými výrazy, případně s grafickými prvky, které lze převést na prvky uložitelné v mozku. Je možné, že jednou začneme pracovat i na jiných principech, ale dnes si je ani neumíme představit. Proto se naše práce bude stále zabývat slovy, jejich kombinacemi a významy těchto slov a kombinací.

2.4.10. Identifikátor významu

Zvukové znění slov nelze sjednotit, jejich sémantický význam však ano. Slova chlapec, hoch, boy, Knabe ve všech jazycích označují totéž. Tento význam lze zachytit. Základní směr prací Carnapa, Vídeňské školy a dalších jazykových škol byl zaměřen na sjednocení jazyka vědy, a třebaže dnes existují vědecká periodika, které tomuto problému věnují značnou pozornost, je tento problém stále aktuální nejen v politice, náboženství a ideologii, ale i v exaktních vědách.

Na jisté, často dostačující úrovni významové přesnosti, jednotné chápání významu ovšem možné je. Kdyby to nebylo možné, nemohly by existovat překladové slovníky, nemohli by se lidé jedné kultury a jednoho jazyka naučit jinému jazyku a pochopit jinou kulturu a civilizaci.

Je velmi pravděpodobné, že význam je do jisté míry závislý na kontextu (lze také říci: na způsobu života společnosti, v níž se objevil a v níž se používá), to platí především pro složité významy, např. náboženské, filosofické, společenskovědní. Vždy však lze najít význam, který je dosti přesně definovatelný. Např. slova chlapec, boy, Knabe, označují člověka mladšího než určitý věk a rodu mužského.

Identifikátor sémantického významu jakéhokoliv slova (slovy vyjádřeného výrazu) je zde zaměňován pětimístným kódem. Viz soubory (005, 006).

Pravidla tvorby pětiznakového kódu jsou vložena do příslušného programu Kód identifikátoru významu. Pro případ, že pracovník Semanu bude chtít vložit nový výraz do souboru všech výrazů systému Seman, který se označuje kódem all, a chce mít jistotu, že nepoužije kód již se vyskytující, je nejlépe vyjít z data, kdy tento krok provádíme a přidělit kód s datem spojený.

Úlohu přidělování kódů na základě dat nalezneme v souboru (007).

Student by si měl zkusit vytvořit kód identifikátoru významu podle několika různých dat. Tato úloha je uvedena v souboru (008).

Kontrola k této úloze je v souboru (008a).

Systém Seman před všechny nové položky (řádky) v českém jazyce připojuje identifikátor významu. Před spuštěním programu je nutno vybrat soubor, k jehož položkám v češtině se budou identifikátory významu připojovat. Program si pamatuje poslední přidělený identifikátor a začíná pracovat na přidělení nejbližší volné kombinace znaků.

V souboru (009) se student může seznámit s slovy a výrazy v češtině s připojenými identifikátory významů.

Kódy identifikátoru významu se objevují u všech položek (řádků), k nimž přísluší. Zatím neumíme přidělení identifikátorů významu provádět zcela automaticky.

2.4.10.1. Přidělování identifikátorů významů

Slova a výrazy, které mohou obsahovat i více slov, jsou převoditelné na číselný a abecední výraz, který má určitou délku (dnes pěti znaků). Při 36 samostatných znacích (26 malých písmen mezinárodní abecedy a 10 číslic) to dává možnost uložit v systému Seman něco přes šedesát milionů výrazů. V případě potřeby bude možno tento výraz rozšířit (např. jen prodloužení kódu z pěti na šest znaků dovoluje rozšířit kapacitu Semanu na více než miliardu jednoznačných výrazů). A tento počet lze velmi jednoduchými úpravami dále prudce zvyšovat. Stále rostoucí kapacity výpočetní techniky slibují, že se do souborů vměstná kompletní soubor všech existujících výrazů a jejich významů v mnoha jazycích. Zvyšující se rychlost operací bude stále více umožňovat provádění stále složitějších intelektuálních operací. A stále se zmenšující rozměry zaručují, že výpočetní technika je a stále více bude moci být zapojována do jiných lidem sloužících zařízení a dokonce do lidského těla.

Ukázky znění pětiznakového kódu identifikátoru významu tvořeného malými písmeny mezinárodní abecedy a číslicemi jsou uvedeny v této práci na více místech.

Identifikátor nikdy v systému Seman nevystupuje sám. A to proto, že identifikátor významu odpovídá významu, který si člověk uvědomuje díky k tomu uzpůsobenému vědomí. Význam je spojen vždy se zněním slova a slova mají v různých jazycích různé znění, i když mají stejný, nebo velmi blízký význam. Slova, která jsou synonymní, znějí rozdílně, i když mají stejný, nebo jen nepatrně se odlišující význam. Např. téměř stejný význam, často odlišitelný je v rámci rozdílných stylů mají dvojice slov hoch- chlapec, žena- manželka, pták- opeřenec .
V systému Seman se musí vždy pracovat se znaky, s nimiž může pracovat výpočetní technika. Aby byla otázka kódu identifikátoru pochopitelná, uvádíme v souboru (006) příklady identifikátoru spolu s kódem českého jazyka a českým slovem, které nejlépe odpovídá významu a nachází se za znakem =.

2.4.11. Změna z databázového přístupu v seznamový přístup

Většina podobných systémů je budována na principu databází. Princip databází je v současné době velice propracovaný. Jestliže autor pracuje primárně s principem seznamovým, je to proto, že v současné době tento princip pořádání, vyhledávání a reorganizace lingvistických prvků je dostatečně propracovaný, aby bylo možno s ním rychle a přehledně pracovat. V budoucnoti snad bude možno více používat databázového principu. Na vlastní podstatě Semanu se tím nic změnit nemusí. Při seznamovém přístupu jsou každá hodnota a každý úkol zapsány nezaměnitelným kódem, který systému říká, co má udělat. Student tento princip pochopí při prostudování této učebnice.

2.4.12. Seman jako seznam významů a k němu se pojících prvků

Základním pojmem Semanu je význam lexikálního výrazu, tedy nikoli jeho znění, ale jeho sémantický význam. Význam českého slova, jehož znění je Sněžka, je „hora v pohoří nazývaném Krkonoše, nacházející se na území dnes osídleném národem nazývaném Češi“, význam německého slova Vogel je „živočich, schopný létat a mající peří“, význam anglického slova queen je „označení člověka ženského rodu, který ve společenském upořádání zastává díky narození nebo volbě místo vedoucí osoby“.
V každém jazyce se tento význam může pojit s jinými zvuky (písmeny, číslicemi). V češtině chlapec, v angličtině boy, v němčině Knabe, v ruštině (transkribovaně) mal ´čik, atd.
Kdyby lidé nepřisuzovali slovním výrazům totožné nebo alespoň velmi blízké významy, nemohla by mezi lidmi existovat komunikace a porozumění. Právě přisuzování rozdílných významů slovním výrazům, o nichž určité skupiny lidí soudí, že jejich význam je jasný, je jedním ze základních důvodů vzájemného nepochopení mezi jednotlivci i skupinami lidí (sociální nepokoje, politické spory a války). Druhá skupina prvků, které vedou k vzájemnému nepochopení, je boj o využívání zdrojů k životu lidmi. Tato skupina prvků nevychází z použití jazyka a Seman je nemůže řešit. Protože však naprostá většina lidmi provozovaných činností je do značné míry spojena s použitím jazyka, odráží se tyto problémy také v jazyce a Seman může napomáhat k jejich odhalení již v prvních stádiích jejich vniku.

2.4.13. Kód jazyka

S využitím kódu jazyka lze za kódem identifikátoru připojit vyjádření toho, o jaký přirozený jazyk se jedná. Kód má čtyři znaky. Na začátku kódu je vždy znak +. Za tímto znakem následuje písmeno označující jazyk. Dvě poslední písmena označují jisté vlastnosti v rámci jazyka. Kódů označujících přirozené jazyky může být až 15 tisíc. Viz ukázku označení jazyků kódy v souboru (010). Systém je připraven k tomu, aby bylo možné postupně zvládnout lexikální výrazy v mnoha jazycích, a to díky tomu, že se za identifikátor významu výrazu připojí identifikátor jazyka a systém si znění výrazu pamatuje. Vyvolejte ukázku v souboru (011), kde uvidíte kód jazyka předcházený kódem identifikátoru významu a následovaný za znakem = slovem (výrazem) v příslušném jazyce.

2.4.14. Kód módu

Pojem mód označuje charakteristické vlastnosti významů a znění slov a sémů a dalších skutečností, s nimiž pracuje systém Seman.
Kód módu je také dlouhý čtyři znaky a má také vždy na prvním místě kódu znak +. Tento kód zachycuje podstatný údaj o identifikátoru významu. Na řádku vždy následuje po mezeře za kódem identifikátoru významu a na tomto místě se střídá

s kódem jazyka, případně s kódy označujícími obsahy semů a vlastní kódy semů. Další nejdůležitější módy jsou zatím především signatury dokumentů v knihovnách. Každá knihovna, každé třídění, každý soubor PSČ má vlastní kód. Například signatury dokumentů v knihovnách jsou označovány jako +sg1 pro fond Národní knihovny, Praha; +sg2 pro fond Brněnské městské knihovny; +sg3 pro fond ostravské krajské knihovny, atd. Délka kódu módu je (včetně znaku plus) 4 znaky.
V souboru (012) najdete ukázku záznamu signatur slovesných děl uložených knihovních fondech . Díky tomuto způsobu zápisu si lze představit využití Semanu i pro všeobecné informační systémy – např. knihovní katalog, v budoucnu dokonce jako společný katalog pro více knihoven.

2.4.15. Seznamy výrazů pro překlad z jazyka do jazyka

Výrazy označené totožným identifikátorem a různými kódy jazyků, jsou významově totožné.
Například:
rklma +ccc=Moje maminka odešla dnes ráno z domova
rklma +aaa= My mother went this morning from the home
Tyto významově totožné věty nohou být značně dlouhé. V programu Preklad zvolíme výchozí a cílový jazyk (tedy např. výchozí= +ccc, cílový = +aaa). Díky tomuto postupu lze připravit systém překlad z výchozího do cílového jazyka. Díky velkým kapacitám a velkým rychlostem lze postupně vybudovat velmi kvalitní systém překladu.
Ukázku výrazů vhodných k překladu viz v (013).

2.4.16. Doplňování vhodných výrazů do překladových výrazů

Jakoukoli část těchto překladových výrazů lze nahradit jinými vhodnými výrazy. Na místo příslušných slov lze vložit výrazy označené třímístným kódem, kde na prvním místě je znak *, na druhém číslice označující počet slov, které se do výrazu ve výchozím jazyce vloží (tento počet slov může být 1 až 9), na třetím místě je číslicí uvedeno pořadové místo, na které má být v obou výrazech zařazen doplňkový výraz.
9ublo +ccc= *11 nejbližší *12 by rád přišel *13 domů
9ublo +aaa= *11 closest *12 would like to come *13 home
Systém najde vhodný výraz a vloží ho na správné místo.
Ukázku vět v češtině a angličtině s vložitelnými místy pro překlad viz v (014).

