

Vizualizace sémantické sítě

edison

27. dubna 2017

Obsah

I	K sémantickým sítím	3
1	Princetonský WordNet	5
1.1	Motivace vzniku	5
1.1.1	Strojově čitelné slovníky	7
1.1.2	Od slovníků k WordNetu	7
1.2	K vlivu psycholingvistiky na organizaci WordNetu	8
1.3	Organizace WordNetu	9
1.3.1	Synsety a vztahy mezi nimi	10
1.4	Sémantické vztahy WordNetu	11
1.4.1	Hyperonymie a hyponymie	12
1.4.2	Meronymie a holonymie	12
1.5	Lexikální vztahy ve Wordnetu	13
1.5.1	Synonymie	13
1.5.2	Antonymie	15
2	Další wordnety	17
2.1	EuroWordNet	17
II	Přehled a porovnání existujících vizualizací sémantických sítí	21
3	Metodologie porovnání	23
3.1	Výběr rozhraní	23
3.2	Strukturalizace přehledu a kritéria hodnocení	24
3.3	Podmínky testování	25
3.4	WordNet Search jako základ porovnání	25
4	Vizualizace s webovým rozhraním	27
4.1	An interactive visualization of the Princeton WordNet database	27
4.2	WordNET Editor	29
4.3	Cornetto Demo	30
4.4	WordVis	33
4.5	sloWTool	36

<i>OBSAH</i>	2
4.6 BabelNetXplorer	38
Bibliografie	43

Část I

K sémantickým sítím

Sémantické sítě, neboli wordnety, jsou lexikální databáze vytvořené s rozličnými záměry, mezi něž patří například i strojová inference informací v počítačovém zpracování přirozeného jazyka. Ve wordnetu jsou slova slučována podle významů do synonymických okruhů a tyto okruhy propojovány sémantickými vztahy, čímž dostávají svému označení sémantické sítě.

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 sa-
mozrejme
bych to
actually
mohl do-
drzovat,
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 [PŠ13], 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 [PŠ13]. 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 vlastně 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 [Mil⁺90].

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 popsanych významů propojených sémantickými vztahy může být nazývána ontologií. [Gar]

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.

Naznačeny tedy byly vlastnosti, jež by lexikální databáze měla oproti tradičnímu slovníku mít, aby byla použitelná pro počítačové zpracování při-

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.

rozeného 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 8). [PŠ13]

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 vu-
bec nero-
zumím,
ale přijde
mi to rele-
vantní

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*. [CQ69]

Jak Collins; Quillian [CQ69] 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 [CQ69] 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 nejpresněji 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 [Vos98]. 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 [Gar82] 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 [FJ65]. 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, 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

⁷ angl. A canary can fly

⁸ angl. A canary can sing

⁹ angl. A canary has skin

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. [FJ65; Mil⁺90]

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 13). Synset je ve WordNetu reprezentací 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. [KM02]

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 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 vyzorovat 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. [Mil⁺90]

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 11, 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. 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 propojeny. Substantiva mají své vlastní vztahy, stejně tak adjektiva, verba a adverbia. Pojmenování těchto vztahů vychází z lingvistických termínů k

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

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 word-
net 3.1
a možná
jestě
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 transitivity a asymetrický. [Mil⁺90] 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 podtřidu neplatí, je tento fakt u ní explicitně uložen (podrobněji v kapitole 1.2 na straně 8). System, v němž jsou atributy takto děděny se nazývá *dědičný systém*¹⁸ (Touretzky, 1986 - The mathematics of inheritance systems).

dohledat
actual
knížku

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.randomhacks.net/2009/12/29/visualizing-wordnet-relationships-as-graphs/>
a cit neco verohodnějšího, ale rozhodne bychom mohli namalovat s tím skriptem nejake pekne obrázky..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* je definována tak, že A je *meronymem* B, pokud A je částí B. *Meronymie* je vztahem stejně jako *hyperonymie* transitivity a asymetrickým. [Cru86]

¹³čes. *běžet*

¹⁴čes. *cestovat rychle, uhánět, ...*

¹⁵čes. *dům*

¹⁶čes. *stavba*

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

¹⁸inheritance system

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.

Vztah části a celku je vlastní výhradně substantivům.

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

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 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

nejakou citaci na ekonomii jazyka...

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

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 *interesanťní*)). 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.

najít české příklady

Bylo by nanejvýš logické považovat synonymii za vztah diskretní, 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, textile: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. [Mil⁺90]

cit. český 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, nekecam?

Jakkoliv to není přímo svázané se synonymií, je na místě poznámka o výskytu stejné formy v různých synstetech. Slovo je kombinací slovní formy

ne, není, ale nechcelo se mi mazat 2k napsaných znaku xD

²⁰běžet

²¹běh

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 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 vypořádat 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

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.

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

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. [Mil⁺90] 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 [PW06]

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

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 [PW06]

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*. [LL82]

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. [Theb]

nakou
kurva ci-
taci

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
dobře? vu-
bec jim
nerozu-
mím

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	Slovnědruhov \acute{e} kombinace	Př \acute{e} klad
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). [Vos97]

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). [Thea] 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). [Vos97]

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 18.

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

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

nekde jsem to cetl, ale nemuzu to dohledat (a slovník to popírá, tak to možná bude jiný jazyk... dunno, TB fixed

nikdy jsem to neviděl, pls je to nekde v našich WN na debdictu?

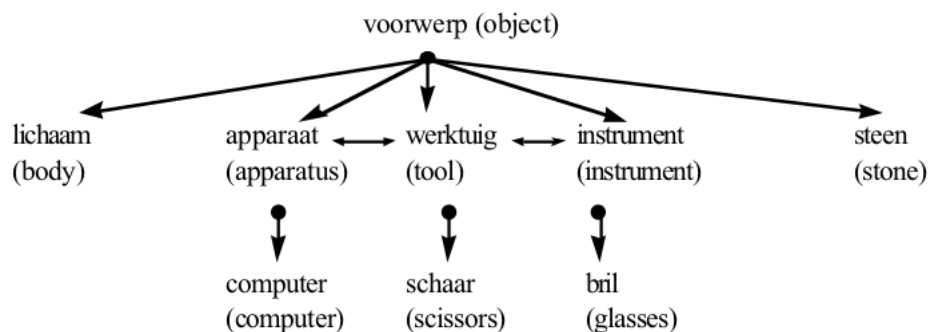
Použití konjunktivního a disjunktivního štítku spočívá v myšlence, že například u meronymie by bylo vhodné rozlišovat, zda jde o části, které dohromady tvoří celek, nebo jde o podčásti částí (např. *nůž* má meronyma *čepel*, *rukojeť*, *ostří*, ale *ostří* je ve skutečnosti meronymem až *rukejeti*, nikoliv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

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

příčina	vztah	následek	non-factive?	nutně vyplývá?
zabít	<i>vyúští v</i>	zemřít	–	ano
hledat	<i>vyúští v</i>	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 10). 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:



Obrázek 2.1: Blízká synonyma (k překreslení)

Jak je z obrázku 2.1 na straně 19 zřejmé, všechna hyponyma slova vo-

vubec netuším, jestli to chapu správně, spis mi přijde, že ne..

jak to prelozit? nezda se mi, ze by to melo s timhle cokoliv spolecneho: <https://google/7Zw1SG>

hmm, a co near antonym? to be investigated

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. [Vos97]

smim v diplomce delat reference na kvalitni literaturu? xD

poznámky

Jak moc mam rozebírat 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...

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

co nechapu:

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

Část II

Přehled a porovnání existujících vizualizací sémantických sítí

Tato část textu bude zaměřena na zmapování existujících dostupných rozhraní a jejich zhodnocení. Výčet jednotlivých rozhraní rozhodně není vyčerpávající, jelikož některé vědecké insituce mohou vyvíjet své nástroje pro práci se sémantickými sítěmi neveřejně, některé nástroje už nejsou dostupné, etc. Nástroje pro vizualizaci² dat wordnetů byly vybírány pro přehled v této práci podle několika kritérií relevantních především pro koncového uživatele, avšak i pro případné vývojáře dalších wordnetů (tedy zda dané rozhraní mohou použít pro svá data).

