Magi

edison

9. dubna 2017

Obsah

I Teoretické pojednání **2**

1 Princetonský WordNet **3**
1.1 Motivace vzniku 3
1.1.1 Strojově čitelné slovníky 5
1.1.2 Od slovníků k WordNetu 5
1.2 K vlivu psycholingvistiky na organizaci WordNetu 6
1.3 Organizace WordNetu 7
1.3.1 Synsety a vztahy mezi nimi 8
1.4 Sémantické vztahy WordNetu 9
1.4.1 Hyperonymie a hyponymie 10
1.4.2 Meronymie a holonymie 10
1.5 Lexikální vztahy ve Wordnetu 11
1.5.1 Synonymie 11
1.5.2 Antonymie 13

2 Další wordnety **15**
2.1 EuroWordNet 15

Seznam literatury **20**

Část I

Teoretické pojednání

Kapitola 1

Princetonský WordNet

Princetonský WordNet je prvním wordnet vůbec. Vznikal na univerzitě v Princetonu pod G. A. Millerem od poloviny 80. let 20. století. Vzhledem k tomu, že byl prvním wordnetem, bylo k němu referováno jako k WordNetu, bez přívlastku. Ačkoliv tento stav v podstatě přetrvává dodnes, oproti době jeho vzniku se situace změnila, vzniklo několik dalších wordnetů a nastala tudíž potřeba je rozlišit. V anglickém prostředí se obvykle pojmem WordNet míní ten princetonský a všechny ostatní wordnety mají přívlastek či vlastní jméno. Příkladem nechť je Balkanet či Eurowordnet. Ačkoliv v mezinárodním prostředí je obvyklé přívlastek „princetonský“ používat, bude tato práce pracovat s následujícím rozlišením:

a samozrejme bych to actually mohl dodrzovat, tohle jsem si vymyslel az po napsani tehle kapitoly, lol

- *WordNet* (ve významu princetonský WordNet)
- *wordnet* (ve významu obecné sémantické sítě založené na WordNetu)
- konkrétní wordnety, např. *Balkanet*

1.1 Motivace vzniku

Od počátků snah o zpracování přirozeného jazyka (NLP, natural language processing) bylo nutné poskytnout programu data o lexiku ve zpracovávaném textu, ať už ona data byla jakákoliv. Kupříkladu pro překlad se mělo za to, že stačí ekvivalentní dvojice ve zdrojovém a cílovém jazyce, později se přidal kontext v případě statistického strojového překladu spolu s dalšími informacemi, jako je například slovní druh. Tradičně se lexikální materiál ukládá způsobem nikoliv diametrálně odlišným od papírových slovníků určených pro lidské uživatele. Ty obvykle obsahují abecenedně (či podle jiného indexu) seřazené jednotlivé záznamy s potřebnými informacemi o slovech, z nichž pak program může čerpat při zpracování textu.

cit?

Jak uvádí Pala; Ševeček, (2013), uspořádání lexikálního materiálu v takovéto formě je sice vhodné pro člověka, ale nikoliv pro strojové zpracování,

a to z několika důvodů. Kromě toho, že vyhledávání v abecedním seznamu je relativně pomalé , struktura tradičního slovníku kvůli onomu abecednímu řazení inherentně vzdaluje slova, jež člověk chápe jako nějakým způsobem blízká (Pala; Ševeček, 2013). Tato blízkost může vyplývat ze vztahu volné synonymie, antonymie, podřazenosti, nadřazenosti, etc. Pokud si tedy například uživatel výkladového slovníku nepříliš obeznámený s daným jazykem vyhledá určité heslo, dozví se sice pravděpodobně jeho význam, ale nebude schopen své znalosti prohlubovat dále kupříkladu zjištěním, jaké slovo odpovídá opačnému významu.

nejaka citace?, dohledat neco, jak takovy slovniky byly ulozeny...

Dalším všeobecným problémem při využití tradičních slovníků k počítačovému zpracování jazyka je fakt, že lexikografové předpokládají u uživatele slovníku značné encyklopedické znalosti. Zařazují tak do slovníku jen informace dle jejich názoru důležité pro rozlišení (differentia specifica) a zařazující do kontextu či přiřazující k určité nadřazené třídě objektů (genus proximum). Vyhledá-li si tedy člověk ve Slovníku spisovného jazyka českého heslo vlk, zjistí následující:

citace

vlk: psovitá šelma šedě (n. šedožlutě) zbarvená, žijící v Evropě, Asii a v Sev. Americe

Definice a priori předpokládá, že uživatel je obeznámen s tím, co je šelma a co je pes. Pokud takovou znalostí neslyne (což je vcelku představitelné například u cizince), je nucen si tato slova ve slovníku najít a podívat se na jejich definice (pomiňme nyní netriviální úkol převést slovo psovitá na základní tvar pes). Pokud nerozumí definicím ani nadřazených slov, musí pokračovat v hierarchii dále a dále.

Z uvedeného případu plyne, že jakkoliv je možné správným vyhledáváním hyperonym[1] dospět k tomu, že vlk je konkrétní entita našeho vesmíru, živá bytost o čtyřech končetinách, savec nějakým způsobem příbuzný se psovi, má šedou srst etc., je takový proces dosti komplikovaný. Případ s cizincem se sice nemusí zdát zcela relevantní, protože se dá předpokladat, že daný člověk má, byť v jiném jazyce, stejné základní znalosti předpokládané lexikografy jako člověk, jehož mateřštinou je čeština. Situace je však dramaticky jiná u počítače (přesněji u počítačového programu). Na rozdíl od člověka totiž počítač nemá žádné předchozí znalosti, tudíž musí projít celým procesem popsaným výše, aby byl schopen kupříkladu určit, že vlk může umřít (ježto je živá bytost) . Protože však tradiční slovníky typu SSJČ byly vytvářené pro papírové médium, neobsahují žádné propojení ve stylu *toto je odkaz na hyperonymum*, a počítač tudíž jen těžko může zjišťovat, na které vlastně slovo se to má podívat, aby se dobral podstaty pojmu vlk.

tohle je celkem myslenkovej skok a nevim, jestli to lze vubec vyvodit z dat wordnetu

[1] nadřazené slovo

1.1.1 Strojově čitelné slovníky

V zájmu automatizace vyhledávání ve slovníku vznikaly tzv. strojově čitelné slovníky[2], což je pojem souhrnně označující lexikální databáze. Podle množství informací, které taková databáze obsahuje, pak lze tyto dělit na slovníky, taxonomie a ontologie. Je evidentní, že obyčejný slovník neobsahuje oproti tradičnímu papírovému slovníku navíc žádné metainformace, takže je počítač při jeho užívání v podstatě omezen na elektronický listovač (Miller et al., 1990).

