

DeepDict: kunteksta reta vortaro de vera lingvouzo



Eckhard Bick
Suddana Universitato & GrammarSoft
eckhard.bick@mail.dk

Resumo: *DeepDict estas sistemo por ĉerpi vortkuntekstojn (precipe dependecajn ligojn) el grandaj tekstaroj kaj krei grafikaĵojn (t.n. leksikogramojn) pri la eblaj kaj tipaj komplementoj kaj atributoj de serĉvorto. Eblas trakti malsamajn lingvojn kaj malsamajn ĝenrojn uzante malsamajn korpusojn, kaj agordebla retinterfaco permesas ĝin uzi al tiel malsamaj grupoj kiel vortaristoj kaj lernantoj. La sistemo mem rangigas rilatojn laŭ korelacioforto, kaj prezentas konkordancojn de ekzemplofrazoj. Aktuale DeepDict traktas 12 lingvojn, inter ili Esperanto.*

1 Fono

Eblas distingi du paralelajn padojn en la evoluo de vortaroj, koncerne unuflanke la vortaroformon, aliflanke la vortaro-materialon. Por krei tradician papervortaron, oni uzis citaĵojn, introspektadon kaj, kompreneble, aliajn vortarojn, ofte kopiante kopiaĵojn sen kontrolo pri vera uzo. En la okdekaj jaroj komenciĝis nova fazo, jam kun statistikaj datenoj el realaj tekstokolektoj (korpusoj), garantiante pli altan nivelon de aŭtentikeco. Samtempe oni eksperimentis pri elektronikaj vortaroj, kiuj tamen unue simple estis elektronikaj versioj de paperaj vortaroj, pli facile serĉeblaj, sed ne esence aliaj. Nur poste venis tria fazo kun dinamikaj elektronikaj vortaroj, eluzante la fakton ke en tiu medio vortartikolo ne devas esti fiksforma, ke eblas perklake aldoni pli da profundeco, ekzemplojn ktp. La aktualaj defioj estas kunlaboraj vortaroj (viki'oj) kaj, koncerne lingvomaterialon, viva fonto-ĉerpado kaj rilatiga uzo de korpusdatenoj.

Por vortaro-uzanto tiu evoluo signifas pli facilan aliron al vortaroj, kombineblajn serĉojn (difino, traduko, sinonimoj) en la sama interfaco, kaj la ppp-principon: profundo [nur] post peto. Eĉ personaj interfac-agordoj eblas, minimume en la formo de serĉmemoro kaj agordomemoro. Tamen eĉ la plej modernaj interfacoj ankoraŭ luktas kun unu praproblemo de vortariko - kiel distingi inter pasiva-komprena vortaro unuflanke, kaj aktiva-produktiva aliflanke, aŭ - sankta gralo de vortareldonejoj - kiel zorgi pri tiuj du aspektoj en unu sama publikaĵo. Fakte, dum en pasiva vortaro, kiu ja celas gepatralingvan uzanton, sufiĉas difini vorton aŭ doni neordigitan liston de pli-malpli ajn sinonimaj tradukoj, la problemo de aktiva vortaro, kiu ja celas fremdlingvan leganton, estas ke ne sufiĉas tia informo - sen denaska intuo ne eblas por la uzanto decidi en kiu kunteksto elekti kiun tradukalternativon, kiel fleksii la cellingvan vorton ktp. Tial aktivaj vortaroj bezonas pli (a) da gramatikaj informoj, (b) ekzemplojn de sintaksa-semantika kompletigo, kaj (c) senco-distingajn kriteriojn (distingilojn). Temas do pri en-kunteksta vortouzo, kaj multe helpas povi montri al la uzanto la plej tipan uzokuntekston de vorto. Skema ekzemplo el papera vortaro povas esti ekzemple 'A donas x al B', kie majuskloj (A,B) signifas homajn rolantojn, kaj minuskloj (x,y) aĵajn. El tia ekzemplo aŭ skemo rekte klariĝas valento kaj kombinebloj (b), sed ankaŭ ofte - nerekte - sencaj distingoj (c), ĉar la kombinortoj limigas la eblan signifon de la serĉvorto. Depende de la lingvo, ankaŭ gramatika trajto (a) povas esti deduktita, ekz. genro, nombro aŭ kvanteco de substantivo el la uzo de artikolo aŭ adjektivo.