²Terminologická poznámka: v této práci je vizualizací dat wordnetů míněna jak textová reprezentace, tak grafická.

Kapitola 3

Metodologie porovnání

3.1 Výběr rozhraní

Hlavním kritériem pro zařazení do rozhraní do výběru byla jeho dohledatelnost podle vyhledávacího dotazu na indexovací a vyhledávací webové stránce <https://www.google.com>. Dotazy byly voleny tak, aby bylo dosaženo co nejvyšší relevance vyhledaných výsledků:

- wordnet visualization
- wordnet visualisation
- wordnet graph

Z přehledu byly samozřejmě vyřazeny implementace, které k době hledání¹ už nebyly dostupné. Také nebyla zahrnuta rozhraní, která jsou funkčně podobná jiným. Ačkoliv většina existujících rozhraní zřejmě pracuje buďto výhradně s princetonským WordNetem, či jej jako zdroj dat nabízí jako jednu z možností, nebylo na toto bráno ohled. Podstatné z hlediska použitelnosti však je, zda je zdrojový kód rozhraní dostupný a je možné jej použít i pro vizualizaci jiné sémantické sítě než kupříkladu princetonského WordNetu. Výsledkem tohoto průzkumu dostupných implementací rozhraní je XX vizualizací dat wordnetů:

- An interactive visualization of the Princeton WordNet database (<http://mateogianolio.com/wordnet-visualization/>)
- WordNET Editor (<http://wordventure.eti.pg.gda.pl/wne/wne.html>)
- A Graphical WordNet Browser (<http://homepages.inf.ed.ac.uk/adubey/software/wnbrowser/index.html>)

¹20. dubna 2017

- Visual Browser (<https://nlp.fi.muni.cz/projekty/visualbrowser/>)
- Cornetto Demo (http://cornetto.clarin.inl.nl/wordnet.xql?ssID=&word_form=&pos=&sense_pos=1)
- WordVis (<http://wordvis.com/>)
- sloWTool (<http://nl.ijs.si/slowtool/>)
- BabelNet (<http://babelnet.org>)
- GRAPH WORDS online thesaurus (<http://graphwords.com/>)

Pro účely porovnání (určení východiska) budou do popisu zahrnuta ještě jedno rozhraní, a to oficiální vyhledávací rozhraní princetonského WordNetu (<http://wordnetweb.princeton.edu/perl/webwn>).

well, base-line

3.2 Strukturalizace přehledu a kritéria hodnocení

Výše vypsaná rozhraní budou v dalších kapitolách rozdělena podle toho, zda umožňují přístup z webového prohlížeče (nativně, bez nutnosti mít nainstalované prostředí Java), budou zhodnoceny jejich positiva a negativa z hlediska použitelnosti pro získání informací o hledaném výrazu a v neposlední řadě také podle toho, jak kvalitní jejich uživatelské rozhraní je.

neja-
kej link,
možna vy-
svetlení?

K rozdělení podle dostupnosti bylo přistoupeno jednak proto, že v době psaní této práce je pokročilost webových technologií dostatečná k tomu, aby podobné vizualizace byly tvořeny jako webové stránky, a jednak proto, že cílem této práce je vytvořit všeobecně dostupné a použitelné rozhraní k wordnetům. Všeobecně použitelným je míněna použitelnost nejen na osobních počítačích, ale také na mobilních zařízeních, což v podstatě vyřazuje použití technologií jako jsou zásuvné moduly Java používané k realizaci různých existujících rozhraní (příkladem budiž Visual Editor).

Rozhraní jsou hodnocena na základě několika kritérií. Cílem je porovnat jejich přínos v kontextu ostatních existujících rozhraní a v kontextu současných trendů webových aplikací a ilustrovat, s jakými problémy se všeobecně rozhraní potýkají. Z tohoto důvodu právě nebyla do přehledu zařazována rozhraní, která se podobají svou funkcionalitou a ovládáním rozhraním již zařazeným. Kritéria hodnocení v této práci jsou následující:

- přínos oproti základnímu oficiálnímu rozhraní princetonského WordNetu
 - originální vizualizace dat vedoucí k možnosti identifikovat trendy v datech, které zůstávají uživateli skryty případně základní textové reprezentace dat

- reprezentace dat vhodnější z hlediska zásad přístupnosti webu²
- responsivita rozhraní – kvalitní použitelnost rozhraní i na zařízeních s menší obrazovkou, jako jsou mobilní zařízení typu chytrý telefon či tablet
- v případě webových aplikací ukládání stavu aplikace v adresním řádku prohlížeče (umožňuje sdílení nebo založení odkazu ke konkrétnímu hledání a zachování nastavení aplikace)

3.3 Podmínky testování

Pro testování rozhraní, které bylo pro účely této práce provedeno, byl použit operační systém Ubuntu 16.04 LTS v 64bitové verzi, webový prohlížeč Pale Moon ve verzi 27.1.1 (rovněž 64bitový) s nastaveným uživatelským agentem na Firefox (z důvodů kompatibility). Ačkoliv se počítalo, vzhledem k nepříliš rozšířenému vykresovacímu jádru toho prohlížeče, s možnými problémy v zobrazení některých rozhraní, nebyly tyto zjištěny, a tak nebylo nutné provádět testování rozhraní ještě v dalších prohlížečích. Lze předpokládat, že v rozšířených prohlížečích, jako jsou Firefox či Chrome, budou jednotlivá rozhraní fungovat a vypadá podobně, jako v prohlížeči Pale Moon.

Testování na mobilním chytrém telefonu bylo prováděno na zařízení Nexus 5 (displej s úhlopříčkou 5" a rozlišením 1080 × 1920 px) s operačním systémem Android 6 v prohlížeči Chrome (sestavení 57.0). Možnost *request desktop site* byla nastavena na *vypnuto*.

3.4 WordNet Search jako základ porovnání

Oficiální rozhraní k princetonskému WordNetu je koncipováno jako textová reprezentace dat, přičemž, byť volitelně, umožňuje zobrazovat veškeré k hledanému výrazu relevantní informace, jež WordNet obsahuje. Data jsou vizualizována jako soubor seznamů.

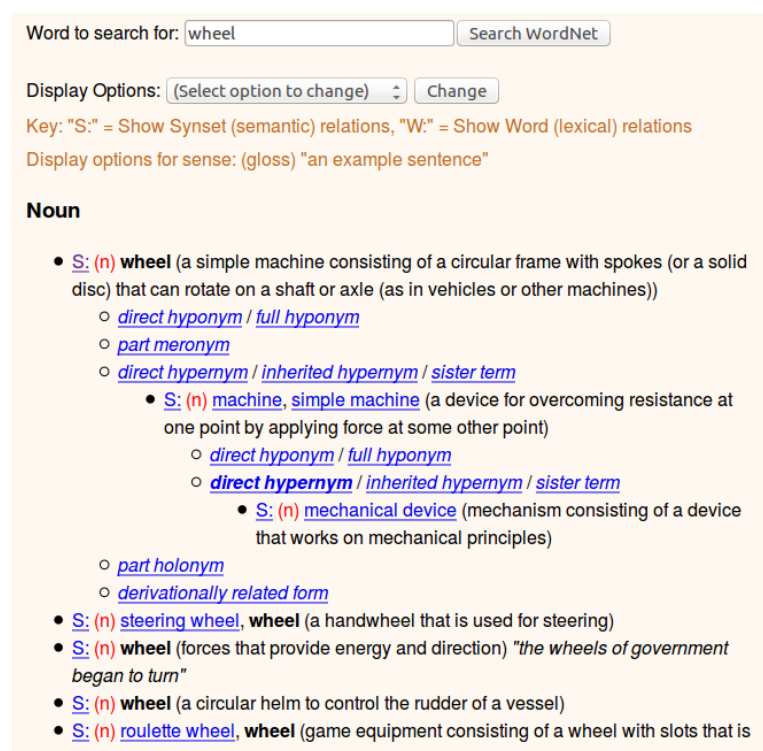
Po vyhledání zadaného výrazu jsou uživateli prezentovány v bodech jednotlivé synsety obsahující hledanou slovní formu rozdělené podle slovních druhů, k nimž přináležejí. V základním nastavení každý bod sestává ze slovních forem tvořících daný synset, jeho definici, případný příklad užití a odkaz na detaily synsetu. Kliknutím na tento odkaz se zobrazí v podseznamu sémantické a lexikální relace, v nichž je daný synset přítomen. Samotné relace jsou opět odkazy, jejichž otevřením je uživateli prezentován seznam synsetů, které jsou spolu s daným synsetem v dané relaci přítomny. Postupným otevíráním detailů synsetů v podseznamech se lze zanořovat hlouběji a hlouběji

²web accessibility

do struktury, resp. se ve struktuře vzdalovat původnímu synsetu po hranách relací.