Míru, jakou se strojově čitelný slovník odliší od pouhé zdigitalizované formu papírového slovníku a přiblíží se k pokročilé lexikální databázi, lze vyjádřit v několika stupních. V případě, že slovník má jednotlivé významy uspořádány v hierarchii dle nadřazenosti–podřazenosti, lze jej označit za taxonomii, tedy systém s hlubší strukturou než pouze abecedním řazením hesel.

nejakej link, kde budou významy/senses vysvetleny

Dalším stupněm je již komplexní lexikální databáze, která má jednotlivé významy propojeny rozličnými vztahy, počínaje onou základní hyperonymií a hyponymií a pokračuje kupříkladu vztahy meronymie[3] či antonymie[4]. Kromě vztahů mezi významy bude taková lexikální databáze obsahovat zřejmě i další informace, například o syntaktických kategoriích slov, definice jejich významů, etc. Databáze tak popsaných významů propojených sémantickými vztahy může být nazývána ontologií. (Garshol, n.d.)

1.1.2 Od slovníků k WordNetu

Výše uvedená opozice papírového slovníku a ontologie ilustruje rozdíly tradičního slovníku a počítačově zpracovatelné lexikální databáze. Jedním z klíčových rozdílu je propojenost jednotek v lexikální databázi – tradiční slovníky, byvše v době svého vzniku většinou určeny pro distribuci v papírové formě určené pro lidského uživatele, neobsahují důsledné propojení sémanticky souvisejících slov. Příkladem budiž kostra a její části, např. lebka. V SSČ[5] i SSJČ se u lebky uvádí, že jde o kostru hlavy. Lze tedy s jistou rezervou tvrdit, že heslo obsahuje své holonymum[6], opačný odkaz však již ani jeden z oněch dvou slovníků neobsahuje. Z celkem evidentních ekonomických důvodů nejsou u hesla kostra uvedeny všechny její části. Tento příklad příhodně ukazuje i jistou nesystémovost tradičních slovníků, která je pro počítačové zpracování fatální, jelikož, jak bylo zmíněno výše, znemožňuje systémové procházení hierarchie slovní zásoby a zjišťování podstaty jednotlivých významů.

[2] machine readable dictionary
[3] vztah *je částí*, tedy např. dveře je meronymem trolejbusu
[4] protikladu
[5] Slovník spisovné češtiny
[6] vztah opačný k meronymii; tedy např. dům je holonymem pro okno, dveře, práh etc.

Naznačeny tedy byly vlastnosti, jež by lexikální databáze měla oproti tradičnímu slovníku mít, aby byla použitelná pro počítačové zpracování přirozeného jazyka. Především jde o systémovost vztahů. Hypero–/hyponymie je vztah oboustranný, tudíž by mělo být možné se stejnou cestou dostat od nadřazeného slova k podřazenému a naopak. Dále je podstatné, aby sémantické vztahy mezi významy byly přesně definované, a tudíž algoritmy zpracovatelné. Jedině tak je totiž možno jednoznačně určit, které slovo (či slova) je v takové databázi konkrétnímu slovu nadřazené, které je jeho specifikací, označením jeho částí, etc.

S touto myšlenkou vznikl WordNet – lexikální síť provázaná sémantickými vztahy, která dle poznatků psycholingvistiky odráží uspořádání lexikálního materiálu v lidském mozku (více v kap. 1.2 na straně 6). (Pala; Ševeček, 2013)

Zde by bylo na místě poznamenat, že ačkoliv se tak z odstavců výše může čtenáři jevit a i všeobecně je to často tvrzeno, WordNet není ontologií v pravém slova smyslu, protože něco něco.. `https://en.wikipedia.org/wiki/WordNet#WordNet_as_a_lexical_ontology`

a tady tomu vubec nerozumim, ale prijde mi to relevantni

1.2 K vlivu psycholingvistiky na organizaci WordNetu

Jelikož G. A. Miller, který byl koordinátorem projektu WordNet, byl svým zaměřením psycholog a přispěl k vzniku psycholingvistiky, ubíral se projekt Wordnetu podobným směrem. Společně s Johnson-Lairdem se Miller zaměřil na výzkum, jakým způsobem je lexikální materiál uložen v lidském mozku. Tento vědní směr je označován právě jako psycholingvistika a jeho počátky jsou spojeny s průzkumem asociací a modelem budování modelu mentálního slovníku člověka. Výchozí myšlenka, jež se odráží i ve způsobu organizace WordNetu, spočívá v tom, že slovní zásoba je konceptuálně (tedy že slova se stejným významem jsou seskupena u sebe) a pro některé slovní druhy (zejména substantiva) hierarchicky.

Jednou z otázek tohoto směru bylo, jakým způsobem je v hierarchickém modelu paměti řešeno získávání vlastností pro význam, které jsou „poděděné“ po významech hierarchicky výše umístěných. Aby člověk byl schopen například určit pravdivostní hodnotu výroku Kanárek může létat, musí použít svou dlouhodobou pamět. Její organizace je pak možná (minimálně) dvěma způsoby. První, redundantní, by vypadal tak, že by u každé podtřídy ptáků bylo uloženo, že její instance jsou schopny létat. Druhý, již na první pohled výrazně méně náročný na úložný prostor, by příznak schopnosti létat měl uložený pouze u třídy pták. Pro zjištění, zda kanárek létá, by pak bylo nutno zapojit inferenční proces ve stylu *kanárek je pták, tudíž může létat.* (Collins; Quillian, 1969)

Jak Collins; Quillian, (1969) dále uvádí, lze předpokládat, že v případě

prvního způsobu organizace paměti by člověk mohl kteroukoliv informaci o příznacích (vlastnostech) z paměti vyvolat za konstantní čas. Naproti tomu v případě způsobu druhého by extrakce příznaku z významu v hierarchii položeného výše měla trvat delší čas než extrakce příznaku přítomného přímo u významu, jenž je subjektem věty. Důvodem by měla být nutnost zapojení inferenčního procesu.