2.4.16.1. Významově totožné výrazy

Takové výrazy jsou velmi časté a při přípravě překladu je nutno se s nimi vyrovnat. Např. český výraz „ pracoval jsem „ lze překládat významově téměř stejně s použitím anglických výrazů I worked, I have worked, I have been working, I was working. Slova, která nemusí v překladu být, dáme do závorky se znakem % . V programu je chod, který zaručuje, že výraz v závorce bude z překladu v případě potřeby vyloučen.
Ukázku vylučovaných a nahrazovaných výrazů při překladu z češtiny do angličtiny viz v (014a).

2.4.17. Identifikační část řádku

Identifikátor a kód jazyka nebo módu tvoří základní dvojici kódů první části záznamu (řádku). Za nimi následuje znak = a za = následují slova, číslice, kódy sémů a podobně.
Příklad se slovy v několika jazycích a jejich sémové ekvivalenty je uveden v (015).

2.4.18. Používané termíny Semanu

Zde uvedený seznam termínů má čtenáři usnadnit orientaci v textech Semanu. Seznam neobsahuje definice, odkazy na příklady nebo úlohy, které by měli studenti řešit. Bližší a (i opakované) definice a vysvětlení termínů se mohou vyskytovat na odpovídajících místech textu.

2.4.18.1. Identifikátor obsahu jazykového výrazu.

Pětiznakový kód (26 malých písmen mezinárodní abecedy a deset číslic), který zastupuje význam lexikálního výrazu v libovolném jazyce. Těchto kódů může být více než šedesát milionů.

2.4.18.2. Identifikátor jazyka

Čtyřznakový kód začínající znakem + identifikující jazyk obsahu jazykového výrazu. Např. +ccc pro češtinu.

2.4.18.3. Identifikátor módu

Čtyřznakový kód určující, jak a případně zda existují určité výrazy, postupy a kódy, které jsou na příslušném řádku identifikátoru modu umístěny za

znakem = . Identifikátory říkají, co je na daném řádku uvedeno. Např. sém, slovní vyjádření, sémový ekvivalent ad.

2.4.18.4. Sém

Nejmenší abstraktní sémantická jednotka, skládající se z kódu a slovně zachyceného významu. Např. pracování = aadd.

2.4.18.5. Kód sému

Čtyř a víceznakový kód nahrazující obsah semu. Např. aadd.

2.4.18.6. Obsah sému

Slovně vyjádřené znění sému. Obvykle jedno české slovo ve tvaru přídavného jména slovesného (pracování) nebo podstatného jména označujícího vlastnost.

2.4.18.7. Sémový ekvivalent

Skupina kódů sémů vyjadřující sémantický obsah slova či výrazu, kombinace sémů.

2.4.19. Sémy a práce s nimi

Práce člověka s jazykem je velmi komplikovaný soubor procesů. V současné době množství systémů, které pracují s jazykem začíná být obrovský. Výpočetní technika tuto práci umožňuje především kvůli velkým paměťovým kapacitám, rychlosti zpracování, a možnosti zadávat velmi složité algoritmy. Bez těchto předpokladů by o semovém přístupu k řešení textů nebylo možno vůbec uvažovat. Rychlý vývoj výpočetní a komunikační techniky pravděpodobně zaručují, že sémový nebo jemu blízký přístup bude poskytovat lidem stále více možností.

2.4.20. Sém

Sém je vyjádřením dále nedělitelného významného sémantického obsahu, např. býti člověkem, býti stromem, býti červený, býti suchý, atd. Každé slovo i

víceslovný výraz lze zapsat abecedním, abecedně číselným nebo jen číselným kódem. Každý z těchto kódů odkazuje na slovně vyjádřený sémantický obsah. sému.
V kódech semů se objevují pouze malé znaky mezinárodní abecedy a číslice. Sémy tvoří slovníkové položky, které mají tento tvar zapsaný ve dvou položkách, každá na samostatném řádku.
Semy jsou vázány v současné době na češtinu (pracovní jazyk systému).
(+sem kód sému, +dig kód výrazu vyjádřeného číslicemi, +ccr kód slovního vyjádření obsahu sému). Příklady sémů v češtině jsou uvedeny v souboru (016) a v angličtině v souboru (017). Přitom je dobré mít na paměti, že v angličtině a češtině se při zápisu číslic využívají odlišným způsobem znaky (.) (,) a (mezera). Bude-li to třeba, tyto rozdíly budou doplněny.

2.4.21. Sémy s významem obecnějším a specifičtějším

Sémy, stejně jako slova a jazykové výrazy, lze seřadit do hierarchických řad od nejobecnějšího významu k nejspecifičtějšímu významu. Řada slov: Evropa-Německo-Braniborsko-Berlin. Řada sémů evrop- nemec- branb- berlin.
V sémových ekvivalentech je zachováván princip, že specifičtější sém je vždy doprovázen vhodným semem dostatečně obecným. (V souboru all za znakem +sem).
Příklad je uveden v souboru (018).

2.4.22. Druhy semů

Sémů je více druhů. Každý sém má slovy nebo číslicemi vyjádřený věcný (sémantický) obsah a kód obvykle 4 a více znaků. Sémantický obsah kódu sému je nezaměnitelný. Sémy lze kombinovat a tím lze poměrně velmi správně vyjádřit věcný (sémantický) obsah každého slova nebo výrazu v libovolném přirozeném nebo umělém jazyce.
Každý sém je od sémů předešlých i následujících v sémovém ekvivalentu (viz dále), i od jakéhokoli jiného znaku, oddělen mezerou.
Intelektuálně tuto činnost vykonáváme neustále, aniž bychom si ji plně uvědomovali. Sem vlastně vyjadřuje a formalizuje pro potřeby výpočetní techniky tuto zvláštní schopnost lidského intelektu (mozku) být si vědom jednotlivých prvků, obsažených ve slovech, obrazech, tónech, barvách, čárách, intenzivních jevů apod. Nikdo není schopen říci, jak to je možné, i když stále vznikají pokusy to objasnit. Abychom mohli sémantické pochody v práci výpočetní techniky co nejvíce přiblížit lidskému intelektu, formulujeme semy a procesy, které s nimi lze provádět, způsobem, který umožňuje sémy co nejvíce přiblížit prvkům objevujícím se v lidském myšlení.

2.4.22.1. Semy vyjadřující číselné hodnoty

Sémy zachycující číselnou hodnotu jsou dvou druhů: pro číselné hodnoty vyšší než nebo rovné 1 (první dva znaky kódu nn)a číselné hodnoty nižší než 1 (první dva znaky kódu pp).
Sémy zachycující čísla zachycují řád číslice a její velikost. Řád umožňuje záznam číslice od jednotek do desítek milionů, záznam číslice od 0 do 9.
Ukázka sémů zachycujících číselné hodnoty celých číslic je uvedena v souboru (019). Záznam každé číslice čísla má tvar: nn-řád-číslice. Například: pro číslo 2: nn12, pro číslo 23: nn22 nn13, pro číslo 235: nn32 nn23 nn15.
Kódy pro čísla menší než 1 jsou tvořeny písmenem pp a dvěmi číslovkami, z nichž první označuje pozici za desetinnou čárkou, druhá hodnotu na této pozici. Ukázku viz v souboru (020).
Další příklady čísel větších a menších než 1 jsou v souboru (021).

2.4.22.2. Sémy zachycující rok

Kódy těchto sémů jsou tvořeny písmeny lk pro křesťanský letopočet, písmeny la pro arabský letopočet a písmeny lz pro hebrejský letopočet. Bude-li se pracovat ještě s jinými letopočty, budou jejich kódy zavedeny. Za dvoupísmenovým kódem letopočtu je uvedeno vročení dvěma číslicemi, z nichž první udává řád, druhá hodnotu číslice roku na škále 0 až 9.
Každý reálný rok v křesťanském letopočtu je tedy tvořen čtyřmi sémy, např. narodil jsem se roku 1951 se sémově vyjádří takto: narodil jsem se roku lk41 lk39 lk25 lk11.
Příklad je uveden v (021a) (identifikátory významu nejsou totožné se souborem all).

2.4.22.3. Sémy označující jména a názvy

Kód takového sému je obvykle odvozen od názvu nebo jména a má 5 a více písmen. Počet těchto sémů může být velmi velký. V rámci strojového zpracování to však nevede k žádným podstatným problémům, neboť všechna slova a výrazy v jazyce jsou vyjadřovány sémovými ekvivalenty, tedy kombinacemi sémů. V následující tabulce jsou uvedena slova týkající se lokalit na území Čech (věcné sémy udávají sémantické obsahy pojmů). (Identifikátory významu neodpovídají souboru all). (022).

2.4.22.4. Sémy označující věcné sémantické obsahy pojmů

Kód sému má čtyři znaky (pouze abecední znaky mezinárodní abecedy). Těchto sémů je v současné době formulováno asi tří a půl tisíce. Stále přibývají, i když se růst počtu sémů zpomaluje (023).

2.4.22.5. Sémy označující gramatické kategorie.

Tyto sémy určují, jak bude slovo, jehož obsah je vyjadřován sémovým ekvivalentem skloňováno nebo časováno, do kterého slovního druhu patří. Tento sém je obvykle také čtyřznakový a jeho ukázky jsou uvedeny v (024).

2.4.22.6. Sémy označující příslušnost výrazu do selekčního jazyka nebo třídění.

Kód sému je vždy tří a vícemístný, kde první dva znaky označují selekční jazyk nebo třídění, následující znaky umístění výrazu v daném souboru.
Ukázku viz v (025).

2.4.22.7. Sémy označující příslušnost slova či výrazu k jazyku

Seman je budován tak, aby mohl pracovat se slovy a výrazy v mnoha přirozených jazycích. V současné době jde především o češtinu, angličtinu a latinu a jejich kombinace. V budoucnu bude moci zpracovatel zvolit i jiné jazyky. Sém má vždy jako první znak +.
V současné obě se pracuje především se třemi jazyky: češtinou (+ccc), angličtinou (+aaa) a s latinou kvůli velkému počtu termínů (+lll).

2.4.22.8. Sémy označující symboly.

Všechny symboly lze vyjádřit jako sémy. Základním problémem je, jak s jistotou identifikovat, že v daném případě jde o symbol určitého významu. To vyžaduje v řadě případů provést inteligentní identifikaci.
Sémy zachycující symboly mají vždy na prvních dvou místech svých kódů znaky yx. Za nimi následuje zkrácené slovní pojmenování symbolu.
Přehled kódů častých symbolů je uveden v souboru (027). Jisté, že další kódy symbolů budou doplňovány.

2.4.23. Sémový ekvivalent

Obsahově lze každé slovo, nebo i víceslovný výraz převést na tzv. sémový ekvivalent.
Sémový ekvivalent je soubor sémů vyjadřujících obsah slova (víceslovného výrazu, symbolu, vzorce). Sémy v ekvivalentech nemusí být stejně důležité.
Ukázka je uvedena v souboru (028).