Uživatel má možnosti si zvolit množství informací, jež jsou mu prezentovány, a to od úplného minima (pouze slovní formy nalezených synsetů a jejich syntaktické kategorie) až po vše, co se ve WordNetu vyskytuje, včetně identifikátorů a dalších technických detailů. Rozhraní svůj stav ukládá jako parametry URL³, takže je možné jej později obnovit z adresy.

Rozhraní není responsivní, ale vzhledem k tomu, že je v základu relativně úzké (pod 700 px), je na chytrém telefonu použitelné.



Obrázek 3.1: Ukázka rozhraní: WordNet Search 3.1 v základním nastavení a s otevřenými detaily synsetů

³Uniform Resource Locator

Kapitola 4

Vizualizace s webovým rozhraním

Jako bylo naznačeno v úvodu této části, vizualizací s webovým rozhraním se v této práci míní taková implementace, která ze strany uživatele nevyžaduje instalaci žádných doplňujících aplikací či aplikačních prostředí kromě samotného webového prohlížeče.

lze takhle nazvat Javu?

4.1 An interactive visualization of the Princeton WordNet database

Jeden z projektů programátora jménem Mateo Gianolio z Lundské univerzity (Švédsko). Jde o jednoduché rozhraní napojené na princetonský WordNet, které po zadání hledaného výrazu zobrazí synsety obsahující daný výraz. Jednotlivé synsety jsou barevně odlišeny podle syntaktických kategorií (slovních druhů), k nimž náležejí, a uspořádány kruhově kolem hledaného výrazu. Z každého synsetu je vždy zobrazena jen první slovní forma. Pokud uživatel najede kurzorem myši na některý ze synsetů, zobrazí se další případné slovní formy náležející do daného synsetu a jeho glosa (definice), je-li dostupná.

src:
<https://www.linkedin.com/in/mateo-gianolio-a89558b6/?trk=profile-badge-name>

Další slovní formy v nalezených synsetech jsou klikatelné, což umožňuje dostat se přes ně na synsety obsahující onu slovní formu, na níž bylo uživatelem kliknuto (s identickým výsledkem, jako kdyby dané slovo uživatel zadal do vyhledávacího pole).

Ačkoliv úvodní odstavec na stránce s rozhraním vybízí uživatele k „prozkoumání desambiguace slov“, zobrazují výsledky hledání pouze synsety a slovní formy do nich náležející, přičemž synsety připojené sémantickými vztahy hyponymie, meronymie etc. nejsou zobrazitelné (jinak než případným náhodným výskytem jedné slovní formy ve více synsetech). To použití WordNetu omezuje na obyčejný thesaurus.

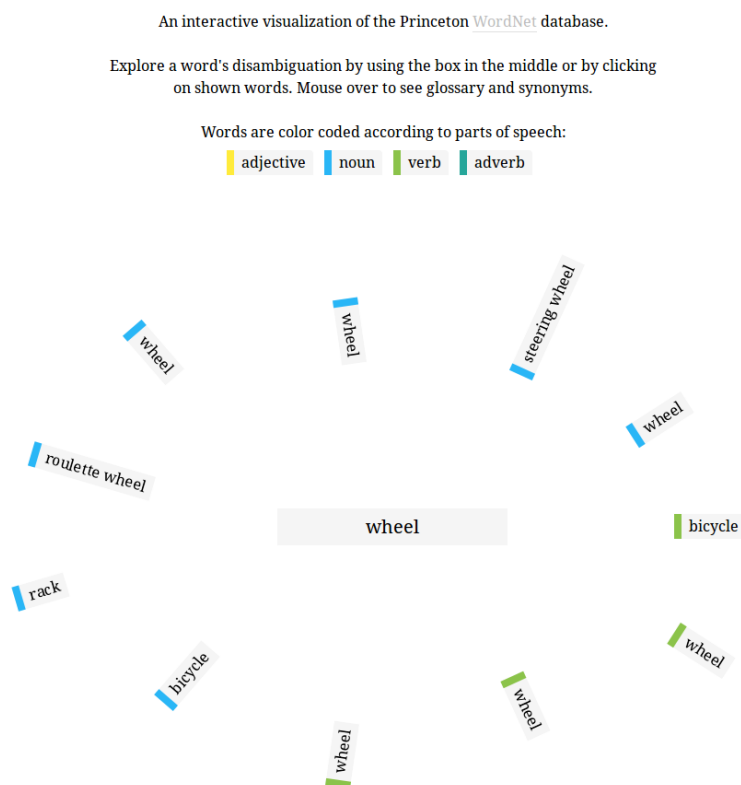
cit web

Samotný design grafického rozhraní je řešen poněkud nešťastně a působí dojmem, že cílem autora bylo prokázat své schopnosti používat různé standardní transformační funkce CSS. Pro indikaci běžícího procesu vyhledávání byl použit efekt rotační animace pro vyhledávací pole a po dokončení vyhledávání jsou výsledky zobrazeny paprskovitě okolo vyhledávacího pole. To způsobuje, že některé texty jsou zobrazeny pod úhlem až 90 stupňů, což může pro některé uživatele činit jejich přečtení obtížnějším.

Rozhraní není tzv. responsivní, tedy nepřizpůsobuje se velikosti obrazovky, na níž je zobrazeno.

Toto rozhraní tedy poskytuje velice omezený přístup k datům WordNetu a neposkytuje žádné výhody oproti základnímu oficiálnímu rozhraní k princetonskému WordNetu. Jeho grafické pojetí je čistě arbitrární, neslouží žádnému účelu a naopak zhoršuje jeho použitelnost.

odkazat na kapitulu o CSS nekde dal a možná to, že to je fakt v CSS...



Obrázek 4.1: Ukázka rozhraní An interactive visualization of the Princeton WordNet database

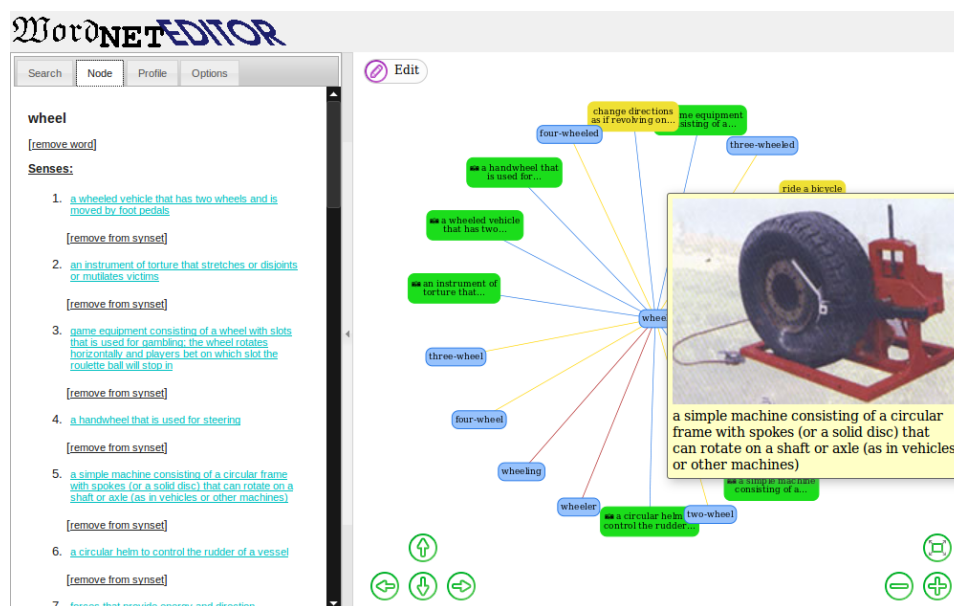
4.2 WordNET Editor

Webová aplikace WordNET Editor byla vytvořena s cílem poskytnout internetové komunitě praktický a účinný nástroj pro rozvoj WordNetu. Její autoři argumentují, že jelikož vývoj WordNetu vyžaduje velké množství lidské práce, a je proto důležité jej tvořit tak, jako je tvořena internetová encyklopedie Wikipedie, tedy kolaborací nezávislých uživatelů. [SDB07] K tomu je však potřeba nástroj, který takovou spolupráci umožní, což podle nich oficiální nástroje princetonského WordNetu nejsou (mimo jiné pravděpodobně proto, že WordNet byl vyvíjen uzavřenou skupinou lidí [Fel05]). Vedle editoru vztahů a synsetů však obsahuje WordNET Editor i prohlížeč wordnetu, pro který bylo rozhraní do tohoto přehledu zařazeno. Data, jež aplikace používá, jsou verzi princetonského WordNetu odvětvenou v jeho verzi 2.1. [SDB07]