Por fidine krei tiajn skemojn por ia serĉvorto, necesas amaso da aŭtentikaj tekstdatenoj, kaj nia projekto, DeepDict, temas ĝuste pri dinamika vortaro pri vortrilatoj ĉerpitaj statistike el grandaj korpusoj.

2 La korpusa bazo

Pro kopirajtaj kialoj grandaj tekstokvantoj malfacile haveblas en la formo de tradicia literaturo, kaj la plej ofte uzata materialo estas gazetaro aŭ politikaj tekstoj (ekz. Europarl, Koehn 2005). Ankaŭ Vikipedio estas interesa fonto, kaj pro sia liberfonteco, kaj pro sia tre bona leksika varieco. Preskaŭ senlima fonto de lingvodatenoj estas la Reto, sed necesas multe purigi la ĉerpaĵojn por forigi lingve miksitajn tekstojn, binarajn datenojn ktp, kaj ankaŭ - en la kazo de Esperanto - por normigi la dekojn da uzitaj enkodigoj. Gravus bona mikso de fontoj, ĉar alikaze statistikaj metodoj montros malveran bildon de la lingva pejzaĝo, kaj povas eĉ manki multaj vortoj aŭ strukturoj entute. Ankaŭ helpas al la statistika aliro havi lemigon (*lemo*=bazformo), kio permesas rekoni fleksiitajn formojn de la sama vorto, tiel plibonigante la statistikan fidindecon per pli da ekzemploj. Fig. 1 montras la fontojn kiujn ni uzis por la Esperanta DeepDict-vortaro, en milionoj (M) da vortoj, 58,4 M entute.

Krom grandeco kaj balancitigo de tekstotipoj, tre gravas por vortaristo kiom kaj kiel la strukturaj informoj en korpuso estas alireblaj. En neprilaborita, kruda tekstokorpuso, ekzemple, estas tre malfacile studi la rilaton inter verbo kaj objekto, ĉar ne eblas scii kio estas verbo, kaj kio objekto, kaj ĉar povas esti multaj vortoj inter verbo kaj ĝia objekto. La lingvistika solvo por tiu problemo estas gramatika analizado kaj markado (etikedizo). En la DeepDict-projekto ni uzis la EspGram-analizilon (Bick 2005-1), tiel-nomatan parsilon, kiu tranĉas tekston en vortojn kaj frazojn, kaj aldonas al ĉiu vorto gramatikajn etikedojn:

1. bazformo (*lemo*), ekz.: *manĝis, manĝu, manĝanta ktp.* -> "*manĝi*"
2. vortklaso, ekz.: V (verbo), N ("nomen", substantivo), ADJ (adjektivo)
3. fleksio, ekz.: ACC (akuzativo), P (pluralo), PR ("prezenco", nuntempo)
4. afiksoj, ekz.: <AFF:il>, <AFF:ej>
5. semantika klaso, ekz. <Hprof> (homo -> profesiulo), <food> (manĝaĵo)
6. sintaksa funkcio, ekz.: @SUBJ (subjekto), @ACC (akuzativa objekto)
7. dependecaj vortligoj, ekz. #2->5 en frazo kiel "*la viro el Parizo venis.*" (la dua vorto [subjekto *viro*] estas la "filino" de la kvina [verbo *veni*])
8. semantika rolo, ekz. §AG (aganto), §REC (ricevanto), §LOC (loko)