2.4.24. Slovník sémových ekvivalentů

Slovník sémových ekvivalentů je řazen podle znění identifikátoru. Lze jej snadno vybrat do samostatného souboru pro každý jazyk, případně pro každou klasifikaci a selekční jazyk a setřídit záznamy podle pravidel platných pro

příslušné případy. Příslušnost dané položky k jazyku, klasifikaci atd. je vyjádřena příslušným kódem jazyka a módu.
Úplný slovníkový seznam má jméno all. Od určité délky bude seznam rozdělen do několika souborů, z nichž každý bude označen názvem all a pořadovým číslem. (all-1, all-2, atd.). Každý výraz s identifikátorem a kódem jazyka nebo módu je zapsán na samostatném řádku.
Příklad řádků s opakovanými identifikátoru s rúznými kódy jazyků nebo módů je uveden v souboru (029).

2.4.24.1. Doplňování slovníku sémových ekvivalentů.

Slovník lze doplňovat převzetím slov a výrazů z jakéhokoli slovníků, zapsáním slova a výrazu náhodně získaného, popřípadě dokonce nově vytvořeného.
Nová slova a výrazy jsou přiřazeny na konec souboru all a následně s využitím třídícího programu zařazeny podle abecedy.
Při doplňování slovníku sémových ekvivalentů je třeba dbát na to, aby v témže sémovém ekvivalentu byly kromě sémanticky nejspecifičtějšího výrazu vždy také obsaženy také sémy obecnějšího výrazu. Proto je nutno zachovávat tento postup:

- odhadnut sémanticky obecnější výraz. Ten je třeba vyhledat v seznamu all. Vyhledaný výraz je zapsán do řádku nového výrazu a na jeho konec je připsán specifický sém. Celý nový výraz je vložen do souboru all a zařazen abecedně. Příklad viz v (030).

Informační pracovník, který bude pracovat se systémem Seman, by měl být schopen do systému doplnit mimo jiné i sémantický ekvivalent kteréhokoli slova či výrazu, který ještě není součástí systému Seman.
V souboru (031) je uveden text úlohy, která by měla přispět k nácviku této schopnosti.
Řešení této úlohy lze zkontrolovat v subor (031a).
Semy těchto výrazů spojíme do jediného ekvivalentu.
K těmto ekvivalentům připojíme nově vytvořený, nejlépe pětiznakový sem, po zkontrolování, že sém dosud neexistuje, jej připojíme k vytvořeným sémovým ekvivalentům. Kontrolu řešení této úlohy najdeme v (032).

2.4.25. Označení slov a výrazů podle přirozeného jazyka

Slova a výrazy v souboru all se dělí podle přirozeného jazyka, z něhož pocházejí. To je nutné proto, že znaková podoba slova může mít stejnou podobu, přestože to slovo má v různých jazycích odlišný význam. Např. „by“ v angličtině je předložková částice obvykle s významem u, vedle, spolu, v češtině tato částice označuje podmínečnost (I live by my mother, přišel by tam rád až později).

Kód přirozeného jazyka má tvar tříznakového kódu s předcházejícím +, který se zaznamenává na řádek, kde je už zapsán identifikátor významu.
Nejvíce se v této době pracuje se třemi jazyky- češtinou (+ccc), angličtinou (+aaa) a latinou (+lll).

2.4.26. Česká slova a výrazy

Čeština je pracovním jazykem systému Seman. Kódy jazyka a módů pro česká slova a výroky mají v sobě vždy nejméně dva znaky c,c:
Čeština byla zvolena jako pracovní jazyk systému Seman nejen proto, že ho tvůrci systému dobře ovládají, ale také proto, že v českém jazyce existuje řada prací a odborníků zabývajících se sémanticko-informační analýzou jazyka.
V jiných jazycích se nepracuje s prefixy, kmeny a sufixy, a to proto, že každý výraz v češtině je významem totožný s identifikátorem a tím i odpovídajícími výrazy v jiných jazycích.
Příklad nalezneme v souboru (033)

2.4.27. Rozklad slov do části slov

Téměř každé slovo se skládá ze tří částí: prefixu, kořenu (někdy se hovoří o kmeni a sufixu). Dělení slov do těchto částí může pomoci při jejich pochopení a zpracování. Otázky dělení slov do částí jsme využili především v pracovním jazyce Semanu, češtině, ale tuto problematiku lze rozpracovat pro jiný jazyk, bude-li rozhodnuto, že on se má stát pracovním jazykem systému Seman.
Kořen slova je ta část slova, která v co největší míře odpovídá znakovému znění slova i jeho sémantickému obsahu. Program práce se slovy textu dovoluje automaticky dělit každé slovo do tří částí, které se samostatně překládají do sémového ekvivalentu. Se sémovým ekvivalentem se pak pracuje běžným způsobem.

2.4.27.1. Kořen slova

Kořen slova je ta část slova, které ve znění zůstává neměnná. Stejný sémový ekvivalent může být připojen k různě tvořeným kořenům se stejným sémantickým významem souboru. V identifikační části řádku bývá mód +ccq, za = bývá kořen slova.
Viz soubor (034).

2.4.27.2. Prefix slova

Prefixem označujeme první část slova. Prefixy jsou soustředěny za označením ccp. Sémové ekvivalenty prefixů jsou dvou typů:

- Prefix převoditelný do sémů - viz soubor (035).
- V některých případech prefix zatím nepřevádíme do již existujících sémů, ale vytvoříme sém nový, pracovní. Každý takový pracovní sém je tvořen

ze znění (někdy zkráceného) prefixu a znaku g na prvním místě. Viz soubor (036).

2.4.27.3. Sufix slova

Sufix je v lingvistice obvykle chápán jako koncovka. V systému Seman nejde vždy jen o koncovku, ale o závěrečnou část slova. Soubor vybraných závěrečných částí slov (koncovek) pro český jazyk viz v souboru (037). Ukázka všech tří součástí tvořících ve svém celku každé české slovo je uvedena v souboru (507). Stroj je díky speciálnímu programu schopen nalézt ve zpracovávaném textu části slov, tyto části nahradit jejich semovými ekvivalenty a z nich sestavit celkový semový ekvivalent slova.

2.5. Překlad z jazyka do jazyka

Systém Sémantického analyzátoru umožňuje i překlad z jednoho jazyka do jazyka jiného. Při překladu se pracuje s identifikátory dvojího druhu: číselným ekvivalentem výrazu a módy příslušných jazyků. Postup je tento:
Je zvolen text, který chceme přeložit. V textu hledáme postupně všechny výrazy přesně odpovídající výrazům nacházejícím se v seznamu sémových ekvivalentů. Programový systém umožní vybrat nejdelší výraz z textu a zaměnit jej významově identickým výrazem v cílovém jazyce.
Překlad je kontrolován a opraven speciálními programy
Uživatel má možnost doplňovat nové vhodné výrazy, případně dokonce opravovat již zabudované výrazy.
Jako první překladový modul je budován překlad z angličtiny do češtiny a z češtiny do angličtiny. Podobných systémů bylo ve světě i nás již vytvořemo více, uživatelé však k nim mají často oprávněné námitky, neboť ani jeden známý systém zatím nedosahuje takové kvality přeloženého textu, aby nestálo za to do něj intelektuálně zasahovat. Nejlepších výsledků v rutinním provozu zatím dosáhly systémy prcující s velmi omezeným slovníkem a popisem situací, které se neustále opakuji. K nim např. patří kanadský systém překladu informací o počasí pracující s angličtinou a francouštinou. Velmi zajíavé a drahé experimenty překladu novinoívých článků z ruštiny do angličtiny byly omezeny v době, kdy Sovětský Svaz přestal být Spojeným Státům vážnou vojenskou protiváhou.
Hotový překlad je možno opravovat a doplňovat.
Ukázku přípravy překladu viz v (038).
Zvolíme z internetu nebo jiného média text v angličtině nebo jiném jazyce a jeho překlad do češtiny.

2.5.6. Doplňování vět z textů

V současné době je k dispozici velké množství přeložených textů v různých jazycích a z různých jazyků. Velké množství z nich je k dispozici ve strojem čitelné formě. Základním

problémem je zajistit, že správné části výchozího textu budou významově totožné s správnými částmi cílového textu.
Zvolíme z internetu nebo jiného média text v angličtině nebo jiném jazyce a jeho překlad do češtiny.
Dosažení stavu potřebné totožnosti nemůže být dokonalé, lze však dosáhnout velmi dobré kvality. Bude třeba alespoň výběrově provádět kontroly a úpravy souborů.
Nejvhodnější pracovní postup je:

- zvolíme text ve výchozím jazyce
- vybereme přeložený text v cílovém jazyce
- předběžně zkontrolujeme, že věty ve výchozím a cílovém jazyce jsou přibližně stejně dlouhé
- přidělíme pořadové číslo každé větě v obou textech (věta začíná po znacích . (nebo ?, nebo !) ,mezera a majuskulou psané písmeno)
- překontrolujeme, že pořadová čísla jsou přidělena správně
- věty se stejnými pořadovými čísly jsou zapsány do dvou za sebou následujících řádků
- u vět přesahujících délku řádku provedeme další rozdělování délky

Ukázku viz v (038).

2.5.7. Doplňování částí věty do překladu

V každé větě lze jisté části nahradit gramaticky a věcně vhodnými výrazy. Toto místo je označeno třímístným kódem, kde je na prvním místě znak *. Na druhém je číslice označující počet slov, která mají být ve výchozím jazyce nalezena a zaměněna slovy cílového jazyka, na třetím místě je číslice označující pořadové číslo výrazu ve větě.
Příklad: Můj nejlepší přítel se narodil ve Francii.
My best friend was born in France.
Můj nejlepší *11 se narodil ve *12
My best *11 was born in *12.
Systém nalezne v souboru all výraz, který se vyskytuje na tomto místě v překládaném textu. Pravděpodobnost chyby je velmi nízká.

2.6. Zpracování textů a obecněji pramenů.

V následujících bodech jsou uvedeny některé zajímavé aplikace. Další budou postupně doplňovány.

2.7. Obecně o textech

Sémantický analyzátor pracuje neustále s texty. Ty překládá z výchozího jazyka do jazyka cílového, různě je upravuje z výchozího znění do znění cílového, zjišťuje četnosti různých sémantických jevů a z nich vyvozuje odhad pravděpodobného vývoje skutečnosti, kterou texty jako produkt lidského myšlení odrážejí.