Rozhraní je rozděleno do levého sloupce a pravého (většího) „plátna“, jež slouží pro grafické vyobrazení vztahů. Po vyhledání zadaného výrazu nabídne aplikace uživateli seznam synsetů, v nichž se výraz nachází. Nabídka synsetů je klikatelná a po uživateli je po zvolení synsetu kliknutím prezentováno grafické vyobrazení daného synsetu se všemi jeho vztahy. V levém sloupci si pak uživatel může vybírat, další synsety, které chce do pravé sekce přidat, může také otevírat dvojklikem již zobrazené synsety a lze i v nastavení filtrovat, které vztahy a slovní druhy se mají zobrazovat. Aplikace však zřejmě neobsahuje intuitivní způsob jak zobrazené synsety z plátna vpravo odebírat zobrazené synsetu a při otevírání dalších plátno nepřepisuje, ale synsety přidává k existujícím (neprovázaně). To může vést ke značné nepřehlednosti reprezentace, avšak aplikace umožňuje přibližování a oddalování plátna a vizuální přesouvání zobrazení po plátně, což s jistým úsilím uživatele nepřehlednost může eliminovat. Také nutno vytknout absenci textového pojmenování vztahů mezi synsety. Barevné rozlišení je v tomto případě nevhodně zvolené, jelikož vztahů je v rozhraní hodně a některé barvy jsou obtížně odlišitelné.

Grafické zobrazení vztahů mezi koncepty je řešeno hvězdčovitě, přičemž ve středu je zvolený synset, jenž byl výsledkem hledání, a od něj jsou vedeny hrany představující vztahy k příbuzným synsetům. Zajímavý je originální koncept zobrazení synsetů příbuzných k hledanému (otevřenému) synsetu. Ten je v grafu reprezentován svými členskými slovními formami, ale příbuzné koncepty jsou reprezentovány svými definicemi. Trochu na závalu však je, že se zřejmě nedá (minimálně v prohlížečím režimu) zobrazit detail příbuzného synsetu a uživatel se dozví pouze jeho definici, nikoliv slovní formy do něj náležející. Jediný detail, který je uživateli dostupný, je u některých synsetů obrázek (o jeho zdroji se však v rozhraní nepíše).

Test chování tohoto rozhraní na mobilním zařízení nebyl proveden, jelikož je zjevné, že je určeno především k editacím a podle toho je také navrženo. Rozhraní poskytuje grafický náhled na strukturu dat, ale pro nevhodné



Obrázek 4.2: Ukázka rozhraní WordNET Editor: zvolený synset ve středu grafu a zobrazený obrázek k příbuznému synsetu

kódování vztahů do barev a nezobrazování dostatečného množství informací je jeho ovládání neintuitivní. Nutno podotknout, že rozhraní tak, jak je vyobrazeno v Szymanski et al. [SDB07], vypadá značně odlišně a například obsahuje pojmenované vztahy.

Podle webové stránky projektu [je rozhraní vyvinuto s otevřeným zdrojem](http://wordnet-structure.eti.pg.gda.pl/), pro jeho obdržení je však nutno kontaktovat vývojáře. Lze ale předpokládat jeho použitelnost i pro ostatní wordnety.

cit
http://wordnet-
structure.eti.pg.gda.pl/

4.3 Cornetto Demo

Cornetto Demo je prohlížeč pro nizozemský wordnet a jako jedno z mála dostupných webových aplikací kombinuje grafickou a textovou vizualizaci dat. To je dáno zřejmě mimo jiné tím, že Cornetto Demo není pouze rozhraním pro wordnet, ale kombinuje data ze dvou zdrojů. Jednak z Referentie Bestand Nederlands [MM05], a jednak z nizozemského wordnetu. Referentie Bestand Nederlands je slovník obsahující informace podobně strukturované jako FrameNet [FBS04] spolu s rozšířením o kombinatorické chování slov v určitém významu [HVR].

Rozhraní je rozděleno na tři moduly, základní vyhledávání, pokročilé vyhledávání a vizualizace synsetů.

Základní vyhledávání slouží k prostému vyhledávání lexikálních jednotek. Dotaz je možné základním způsobem omezovat či rozšiřovat, a to pomocí

zástupných znaků (*, ?¹) a volby slovních druhů, v jakých se má vyhledávat. [Cor]

Pokročilé vyhledávání umožňuje v nizozemském wordnetu najít lexikální jednotky, které mají společné specifické parametry. Jako kritéria pro vyhledávání lze zvolit kterékoliv kombinace všech příznaků, jež jsou ve wordnetu u lexikálních jednotek přítomny. Takto je možné například vyhledat všechna slovesa, která jsou řazena do domény tance a jsou označena jako archaická.

Pokud hledaný výraz či zvolená kritéria odpovídají některým lexikálním vyskytujícím se v databázi, je uživateli prezentován seznam nalezených jednotek, přes něž se lze prokliknutím dostat na detail té které jednotky. Detail je textovou reprezentací dostupných dat, tedy syntaktických a sémantických informací o vyhledaném slově (nutno podotknout, že značná část informací v této části zřejmě pochází spíše Referentie Bestand Nederlands, nikoliv z wordnetu) a hierarchického zařazení konceptu, do něž slovo patří, ve wordnetu. Hierarchické zařazení je vyvedeno ve formě podobné tradičnímu znázornění adresářového stromu a obsahuje, zřejmě z úsporných důvodů, pouze přímé nadřazené a přímé podřazené koncepty a v rozhraní chybí úplné zobrazení cesty od kořene wordnetu k zobrazenému synsetu. I tak může strom být relativně dlouhý vzhledem k tomu, že u obecných hyperonym bývá seznam hyponym rozsáhlý.

Simple Search	Advanced Search	Synset Visualization	Help
---------------	-----------------	----------------------	------

Simple Search	Matching Lexical Entries	Selected LE	Related LE or Synset
---------------	--------------------------	-------------	----------------------

Details of Related Lexical Entry (rad-n-2)




General	Word Forms	Sense
Lemma: rad	Form	Definition: wiel v.e. voertuig
Part of Speech: noun	rad	het singular
Morpho-syntax: n	raderen	de plural
	raden	de plural

Semantics	Syntax	Relations
Countability: count	(No syntactics)	(No sense relations)
Reference: common		(No form relations)
Semantic Type: artefact		

Examples

Type/Phrase	Canonical Form
freeCombination / np	de raderen van een treinstel/fiets

Hierarchy

 schijf-n-1
 rad-n-2 wiel-n-1
 looswiel-n-1

Obrázek 4.3: Ukázka rozhraní Cornetto Demo: textová reprezentace dat

Z odkazů v hierarchii lze kliknutím na ikonu wordnetu u synsetu přejít do

¹ přičemž * zastupuje posloupnost kterýchkoliv znaků, ? zastupuje jeden kterýkoliv znak

vizualizačního režimu, čímž se zobrazí grafické znázornění daného synsetu.