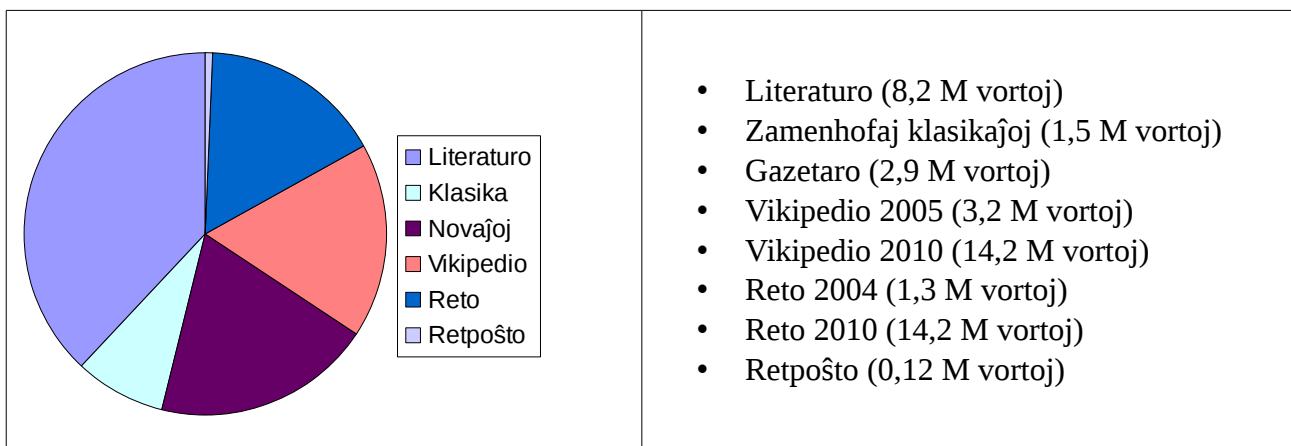


Fig. 1: Korpusaj fontoj

El la etikeditaj korpusoj eblas facile ĉerpi ekzemplojn pri certaj strukturoj, ekz. statistikon pri tio, kiujn kap-substantivojn permesas la adjektivoj 'granda', 'alta' kaj 'larĝa', por esplori la semantikajn limigojn de ilia uzo. La CorpusEye-interfaco ĉe Suddana Universitato (Bick 2005-2), kie loĝas la Esperantaj korpusoj, permesas ne nur fari tion per grafika interfaco, sed ankaŭ prezentas ekz. relativan frekvencordigon de la trovitaj substantivoj.

3 La kerna ideo de DeepDict

DeepDict volas esti kunteksta vortaro, kiu klarigas vortojn kaj iliajn uzon kaj lingvistikajn rilatojn per la plej tipaj aŭtentikaj ekzemploj. Sed kiel montri - por ia serĉvorto - ĉiujn strukturojn kaj ekzemplojn samtempe? Kiel oni evitas forgesi iun aspekton, kaj - fine-ne-forgese - kiel oni faru ĉi ĉion aŭtomate?

La kerna ideo realigita en DeepDict estas ke eblas fidinde ĉerpi strukturojn kaj rilatojn el dependecmarkita korpuso, ke statistikaj metodoj povas identigi la plej tipajn komplementojn, kaj ke grafika prezento povas ebligi samtempan montron de multaj rilatoj. Por prepari la necesajn datenojn, necesis la jenaj paŝoj:

1. etikedizi la korpusojn, uzante la EspGram-parsilon

Petro [Petro] @SUBJ manĝis manplenon da nuksoj [nukso] @ACC

Katoj [kato] @SUBJ manĝas musojn [muso] @ACC

2. ĝeneraligi kolektante kaj nombrante dependecparojn lemo-1 --> lemo-2 per aŭtomata programo. En simpligita formo, la rezulto el la frazoj (1) estas la jena:

ekz. subjekto-verbo-rilato: PROP_SUBJ --> manĝi, kato_SUBJ -> manĝi

ekz. akuzativo-verbo-rilato: nukso_ACC --> manĝi, muso_ACC -> manĝi

(PROP estas ĝeneraligo de propraj nomoj)

3. prezenti rezultojn en listoformo, aŭ stoki en datumbazo, kune kun statistikaj informoj:

{PROP, kato} SUBJ --> manĝi <-- {nukso, muso} ACC

La rezulto montras la uzon kaj lingvistikan funkcion de la verbo *manĝi* - ne nur la fakton, ke ĝi estas transitiva kun objekto kaj eksplicita subjekto, sed ankaŭ la semantikan informon pri tio, kion oni povas manĝi, kaj kiu povas esti manĝanto.