Texty jsou různého druhu a zde pojednáme jen o těch, které jsou v současné době nejčastěji zpracovávány.
V řadě případů jde o prameny, s nimiž informatika běžně pracuje.
Všechny prameny a texty musí být ve strojem čitelné formě, ve formátu přistupném editoru Windows. Pokud jsou v jiném editoru, musí být do tvaru tohoto editoru převedeny, pokud existují pouze v písemné nebo jiné podobě, musí být v tomto tvaru uloženy. Protože v současné době už většina pramenů v tomto editoru existuje, lze počítat s obrovským souborem pramenů.
Každý pramen (text) má název. Název má tvar sedmimístného čisla k němuž je připojen sém označující o jaký druh textu se jedná. Další sémy v řádku sedmimístného identifikátoru označují, které operace byly s pramenem prováděny, které se chystají a kde jsou jejich výsledky uloženy
Sémy těchto informací jsou zapsány v souboru sémů.

2.7.6. Překladové jazykové slovníky

Pro potřeby Semanu se využívají především soubory, které lze zakoupit na CD-ROMech, nebo získat z počítačů, slovníky v papírové formě mohou sloužit pouze jako pomocné, neboť jejich uložení je pracné. Dokonce i tam, kde existuje v elektronické podobě pouze seznam slov jednoho jazyka, našem případě češtiny, je možno od něj odvíjet přípravu dalších jazyků, a to včetně přidělování těmto slovům v dalších jazycích semových ekvivalentů. Množství elektronických souborů obsahujících soubory dvou a více jazyků ve tvaru překladového slovníků je značný. Před jejich použitím musíme zajistit, aby záznam sov obou jazykových mutací vyhovoval potřebám Semanu. To prakticky znamená, zkontrolovat, případně odstranit rúzné poznámky, doplňky a různočtení. Tuto činnost je nutno, přes její obtížnost, vykonat. Pracovníci zjistí, že tato práce při jisté dovednosti je mnohem snazší, než by se dalo očekávat. V Souboru. Lze vzít z česko-anglický slovník a českou část zaměnit za výrazy s jejich sémovými ekvivalenty. Potom lze českou část zrušit a přihrát sémové ekvivalenty k části anglické.

2.7.7. Bibliografické knihovnické záznamy

Analýzou bibliografických záznamů lze zjistit velmi mnoho o stavu a vývoji nakladatelské činnosti a tím i o stavu a vývoji společnosti. Data potřebná k takové analýze jsou snadno dostupná.

Význam kódu módu	kódu
název knihy nebo jiného dokumentu=	+awc
jméno autora (jména autorů)=	+awd
délka v počtu stran=	+awe
rok vydání=	+awf
město vydání=	+awg
nakladatel nebo vydavatel=	+awh
ilustrátor=	+awi

editor= +awj
sigma nebo jméno knihovny= +awk
signatura dokumentu= +awl
aplikaci těchto sémových kódů viz v (041a).

2.7.8. Analýza obsahu textů s pomocí četnosti sémů

Sémy podstatně zpřesňují sémantickou vypovídací hodnotu slov.
Vvtvoříme frekvenční seznam sémů v textu, části textu nebo dokonce v souboru textů a z tohoto seznamu usuzujeme na obsah textu, části textu nebo celého souboru. Pomocí této analýzy můžeme kvalifikovaně usuzovat na to, co se pravděpodobně děje ve skutečnosti, o čem jsou vědecká , případně i umělecká díla apod. V současbé době stojíme na samém začátku podobných prací, jejich vypovídací hodnota i postupy budou teprve formulovány, ale je pravděpodobné, že analýza četností sémů je schopna poskytnout rychlé a dosud neznámé pohledy nejen na texty a jejich soubory, ale i na skutečnost, jíž se te xty zabývají. Studenti by měli uvažovat o způsobech využití analýzy četností sémů. Zde uvádím jeden z relativně úspěšných příkladů takové analýzy. Jde o analýzu pěti povídek významných světových autorů. Tyto povídky jsou v různých jazycích, různě dlouhé, a různě zaměřené.
Jde o:

1. Bjornstjerne Bjornson: Nebezpečné námluvy
2. Lion Feuchtwanger: Wollstein
3. Guy de Maupassant: Závěť
4. Jorge Luis Borges: Sekta ptáka Fénixe
5. Anton Pavlovič Čechov: Štába Prišibejev.

Každou povídky jsme mechanicky rozdělili na tři stejně dlouhé části a provedli výpočet četností sémů v těchto třetinách, a vytvořili seznam sémů s četností vyšší než stanovená hranice. Pro každou povídku jsme stanovili hranice jinak, v závislosti na její délce a počtu semů. V příkladech text povídek neuvádíme, student se s nimi může seznámit v uvedených pramenech. Délka každé povídky byla zjišťována podle překladu do češtiny, měření poču slov v púvodních jazycích, se může trochu lišit, ale pro analýzu opírající se o sémy, na tom v zásadě nezáleží, neboť sémy jsou v každé jazykové verzi stejné, liší-li se kvůli způsobu překladu, je to pro naší analýzu málo podstatné.
Pro naší analýzu je důleřité, od jaké četnosti jsme sém zařadili do analýze. Platí, že velmi málo četné sémy nestojí za to, analyzovat. Stejné pravidlo platí i i pro četnost slov v teorii pochopitelnosti, působivosti a obtížnosti textu, kterými se zde nezabýváme, s nimiž by se studenti mohli s prospěchem pro své budoucí povolání seznámit.
Lze uvést jedinou skupinu důvodů, proč může být velmi nízký výskyt slova nebo sému v textu významný, a tím je snaha emovionálně zapůsobit na vnímatele. Tyto důvody s však obvykle vyskytují v uměleckém, citově

zaměřeném textu, nikoli v textu odborném, s nímž informační pracovníci obvykle pracují.
Počty sémů, od nichž byly semy v povídce brány v úvahu jsou uvedeny v této tabulce:

Číslo povídky	hraniční počet semů
1	8
2	11
3	7
4	5
5	7

V souboru (042) jsou uvedeny sémy s četností pro každou třetinu textu povídky 1. Z těchto statistických údajů lze velmi dobře uvažovat o obsahu a průběhu děje povídky. Sémy nejsou zapsány svými kódy, ale jen slovními významy.

2.7.9. Porovnávání textů s pomocí sémů

Všechny texty lze převést ze slovní podoby do podoby sémové.
Tento převod se provádí následujícím způsobem:
Všem slovům textu je přiděleno pořadové číslo. Toto pořadové číslo je přiděleno všem jednotlivým sémům sémového ekvivalentu tohoto slova.
Vytvoří se dvojice sémů.
Těchto dvojic je obrovské množství. Aby byla zachována jednota je dvojice tvořena vždy tak, že sem s nižší abecední hodnotou je vždy na prvním , za ním následuje znak : , za tímto znakem druhý sem dvojice. Dvojice sému je v textu vždy v jisté vzdálenosti slov. Počet mezer slovy určuje vzdálenost. Čím si jsou slova blíže, tím větší má dvojice váhu. A to podle této tabulky:

Vzdálenost mezi slovy	váha
Sémy jsou ve stejném semovém ekvivalentu	9
1 mezera	8
2 mezery	7
3 mezery	6
4 mezery	5
5 mezer	4
6 mezer	3
7 mezer	2
8 mezer	1

Tuto tabulku a odpovídající programy lze v budoucnosti rozšířit a doplnit.
Ke každé dvojici je připojeno číslo váhy. Aabd:aahg 6 aaba:aahg 2, atd.
Váhy všech stejných dvojic sémů se sečtou. Nevíce zastoupené dvojice ve dvou textech jsou Aaba:aahg 76 x aabd:aahg 124. Pouhý pohled na tyto dva údaje ukazuje, že druhý zkoumaný text se více zabývá muži než ženami.
Vytvoříme seznam všech dvojic sémů s konečným číslem.

Ukázka takových dvojic sémů s hodnotou odpovídající jejich vzdálenostem je uvedena v souboru (047).

Z tohoto seznamu vytvoříme vážený, nejlépe procentuální seznam.Tento seznam nazýváme Sémová konzerva textu. Když provedeme zjištění součtu velikosti absolutních rozdílů mezi dvěma texty a sečteme je, pak texty s nulovou hodnotou součtu rozdílů jsou obsahově i strukturálně totožné.Čím je číslo rozdílu větší, tím jsou si texty obsahově méně blízké.

Tento přístup lze uplatnit při

a) předtřiďování textů

b) vyhledání textů příbuzných k předvolenému textu (v souboru koeficientů podobnosti stanovíme nám vyhovující hranici platnosti podobnosti)

c) Rozdělení textu do částí podle závažnosti částí

2.7.10. Výpočty sémantických vztahů

Vyjádření slov a výrazů v sémových ekvivalentech umožňuje práci se sémy jako se symboly obsahových obrazů slova, výrazu, věty, textu. Obsah sémantických úseků lze tedy vypočítat. Je nepochybné, že dnes stojíme na samém počátku teorie takových výpočtů, a že bude trvat ještě dlouho, než taková teorie bude fungovat opravdu uspokojivě. Již dnes však v řadě oblastí mohou tyto postupy vést k velmi dobrým výsledkům.

Slovo, výraz lze chápat jako množinu sémů, mezi těmito množinami existují vyčíslitelné vztahy, např.

$$P_{1,2} = \frac{2s_0}{s_1 + s_2}$$

kde $P_{1,2}$ je sémantická podobnost slov (výrazů) 1 a 2

s_0 je soubor sémů společných pro ekvivalenty 1 a 2

s_1 s_2 je soubor sémů ekvavalentů 1 a 2

6.5.Zachycování změn v četnostech sémů

Každý text (informace) zachycuje pohled jednotlivého člověka, skupiny lidí, nebo instituce na určitý úsek skutečnosti. Vyjádření textů v sémových ekvivalentech umožňuje vytvářet strojové postupy, které pravděpodobně povedou k nalezení metod sledování změn ve skutečnosti a hodnocení těchto

změn lidmi. V této době ještě nehledáme význam změn v četnostech sémů a jejich kombinací, ale jen postupy, které dovolí porovnávání těchto změn.

2.7.10.1. Zjišťování četnosti sémů za určité období

Četnost sémů za určité období je možno získat pouze z textů za určité období. Soubor textů je možno vymezit časem jejich publikace, prameny, zdroji, vydavatelstvím, obsahem (např. klíčovými slovy, označením profilu internetu a podobně).
Soubor textů je pak zpracován takto:
- všechna slova v něm se vyskytující
  jsou chápána jako soubor sémových ekvivalentů
- spočítá se četnost těchto sémů
- tyto četnosti se případně převedou na četnosti relativní

Z absolutních či relativních četností vybraných semů lze usuzovat na stav souboru a společnosti.

2.7.10.2. Srovnání četností sémů ve dvou časových souborech

Tento postup je základní pro odhad vývoje. Stanovíme dvě časová období, která chceme srovnávat a prameny, které v nich chceme zpracovávat.Např. čas omezíme na rok 1988 a rok 1989. Budeme chtít zjistit rozdíl v četnostech sémů v těchto dvou letech a v prameni Lidové noviny, jejichž texty jsou k dispozici v databázi novinových textů v bibliografickém oddělení České národní knihovny. S použitím vhodných programů vytvoříme seznamy sémů: (Kódy semů neodpovídají kódům použitým v systému Seman).