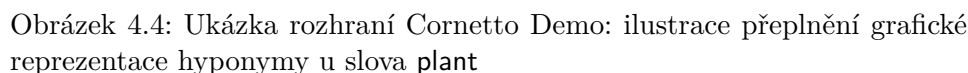
Modul vizualizace synsetů je v praxi pouze pseudomodul, jelikož jeho vyhledávání se od základního vyhledávání liší pouze tím, že vybírá jeden výsledek vyhledávání a rovnou jej zobrazí v grafické znázornění. Který výsledek se má vybrat, je možné zvolit při inicializaci hledání. Je tak eliminována nutnost zvolit synset, otevřít jeho detaily a kliknout na výše zmíněnou ikonu, aby se uživatel dostal na grafické znázornění daného synset, jedná se však pouze o zkratku v rozhraní, jež nezavádá nic nového.

Samotná vizualizace je realizovaná pomocí javascriptové knihovny d3 [BOH11] a umožňuje uživateli prohlížet všechny synsety svázané nějakým vztahem s vyhledaným synsetem. Zobrazení rozlišuje syntaktické kategorie i sémantické vztahy barevně, přičemž vztahy jsou navíc ještě popsány slovně. Autoři použili rozšířený model vizualizace synsetů tak, že synset je představován uzlem, k němuž vede sémantický vztah, a z tohoto uzlu pak vedou nepojmenované hrany představující vazbu slovní formy na daný synset (tedy v zásadě vztah synonymie). Tento způsob je alternativní k zobrazení, v němž sémantické vztahy vedou k jednotlivým slovním formám a není tak na první pohled zřejmé, které navázané formy jsou součástí kterých synsetů. Zde použité zobrazení vhodně omezuje množství hran, které jsou nutné k vykreslení grafu, byť možná na úkor srozumitelnosti pro neznalého uživatele, jelikož body označující synset a hrany označující synonymii nejsou nijak označeny v grafu (pouze v legendě a při přejetí myši). Všechny uzly mají přiřazenou i vlastní legendu ve formě informační bubliny, která obsahuje například definici, příklady, slovní druh či identifikátory. Vizualizaci je možné kolečkem myši přibližovat a oddalovat a tažením myši se po zvětšeném grafu přesouvat. Funkčnost přibližování je možná poněkud diskutabilní, jelikož se při jejím použití nemění proporce zobrazovaných informací (cf. elektronické mapy na WWW, které při přiblížení sice zvětšují detail struktury, ale zachovávají velikost písmen v nápisech). Její smysl by v opačném případě tkvěl například v použití při vyhledání slov jako je *plant*², které mají velké množství hyponym, a tudíž jejich graf je nepřehledný až do rozměrů naprosté nepoužitelnosti (obrázek 4.4 na straně 33). Ovládání klávesnicí možné není.

Rozhraní je neresponsivní, takže je na mobilním zařízení poněkud nepohodlné jej používat, byť to není nemožné. Šířka bloku s informacemi je natolik velká, že je nutné horizontálního rolování při čtení textu, což se považuje z hlediska použitelnosti webu za velmi negativní [; RH04]. Reprezentace grafem je vzhledem k ovládání, které používá, ještě hůře použitelná. Dotyk prstem do plochy s vykresleným grafem je totiž zároveň interpretován jako rolování celé stránky a zároveň jako tažení plátna s grafem do stran, které je běžně prováděno tažením myši po plátně. Přibližování a oddalování grafu, nikoliv celé stránky, v testovacím prostředí nebylo možné vůbec, ale vzhledem k vlastnostem této funkce popsaným výše to použitelnost nijak neomezuje.

vizualizuje
vizuali-
zační vi-
zualizace,
mwhahaha

²niz. rostlina



Oproti základnímu rozhraní WordNet Search 3.1 přináší toto rozhraní přehlednější textovou reprezentaci dat spojenou s vizualizací vztahů vyhledaného synsetu. Textové rozhraní však zobrazuje poněkud omezené množství informací z wordnetu a soustřeďuje se především na data ze slovníku Referentie Bestand Nederlands. Vztahy mezi slovy jsou zobrazovány pouze v grafické reprezentaci dat, což vzhledem k její nepříliš vysoké použitelnosti na mobilních zařízeních s menší obrazovkou může být omezující. Rozhraní uchovává svůj stav podrobně v URL a umožňuje jej z ní plně obnovit.

WordVis je rozhraní zaměřené na vizualizaci dat princetonského WordNetu grafem. Sestává z vyhledávacího pole, levého sloupce s výběrem synsetů a plátnem, na němž je vykreslen graf vztahů zvoleného synsetu nebo slovní formy.

Ačkoliv to na první pohled uživateli nemusí být zřejmé, vizualizace fun-

guje ve dvou režimech. Prvním je zobrazení slovní formy a synsetů, v nichž se tato slovní forma vyskytuje, druhé je zobrazení synsetu jako centrální jednotky a vztahů a slovních forem, které k danému synsetu patří. První zobrazení je užíváno například při prvotním zobrazení výsledků hledání. Vizualizace zobrazí ve středu grafu vyhledanou slovní formu a připojí k ní nepojmenovanými hranami synsety, jež danou slovní formu obsahují. Pokud pak uživatel zvolí kliknutím myši některý konkrétní synset, zobrazí se ve středu vizualizace jeho značka, k níž jsou připojeny pojmenovanými hranami představujícími sémantické vztahy další synsety. Stejně jako v prvním zobrazení, i v tomto se slovní formy náležející k určitému synsetu připojují nepojmenovanými hranami jako textové uzly. Pokud jedna slovní forma náleží k více synsetům (např. substantivum *bicycle* a sloveso *bicycle*), je v grafu její textový uzel pouze jednou a vedou od něj dvě nepojmenované hrany k náležitým synsetům.

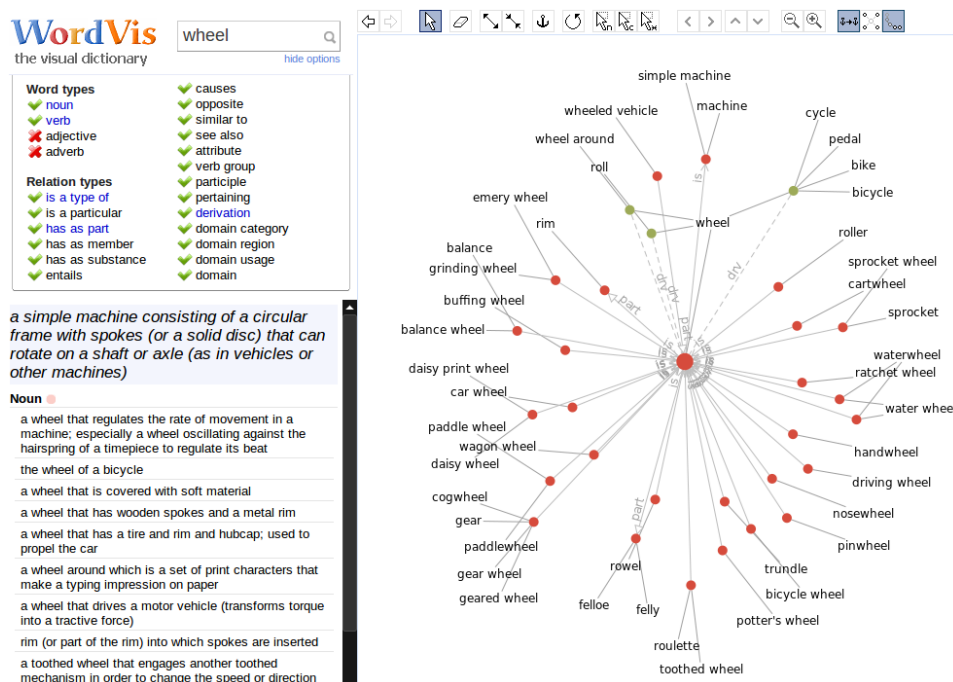
Hrany mezi synsety jsou orientované (znázorněné jako šipky) a jejich k jejich pojmenování zvolil autor sémantické významy daných vztahů, tedy např. hyperonymie je značena slovem *is*³; např. *apple tree is fruit tree*⁴.