Tamen, por la plej multaj serĉvortoj, tiaj rilato-listoj fariĝus neuzeble grandaj se oni simple kompilis ilin el tuta korpuso, kaj eĉ frekvenca ordigo estus problema - ekz. la oftaj kolor- kaj grandecadjektivoj kombineblas kun la plej multaj substantivoj kaj tial ne tiom helpas difini kapsubstantivon. La demando estas kiel distingi *tipajn* de ne-informivaj komplementoj kaj atributoj - alivorte, pli gravas doni tipan ekzemplon ol oftan. Por realigi tiun distingon, ni uzis specialan mezuron, *korelacioforton*, kiu mezuras la interrilatecon de du vortoj. Por tiu mezuro oni normigas la absolutan korpusan oftecon (frekvencon) de iu vortkombino laŭ la baza sendependa ofteco de ĝiaj partoj. Konkrete, oni dividas la kuran oftecon per la produto de la unuopaj oftecoj (Church 1990). La formulo de DeepDict aplikas tiun metodon en logaritma skalo, kaj pezigas la kuran oftecon uzante ĝian kvadratonombbron:

Korelacioforto: $K = C * \log(p(a \rightarrow b) ^2 / (p(a) * p(b)))$

Ni rigardas kiel tipajn tiujn vortligojn kiuj havas pozitivan korelacioforton super difinita sojlo. Krome, DeepDict markas la gravecon de ligita vortparo per logaritma frekvencklaso. Frekvencklaso 0 tiel signifas ke la korpuso enhavis nur unu ekzemplon de la koncerna rilato, frekvencklaso 4 signifas 9-16 ekzemplojn ($16 = 2$ je la 4a potenco). Raraj ekzemploj estas danĝeraj - unue ili riskas esti tajperaroj aŭ analizeroj, kaj due ili riskas aperi en la listo ĉar nia metriko povas miskompreni rarecon kiel tipecon (simple ĉar la izolita ofteco de tajperaro estas same malofta kiel la kunokazo kun la serĉvorto kaj tial "normaligas" ĝin).

Certajn vortklasojn necesis trakti speciale. Tiel, nomoj estas simple PROP, kaj nombro NUM - por ambaŭ kategorioj subdistingo estus sen reala valoro kaj nur diluus la statistikon de DeepDict. Pli

problemaj estas pronomoj - unuflanke ili estas tiom oftaj ke ili ankaŭ "diluus" nian metrikon, sed aliflanke ili enhavas interesajn semantikajn informojn, ekz. pri sekso (li - ŝi - ĝi), kvanteco (iom, tiom) aŭ informon pri adverbialaj (PAG: adjektaj) komplementoj (tie, tien). Pro tio ni konservas pronomojn individue, sed rangigas ilin aparte, ekster la listo de aliklasaj vortligoj. Finfine, eksperimente, DeepDict prezentas ĝeneraligon per semantikaj prototipoj - do ĉe la vorto *manĝi* ĝi anstataŭ (aŭ kune kun) la listo de manĝaĵobjektoj montras simple la kategoriojn <food> (manĝaĵo), <fruit> (frukto), <Aich> (fiŝo) ktp. Entute ekzistas ĉirkaŭ 200 tiaj prototipoj por substantivoj, kiujn liveras la etikedito de EspGram.

4 Projekta kadro

En sia grafika, reta versio, DeepDict loĝas ĉe www.gramtrans.com, kie ĝi estas parto de integrita aro de tradukhelpiloj, kaj funkcias kiel profunda, cellingva nivelo de la dulingva vortaro *QuickDict* (Fig. 2, kvarlingva), aŭ kiel suplemento al la unulingva difino- kaj sinonimo-vortaro *Lexicator* (nur danlingva).

