

Magi

edík

3. 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	10
1.5.1	Synonymie	11
1.5.2	Antonymie	13
2	Další wordnety	15
2.1	EuroWordNet	15
2.2	BalkaNet	15
	Seznam literatury	16

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 přívlasktu. 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 přívlasktu či vlastní jméno. Příkladem necht je Balkanet či Eurowordnet. Ačkoliv v mezinárodním prostředí je obvyklé přívlasktu „princetonský“ používat, bude tato práce pracovat s následujícím rozlišením:

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

a samozřejmě bych to actually mohl do-
držovat, tohle jsem si vymyslel az po na-
psani tehle kapitoly, lol

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 abecedně (či podle jiného indexu) seřazené jednotlivé záznamy s potřebnými informacemi o slovech,

cit?

z nichž pak program může čerpat při zpracování textu.

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 něco, jak takový slovníky byly uloženy...

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 hyperonym¹ 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. Příklad 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 vlastné slovo se to má podívat, aby se dobral podstaty pojmu vlk.

tohle je celkem myšlenkový skok a nevím, jestli to lze vůbec vyvodit z dat wordnetu

¹nadřazené slovo

1.1.1 Strojově čitelné slovníky

V zájmu automatizace vyhledávání ve slovníku vznikaly tzv. strojově čitelné slovníky², 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 odlišuje od pouhé zdigitalizované formy 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 vý-
znamy/sen-
ses
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 meronymie³ či antonymie⁴. 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 rozdílů 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 SSČ⁵ i SSJČ se u *lebky* uvádí, že jde o *kostru hlavy*. Lze tedy s jistotou rezervou tvrdit, že heslo obsahuje své holonymum⁶, 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.

²machine readable dictionary

³vztah *je částí*, tedy např. *dveře* je meronymem *trolejbusu*

⁴protikladu

⁵Slovník spisovné češtiny

⁶vztah 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
vůbec
nerozumím,
ale přijde
mi to
relevantní

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 „podděně“ po významech hierarchicky výše umístěných. Aby člověk byl schopen například určit pravdivostní hodnotu výroku *Kanárek může létat*, musí použít svou dlouhodobou paměť. 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 nej přesněji 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.

Aby bylo možno WordNet použít k inferenčnímu vyvozování závěrů (získávání informací) o slovech, a to strojově, což znamená bez nutnosti mít jakékoli předchozí encyklopedické znalosti, které má obvykle uživatel tradičního slovníku k dispozici, 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 vypožorovat 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)

Dalším argumentem pro striktní rozdělení slovní zásoby dle slovních druhů je fakt, že různé slovní druhy mají různou hierarchizaci. Jak bude

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

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 nevyskytuje¹⁰. Verba jsou zase provázána vztahy vyplývání, který u substantiv není příliš evidentní a intuitivní¹¹, 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 formy¹².

cit:
<https://wordnet.princeton.edu/wordnet/man/wn-g-loss.7WN.html>

1.4 Sémantické vztahy WordNetu

V této kapitole budou podrobněji rozebrány sémantické vztahy konstituující WordNet. Struktura těchto vztahů není, jak by se na první pohled mohlo zdát, plochá, ale organizovaná podle syntaktické kategorie významů, jež jsou jimi propojeny. Substantiva mají své vlastní vztahy, stejně tak adjektiva, verba a adverbia. Pojmenování těchto vztahů vychází z lingvistických termínů 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 možná
jeste
neco...

¹⁰Lze 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:dobry*. Stejně tak např. *Bůh* a *Ďábel* jsou sice proti sobě pokládány 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.

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

¹²derivationally related form

¹³čes. *běžet*

¹⁴čes. *cestovat rychle, uhánět, ...*

¹⁵čes. *dům*

¹⁶čes. *stavba*

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 elektřinou¹⁷. Pokud některý ze zděděných příznaků pro podtřidu 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 systém¹⁸ (Touretzky, 1986 - The mathematics of inheritance systems).

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 Word-Netu činí nikoliv stromovou strukturu, jak bývá často vizualizován, ale cyklický graf.

dohledat
actual
knížku

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 je definována tak, že A je meronymem B, pokud A je částí B. Meronymie je vztahem stejně jako hyperonymie transitivním a asymetrickým. (Cruse, 1986)

src:
http://www.ran-
do-
mhacks.net/2009/12/29/vi-
wordnet-
relationships-
as-graphs/
a cit neco
verohod-
nejsiho,
ale
rozhodne
bychom
mohli
namalovat
s tím
skriptem
nejake
pekne ob-
razky..btw,
je to
cyklicky,
ze jo?

1.5 Lexikální vztahy ve Wordnetu

Lexikální vztahy jsou, na rozdíl od vztahů sémantických, které jsou vztahy mezi významy (koncepty), vztahy mezi slovními formami. Rozdíl nejlépe ilustruje protipříklad – kdyby synonymie 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.

¹⁷a wheeled vehicle that runs on rails and is propelled by electricity

¹⁸inheritance system

cit Cruse,
1986, ale
vubec
tomu ne-
rozumim...
http://i.im-
gur.com/h6NcRVJ.png

1.5.1 Synonymie

Synonymie je centrálním organizačním vztahem pro substantiva ve Word-Netu. 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 synsetů¹⁹ 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 zdánlivé 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 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

nejakou citací na ekonomii jazyka...

najít české příklady

¹⁹<https://www.englishforums.com/English/AdjectiveSatellite/nwzhv/post.htm#1126701>

jedna za druhou, ovšem v případě, že je forma **board** použita ve významu **committee**, těžko ji lze nahradit formou **plank**, neboť by se věta obsahující takové nahrazení stala zcela nesmyslnou.

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“ (nechtě nyní čtenář pomine úvahy, jakým způsobem lze *vynutit* slovní druh formy), stala by se výpověď zcela negramatickou a nesmyslnou.

check, ne-
kecam?

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 homonymií²². S polysémií souvisí ještě homonymie, což ve své

ne, není,
ale nech-
telo se mi
mazat 2k
napsaných
znaku xD

²⁰běžet

²¹běh

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

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 vypořizovat 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 odlišného základu – *kolej* jako ubytovací zařízení pochází z latinského *collegium*, kdežto výraz pro dráhu je odvozeno od českého *kolo*.

cit. SSJC

cit. SSJC

cit: etymolog. slovník, ale jeho online verze to neuvádí

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 účinně, a nahradí tak v případě, že WordNet používá člověk, neúčinné lidské procházení restříkovaného obsahu.

psáno v chvatu, mohlo by se to možná 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 prvního 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, 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 významu²³), verb (**open:close**) i adverbií (**well:ill**).

²³cf. také antonymní vztah **Caucasian: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 .

nakou
kurva ci-
taci

<http://globalwordnet.org/wordnets-in-the-world/>

2.1 EuroWordNet

2.2 BalkaNet

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.htm>. Dostupné také z: <http://www.ontopia.net/topicmaps/materials/tm-vs-thesauri.html#N773>.
- 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, 1998. Introduction to eurowordnet. *Computers and the Humanities*, 32(2-3): s. 73–89.