Vyhledávací formulář umožňuje filtrování výsledků hledání podle slovních druhů a typů vztahu. Graf je navíc aktualizován v reálném čase podle toho, jak jsou podmínky filtrování uživatelem měněny. Nevýhodou však je absence funkcí *vybrat vše* a *nevybrat nic*, která je citelná zvláště v případě, že uživatel chce vybrat pouze jeden druh relace (kterých je mnoho, tudíž odebrání všech relací z výběru kromě jedné může být časově náročné). Hledání také nabízí funkci napovídání poté, co uživatel zadá několik znaků z počátku slova. To umožňuje vybrat hledané slovo z nabídky a ušetřit několik úhozů, což je zvláště užitečné u delších slov či pro uživatele, který nevyhledává ve svém rodném jazyce a není si jist pravopisem hledaného slova.

Technologicky je vizualizace řešena javascriptovou knihovnou, jež vykresluje graf na HTML5 plátně [W3S]. Uspořádání prvků na plátně je řešeno modelováním fyzikálních vlastností našeho světa. Jednotlivé uzly na plátně mají zadefinováno, že se navzájem odpuzují (podobně jako elektrony), naopak vazby mezi nimi fungují jako pružiny, jež mají nějakou ideální délku, v níž se snaží setrvávat. Díky těmto dvěma soupeřícím silám jsou body na plátně rozmístěny tak, aby využívaly prostor co nejúčinněji. Hrany navíc mohou mít zadanou preferovanou orientaci, což umožňuje orientovat například hyperonymní synsety směrem nahoru, hyponymní směrem dolů etc. [Ver10] Vzhledem k tomu, že však zřejmě v prostředí neexistuje žádná penalizace za křížení hran, je relativně běžné, že se hrany překříží. To značně snižuje přehlednost výsledného grafu, vzhledem k tomu, že pak není jasné, zda jsou hrany překříženy z nějakého hlubšího důvodu (například jedna slovní forma náležející k více synsetům), či nikoliv. Knihovna vykreslující graf umožňuje

³angl. *je*

⁴angl. jablono je ovocný strom



Obrázek 4.5: Ukázka rozhraní WordVis (po zvolení konkrétního vyhledaného synsetu)

uživatelé také modifikovat tažením jednotlivých uzlů myši, prodlužovat či zkracovat hrany a podobně. V aplikaci pro zobrazení dat z WordNetu toto nehraje příliš zásadní roli a implementace této funkcionality je zřejmě dána tím, že knihovna je určena pro univerzálnější použití. [Ver10]

Rozhraní je neresponsivní a na mobilním zařízení je nutné k zajištění čitelnosti textů na stránce obsah přiblížit, což vede k nutnosti rolování jak vertikálnímu, tak horizontálnímu. Také není možné využít funkcionality přesouvání prvků na plátně, což ale při použití na vizualizaci dat z wordnetu není velkým omezením funkčnosti. Co se udržování stavu aplikace v URL týče, je implementováno pouze čtení parametru *q* (od *query*), zpětně do něj ale už zapisováno není. Odkazy na synsety na levé straně tento parametr obsahují, stejně tak jej lze zkopírovat přes kontextové menu pro každý uzel v rozhraní, takže teoreticky je možné obnovit kterýkoliv stav aplikace, byť dostupnost této funkcionality není příliš intuitivní.

Přínos tohoto rozhraní tkví především v relativně přehledné vizualizaci dat princetonského WordNetu, nutno však podotknout, že pro některé své vlastnosti nemusí pro nezkušeného uživatele být zpočátku příliš intuitivní (např. zmíněné křížení hran).

Na knihovně, na níž je toto rozhraní postaveno, stojí ještě některá další

rozhraní k princetonskému WordNetu, například VisuWords⁵ či Visual Thesaurus⁶. Ta nebyla do hodnocení v této práci zahrnuta, protože jsou funkčně podobná tomuto rozhraní.

4.5 sloWTool

Rozhraní sloWTool bylo vyvinuto pro potřeby slovinského wordnetu, jenž je založen na princetonském WordNetu a vznikl skombinováním několika zdrojů, například Wikipedie, dvojjazyčných slovníků či paralelních korpusů. Je víceúčelovým nástrojem, který umožňuje textovou reprezentaci, vizualizaci a editaci dat z wordnetu. Tvůrcům sloužilo k revizím slovinského wordnetu po jeho rozšiřování automatickými nástroji. [FNE12] Cílem při vývoji tohoto rozhraní podle Fišer; Novak [FN11] bylo překonat nevýhody tehdejších dostupných rozhraní a vyvinout nástroj, který by mezi jiným umožňoval prohlížení i editaci, spolupráci mnoha autorů včetně anonymních editací (záměr tvůrců byl využít náhodných návštěvníků k opravování chyb, jež ve wordnetu naleznou) či možnost jednoduché registrace uživatelů. Mezi dalšími podmínkami byla možnost přidávání dalších wordnetů do systému a s tím spojená schopnost rozhraní zobrazovat vícejazyčná data. V neposlední řadě se autoři v kontextu tehdejších nástrojů také snažili zaměřit na platformní nezávislost a přenositelnost rozhraní.

Rozložení stránky je vzdáleně podobné tomu u základního oficiálního rozhraní k princetonskému WordNetu, a to v tom smyslu, že neobsahuje relativně rozšířený levý sloupec na výběr synsetů, ale jednotlivé významy, které jsou nalezeny po zadání hledané slovní formy do vyhledávacího pole, seskupuje do sekcí v hlavním bloku s textovou reprezentací dat. Vyhledávací pole podporuje napovídání existujících významů, což může usnadnit práci s rozhraním. V textové reprezentaci rozhraní zobrazuje zřejmě všechny dostupné informace o synsetech, tedy se chová podobně jako základní rozhraní princetonského WordNetu, je-li nastaveno tak, aby nefiltrovalo žádné informace. U jednotlivých vztahů, v nichž je daný synset přítomen, pak je možné kliknutím otevřít synset nacházející se pomyslně na druhé straně daného vztahu (např. *bicykle* má meronymum *sedlo*, takže lze otevřít detail synsetu *sedlo*). To trpí stejným neduhem, jako základní rozhraní princetonského WordNetu, to jest, že lze ve smyčce otevírat nadřazený a podřazený synset a stále se zanořovat do smyčky hlouběji. To je nejen nesmyslné, ale zároveň to v implementaci tohoto rozhraní postupně může začít zpomalovat rychlost reakcí prohlížeče pro nadměrné množství uzlů v DOM⁷.

Po levé straně se nachází lišta s ikonami odkazujícími na dalšími moduly, které se otevírají v emulacích nových oken (v témže panelu). Mezi tyto

⁵<http://www.visuwords.com/>

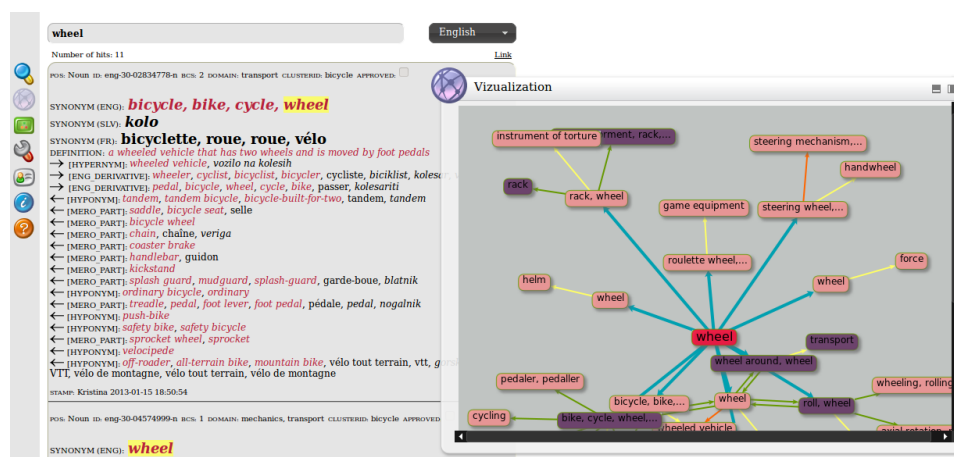
⁶<https://www.visualthesaurus.com>

⁷document object model

moduly patří mimo jiné i pokročilé hledání, vizualizace a nápověda.