<p>Angla-Esperanta <i>QuickDict</i></p> <ul style="list-style-type: none"> • 213 000 leksemoj (vortoj) <ul style="list-style-type: none"> ◦ [116 000 plurvortaĵoj] • 257 000 tradukoj • 238 000 distingitaj signifo • el tio 43 800 nomoj <ul style="list-style-type: none"> ◦ [24 400 plurvortaĵoj] • (+ 360 000 nekontrolitaj plurvortaĵoj) • (+ fiksjaj esprimoj) 	<p>give</p> <p>give (n) — elasteco ; cedemo give (v) — doni</p> <p>give {window} on to — havi vidon al give lecture — doni prelego give {<Lpath>_SUBJ} into — konduki en give toast — proponi tosto give rise to — kaŭzi give birth — naski give {<unit>} — evolui give up {<refl>_ACC} to — cedi al give up — rezigni give lurch a — ŝanceliĝi give {<H>_ACC} away — perfidi give {<sem.*>_ACC} away — malkaŝi give {<refl>} away — malkaŝi give away — fordoni give off {ACC} — dissendi give way — kolapsi give priority to — prioritati give up {ACC} — prirezigni give in {!ACC} — malinsisti give {<H>_SUBJ} {<refl>_ACC} away — misparoli</p>
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Fig. 2a: QuickDict

QuickDict estas filtrita "ekstrakto" el maŝintraduka vortaro, kaj dum la lastaj jaroj ĝuste la angla-Esperanta vortaro alfrontis grandajn defiojn, ĉar ĝi estis uzata en la *WikiTrans*-projekto (Bick 2011), por tradukado de la angla Vikipedio al Esperanto. Ĉar Vikipedio tuŝas ĉiujn temojn en la mondo, traduki ĝin postulas grandan terminologian/vortaran laboron. Tial, por permesi al la Esperanto-komunumo influi la decidojn, QuickDict ricevis ĝemelan interfacon (<http://xl.wikitrans.net/>), kie eblas ne nur serĉi tradukojn, sed ankaŭ mem proponi aldonojn aŭ ŝanĝojn. Aparte problemaj estas esprimoj kiuj ne estas kunmete tradukeblaj el partradukoj, sed postulas freŝan kaj kreeman plentradukon, ĉar ili:

- estas sufiĉe oftaj en la angla por sentiĝi kiel tutaĵo eĉ kiam ili estas plurvortaj
- ne estas latin-grekaj internaciismoj
- ne laŭparte tradukeblas

- kaj ofte tial eniras aliajn lingvojn senŝanĝe anglaforme (sed en Esperanton tiel prefere ne)

Konsideru ekz. la jenajn ekzemplojn, inter kiuj verŝajne ĉiu leganto trovos ion por li malfacilan:

advocacy group, dropout rate, law-and-order, whistle blower, tax cut, write-off, spinoff, sedan heartland, forfeiture, crawl the net, betting parlor, bandwagon, to backspace, bunkmate, fiche, SUV, computer hacker, birdie

(vidu "solvojn"/proponojn fine de la artikolo)

WikiTrans Translation Game

Does the English term translate to the following Esperanto term?					You			
Problem	bean sauce <small>n</small>	fabsaŭco <small>n</small>	50%	Yes	Also...	Don't know	No/Replace...	93.160.65.38 Points: 1 Sign in with Facebook Sign in with Google Today Weekly Monthly
Problem	fossil show <small>n</small>	fosilia spektaklo <small>n</small>	50%	Yes	Also...	Don't know	No/Replace...	
Problem	dowry murder <small>n act</small>	dotmurdo <small>n</small>	50%	Yes	Also...	Don't know	No/Replace...	
Problem	checkpoint operation <small>n</small>	transirejoperacio <small>n</small>	50%	Yes	Also...	Don't know	No/Replace...	
Problem	gold single <small>n</small>	orunuopaĵo <small>n</small>	50%	Yes	Also...	Don't know	No/Replace...	
Problem	convenience-foods division <small>n</small>	facilmanĝaĵa dividado <small>n</small>	50%	Yes	Also...	Don't know	No/Replace...	
Problem	test task <small>n act</small>	testtasko <small>n</small>	50%	Yes	Also...	Don't know	No/Replace...	
Problem	alcohol brand <small>n</small>	alkoholmarko <small>n</small>	50%	Yes	Also...	Don't know	No/Replace...	
Problem	monster hurricane <small>n wea-wind</small>	monstruragano <small>n wea-wind</small>	50%	Yes	Also...	Don't know	No/Replace...	
Problem	court psychiatrist <small>n Hprof</small>	tribunalpsikiatro <small>n</small>	50%	Yes	Also...	Don't know	No/Replace...	