Pokus, kterým podpořili Collins; Quillian, (1969) druhý, neredundantní, způsob ukládání příznaků v paměti, spočíval v tom, že testovací subjekty, dobrovolníci z řad zaměstnanců společnosti Bolt Beranek and Newman, měly určovat, zda je jim předložený výrok pravdivý, či nepravdivý. Měli tak činit co nejpřesněji a v co nejkratším čase, přičemž byla měřena rychlost jejich reakce. Ukázalo se, reakční doba při určování pravdivosti výroku Kanárek umí létat[7] je delší než při určování pravdivosti výroku Kanárek umí zpívat[8] a ještě delší při určování výroku Kanárek má kůži[9]. Důvodem pro tyto progresivní prodlevy podle nich právě byla zvětšující se vzdálenost od významu kanárka ke významu, u něhož byl uložen příslušný příznak, tedy umí zpívat, umí létat, resp. má kůži. Příznak umí zpívat totiž je pravděpodobně uložen přímo u kanárka, jelikož jej odlišuje od ostatních ptáků, zatímco příznak umí létat je obecným znakem ptáků, tudíž je uložen u významu pták. V poslední řadě pak příznak má kůži bude patrně uložen u významu zvíře, který je oněch tří v hierarchii nejvýše, a ze všech tudíž od významu kanárek nejdále.

WordNet se svou hierarchickou organizací substantiv a verb pravděpodobně konceptuálně blíží organizaci lexika v lidské paměti.

1.3 Organizace WordNetu

Ve WordNetu lze nalézt informace autosémantikách, tedy substantivech, adjektivech, slovesech a příslovcích (Vossen, 1998). Synsémantika (např. předložky, spojky etc.) nebyla zahrnuta, jelikož se zdá, že jsou uložena odděleně od slov plnovýznamových. Teorii, že jsou funkční slova uchovávána jako součást syntaktikonu, podpořil kupříkladu Garrett, (1982) při svém pozorování afatických pacientů.

Vůbec první podnět k uvědomění, že různé slovní druhy podléhají různé strukturalizaci v paměti, vyvolal asociační test, který provedli Fillenbaum; Jones, (1965). Tomuto asociačnímu testu byli podrobeni anglicky mluvící subjekty, kteří měli za úkol uvést první slovo, které je napadne při myšlence na předložené slovo. Předkládána jim byla dobře známá a často používaná slova náležející k různým slovním druhům. Ukázalo se, že ve většině případů náleží asociované slovo ke stejnému slovnímu druhu jako slovo, které asociaci vyvolalo. Substantiva vyvolala asociaci na substantivum v 79 % případů,

[7] angl. A canary can fly
[8] angl. A canary can sing
[9] angl. A canary has skin

adjektiva v 65 % případů a slovesa v 43 % případů.

Ačkoliv není zřejmé, jak je znalost o slovním druhu určitého slova získávána, lze z uvedených dat předpokládat, že slovní druh je vskutku primární organizační vlastností lexikálního materiálu v lidském mozku a informace o něm je snadno dostupná (alespoň intuitivně). Jelikož správné tvoření vět vyžaduje alespoň intuitivní povědomí o tom, které slovo náleží ke které syntaktické kategorii, není s podivem, že tato informace je dostupná lidskému uvažování velmi jednoduše. Jelikož se však slova stejného slovního druhu příliš často nevyskytují pohromadě, není evidentní, jak tyto znalosti člověk získává. (Fillenbaum; Jones, 1965; Miller et al., 1990)

1.3.1 Synsety a vztahy mezi nimi

Slova (slovní formy) jsou ve WordNetu seskupována podle svého významu a slovního druhu, k němuž náležejí. Takové řadě slov se v terminologii WordNetu říká synset (synonym set), neboli synonymická řada. Každý synset reprezentuje jeden význam, ale je nutno mít na paměti, že granularita synsetů nemusí být konsistentní a v podstatě záleží na tom, jak si tvůrci zadefinovali synonymum (více v kap. 1.5.1 na straně 11). Synset je ve WordNetu reperezentací významu a je definován slovy (formami), které obsahuje. Jelikož význam slov je definován tím, v jakém synsetu se vyskytují (ke kterému konceptu náleží), jde v podstatě o kruhovou definici, a tudíž je zřejmé, že definice významů musí být rozšířena. Lze říci, že význam konceptu reprezentovaného synsetem je založen na jeho pozici v celé struktuře. Význam konceptu je tedy definován jeho kontextem, to znamená nadřazenými a podřazenými koncepty. (Kamps; Marx, 2002)

Aby bylo možno WordNet použít k inferenčnímu vyvozování závěrů (získávání informací) o slovech, a to strojově, což znamená bez nutnosti mít jakékoliv předchozí encyklopedické znalosti, které má obvykle uživatel tradičního slovníku k disposici, jsou synsety ve WordNetu propojeny vztahy, z nichž je zřejmé, jakou informaci inferenční stroj získá, přejde-li po onom vztahu k dalšímu konceptu.

Vztahy mezi koncepty jsou vztahy sémantické, jelikož se týkají významů slov (cf. lexikální vztahy níže).