Pokročilé hledání funguje podobně jako u ostatních rozhraní, tedy umožňuje používat zástupné znaky (* a ?), filtrovat slovní druhy, nebo vyhledávat v jiných polích než jsou slovní formy patřící do synsetů (například v definicích či podle identifikátoru synsetu).

Vizualizace, modul z hlediska této práce nejpodstatnější, je určen k zobrazení grafické interpretace vztahů vyhledaného synsetu. Na rozdíl od ostatních rozhraní tato vizualizace zobrazuje všechny vyhledané synsety (s kořenem grafu označeným jako *search*⁸) a neumožňuje jejich filtrování. Také nepodporuje zřejmě žádnou formu interakce kromě přesouvání prvků na plátně a zvýraznění příslušného synsetu v hlavním bloku textové reprezentace poté, co je na něj (nebo jeho členskou slovní formu) ve vizualizaci kliknuto. Pokud uživatel potřebuje zobrazit vizualizaci jednoho konkrétního synsetu, je nucen použít pokročilé vyhledávání a vyhledat daný synset nejlépe podle jeho identifikátoru, který je vždy unikátní. Vizualizace je zatížena také dalšími problémy, jako jsou nedostatečné možnosti přizpůsobování velikosti zobrazovací plochy (zvětšování okna je podporováno, ale plátno s grafem zůstává konstantně velké), občasnou desynchronizaci obsahu vizualizace s výsledky hledání, náročností na uživatelské technické vybavení počítače (především výpočetní jednotku) a značnou nepřehledností způsobenou nedostatečně či nevhodně řešenou penalizací překrývání prvků na zobrazovací ploše.



Obrázek 4.6: Ukázka rozhraní sloWTool (s otevřenou grafickou vizualizací)

Rozhraní je svými možnostmi zprostředkování informací z wordnetu poměrně inovativní, reálná použitelnost ovšem trpí zmíněnými nedostatky a odchyluje jej tím od původního záměru autorů, který minimálně zčásti zůstává nenaplněn. Je sice pravdou, že *s aplikací je možné pracovat v každém moderním přehlížeči, ať už na počítači, tabletu, či dokonce mobilním tele-*

⁸hledání

fonu [FN11], nutno ale podotknout, že rozhraní je zcela neresponsivní a jeho ovládací prvky jsou zcela nevhodné pro ovládání na menší obrazovce a dotykem.

Stav aplikace lze do jisté míry uložit pomocí odkazu *Link*, který vede na adresu, z níž lze obnovit hledané slovo (konfiguraci otevřených nástrojů však už nikoliv).

Navzdory všem nedostatkům této aplikace je dlužno uznat, že se svou univerzalitou a funkcionalitou značně přibližuje rozhraní, které je praktickým cílem i této práce.

Slowtool bylo vytvořeno pod licencí Creative Commons⁹ a jeho zdrojový kód je dostupný na platformě pro aplikace s otevřeným zdrojem Launchpad¹⁰.

4.6 BabelNetXplorer

BabelNet je rozsáhlý projekt vícejazykové sémantické sítě, který čerpá data z více zdrojů. Záměrem autorů bylo při tvorbě této sémantické sítě eliminovat základní faktory definující nevýhody tehdejších (a potažmo i aktuálních) mezinárodních projektů zabývajících se sémantickými sítěmi. Těmi jsou manuální tvorba dat, které síť konstituují, a s tím související nerovnoměrnost množství dat přes jednotlivé jazyky. Jazyky s vysokou hustotou zdrojů, jakým je například angličtina, tak ve výsledku mají více dat i v sémantické síti. Autoři BabelNetu se pokusili tento problém překonat kombinací několika metod, jejichž společným jmenovatelem je automatizace. Informace o významech jsou v BabelNetu doplňovány z Wikipedie, která je podle autorů díky mnohačetným zásahům expertů z různých oborů ve výsledku přesným a informacně bohatým zdrojem. Druhá důležitá metoda, zaměřující se především na nerovnoměrnost dat v různých jazycích, je automatický překlad zdrojů. [NP10]

BabelNetXplorer je webové grafické rozhraní vytvořené pro vizualizaci dat z BabelNetu. [NP12] uvádějí, že rozhraní slouží k vizualizaci vztahů pro slova nalezená v BabelNetu a ilustrují vzhled rozhraní dvěma snímky obrazovky. V době psaní této práce však rozhraní BabelNetXploreru vypadá výrazně odlišně, což je vzhledem k tomu, že od vzniku práce [NP12] uběhlo pět let, pochopitelné. V této práci bude z evidentních důvodů rozebrána použitelnost a funkcionalita současného rozhraní.

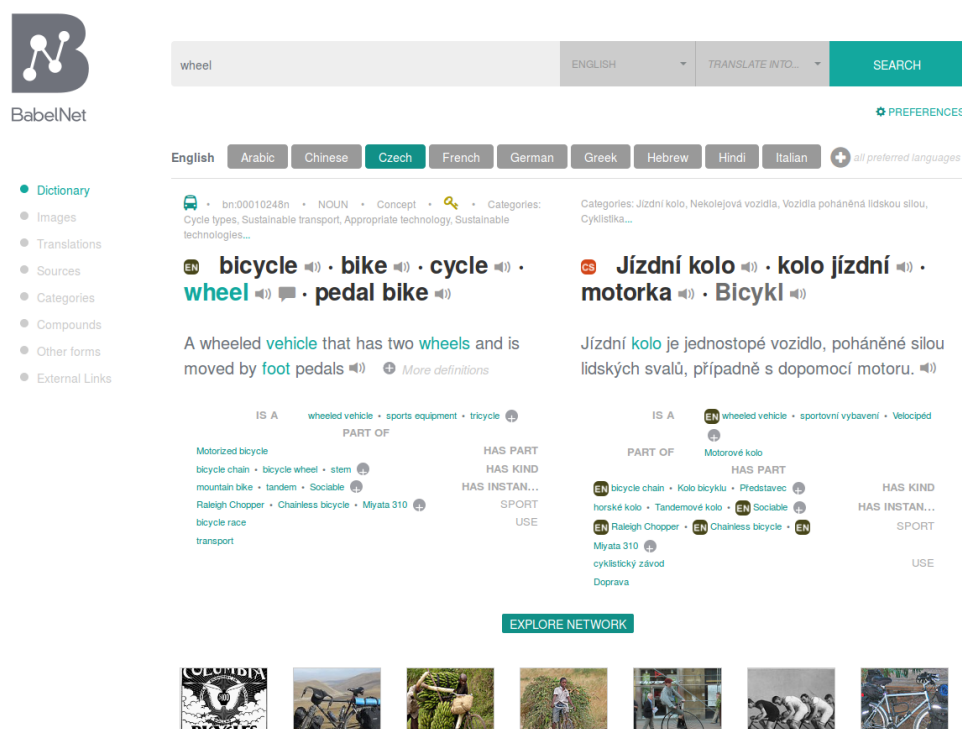
Rozhraní obsahuje klasicky vyhledávací pole, vedle něhož se nacházejí rozbalovací menu, v nichž si uživatel může vybrat, ve kterém jazyce chce vyhledávat a, případně, do kterého jazyce chce slovo přeložit. Po úspěšném dokončení vyhledávání jsou uživateli zobrazeny odkazy na jednotlivé synsety v seznamu, který není nepodobný ostatním rozhraním. Podstatným rozdílem

⁹<https://creativecommons.org/licenses/by-nc-sa/3.0/>

¹⁰<https://launchpad.net/slowtool>

však je, že se v něm zobrazují i ilustrační obrázky, které, byť nejsouce vždy velmi informativní, mohou napomoci uživateli v orientaci, který synset jej zajímá.

V detailu synsetu, který se zobrazí po kliknutí na příslušný odkaz, je uživateli zobrazen seznam slovních forem, které k danému synsetu náleží, jeho definice (s možností zobrazit i jeho definice nejen z daného wordnetu, v němž uživatel vyhledává, ale i z Wikipedie a dalších zdrojů), jeho případný překlad a sémantické vztahy, do nichž tento synset náleží. K dispozici má uživatel i možnost zobrazit si daný synset paralelně v dalších jazycích pomocí nabídky pod vyhledávacím polem (viditelné na snímku obrazovky 4.7 na straně 39). Níže na stránce jsou uživateli prezentovány informace z dalších zdrojů napojených na BabelNet, jako jsou obrázky, překlady, odkazy na Wikipedii, etc. a odkaz na vizuální reprezentaci sémantických vztahů otevřeného synsetu.