Add new terms		
English	Esperanto	
value-added tax	<input type="text"/>	<input type="button" value="Add"/>
Part of Speech: ? Semantic: ?		
value-added tax <small>?</small>	aldonvalora imposto <small>?</small>	<input type="button" value="Correct"/> <input type="button" value="Wrong"/>

Fig. 2b: Interaga vortaro

DeepDict mem estas plurjara esplorprojekto de la dana firmao GrammarSoft, kunlabore kun Suddana Universitato, kaj kovras nun 12 lingvojn - krom Esperanto ankaŭ la grandajn romanidajn (latinidajn) lingvojn, la skandinavajn lingvojn, la anglan, la germanan kaj eksperimente la ĉeĥan. La celgrupo estas vortaristoj, tradukistoj, lingvistoj kaj komercaj eldonejoj, sed ni ankaŭ antaŭvidas certan uzon de ĝeneralaj vortarouzantoj, kiuj ĉe iu aŭ alia vorto bezonas pli da kuntekstaj informoj ol ordinara vortaro proponas.

DeepDict ne estas la sola projekto pri rilatiga leksikografio. La plej proksima "parento" estas *Sketch Engine* (Kilgariff et al. 2004), kaj la Leipzig Wortschatz-projekto (Biermann et al. 2004). Ankaŭ tiuj projektoj estas plurlingvaj kaj havas grafikajn retinterfacojn, sed DeepDict distingiĝas en sia uzo de profunda sintaksa analizo kaj kapablo utiligi eĉ longdistancajn dependeco-ligojn. El historia vidpunkto, DeepDict eblis nur surbaze de multaj antaŭaj jaroj da parsilo-verkado en pluraj lingvoj. Ĝi ankaŭ eluzas spertojn el la CorpusEye-projekto (corp.hum.sdu.dk) kaj laboron faritan por diversaj leksikografaj kaj maŝintradukaj esploroj. Kelkaj el la defioj dum la kreado de DeepDict estis pure teknikaj, ne lingvistikaj. Ĉar la kvanto de eblaj vortkombinoj kreskas proporcie al la

kvadrato de vortnombro en la leksiko de iu lingvo, problemis precipe la ekstreme granda datenspaco (ĝis 100 GB), kiu necesigis tute novan solvon por serĉfunkcio¹.