Zmíněné kritérium, že slovní formy jednoho synsetu musí náležet k jedné syntaktické kategorii (slovnímu druhu), je podloženo jednoduchým závěrem o nezaměnitelnosti slov přináležejících různým slovním druhům.Seskupování konceptů podle slovního druhu a zřejmě navzdory snaze o ekonomii ukládání informací, kterou se lidský mozek vyznačuje, zprostředkovaně vede k jisté redundantnosti systému. Existuje totiž mnoho slov (zvláště např. v angličtině), která zastupují jak substantivum, tak verbum (např. angl. show, popř. české stát). Míra sémantické podobnosti takových slov může být značně odlišná. V angličtině je relativně běžné, že substantivum popisuje činnost, k jejímuž vyjádření se užívá sloveso stejné formy (např. run vyjadřuje běh a

běžet). U zmíněného českého stát sice lze vypozorovat poněkud vzdálenou sémantickou příbuznost (pojmenování pro stát jako základní územní mocenskou jednotku je zřejmě motivováno jako něco stálého, co dlouho *stojí*), ale není to příliš intuitivní a takové dva výrazy nemohou být zařazeny do stejného konceptu. Slova náležející do odlišných syntaktických kategorií se rovněž syntakticky chovají zcela rozdílně a rozhodně v žádném kontextu nemohou být zaměněna jedno za druhé, což také znemožňuje jejich zařazení do stejného synsetu. (Miller et al., 1990)

cit https://cs.wikipedia.org/wiki/Stát

Dalším argumentem pro striktní rozdělení slovní zásoby dle slovních druhů je fakt, že různé slovní druhy mají různou hierarchizaci. Jak bude popsáno v kapitole 1.4 na straně 9, například substantiva jsou hierarchizována podle vztahu hyperonymie a hyponymie, přičemž u nich existují další vztahy jako meronymie, která například u sloves existovat nemůže. Naopak vztah antonymie, který je relativně běžný u adjektiv, se u substantiv téměř nevysktuje[10]. Verba jsou zase provázána vztahy vyplývání, který u substantiv není příliš evidentní a intuitivní[11], ale u sloves je vcelku hojný – například z činnosti zírat vyplývá i *nadřazená* činnost hledět.

Sémantické relace mezi slovy různých kategorií ve WordNetu neexistují, avšak pro tyto případy jsou definovány relace lexikální. Oproti sémantickým relacím, které provazují celé koncepty, jsou lexikální relace definovány na úrovni jednotlivých forem. Dvě stejné formy, například run, jedna náležející k substantivům, druhá k verbům, budou propojeny vztahem derivačně příbuzné formy[12].

cit: https://wordnet.princeton.edu/wordnet/man/wngloss.7WN.html

1.4 Sémantické vztahy WordNetu

V této kapitole budou podrobněji rozebrány sémantické vztahy konstituující WordNet. Sémantické vztahy jsou, na rozdíl od vztahů lexikálních, které jsou vztahy mezi slovními formami, vztahy mezi koncepty (významy). Rozdíl nejlépe ilustruje protipříklad. Synonymie je typickým vztahem lexikálním; kdyby byla vztahem sémantickým, znamenalo by to, že dva různé významy (koncepty) mají stejný význam, což je nesmysl, jelikož v tom případě to nejsou dva významy, ale jeden.

Struktura těchto vztahů není, jak by se na první pohled mohlo zdát, plochá, ale organizovaná podle syntaktické kategorie významů, jež jsou jimi

[10] Lze argumentovat, že např. život je antonymem pro smrt, faktem ale je, že jde o velmi volnou antonymii – život popisuje stav či průběh doby, kdy je bytost živá, smrt referuje pouze k okamžiku, kdy se z živé bytosti stává mrtvá bytost, tedy rozhodně nejde o přímý protiklad jako například u adjektiv světlý:tmavý nebo špatný:dobrý. Stejně tak např. Bůh a Ďábel jsou sice proti sobě pokládané bytosti, ale jejich antonymie spočívá spíše ve vlastnostech jim připisovaných, tedy subjektivních, a uživatel jazyka může prohlásit, že obě tyto bytosti jsou špatné, čímž ztratí svou protikladnost.

[11] Asi lze tvrdit, že z života vyplývá smrt, ale pravděpodobně takto provázaných substantiv nebude mnoho.

[12] derivationally related form

propojeny. Substantiva mají své vlastní vztahy, stejně tak adjektiva, verba a adverbia. Pojmenování těchto vztahů vychází z lingvistických termínu k nim relevantních (např. hyperonymie) a v některých pojmenování některých se napříč různými syntaktickými kategoriemi překrývá, ačkoliv jde o vztahy různé. Například angl. sloveso run[13] má ve WordNetu jako hyperonymum uveden synset s významem pohybovat se velmi rychle obsahující slovesa travel rapidly, speed, hurry, zip[14]. Je evidentní, že tento vztah hyperonymie není identický se vztahem hyperonymie u substantiv, kde house[15] má jako přímé hyperonymum uveden synset building, edifice[16]. Z činnosti běžet vyplývá činnost rychle se pohybovat, ale budova je pro dům nadřazenou třídou. Jde tedy o vztah nikoliv nepodobný, ale ne identický.

cit wordnet 3.1 a mozna jeste neco...

1.4.1 Hyperonymie a hyponymie

Vztah nadřazenosti a podřazenosti strukturuje především slovní zásobu substantiv. Hyperonymie je vztahem třídy k podtřídě, hyponymie vztahem podtřídy k třídě. Jde o vztah transitivní a asymetrický. (Miller et al., 1990) Díky této hierarchizaci se lze například vyhnout redundanci ukládání informací v paměti, jelikož příznaky třídy není nutné ukládat u každé podtřídy. Podtřída dědí všechny příznaky své mateřské třídy a přidává minimálně jeden další. Například tramvaj je pouličním kolovým přepravníkem, který jezdí po kolejích a je poháněn elektřinou[17]. Pokud některý ze zděděných příznaků pro podtřídu neplatí, je tento fakt u ní explicitně uložen (podrobněji v kapitole 1.2 na straně 6). System, v němž jsou atributy takto děděny se nazývá dědičný systém[18] (Touretzky, 1986 - The mathematics of inheritance systems).

dohledat actual knizku

Substantiva jsou ve Wordnetu organizována tak, že každý význam má svůj mateřský význam (hyperonymum), kromě jednoho jediného, a tím je entity, tedy uměle vytvořený pojem sloužící jako kořen celé sítě. Jeden koncept může mít hyperonymních významů více, například house má jako svá hyperonyma uvedeny synsety (n) dwelling, home, domicile, abode, habitation, dwelling house a (n) building, edifice. Tato vlastnost mimochodem z WordNetu činí nikoliv stromovou strukturu, jak bývá často vizualizován, ale cyklický graf.

src: http://www.random-mhacks.net/2009/12/29/vi wordnet-relationships-as-graphs/ a cit neco verohodnejsiho, ale rozhodne bychom mohli namalovat s tim skriptem nejake pekne obrazky..btw, je to cyklicky, ze jo?