Obrázek 4.7: Ukázka rozhraní BabelXplorer (detail synsetu v textové reprezentaci)

Vizuální reprezentace je řešena tradičním hvězdicovitým grafem, který má dva režimy. Jeden zobrazuje vedle textových názvů jednotlivých synsetů přítomných v grafu i jejich zástupné obrázky, druhý je čistě textový a, nutno podotknout, značně přehlednější. Uzly reprezentující synsety jsou provázány barevně odlišenými hranami, kde barvy značí druh vztahů, který dva synsety spojuje. Uzly jsou klikatelné, přičemž po kliknutí na některý z příbuzných

synsetů se zobrazí tento synset a opět hvězdovitě další synsety, s nimiž je zkoumaný synset provázán některým ze vztahů. Při najetí kursoru myši na určitý synset se zobrazí jeho detail; tyto informační bubliny jsou však vázány na pozici kursoru myši a nelze tak kursorem najet do detailu tak, aby bylo možné kliknout na odkazy, které jsou v informační bublině obsaženy, a dostat se tak na textovou reprezentaci daného synsetu. Zdá se tedy, že není možné přejít z grafické reprezentace do textové.

Poněkud nešťastně je řešen design vizualizace, jelikož jednak není možné graf přibližovat a oddalovat, a jednak pokud uživatel nenajede na konkrétní synset kursoru myši, graf se samovolně neustále otáčí, což ztěžuje orientaci v něm.

Rozhraní všeobecně responsivní, textová reprezentace je dobře použitelná i na mobilním rozhraní s malou obrazovkou. Grafická reprezentace je však omezena na větší obrazovky, jelikož v ní není možné graf posouvat po obrazovce, a tudíž jeho značná část (byť se částečně přizpůsobuje zobrazovací ploše), zřejmě v závislosti na velikosti obrazovky zařízení, zůstává uživateli skryta.

V době vzniku této práce mělo rozhraní dvě verze, z nichž jedna byla označena jako *živá beta*, druhá jako *současná* (3.7), avšak při testování nebyly zjištěny zásadní rozdíly mezi těmito verzemi. Nutno ovšem podotknout, že na testovacím desktopovém prohlížeči vykazovala verze beta minoritní chyby v rozložení stránky.

Bibliografie

- [] [online] [cit. 22. dub. 2017]. Dostupné z: <https://www.nngroup.com/articles/scrolling-and-scrollbar/>.
- [BOH11] BOSTOCK, Michael et al. D3 Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics* [online]. 2011, roč. 17, č. 12, s. 2301–2309 [cit. 21. dub. 2017]. ISSN 1077-2626. Dostupné z DOI: 10.1109/TVCG.2011.185.
- [CQ69] COLLINS, Allan M; QUILLIAN, M Ross. Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*. 1969, roč. 8, č. 2, s. 240–247.
- [Cor] CORNETTO. *Getting started with the Cornetto demo*. Dostupné také z: <http://cornetto.clarin.inl.nl/help/GettingStarted.pdf>.
- [Cru86] CRUSE, D Alan. *Lexical semantics*. Cambridge University Press, 1986.
- [Fel05] FELLBAUM, Christiane. WordNet and wordnets. *Encyclopedia of Language and Linguistics*. Second. 2005, s. 665–670.
- [FJ65] FILLENBAUM, Samuel; JONES, Lyle V. Grammatical contingencies in word association. *Journal of Verbal Learning and Verbal Behavior*. 1965, roč. 4, č. 3, s. 248–255.
- [FBS04] FILLMORE, Charles J et al. FrameNet as a”Net”. In: *LREC*. 2004.
- [FN11] FIŠER, Darja; NOVAK, Jernej. Visualizing sloWNet. *Proceedings of the Electronic Lexicography in the 21st Century (eLex 2011)* [online]. 2011, s. 76–82 [cit. 25. dub. 2017].
- [FNE12] FIŠER, Darja et al. sloWNet 3.0: development, extension and cleaning. In: *GWC 2012 6th International Global Wordnet Conference* [online]. 2012, s. 113 [cit. 23. dub. 2017].
- [Gar82] GARRETT, Merrill F. Production of speech: Observations from normal and pathological language use. *Normality and pathology in cognitive functions*. 1982, s. 19–76.

- [Gar] 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>.
- [HVR] HORÁK, Aleš et al. The development of a complex-structured lexicon based on WordNet. In: [online] [cit. 17. dub. 2017].
- [KM02] KAMPS, Jaap; MARX, Maarten. Visualizing wordnet structure. In: *Proc. of the 1st International Conference on Global WordNet*. 2002, s. 182–186.
- [LL82] LEHRER, Adrienne; LEHRER, Keith. Antonymy. *Linguistics and philosophy*. 1982, roč. 5, č. 4, s. 483–501.
- [MM05] MARTIN, Willy; MAKŠ, I. Referentie Bestand Nederlands. *Met medewerking van S. Bopp en M. Groot*. 2005.
- [Mil+90] MILLER, George A et al. Introduction to WordNet: An on-line lexical database. *International journal of lexicography*. 1990, roč. 3, č. 4, s. 235–244.
- [NP10] NAVIGLI, Roberto; PONZETTO, Simone Paolo. BabelNet: Building a Very Large Multilingual Semantic Network. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* [online]. Uppsala, Sweden: Association for Computational Linguistics, 2010, s. 216–225 [cit. 27. dub. 2017]. ACL '10. Dostupné z: <http://dl.acm.org/citation.cfm?id=1858681.1858704>.
- [NP12] NAVIGLI, Roberto; PONZETTO, Simone Paolo. BabelNetXplorer: A Platform for Multilingual Lexical Knowledge Base Access and Exploration. In: *Proceedings of the 21st International Conference on World Wide Web* [online]. Lyon, France: ACM, 2012, s. 393–396 [cit. 27. dub. 2017]. WWW '12 Companion. ISBN 978-1-4503-1230-1. Dostupné z DOI: 10.1145/2187980.2188057.
- [PŠ13] PALA, Karel; ŠEVEČEK, Pavel. Česká lexikální databáze typu WordNet. *Feedback*. 2013, roč. 48, č. A47.
- [PW06] PARADIS, Carita; WILLNERS, Caroline. Antonymy and negation—The boundedness hypothesis. *Journal of pragmatics*. 2006, roč. 38, č. 7, s. 1051–1080.
- [RH04] RICHARDS, John T.; HANSON, Vicki L. Web accessibility: a broader view. In: *Proceedings of the 13th conference on World Wide Web - WWW '04* [online]. ACM Press, 2004 [cit. 21. dub. 2017]. Dostupné z DOI: 10.1145/988672.988683.

- [SDB07] SZYMANSKI, J et al. Cooperative editing approach for building wordnet database. In: *Proceedings of the XVI international conference on system science*. 2007, s. 448–457.
- [Thea] THE GLOBAL WORDNET ASSOCIATION. *GWA Base Concepts* [online]. Global WordNet Association [cit. 6. dub. 2017]. Dostupné z: <http://globalwordnet.org/gwa-base-concepts/>.
- [Theb] THE GLOBAL WORDNET ASSOCIATION. *Wordnets in the World* [online] [cit. 25. břez. 2017]. Dostupné z: <http://globalwordnet.org/wordnets-in-the-world/>.
- [Ver10] VERCRUYSSSE, Steven. *About WordVis* [online]. 2010 [cit. 23. dub. 2017].
- [Vos98] VOSSEN, Piek. Introduction to eurowordnet. *Computers and the Humanities*. 1998, roč. 32, č. 2-3, s. 73–89.
- [Vos97] VOSSEN, Piek et al. EuroWordNet: a multilingual database for information retrieval. In: *Proceedings of the DELOS workshop on Cross-language Information Retrieval*. 1997, s. 5–7.
- [W3S] W3SCHOOLS. *HTML Canvas Reference* [online] [cit. 23. dub. 2017]. Dostupné z: https://www.w3schools.com/tags/ref_canvas.asp.