Rok 1988	rok 1989
aaaa 36967	27842
aaab 24	282
aaac 1	2
aaad 453	324
.	
.	
zzzy 789	359

Sémy s nízkou četností ze zpracování vyloučíme.
Zjistíme, o kolik došlo k růstu či poklesu četnosti. Je li výsledné číslo větší než jedna, došlo k růstu četnosti semu, je-li menší než jedna, došlo k poklesu. Čím je číslo větší ne jedna, tím k většímu nárustu četnosti semu, čím je číslo menší než jedna, tím k většímu poklesu došlo. Vytvoříme tak výsledný seznam.
Změny v četnosti semu v souboru roku 1989 v vůči souboru z roku 1988.

Sem	změna
Aaaa	0,751
Aaab	11,75

Aaad 1,4
.
.
aazy 0,71

2.8. Vzory pro doplňování různých tvarů slov

Těchto vzorů je možno vytvořit mnoho a s jejich pomocí lze vytvářet soubor all. Zde uvádíme jen několik příkladů takových vzorů.
Každý zde uváděný vzor je jen ukázkou, a má ve skutečnosti desítky až stovky položek. Hotových vzorů jsou desítky.

2.8.6. Vzor pro slovesa v češtině s infinitívní koncovkou –it

Jeho rozpis je uveden v souboru (048).

Kompletní seznam všech slov může vznikat s pomocí speciálního programu a předpisů. Předpisy použitelné pro prácí se slovy pracovně označuji jako vzory.
Rody a čísla nejsou v ekvivalentu zapsány.
Pomlčky pouze naznačují skladbu výrazu a do výsledného souboru se nepromítají.
Později bude podle potřeby možno tyto informace do vzoru zapsat.
Vzory pro další typy (skupiny) slov v češtině jsou uloženy na chráněných počítačích. Vzory pro angličtinu budou dopracovány.

7.2. Zkuste vytvořit předpis pro slova podle vzoru mužského Vzor pro podstat. jm. v češtině, vzor pán.
Připravený vzor je možno srovnat pro kontrolu se souborem (049).

2.9. Základní druhy nebezpečí využívání Semanu

Všechna nebezpečí si v současné době neumíme představit. Jejich společným znakem je, že povedou člověka k chybnému hodnocení skutečnosti a tím i k chybným zásahům do této skutečnosti.
Největší dnes odhadnutelná nebezpečí jsou:

- skupina lidí si vymyslí představu o optimálním vývoji společnosti a tu ve společnosti prosadí, protože z uskutečnění této představy bude mít skupina výhody. Lze si dokonce představit, že uskutečnění takové představy, i když nemá nic nebo mnoho společného se Semanem, bude po určitou dobu výhodné a úspěšné.Nebezpečí spočívá v tom, že možnost dlouhodobě chybného řešení je vysoká. (Většina totalitních režimů byla vybudována zneužitím velmi dobrých myšlenek.)
- Seman, nebo jiný podobný systém, doporučí řešení, které je vhodné pro působení známých faktorů, ale toto řešení je chybné, nebo alespoň

neúplné, neboť začal působit díky vývoji skutečnosti nový, dosud neznámý nebo dosud nevýznamný faktor. (O teorii a faktorech vývoje skutečnosti bude napsán zvláštní text.)

2.9.6. Možný postup při zneužití sémanu

Lze se obávat, že k zneužití Semanu, ať už náhodnému, nebo promyšlenému, může docházet především po případném úapěšném dokončení faktorového modelu a možném reálném vývoji společnosti, který by vedl ke zhoršení životních podmínek velkých skupin lidí. I když dnes na to nechceme myslet, dějiny poskytují dostatek příkladů, že se to stává. Historické zkušenosti také ukazují, že skutečně dlouhodobě nelze žádnou myšlenku před zneužitím nebo před využitím v rámci akcí, s nimiž se tvůrci myšlenky neztotožňujíní, zcela ochránit. Jediný postup, který může snížit možnost zneužití nebo alespoň zkrátit dobu zneužívání, je umožnit k myšlence přístup co největšího množství inteligentních lidí. To zvyšuje pravděpodobnost, že vzniknou pokusy o zachování myšlenky a systémů ve prospěch většiny lidí, a to i za cenu značných zásahů do systému a do politicko-organizačních opatření s ním spojených. Naštěstí v současné době je již možno jakýkoli soubor vložit do systému Internetu. Nezaručuje to sice ochranu systému Seman, a snižuje to pravděpodobnost získání případných výhod pro tvůrce systému Séman. Ale je to pravděpodobně jediná cesta, která by mohla vést k zajištění aktivního podílu systému Seman na rozumné existenci lidí.
Seman, aby byl opravdu využitelný, musí být doveden do ralizačního stavu, tedy do stavu, kdy se znalosti spojí s jejich hodnocením a praktickým využíváním. Tomu by měla sloužit představa Faktorového modelu světa, jednou v budoucnosti zahrnující i metody a postupy různých věd napomáhajících životu člověka a lidské společnosti.

2.10. Faktorový model světa

Zde uvádíme jen základní infornaci o tvaru a principech faktorového modelu. Skutečnost, která lidi obklopuje a působí na ně, je velmi složitá. Aby ji lidé pochopili a mohli měnit, musí si pochody, objekty a situace, které ve skutečnosti existují, rozložit na menší úseky a působení každého úseku zkoumat samostatně a později v jejich vzájemném působení.
Seman je systém pracující s popisem pochodů, objektů a situací a proto bude možno včlenit do systému podobného systému Séman i faktorový model. Text faktorového modelu světa bude tedy věcným i programovým pokračováním tohoto textu, bude jeho druhým dílem. Základní rozdíl vždy zůstane v tom,že zatímco Seman pracuje a bude pracovat pouze s jazykovými prostředky vyjádřenými poznatky, faktorový model bude pracovat, nejobecněji vyjádřeno, se zásahy do hmotné nebo sociální skutečnosti. Jeho kladné i záporné výsledky mohou být proto pro lidi mnohem citelnější. Oba systémy svou strukturou, funkcemi a cíly na sebe ale tak silně navazují, že by bylo škoda nepokusit se

faktorový model nevystavět. Možnost neúspěchu je tu však ještě větší než v případě budování sémanu. Na druhé straně však má-li lidstvo nejen pochopit skutečnost ve které žije, ale má-li do značné míry nutné k rozumnému přežití, ji dokoncee ovládnout, pak práce na přístupech, které zde označujeme jako faktorový model, budou pokračovat, a stále další skupiny odborníků se o ně budou pokoušet.

2.11. Vybrané programy

Zde jsou zapsány pouze slovní stručné popisy programů.

2.11.6. Abecední řazení řádků

Specifikum: řadí se podle znění identifikátoru a za ním následujícího znaku jazyka nebo modu.

2.11.7. Abecední řazení slov textu

2.11.8. Přidělení pořadových číslic slovům textu

Slova zůstanou v pořadí, v nichž jsou v textu. Za každé slovo (interpunkční znaménka se pokládají za samostatná slova) je připojeno pořadové číslo v rozmezí 000001 až 999999.

2.11.9. Přidělení slovům textu pořadových číslic a kódu způsobu záznamu

(totéž, co v předcházejícím programu. Za každé slovo se připojí kód způsobu zápisu slova.

2.11.10. Přidělení identifikačního kódu textu.

Každý text dostane svůj identifikační kód, v němž jsou dvě části,

- text
- pořadové číslo v rozmezí 000001 – 999999

2.11.11. Výběr řádků s určitou kombinací znaků ze souboru do zvláštního souboru

Do masky programu jsou na příslušná místa zapsány tři údaje:

1. název souboru, z něhož se čerpá
2. znaky v tom pořadí a znění, v němž budou vyhledány
3. název souboru, do něhož budou nalezené řádky začleněny

2.11.12. Včlenění nového souboru řádků do již existujícího souboru

1. Je vybrán soubor řádků a označen jako nový soubor
2. je vybrán soubor, do něhož m být nový soubor připojen
3. oba soubor jsou spojeny do jednoho
4. tento nový soubor řádků je seřazen podle abecedy

2.11.13. Výpočet sémantické podobnosti dvou výrazů s využitím vzorce podobnosti

1. Slova (výrazy), jejichž sémantická podobnost je porovnávána, jsou vypsána i se sémovými ekvivalenty.
1. Oba sémové ekvivalenty jsou vypsány.
2. sémy vyskytující se v obou sémových ekvivalentech jsou označeny
3. označené sémy jsou přeneseny do speciálního souboru, který označíme jako čitatel
4. sémové ekvivalenty obou výrazů spojíme be ohledu na to, že se v nich sémy opakují
5. porovnáme sémy v čitateli se sémy ve jmenovateli, sémy (jen jednotlivé) vyskytující se v čitateli i jmenovateli, zrušíme.

7. Případy, kdy se v čitateli a jmenovateli vůbec nevyskytují stejné semy, z dalšího zpracování vyloučíme
6. spočítáme počet sémů v čitateli a ve jmenovateli
7. počet sémů ve čitateli vydělíme počtem sémů v ve jmenovateli
8. výsledem vepíšeme jako sémantickou podobnost, která bude vždy menší nebo rovná 1 a větší nebo rovná 0

2.11.14. Přidělení pořadových čísel slov sémům v sémových ekvivalentech slov

1 k pořadovým číslem označeným slovům textu (souboru) připojíme jim odpovídající sémové ekvivalenty

9. pořadová čísla slov textu se připojí k sémům odpovídajícího sémového ekvivalentu

2.11.15. Vytvoření struktury dvojsémů textu

Vezmeme výsledek programu 010

1. Nalezení dvojice sémů odpovídajících tabulce vzdáleností slov a jejich vah
2. vytvoření seznamu dvojsémů s vahou vzdálenosti mezi nimi
3. sečtení vah kombinaci pro každý jednotlivý dvojsém
4. vytvoření abecedního seznamu všech dvojsémů s hodnotou sumy

2.11.16. Kontrola, že nově zaváděný sem se stejným kódem se již v souboru neobjevil

1. vytvoření seznamu všech kódů sémů s připojeným záznamem významu sému
2. označení dvou řádků sémů, kde je kód stejný a význam jiný, nebo naopak

2.11.17. Vytvoření speciálního souboru řádků, z nichž každý obsahuje kombinaci znaků

1. Zvolíme výchozí soubor
2. Zvolíme označení cílového souboru
3. zvolíme vyhledávanou kombinaci znaků
4. provedeme

2.11.18. Porovnání dvojsémové struktury dvou textů a výpočet rozdílů mezi nimi

1. nahrání dvojsémového seznamu 1
2. nahrání dvojsemového seznamu 2
3. porovnání hodnot stejných sémů v seznamu 1. a 2.
4. vytvoření seznamu sémů s hodnotou absolutních rozdílů mezi sémy v obou seznamech
5. sečtení všech absolutních rozdílů
6. Stejný postup opakovat u všech dvojic textů
7. seřazení všech kódů dvojic textů podle velikosti součtu rozdílů textů vůči textu zvolenému pro porovnávání jako základní
8. rozhodnutí od jaké velikosti součtu rozdílů, který bude pokládán za dostatečný
9. seřazení kódů textů podle velikosti součtu rozdílů