1.4.2 Meronymie a holonymie

Meronymie (a k ní komplementární vztah holonymie) jsou, navzdory nepříliš rozšířenému názvosloví, dalším vztahem, jenž je pro uživatele jazyka intuitivní a známý. Jde o vztah *být částí*, potažmo *mít část*. Meronymie

[13] čes. běžet
[14] čes. cestovat rychle, uhánět, ...
[15] čes. dům
[16] čes. stavba
[17] a wheeled vehicle that runs on rails and is propelled by electricity
[18] inheritance system

je definována tak, že A je meronymem B, pokud A je částí B. Meronymie je vztahem stejně jako hyperonymie transitivním a asymetrickým.(Cruse, 1986) Tento vztah také hierarchizuje lexikum do určitých úrovní, ale na rozdíl od vztahu nadřazenosti, v němž obvykle jeden význam mívá jeden až dva nadřazené významy, u vztahu části a celku by byla situace složitější. Je totiž na snadě, že jeden význam může být meronymem mnoha holonymům – kupříkladu dveře jsou meronymem u dům, auto, šatník, občas počítačová skříň, etc.

cit Cruse, 1986, ale vubec tomu nerozumim... http://i.imgur.com/h6NcRVJ.png

Vztah části a celku je vlastní výhradně substantivům.

1.5 Lexikální vztahy ve Wordnetu

1.5.1 Synonymie

Synonymie je základním definičním vztahem pro synsety ve WordNetu. Na praktických aplikacích je tento jev nejlépe pozorovatelný, jelikož při vyhledání konkrétní formy je uživateli obvykle nabídnut výběr z jednotlivých významů dané formy. Aby byly od sebe významy oné formy odlišitelné, nabídka běžně se obvykle sestává ze seznamu skupin slovních forem náležejících do nalezených synsetů[19] Kupříkladu při vyhledání slova kolo v českém wordnetu tak je uživatel konfrontován s několika skupinami, které obsahují zhruba následující:

- kolo (1),
- jízdní kolo (1), bicykl (1), kolo (2),
- kružnice (1), cívka (1), kolo (3),

přičemž čísla (zde) v závorce značí index významu dané formy v daném synsetu. Reprezentace v různých aplikacích a různých wordnetech se liší (standardem bývá číslo významu psát za dvojtečku), koncept však zůstává neměnný.

Navzdory zdánilivé jednoduchosti uvedeného konceptu je všeobecnou otázkou, jak synonymii pojímat. Striktní teorie (obvykle připisovaná Leibnizovi) praví, že dvě slova jsou synonymní, pokud se jejich záměnou nikdy nezmění pravdivostní hodnota výroku. Lingvistickou interpretací tohoto poněkud matematicko-logického výroku může pak být, že synonymní dvě slova jsou v případě, že se jejich záměnou nikdy neporuší význam (zhruba ona pravdivostní hodnota) a gramatičnost výroku. Je nasnadě, že takto striktně synonymní slova budou pospolu v jazyce těžko přežívat, jelikož je dokázáno, že jazyk tíhne k ekonomičnosti, která by takovým soužitím dvou slov byla hrubě porušena . Pravděpodobně jedinými obecně uznávanými synonymy

nejakou citaci na ekonomii jazyka...

[19] https://www.englishforums.com/English/AdjectiveSatellite/nwzhv/post.htm#1126701

jsou obvykle dvojice cizího slova a domácího slova, například internacionální a mezinárodní. Jejich záměnou se velice pravděpodobně nikdy pravdivostní hodnota výroku nezmění, stejně tak jako jeho gramatičnost. Stále však zůstává ve hře stylistika, která může být podobnou náhradou narušena (např. z důvodu cílové skupiny čtenářů či stylistické příznakovosti jednoho ze slov (cf. zajímavý a interesantní)). Co se tendence k ekonomičnosti jazyka týče, lze předpokládat, že v těchto případech převládá potřeba synonym k eliminaci opakování určitých slov v textu a tím zajištění jeho stylistické uhlazenosti.

Volnější interpretace synonymie počítá ještě s kontextem. Dvě slova jsou synonymní, jsou-li bez způsobení škod nahraditelná alespoň ve stejném kontextu. Jako příklad mohou posloužit formy board a plank. V kontextu dřevařství mohou tyto dvě formy pravděpodobně bez problému být nahrazeny jedna za druhou, ovšem v případě, že je forma board použita ve významu comittee, těžko ji lze nahradit formou plank, neboť by se věta obsahující takové nahrazení stala zcela nesmyslnou.

najit ceske priklady

Bylo by nanejvýš logické považovat synonymii za vztah diskrétní, tedy že dvě formy buďto synonymní jsou, či nejsou. Z logického hlediska to nepochybně z již uvedeného vyplývá, ovšem lingvisticko-filosofický náhled vycházející z poznatků reálného jazyka na tuto problematiku nahlíží poněkud odlišně. Synonymie v striktním slova smyslu je velice vzácná. Její volnější interpretace je značně častější, ale také výrazně vágnější – kontext, v němž dvě formy synonymní jsou, může být velmi široký, či naopak velice úzký. Záměna některých dvojic (či spíše obecně n-tic, volné synonymní řady mohou být vcelku dlouhé – textil:1, látka:1, textilie:2, plena:1, tkanina:1) může měnit stylistiku a význam výpovědi více či méně, přičemž ony dvě formy stále dle daných kritérií lze považovat za synonymní. Nelze tedy než vyvodit, že synonymie, minimálně z pohledu přirozeného jazyka, je jevem graduálním, a některé formy jsou tak *synonymnější* než jiné. (Miller et al., 1990)