2.11.19. Zjištění četností sému v textu, souboru

1. Ke každému slovu textu (souboru) jsou připojeny sémové ekvivalenty
2. Je vypočtena četnost každého sému
3. je vytvořen abecední seznam sémů s četnosti

2.11.20. Rozdělení textu na stejně velké díly(části)

1. zvolíme soubor k dělení
2. zvolíme velikost výsledných souborů v počtu znaků
3. zvolíme název pro díly souboru
4. provedeme rozdělení, přičemž ke každému dílu bude připojeno za znakem = pořadové číslo

2.11.21. Rozdělení textu (souboru) na části s stejným počtem řádků

1. zvolíme soubor
2. zvolíme počet řádků (např. 500)
3. zvolíme název pro díly souboru
4. provedeme rozdělení, přičemž před každým dílem bude zapsán název a jeho pořadové číslo

2.12. Možnosti některých úloh řešitelných Semanem

2.12.6. Obecně o úlohách řešitelných Semanem.

Úlohy řešitelné Semanem mohou především využívat to, že se nepracuje jen se slovy a jejich kombinacemi, ale také se sémy, tedy s významy obsaženými ve slovech a jejich kombinacích. Vzdělání člověka se mimo jiné projevuje v tom, že vzdělaný člověk je schopen vnímat slova (a jakýkoli lexikální výraz v souvislostech s významovým kontextem. Například výraz Přemysl Otakar II. Podává vzdělanému člověku informaci, že to byl český král žijící ve třináctém století. Při historickém vzdělání se mu může vybavit i jeho finanční a vojenská síla, jeho územní rozpínavost atd. Všechny tyto informace lze vložit do semů a vytvořit tedy počítačový systém umožňující takovéto lidsky intelektuální pohledy na text a práci s ním. V současné době to ještě není možné, neboť systém Seman dosud není dostatečně naplněn. Pokračování v jeho plnění připomíná proces vzdělávání člověka. V současné době je systém Seman na úrovni malého dítěte a bude záležet především na způsobu jeho plnění a práce s ním, jaké úrovně a směru vzdělanosti dosáhne.
Čím častěji se sem označující osobu, případně skupinu osob, určitý předmět, pochod, prvek objevuje, tím je pravděpodobně významnější, čím více se objevuje s jinými sémy, tím je pravděpodobně toto spojení významnější.
Na základě četností sémů a jejich kombinací lze velmi vážně uvažovat o řadě situací které mohou přispět k následujícím poznatkům:

2.12.7. Zjištění významnosti osoby

2.12.7.1. Zjištění posunů ve významnosti osoby ve dvou či více časových obdobích

2.12.7.2. Zjištění rozdílů v hodnocení osoby různými skupinami lidí

2.12.8. Zjištění významnosti státu, země, národa

2.12.8.1. Zjištění posunů ve významnosti státu, země, národa, podniku, výrobku, vědecké metody, léku apod. ve dvou či více časových obdobích

2.12.8.2. Zjištění rozdílů v hodnocení státu, země, národa, podniku, výrobku, vědecké metody, léku apod. různými skupinami lidí

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
(001)

Strojové zpracování textu vyžaduje využití zařízení s dostatečnou kapacitou a

rychlostí , takové programy které toto zpracování umožňují .

(002)

strojové 000001

zpracoání 000002

textu 000003

vyžaduje 000004

využití 000005

zařízení 000006

s 000007

dostatečnou 000008

kapacitou 000009

a 000010

rychlostí 000011

, 000012

takové 00013

programy 000014

které 000015

toto 000016

zpracování 000017

umožňují 000018

. 000019

(003)

Strojové zpracování textu vyžaduje použití zařízení s dostatečnou kapacitou a rychlostí, a takové programy, které toto zpracování umožňují.

(004)

automatické 000001

zpracování 000002

textu 000003

vyžaduje 000004

využití 000005

zařízení 000006

s 000007

dostatečnou 000008

kapacitou 000009

a 000010

rychlostí 000011

, 000012

takové 00013

programy 000014

které 000015

toto 000016

zpracování 000017

umožňují 000018

. 000019

(005)
ukázka možných identifikátorů významu. (V současné době vždy jen jedno-jednoznačná kombinace pěti znaků). Z ukázky je zřejmé, že identifikátor významů může mít jakýkoli tvar, pokud sestává z pěti znaků a je umístěn na počátku řádku. K přidělování jednoznašných kódu je vytvořen speciální program.
abcde
acc12
a1bkm
alvnm

1aaaa
abtk1
lsfun

(006)
Možné znění kódů identifikátorů významu pro slova a výrazy v českém jazyce. (Viz řádky, na nichž je mód +ccc.)
Sémantické ekvivalenty jsou uváděny na dalším řádku se týmž identifikátorem a módem +sem).
Abcde +ccc= žena
Abcde +sem= aahg aaba aazb
Bdlvn + ccc= ženský
Bdlst +sem= aahg aaba aabf
Ankl2 +ccc= muž
Ankl2 +sem= aahg aabd aazb
cnhl2 +ccc= mužský
Ankl2 +sem= aahg aabd aahf
acc12 +ccc= dítě
accl2 +sem= aahg anla aazb
acknu +ccc= dětský
acknu +sem= aahg anla aahf

(007)
Je-li dnes např. 1.1.2008, pak kód bude mít tento tvar:82011 až z. Tedy první číslice vyjadřuje rok v jedenadvacátém století(rok 10= 0,rok 11= a, rok12= b, atd. druhá pozice měsíc (prvních deset měsíců číslicemi 1 až 0, měsíce listopad a prosinec písmeny a , b), třetí číslice zachycuje den od 1 do 10, která se zapíše jako 0, dále za každý den od 11 do 31 a, b, až u.
Tedy k 1.1.2009 = 9101
k 31.12.2009= 9b31
Za tento čtyřznakový kód se připojí na pátou pozici znak z 36 znakového souboru písmen a číslic, tedy např. 9b311, 9b312 až 9b31y,9b31z.

(008)
Vytvořte kódy identifikátorů k datu 23.2.2008, a k datu 19.května 2O16.

(008a)
Kontrola úlohy z (008)
82231
82232
82233
.
.

.
8223y
8223z

f5191
f5192
f5193
.
.
.
f519y
f519z

(009)
České výrazy s identifikátory významu s kódem jazyka.
ag23r +ccc= chlapec
lmnrs +ccc= děvče
krtsd +ccc= dívenka
chs53 +ccc= knihovna
chr21 +ccc= knihovník
mrtal +ccc= kniha
ntsfa +ccc= knihovnice

(010)
Kódy některých jazyků.
český jazyk= +ccc
anglický jazyk= +aaa ,
německý jazyk= +ggg,
latina= +lll
francouzský jazyk= +fff
španělský jazyk= +sss
burgundština= +bur
holandština= +hol
klasická arabština= +ara
svahilština= +sva
rumunština= +rum
italština= +ita
portugalština= +prt
antická řečtina= +gre
ruština= +rrr
běloruština= +bru
makedónština= +mak

albánština= +alb
novodobá řečtina= +gra

(011)
Ukázka řádků se stejným identifikátorem významů a různými kódy jazyků.
2bfgk +ccc= muž
2bfgk +aaa= man
2bfgk +ggg= der Man
ankl1 +aaa= boy
ankl1 +ccc= chlapec
ankl1 +ggg= der Junge
glvno +ccc= hoch
glvno +ggg= der Knabe
glvnr +ccc= hošík
glvnr +aaa= little boy
glvnr +ggg= der kleine Knabe
hrhpo +ccc= hošíček
hrhpo +aaa= little boy

(012)
Záznamy signatur různých knihoven.
3ckjl +ccc= Život a dílo skladatele Foltýna
3ckjl +sg1= ankl257
3ckjl +sg7= b0326
dlkxa +ccc= Válka s mloky
dlkxa +sg1= ankl3571
dlkxa +sh7= b0331
87an2 +ccc= Poslední případ Agathy Chistie
87an2 +sg1= bvc057
87an2 +sg7= ha482
(+sg1 označuje signaturu dokumentů České národní knihovny Praha, Klementinum. +sg7, signaturu dokumentů Bezručovy knihovny v Ostravě, signatury a názvy knihoven nejsou vzaty ze skutečnosti.

(013)
Obsahově totožné výrazy v češtině a angličtině, které lze použít při překladu.
rklma +ccc=Moje maminka odešla dnes ráno z domova
rklma +aaa= My mother went this morning from the home
klda9 +ccc=Můj tatínek s mojí maminkou žijí ve městě, které leží na horním toku řeky Lužnice.
klda9 +aaa=My father lives with my mother in the town situated on the upper part of Luznice river.
alvk1 +ccc= můj otec

alvk1 +aaa= my father
ar1o2 +ccc= můj drahý otec
ar1o2 +aaa= my dear father
cek12 +ccc= můj nejdražší dědeček
cek12 +aaa= my dearest grandpapa

(014)
9ublo +ccc= *11 nejbližší *12 by rád přišel *13 domů
9ublo +aaa= *11 closest *12 would like to come *13 home
překlad pak zní:
můj nejbližší přítel by rád přišel zítra domů
My closest friend wold like to come tomorrow home
Nebo
9ublz +ccc= *11 nejbližší *12 by ráda přišla *13 domů
9ublz +aaa= *11 closest *12 would like to come *13 home
překlad pak zní:
moje nejbližší přítelkyně by ráda přišla zítra domů
my closest girl-friend would like to come tomorrow home

(014)
Překlad z češtiny do angličtiny s potlačením stylisticky velmi podobných výrazů.
b4akn +ccc= pracoval jsem
b4akn +aaa= I worked
b4akn +aaa= (I have worked %)
b4akn +aaa= (I have been working %)
b4akn +aaa= (I was working %)
Překlad pak zní
Včera jsem pracoval na poli
Yesterday I worked (I have worked , I have been working, I was working) at field

(015)
Výrazy v angličtině, češtině a latině se semovými ekvivalenty.
Legenda: +aaa - anglický výraz
+ccc - český výraz
+lll – latinský výraz
+sem – semový ekvivalent
aavt- sém pro dětskost
aaba- sém pro ženskost
aahg- sém pro býti člověkem
aazb- sém pro býti subjektem
Na prvním místě řádku je vždy kód ekvivalentu významu.