cit. cesky WN

Zaměnitelnost forem podporuje ještě jeden koncept, na němž je WordNet postaven, a to fakt, že jednotlivé významy jsou seskupovány podle slovních druhů. Tento systém vede k jisté redundantnosti, jelikož zvláště v syntetických jazycích, jako je kupříkladu angličtina, lze nalézt mnoho případů, kdy identická slovní forma zastupuje více slovních druhů. Významy, které taková slovní forma zastupuje (napříč slovními druhy), mohou být velice blízké, nikdy však nebudou stejné (nelze říci, že význam slovesa run[20] a substantiva run[21] je identický). Jejich záměnou by se sice nestalo vůbec nic, jelikož čtenář či posluchač textu, v němž taková záměna nastala, by automaticky formu interpretoval ve prospěch správného slovního druhu, avšak pokud by slovní druh byl nějakým způsobem „vynucen“ (nechť nyní čtenář pomine úvahy, jakým způsobem lze *vynutit* slovní druh formy), stala by se výpověď zcela negramatickou a nesmyslnou.

check, nekecam?

[20] běžet
[21] běh

Jakkoliv to není přímo svázané se synonymií, je na místě poznámka o výskytu stejné formy v různých synstetech. Slovo je kombinací slovní formy a významu, nebo slovního významu. Slovní forma je projevem „fyzickým“, tedy je to vyřčená či napsaná instance významu. Jak je zjevné z přirozeného jazyka, nelze počítat s tím, že by zobrazení významu na formu bylo bijektivní, tedy každý význam byl namapován jedna ku jedné na slovní formu. V přirozeném jazyce může jedna forma zastupovat více významů a jeden význam může být vyjádřen více formami. Příkladem budiž slovní forma koruna, která může zastupovat význam měny, vrcholku stromu, vladařského odznaku, etc. Toto zobrazení jedné formy na více významů se nazývá polysémií nebo homonymií[22]. S polysémií souvisí ještě homonymie, což ve své podstatě dosti podobný vztah, ale totožnost formy je zcela nahodilá. Kupříkladu formu kolej lze interpretovat jako referenci k stopě po voze, případně dvojici kolejnic jako vodící dráze pro dopravní prostředky a zároveň jako zařízení vysoké školy pro ubytování studentů.. U významů formy koruna lze vypozorovat nějaký společný základ (koruna stromu je nahoře, panovnickou korunu má panovník na hlavě, tedy nahoře, koruna jako mince zase pravděpodobně získala své pojmenování díky faktu, že na mincích bývá vyobrazen panovník). Naproti obě formy kolej pochází z odlišeného základu – kolej jako ubytovací zařízení pochází z latinského collegium, kdežto výraz pro dráhu je odvozeno od českého kolo .

Seskupování významů podle slovních druhů a seskupování forem dle vztahu synonymie se tedy zdá v případě lexikální databáze určené pro strojové zpracování jako vhodným konceptem. Oproti tradičním slovníkům se totiž počítačově zpracovávaná lexikální databáze nemusí potýkat s problémem lidského faktoru – jednotlivé synonymické řady je stroj schopen prohledávat, na rozdíl od člověka, velice účině, a nahradí tak v případě, že WordNet používá člověk, neúčinné lidské procházení restříkového obsahu.

ne, neni, ale nechtelo se mi mazat 2k napsanych znaku xD

cit. SSJC

cit SSJC

cit: etymolog. slovnik, ale jeho online verze to neuvadi

psano v chvatu, mohlo by se to mozna trochu uhladit...

1.5.2 Antonymie

Antonymie, neboli protiklad, je navzdory zdánlivé triviálnosti koncept překvapivě těžce definovatelný. Všeobecně se antonymií rozumí významová opozice, faktem však je, že použití tohoto termínu je velmi široké a druhů antonymie je několik. Nejjednodušším druhem je například antonymie mezi adjektivy živý a mrtvý. Negace prvého automaticky značí druhé a naopak (je-li řeč o živých bytostech), jelikož v reálném světě neexistuje žádný další třetí stav. Tento jednoduchý vztah však nefunguje vždy – například s adjektivy bohatý a chudý je to jiné. Mnoho lidí se nepovažuje ani za chudé, ani za bohaté, a tudíž z toho, že někdo není bohatý, automaticky neplyne to, že by byl chudý. Miller et al., (1990) Zajímavé je, že tento vztah není reflexivní. Pokud někdo není bohatý, tak to nemusí znamenat, že je chudý,

[22] obojí znamená totožnost formy pro různé významy, u polysémie však ony významy mají společný základ (byť může být velmi vzdálený)

ale pokud je o někom tvrzeno, že *je* bohatý, tak to nutně znamená, že *není* chudý. Paradis; Willners, (2006)

Rozdíl mezi výše uvedenými dvojicemi, tedy mrtvý:živý a chudý:bohatý spočívá ve stupňovatelnosti daných adjektiv. Pro ilustraci – lze říci, že někdo je bohatší než někdo jiný, ale nelze říci, že někdo je *mrtvější* než někdo jiný. Pokud jsou adjektiva stupňovatelná, tedy lze říci, že objekt A je více X než objekt B, neoznačují komplementární stav, ale graduální vlastnost. Označované pak může být zařazeno kamkoliv mezi tyto dva póly, přičemž nachází-li se v pomyslné střední šedé zóně, nelze jej označit výrazy odpovídajícími pólům gradientu. Tvrzení, že někdo není ani chudý, ani bohatý, dává smysl, protože tato adjektiva označují extrémní stavy, mezi nimiž je prostor pro normální stav. Paradis; Willners, (2006)

Vztah antonymie ve WordNetu je koncipován tak, aby zřejmě byl co nejpodobnější uvažování široké populace uživatelů jazyka, tedy užívá primitivního konceptu antonymie. Některé studie dokonce za antonymní považují výrazy pouze vágně, intuitivně protikladné, jako například muž:žena či chytrý:hloupý. (A. Lehrer; K. Lehrer, 1982)

Ve WordNetu se antonymie vyskytuje u substantiv (man:woman), adjektiv (rich:poor, a dokonce i white:black v rasovém významu[23]), verb (open:close) i adverbií (well:ill).

[23] cf. také antonymní vztah Caucassian:black ve WN

Kapitola 2

Další wordnety

Podle vzoru princetonského WordNetu začaly postupně vznikat i další sémantické sítě založené na stejném konceptu. Tyto sémantické sítě se samozřejmě svou strukturou do větší či menší míry liší , hlavním kritériem pro to, aby mohly být považovány za wordnet, je to, aby obsahovaly synsety a hyponyma. (*Wordnets in the World*, 2017)

nakou kurva citaci

2.1 EuroWordNet

EuroWordNet je mezinárodní lexikální databáze pro osm evropských jazyků (angličtina, čeština, dánština, francouzština, italština, němčina, španělština). Jde o soubor jednotlivých národních wordnetů, které jsou propojeny takzvaným mezijazykovým indexem (ILI, *inter-lingual-index*). Obecně jsou wordnety Eurowordnetu založené svou strukturou na princetonském WordNetu (verze 1.5), ale z důvodu různorodosti jazyků se v některých aspektech od něj odlišují.

Základní motivací pro vznik EuroWordNetu byla evropská jazyková různorodost a z ní pramenící problémy ve zpracování dat a napomáhání uživateli v přístupu k neanglickým datům. Vossel 1999 (Vossen-Eurowordnet.pdf, pg XX) argumentuje, že uživatel musí umět anglicky a být obeznámen s tím, jak je zdroj, v němž vyhledává napsán, aby byl schopen v něm účinně hledat. Vytvořením wordnetů pro jiné jazyky si slibuje, že se zlepší možnost přístupu uživatelů k neanglickým datům, možnosti inference znalostí z těchto dat a případně i mezijazykové vyhledávání. Poslední je založeno na faktu, že od počátku byly jednotlivé wordnety EuroWordNetu vytvářeny s tím, že budou propojeny na základě základních konceptů (BCS, *Base Concepts*) a mezijazykového indexu.

rikam to dobre? vubec jim nerozumim

Jelikož se jednotlivé jazyky zapojené v projektu EuroWordNetu značně odlišují ve struktuře své slovní zásoby, jsou jednotlivé wordnety nezávislé. To znamená, že se mohou odlišovat například svou hierarchizací. Stejný koncept tak může ve dvou různých wordnetech mít různá hyperonyma, meronyma,

Relace	Slovnědruhové kombinace	Příklad
antonymie	A-A, V-V	open:close
hyponymie	N-N, V-V	car:vehicle, walk:move
meronymie	N-N	head:nose
vyplývání[1]	V-V	buy:pay
následek	V-V	kill:die

Tabulka 2.1: Vztahy přejaté z princetonského WN (N: substantivum, A: adjektivum, V: verbum)

etc., protože například anglické označení pro prst je odlišené, pokud jde o prst na noze (toe), či na ruce (finger). Podobně má v jiném příkladu dánština odlišené označení hlavy u zvířat vyjma koní, tedy kof, a hlavy lidské a koňské (hoofd) . (Vossen, 1997)

nekde jsem to cetl, ale nemuzu to dohledat (a slovnik to popira, tak to mozna bude jiny jazyk... dunno, TB fixed

Národní wordnety jsou vzájemně propojené přes mezijazykový index s anglickým wordnetem, který je obsahově založený na princetonském WordNetu, ale není identický. Anglický wordnet byl přizpůsoben strukturně tak, aby byl použitelný v EuroWordNetu, tedy byly přidány dodatečné metainformace a druhy vztahů (podrobněji dále). V národních wordnetech existuje několik druhů konceptů, které jsou rozlišeny podle příbuznosti s koncepty v ostatních národních wordnetech. Pokud je koncept přítomen ve všech wordnetech EuroWordNetu, jde o koncept tzv. *Global Base Concept* (GBC). Koncept, jenž jen přítomen v alespoň dvou národních wordnetech je označován jako *Common Base Concept* (CBC) a v poslední řadě koncept, který se vyskytuje pouze v jednom národním wordnetu nese označení *Local Base Concept* (LBC). (*GWA Base Concepts*, 2017) Propojení konceptů společných pro více jazyků je zajištěno pomocí jednotných identifikátorů a mezijazykového indexu, který je nadmnožinou všech konceptů v EuroWordNetu. ILI je hierarchicky plochá struktura (proto *index*, nejde o další „všejazykový“ wordnet). (Vossen, 1997)

Jelikož v době, kdy EuroWordNetu vznikal, byl princetonský WordNet poněkud omezený mimo jiné co se vztahů mezi slovními druhy týče, vznikly pro EuroWordNet speciální vztahy umožňující úplnější práci s významy. Základní vztahy přejaté z princetonského WordNetu 1.5 jsou uvedeny v tabulce 2.1 na straně 16.

Navíc k těmto vztahům byly přidány štítky (*labels*), jež relaci konkretizují. Byly použity následující štítky:

nikdy jsem to nevidel, pls je to nekde v nasich WN na debdictu?

- conjunction/disjunction
- non-factive
- reversed
- negation

Použití konjunktivního a disjunktivního štítku spočívá v myšlence, že například u meronymie by bylo vhodné rozlišovat, zda jde o části, které dohromady tvoří celek, nebo jde o podčásti částí (např. nůž má meronyma čepel, rukojeť, ostří, ale ostří je ve skutečnosti meronymem až rukejeti, nikoliv přímo samotného nože).

Štítek *non-factive* je používán u kauzální relace, která nemusí být nutně naplněna:

příčina	vztah	následek	non-factive?	nutně vyplývá?
zabít	*vyústí v*	zemřít	–	ano
hledat	*vyústí v*	najít	non-factive	ne

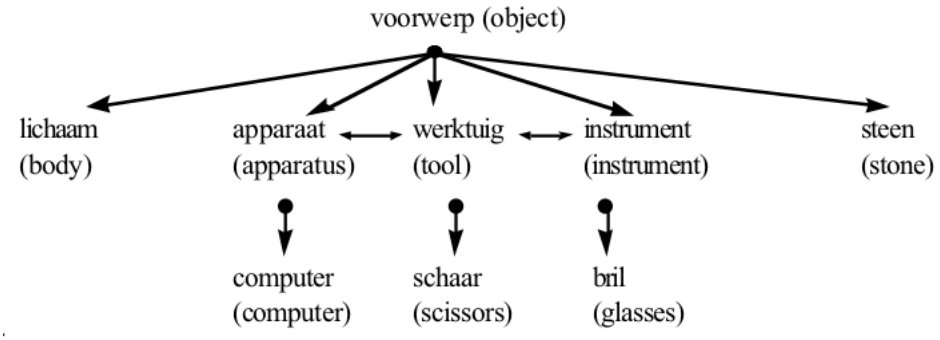
Podobně lze upřesnit pomocí štítků další relace tak, že jsou ve výsledku jednoznačnější a wordnet, v němž jsou takto označené vztahy obsaženy, může poskytovat více informací.