2dkmo +aaa= girl
2dkmo +ccc= děvče
2dkmo +lll= puela
2dkmo +sem= aahg aaba aazb
kmuo1 +ccc= Emanuel Kant
kmuol +sem= emanuel kantx aahg aabd
mldrs +ccc= dítě
mldrs +aaa= das Kind
mldrs +sem= aafg aavt aazb

(016)
Semy v češtině.
abbda +dig=1 (na pozici set)
abbda +sem= nn31
abbjr +dig= 5 (na pozici desítek)
abbjr +sem= nn25
bsklm +dig= 35721
bsklm +sem= nn53 nn45 nn37 nn22 nn11
khvln +ccr= samickost
khvln +sem= aaba
khume +ccr= sameckost
khume + sem= aabd
kruvz +ccr= městskost
kruvz +sem= aada
brtya +ccr= anglickost
brtya +sem= angli

(P17)
Semy v angličtině.
abbda +dig= 1 (hundreds)
abbda +sem= nn31
abbjr +dig= 5 (tens)
abbjr +sem= nn25
bsklm +dig= 35721
bsklm +sem= nn53 nn45 nn37 nn22 nn11
khvln +aaa= female
khvln +sem= aaba
khume + aaa=male
khume+sem= aabd
khuvz +ccc= to be a town
khuvz +sem= aada
brtya +aaa= to be English
brtya +sem= angli

(018)
Sem doprovázený vhodným sémem dostatečně obecným.
Berlín= evrop nemec branb berln aada
Jednotlivé sémy s lexikálním vyjádřením:
Evropa= evrop
Německo= nemec
Braniborsko= branb
Berlín= berln
Město= aada
Berlíňanka= evrop nemec branb berln aaba aahg aazb
Jednotlivé semy s lexikálním vyjádřením:
Evrop= evropa
Nemec= Německo
Branb= Braniborsko
Berlínskost= berln
Městskost= aada
Ženskost= aaba
Člověk= aahg
Subjekt= aazb
Kokršpaněl= acac pesxx kokrs aazb
Jednotlivé semy s lexikálním vyjádřením:
Acac= živošich
Pesxx= pes
Kokrs= kokršpaněl
Aazb= subjekt

(019)
Semantické ekvivalenty celých čísel.
+dig= 359
+sem= nn33 nn25 nn19
+dig= 1489
+sem= nn41 nn34 nn28 nn1
+dig= 2754691
+sem= nn11 nn29 nn36 nn44 nn55 nn67 nn72
+dig= 42154682
+sem= nn12 nn28 nn36 nn44 nn55 nn61 nn72 nn84
+dig

(020)
Semantické ekvivalenty čísel menších než 1.

Kódy pro čísla menší než 1 jsou tvořeny písmenem pp a dvěmi číslovkami, z nichž první označuje pozici za desetinnou čárkou, druhá hodnotu na této pozici.
+dig= 0,78
+sem= pp17 pp28
+dig= 0,5
+sem= pp15
+dig= 0,2374
+sem= pp12 pp23 pp37 pp44

P20
Semy různých čísel vyjádřených sémovými ekvivalenty.
345ab +dig= 345
345ab +sem= nn33 nn24 nn15
045nc +dig= 0,45
045nc +sem= pp14 pp25
231dg +dig= 2,31
031dg +sem= nn12 pp13 pp21
lgn24 +dig= 847,25
lgn24 +sem= nn38 nn24 nn17 pp12 pp25

(021)
Příklad (identifikátory významu nejsou totožné se souborem vsechno) :
Legenda: +dig- číslo, v tomto případě vyjadřující rok
+sem- semový ekvivalent
a1997 +dig= 1997
a1997 +sem= lk41 lk39 lk29 lk17
c2056 +dig= 2056 před n. l.
c2056 + sem= aalg lk42 lk30 lk25 lk16

(022)
kódy semů a především identifikátorů významů neodpovídají vždy souboru all
aklvr +ccc= Praha
aklvr +sem= aada evrop
aklvt +ccc= evropan
aklst +sem= evrop aahg aabd
brtst +ccc= Čechoslovák
brtst +sem= evrop czech slove aahg aabd
drvkl +ccc= Alena
drvkl +sem= aahg aaba aazb alena
drcll +ccc= alenka
drcll +sem= aaba aabf aahg aazb alena
drclm +Alenkou

drclm +sem= aaba aahf aahg aazb alena pj17

(023)
Semy zachycující věcné obsahy pojmů.
Sémy vyjádřené různými slovy mohou mít stejný kód
Sameckost= aabd
Člověckost= aahg
Samickost= aaba
Horskost= aahn
Pracování= bdag
Milování= ckno
Smrt= bkaa
Umírání= bkaa adro
Hmyzovost= alzv
Zdrobnělost= aabf
Některé specifičtější sémy musí mít připojen obecnější sém:
Motýlovost= alzv barh bdtr
Motýlkovost= alzv barh bdtr aabf
Babočka paví oko= alzv barh bdtr pavok
Dětskost= atsv
Ženskost= aahg aaba
Mužskost= aahg aabd
Lvicovost= acac akuc aaba levxx
Levhartovost= acac akuc levht

(024)
Tento sém je obvykle také čtyřznakový, se znakem + na prvním místě.
Legenda: +ccr – český výraz
Aaabd +ccr= přídavné jméno
Aaabd +sem= ahaf
Aaagt +ccr= sloveso
Aaagt +sem= aait
Aaavl +ccr= adverbium
Aaavl +sem= svun
Aaavl +ccr= příslovečné určení
Aaavl +sem= svun

(025)
záznak prvků ze selekčních jazyků a klasifikací v semových ekvivalentech.
Mezinárodní desetinné třídění= mdxxxxxxx
Mezinárodní třídění vynálezů= mpxxxxxx
Poštovní směrovací čísla ČR= msxxxxx
Smerovací čísla americké pošty= maxxxxxx

Poznámka: xxxx označuje konkrétní číselný, abecedněčíselný nebo abecední kód třídění.

(027)
Semy některých symbolů používaných v matematice a výpočetní technice.

Jméno symbolu	kód sému
+ (plus)	yxpl
- (mínus)	yxm
. (krát, násobek)	yxkr
: (děleno)	yxde
= (rovná se)	yxro
(integrál)	yxin
(diferenciál)	yxdi
(summa)	yxsu
--- (podíl ve zlomku)	yxzl
(mocnina, mocnitel)	yxmo
(odmocnina, odmocnitel)	yxod
(faktoriál)	yxfa
(logaritmus)	yxlo

(028)
Semové ekvivalenty:
Legenda: +ccc – slovní výraz v češtině
+aaa – slovní výraz v angličtině
+sem- sémový ekvivalent k výrazu v jazyce se stejným identifikátorem významu
na prvním místě řádku je kód identifikátotu významu
adrsa +ccc= ženský
adrsa +sem= aahg aaba ahaf
brkln +aaa= Scottland
brldn +ccc= Skotsko
brldn +sem= scott evrop
cbsk0 +ccc= Jarmila
cbsk0 +sem= aahg aaba jarmila
cbsk4 +ccc= Jarmilka
cbsk4 +sem= aahg aaba jarmila aabf

(029)
Ukázka ze souboru all (abecední část v češtině .) Uváděný příklad nemusí plně odpovídat současnému stavu souboru all.
Bnrk1 +ccc= hazardování
Bnrkl +aaa= gambling

Bnrk1 +sem= aa9r
Bnrk2 +ccc= házení
Bnrk2 +aaa= throwing
Bnrk2 +sem= anlp
Brnk3 +ccc= helikoptéra
Brnk3 +aaa= helicopter
Brnk3 +sem= aaxu trnp klvt
Brnk4 +ccc= Helsinky
Brnk4 +aaa= Helsinki
Brnk4 +sem= evrop finsk helsn aada
Brnk5 +ccc= herec
Brnk5 +aaa= actor
Brnk5 +sem= aahg aabd aakl aabz
Brnk6 +ccc= herectví
Brnk6 +aaa= dramatic art
Brnk6 +sem= aakl nzvd

(030)
V souboru vsechno nemáme dosud výraz žena. Sémanticky obecnější slovo je člověk. Sémový ekvivalent výrazu člověk zní aahg. Žena je samice od člověka. Sém samice zní aaba.
K sému aahg tedy připojíme sém aaba.
Výsledek: žena = aahg aaba
Při formulaci nových specifických sémů je nutno dbát na to, aby se neopakovaly sémy již existující s jiným významem. Ke kontrole jsou vytvořeny kontrolní programy.
Úkol utvořte nové sémvé ekvivalenty ke slovům klekáníce, stařec, novomanželka, OSN.

(031)
Úloha spočívá v hledání semů, které tvoří sémový ekvivalent slov klekánice, stařec, novomanželka, OSN.
Nalezneme sémové ekvivalenty ke slovům
1. strašidlo, žensský rod, poledne
10. člověk, sameckost. Stáří, subjekt
11. člověk, manželství, ženský rod, začínání, novost
12. institučnost, mezinárodnost

(032)
kontrola řešení čtyř nově vkládaných slov. Protože student nemá přístup k seznamu semů, zde uvádíme seznam semů, umožňujících řešení:

Klekánice: strašidlo= arvd klnm
Ženskost= aaba
Poledne= vlzn
Kledánice: arvd klnm aaba vlzn
Stařec
Novomanželka
OSN

(033)
české prefixy, kořeny (kmeny) a sufixy
+ccp pro české prefixy:
1ab23 +ccp= pro
1ab23 +sem= gpro
+ccq pro české kmeny:
2ndkl +ccq= člověk
2ndkl +sem= aagh
+ccs pro české sufixy:
0absz +ccs= ký
0absz +sem= ahaf
+ccc pro česká slovo a výrazy
2btk3 +ccc= kriminalita
2klb1 +ccc= nejrychlejší pohyb při couvání

(034)
Kořeny (kmeny, radixy) českých slov
Bko2a +ccq= muž
Bko2a +sem= baahg aazb
Bko2g +ccq= mužík
Bko2g= aahg aabd aazb

(035)
Prefixy vyjadřované již zapsanými semy
Prefix sémový ekvivalent
ne aane

(036)
Prefixy předznamenané znakem g
0125a +ccp= pro
0125a +sem= gpro
brwlb +ccc= ležet
brwlb +sem= kder
brz1a +ccc= proležet
brb1a + sem= kder gpro aalz

htfaf +ccc= proležely
hrfaf +sem= kder gpro ahaf

(037)
0bkma +ccs= ský
0bkma +sem= ahaf
0bkmb +ccs= ským
0bkmb +sem= ahaf
0bkmd +ccs= skými
0bkmd +sem= ahaf

(038)
Hotový překlad je možno opravovat a doplňovat.