Jako další vylepšení oproti tehdejší verzi princetonského WordNetu přinesl EuroWordNet také relace mezi slovními druhy a vztah blízkého synonyma . Argument pro zavedení mezislovnědruhových relací je relativně přímočarý, a to, že umožňují „sblížit“ koncepty, které jsou si příbuzné, jen náleží k jinému slovnímu druhu. Nutno podotknout, že v době psaní této práce je princetonský WordNet ve verzi 3.1 a obsahuje už relaci derivationally related form, která zajišťuje přesně toto propojení (více o synsetech v princetonském WordNetu v kapitole 1.3.1 na straně 8). Co se vztahu blízkého synonyma (near synonym) týče, důvodem pro jeho zavedení byl údajně zájem mít možnost přiblížit koncepty, které jsou si významově podobné, ale pouze na své úrovni. U takových konceptů platí, že byť jejich význam je podobný, jejich hyponyma nelze zařadit pod jeden koncept, jelikož se rozdíl mezi oněmi koncepty svázanými vztahem blízkého synonyma prohlubuje. Příkladem budiž trojice nizozemských slov aparaat, werktuig a instrument, jež jsou si významově nepříliš vzdálená:

vubec netusim, jestli to chapu spravne, spis mi prijde, ze ne..

jak to prelozit? nezda se mi, ze by to melo s timhle cokoliv spolecneho: https://goo.gl/7Zw1SG

hmm, a co near antonym? to be investigated



Obrázek 2.1: Blízká synonyma (k překreslení)

Jak je z obrázku 2.1 na straně 17 zřejmé, všechna hyponyma slova vo-

orwerp (objekt) jsou si rovna, avšak některá jsou si rovnější . Tři výše zmíněná slova jsou si navzájem významově výrazně bližší, než jsou si blízká s ostatními hyponymy na jejich úrovni. Právě aby bylo možno tento vztah reflektovat a tím docílit možnosti například nahrazovat za sebe slova, která sice nemohou být ve stejném synsetu, ale jsou si podobná, byl zaveden vztah blízkého synonyma. Lze totiž předpokládat, že uživatel jazyka podobná slova také může zaměnit. (Vossen, 1997)

smim v diplomce delat reference na kvalitni literaturu? xD

cela ta kapitola je v podstate z toho vossena, tak nevim, jak to citovat :/

poznamky

Jak moc mam rozebirat další wordnety (resp. site WN, jako je Eurowordnet, balkanet...)?

Jak dal: Dal bych chtel rozebrat nejak obecne vizualizace WN (ze ne vsechno lze nacpat do jedny), jaky se obecne voli pristupy (fuck everything, delame jen hyponymii, vetsinou), a pak vybrat par nejakejch dobre pristupnejch a/nebo originalnich, ktery rozebrat, proc jsou spatny...

co nechapu:

- adjektiva (jakysi model s kolem (stred kola nejakej core), sprinclikama k obvodu (vztahy k "satelite adjectives"), osa k dalsimu "stredu kola"...)
- nevim, kde najit rozdily mezi Princeton WN verzema (1.5 vs 2.1 vs. 3.0)

Seznam literatury

Collins, Allan M; Quillian, M Ross, 1969. Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*, 8(2): s. 240–247.

Cruse, D Alan, 1986. *Lexical semantics*.

Fillenbaum, Samuel; Jones, Lyle V, 1965. Grammatical contingencies in word association. *Journal of Verbal Learning and Verbal Behavior*, 4(3): s. 248–255.

Garrett, Merrill F, 1982. Production of speech: Observations from normal and pathological language use. *Normality and pathology in cognitive functions*, s. 19–76.

Garshol, Lars Marius. Metadata? Thesauri? Taxonomies? Topic Maps! *2004—10—26)[2010-6—19] http://www. ontopia. net/topicmaps/materials/tm-VS—thesauri. htm1*. Dostupné také z: ⟨`http://www.ontopia.net/topicmaps/materials/tm-vs-thesauri.html#N773`⟩.

GWA Base Concepts [online]. [Cit. 2017-04-06]. Dostupné z: ⟨`http://globalwordnet.org/gwa-base-concepts/`⟩.

Kamps, Jaap; Marx, Maarten, 2002. Visualizing wordnet structure. In: *Proc. of the 1st International Conference on Global WordNet*. S. 182–186.

Lehrer, Adrienne; Lehrer, Keith, 1982. Antonymy. *Linguistics and philosophy*, 5(4): s. 483–501.

Miller, George A et al., 1990. Introduction to WordNet: An on-line lexical database. *International journal of lexicography*, 3(4): s. 235–244.

Pala, Karel; Ševeček, Pavel, 2013. Česká lexikální databáze typu WordNet. *Feedback*, 48(A47).

Paradis, Carita; Willners, Caroline, 2006. Antonymy and negation—The boundedness hypothesis. *Journal of pragmatics*, 38(7): s. 1051–1080.

Vossen, Piek et al., 1997. EuroWordNet: a multilingual database for information retrieval. In: *Proceedings of the DELOS workshop on Cross-language Information Retrieval*. S. 5–7.

Vossen, Piek, 1998. Introduction to eurowordnet. *Computers and the Humanities*, 32(2-3): s. 73–89.

Wordnets in the World [online]. [Cit. 2017-03-25]. Dostupné z: ⟨`http://globalwordnet.org/wordnets-in-the-world/`⟩.