Zvolíme z internetu nebo jiného média text v angličtině nebo jiném jazyce a jeho překlad do češtiny. Zvolený příklad ukazuje, že nejde o klasický překladový slovník, ale o pravděpodobnostně koncipovaný seznam vycházející z překladu.
Například:
1 krok

Český text	Anglický text
vyvstalá vrstva tuku na povrchu mléka,při pomalé aglomeraci emulgovaných tukových kuliček	film of fat which forms naturally on the surface of the milk by slow agglome

Anglický text
film of fat which forms naturally on the surface of the milk by slow agglomeration of emulsifying fat globules
2-krok
Významově stejné úseky v každém jazyce
Český text
vyvstalá vrstva tuku
na povrchu mléka,
při pomalé aglomeraci
emulgovaných tukových kuliček
Anglický text
film of fat which forms naturally
on the surface of the milk
by slow agglomeration
of emulsifying fast globules
3.krok
český text
11vrstva tuku
na povrchu 11
při pomalé 11
emulgovaných tukových 11

anglický text
film of fat 11
on the surface of 11
by slow 11
of emulsifying fast 11

(042)
Četnosti semů v povídce č.1 ve třetinách textu

Význam semu	1.třetina	2.třetina	3.třetina
Jméno	27	23	13
Živočich	6	3	9
Přírodní předmět	2	4	9
Zdrobnělina	10	2	1
Začínání	4	3	11
Ukončení	7	13	8
Mluvnická částicd	15	5	8

(047)
Dvojice A číslo
Dvojice B číslo
.
.
.
dvojice Y číslo
dvojice Z číslo

.
(048)
Vzor pro slovesa v češtině s infinitívní koncovkou –it

kmen-it= sémový+ekvivalent ahl
kmen-iti= sémový+ekvivalent ahl
ne-kmen-ívat= sémový+ekvivalent ahl ane cak
kmen-il= sémový+ekvivalent abv
kmen-ila= sémový+ekvivalent abv
kmen-ívaly= sémový+ekvivalent abv cak
ne-kmen-íval= sémový+ekvivalent abv ane cak
ne-kmen-ili-li= sémový+ekvivalent abv ane caa
ne-kmen-ily-li= sémový+ekvivalent abv ane caa
kmen-íval-li= sémový+ekvivalent abv cak caa
kmen-ívala-li= sémový+ekvivalent abv cak caa

kmen-ívala-li= sémový+ekvivalent abv cak caa
kmen-ívalo-li= sémový+ekvivalent abv cak caa
kmen-ívali-li= sémový+ekvivalent abv cak acaa
kmen-ívaly-li= sémový+ekvivalent abv cak caa
ne-kmen-íval-li= sémový+ekvivalent abv ane cak caa
ne-kmen-ívala-li= sémový+ekvivalent abv ane cak caa
ne-kmen-ívala-li= sémový+ekvivalent abv ane cak caa
ne-kmen-ívalo-li= sémový+ekvivalent abv ane cak caa
ne-kmen-ívali-li= sémový+ekvivalent abv ane cak caa
ne-kmen-ívaly-li= sémový+ekvivalent abv ane cak caa
kmen-ím= sémový+ekvivalent aig cab aca
kmen-íš= sémový+ekvivalent cac aca
kmen-íme= sémový+ekvivalent aig caa acd
kmen-íte= sémový+ekvivalent cac cab acd
kmen-íte= sémový+ekvivalent cad cab cd
ne-kmen-ejí= sémový+ekvivalent aig cab aca ane
ne-kmen-ím= sémový+ekvivalent aig cab aca ane
ne-kmen-íš= sémový+ekvivalent cac cab aca ane
ne-kmen-ímekmen-ání aig cab cd ane
ne-kmen-íte= sémový+ekvivalent cac cab acd ane
ne-kmen-ejí= sémový+ekvivalent cad cab acd ane
kmen-ím-li= sémový+ekvivalent cig cab aca caa
kmen-íš-li= sémový+ekvivalent cac cab aca caa
kmen-íme-li= sémový+ekvivalent aig cab acd caa
kmen-íte-li= sémový+ekvivalent cac cab acd caa
kmen-ejí-li= sémový+ekvivalent cad cab acd caa
ne-kmen-ím-li= sémový+ekvivalent aig cab aca ane caa
ne-kmen-íš-li= sémový+ekvivalent cac cab aca ane caa
ne-kmen-íme-li= sémový+ekvivalent ig ac2 cd ane caa
ne-kmen-íte-li= sémový+ekvivalent c3 ac2 cd ane caa
ne-kmen-ejí-li= sémový+ekvivalent c4 ac2 cd ane caa

(049)
79201 +aaa= automatic content analysis of art
79201 +ccc= automatická obsahová analýza umění
79202 +aaa= automatic content analysis of *11
79202 +ccc= automatická obsahová analýza *11
79203 +aaa= automatic content analysis
79203 +ccc= automatická obsahová analýza
79204 +aaa= content analysis of *11
79204 +ccc= obsahová analýza *11
79205 +aaa= content analysis of art
79205 +ccc= obsahová analýza umění

79206 +aaa= content analysis of art
79206 +ccc= obsahová analýza umění
79207 +aaa= analysis of art
79207 +ccc= analýza umění
79208 +aaa= automatic analysis of art
79208 +ccc= automatická analýza umění
79200 +aaa= automatic content analysis of *12
79200 +ccc= automatická obsahová analýza *12
7920a +aaa= *11 content analysis of *21
7920a +ccc= *11 obsahová analýza *21
7920b +aaa= automatic content analysis of *11 art
7920b +ccc= automatická obsahová analýza *11 umění
70001 +ccr= přistihování
70001 +sem= caaa
70002 +ccr= torpedování
70002 +sem= caab
70003 +ccr= kompostování
70003 +sem= caac
70004 +ccr= sklapování
70004 +sem= caad
70005 +ccr= tuhnutí
70005 +sem= caae
70008 +ccr= čpavkovost
70008 +sem= caaf
70009 +ccr= skalpování
70009 +sem= caag
7000a +ccr= slídění
7000a +sem= caah
7000b +ccr= bočnost
7000b +sem= caai
7000c +ccr= lombardskost
7000c +sem= lombr
7000d +ccc= lombardie
7000d +sem= italy evrop lombr
7000e +ccc= Lombarďan
7000e +sem= lombr aabd aahg italy evrop
7000f +ccr= lombarďanka
7000f +sem= lombr aaba aahg italy evrop
7000g +ccc= Lombarďana
7000g +sem= lombr aabd aahg italy evrop
7000h +ccr= lombarďanky
7000h +sem= lombr aaba aahg italy evrop
7000i +ccc= Lombarďanovi

7000i +sem= lombr aabd aahg italy evrop
7000j +ccc= lombarďance
7000j +sem= lombr aaba aahg italy evrop
7000k +ccr= kompostování
7000l +ccr= sklapování
7000m +ccr= tuhnutí
7000n +ccr= čpavkovost
7000o +ccr= skalpování
7000p +ccr= slídění
7000q +ccr= bočnost
7000r +ccc= co ještě jsi objevila
7000r +aaa= what more did you discovered
7000t +ccc= co ještě jsi *11
7000t +aaa= what more did you *11
7000u +ccc= co ještě jsi *11 *12
7000u +aaa= what more did you *11 *12

(050) Vzor pro podstat. jm. v češtině, vzor pán

kmen-sémový+ekvivalent abd aca paa
kmen-a=sémový+ekvivalent abd aca
kmen-ův=sémový+ekvivalent haf waa
kmen-ova= sémový+ekvivalent haf waa
kmen-vu= sémový+ekvivalent haf waa
kmen-i= sémový+ekvivalent abd acd
kmen-ů= sémový+ekvivalent abd acd

(051)
Ke příjmením přiřaďte sémové ekvivalenty, příčemž víte, že sem býti člověkem je aahg, sém býti mužem je aabd, sem býti ženou je aaba a semy jmen se v tomto případě rovnají znění jména bez háčků, čárek a velkých počátečních písmen (č= c9, š= s9, í= i6,
příjmení osob abecedně od Dan

054ab +ccc= Danyo
054ax +ccc= Danyová
054ad +ccc= Daněček
054ae +ccc= Daníček
054af +ccc= Daněček
054ag +ccc= Daníček
054ah +ccc= Daníček
054ai +ccc= Daněček
054aj +ccc= Danys

054ak +ccc= Danys
054am +ccc= Danys
054an +ccc= Danyś
054ao +ccc= Danyś
054ap +ccc= Danyś
054aq +ccc= Danyš
054ar +ccc= Danyš
054as +ccc= Danyš
054at +ccc= Daníšek
054au +ccc= Daníšková
054av +ccc= Daníška
054aw +ccc= Daníška
054ax +ccc= Daníška
054ay +ccc= Daníška
054az +ccc= Danz
054ba +ccc= Danzová
054bb +ccc= Danz
054bc +ccc= Danzer
054bd +ccc= Danzerová
054be +ccc= Dancer
054bf +ccc= Danzig
054bg +ccc= Danzigová
054bi +ccc= Danzinger
054bj +ccc= Danzingerová
054bl +ccc= Danzmajer
054bo +ccc= Danzmajeroová

(052)
Kontrola správnosti přidělení sémů příjmením při řečení úkolu z (051)

legenda:
člověk = aahg
muž= aabd
subjekt= aazb:
žena= aaba
Příjmení v prvém pádu jednotného čísla je uvedeno za kódy identifikátoru a českého jazyka. První písmeno příjmení je vždy velké.Sémantický ekvivalent příjmení je uveden za kódem identifikátoru a příslušnými sémy uvedenými v legendě a sémem jména s malým začátečním písmenem.
05401 +ccc= Danyo
05401 +sem= aahg aabd aazb danyo
05402 +ccc= Danyová
05402 +sem= aahg aaba aazb danyo

05403 +ccc= Daněček
05403 +sem= Daněček
05404 +ccc= Daníček
05404 sem= aahg aabd aazb dani6c9ek
05405 +ccc= Daníčková
05405 sem= aahg aaba aazb dani6c9ek
05406 +ccc= Daněček
05406 +sem= aahg aaba aazb dane9c9ek
05407 +ccc= Daněčková
05407 +sem= aahg aaba aazb dane9c9ek
05408 +ccc= Danys
05408 +sem= aahg aabd aazb danys
05409 +ccc= Danysová
05400 +sem= aahg aaba aazb danys
0540a +ccc= aahg aabd aazb danye9
0540a +ccc= Danyšová
0540a +ccc= aahg aaba aazb danys9
0540b +ccc= Daníšek
0540b +sem= aahg aabd aazb dani6s9ek
0540c +sem= aahg aaba aazb dani6s9ek
0540d +ccc= Daníška
0540d +sem= aahg aabd aazb dani6s9ka n
0540e +ccc= Daníšková
0540e +sem= aahg aaba aazb Daníška
0540f +ccc= Danz
0540f +sem= aahg aabd aazb danzx
0540g +ccc= Danzová
0540g +sem= aahg aaba aazb danzx
0540h +ccc= Danzer
0540h +sem= aahg aabd aazb danzer
0540i +ccc= Danzerová
0540i +sem= aahg aaba aazb danzer
0540j +ccc= Danzig
0540j= aahg aabd aazb danzig
0540k +ccc= Danzigová
0540k +ccc=aa aDanzig;
0540l +ccc= Danzinger
0540l +sem= aahg aabs aazb danzinger
0540m +ccc= Danzingerová
0540m +sem= aahg aaba aazb danzinger
0540n +ccc= Danzmajer
0540n +sem= aahg aabd aazb danzmajer
0540o +ccc= Danzmajerová

0540o +sem= aahg aaba aazb danzmajer