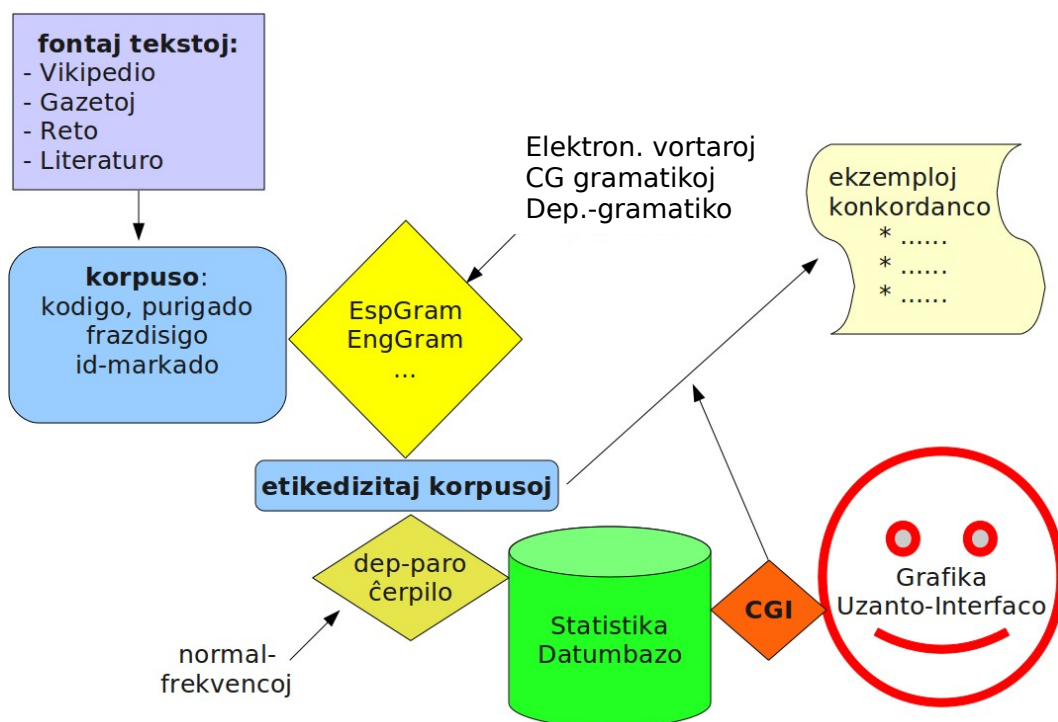


Fig. 3: Paŝoj por produkti (novan) DeepDict

En sia nuna, matura formo, DeepDict permesas aldoni novan lingvon relative facile - necesas nur dependecmarkita korpuso en la koncerna lingvo, kaj eĉ eblas krei apartan DeepDict por individua kliento kun propra korpuso aŭ domajno. Depende de la balanciteco de tia korpuso, necesas 10-20 milionoj da vortoj. Por uzi DeepDict, necesas registriĝi, sed ĝia uzo estas senkosta por Esperanto. Por la plej multaj aliaj lingvoj necesas privata aŭ komerca abonpago, aŭ universitata kunlaborinterkonsento.

5 La DeepDict-interfaco

Word to look up: <input type="text" value="karesi"/> Word class: <input type="radio"/> Noun <input checked="" type="radio"/> Verb <input type="radio"/> Adverb <input type="radio"/> Adjective <input type="button" value="Look up via DeepDict"/>	Lookup language: <input type="radio"/> Danish <input type="radio"/> English <input checked="" type="radio"/> Esperanto* <input type="radio"/> French <input type="radio"/> German <input type="radio"/> Italian <input type="radio"/> Norwegian <input type="radio"/> Portuguese* <input type="radio"/> Spanish <input type="radio"/> Swedish	Lexical frequency threshold: <input type="radio"/> High <input checked="" type="radio"/> Medium <input type="radio"/> Low <input type="radio"/> None Minimum occurrence: <input type="text" value="2"/> Minimum relative frequency: <input type="text" value="0.0"/> Show top: <input type="text" value="25"/> Semantics: <input type="text" value="Exclude"/>
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Fig. 4: DeepDict-portalo

Krom administri entajpadon de vorto, kaj elekton de lingvo kaj vortklaso, la enira interfaco de DeepDict ankaŭ permesas mane ŝanĝi diversajn agordojn, kiel la jam menciitajn frekvencosojlojn.

¹ La teknikan optimumigon de datumbazoj kaj serĉfunkcio prizorgis la programisto de GrammarSoft, Tino Didriksen.

La kvar niveloj de leksika frekvenco (alta, meza, malalta kaj nenia) subtenas uzantojn de malsama lingvonivelo aŭ klereco. 'Alta' signifas ke nur oftaj vortoj montriĝas kiel *kuntekstovortoj*, ĉar oni ekz. al lernantoj volas klarigi maloftan vorton per oftaj (do verŝajne konataj) vortoj, ne per vortoj same aŭ eĉ pli maloftaj ol la serĉvorto. La dua frekvencparametro (minimuma ofteco) rilatas al la ofteco de la ligita vortparo mem, kaj profesia vortaristo eble elektos '0', ĉar li preferas vidi ĉion, dum lernantoj uzu ekz. '4', por nur vidi certajn ekzemplojn kaj forfiltri tajperarojn ktp.

cxevalo (noun)

total of 8257 relations

Hide Frequencies

Premodifiers: 5.05:3 bonrasa · 5.33:2 blankfrunta · 2.32:5 blanka · 3.32:4 galopanta · 5.08:2 okkrura · 2.75:3 troja · 1.7:4 sovagxa · 1.49:4 nigra · 2.47:3 islanda · 1.1:4 ora · 3.94:1 cxara · 3.94:1 trifingra · 3.94:1 trojana · 2.92:2 flugilhava · 2.56:2 bauxmanta · 1.45:3 fajra · 3.25:1 eltena · 1.23:3 hejma · 2.11:2 senspira · 1.04:3 ligna · 2.56:1 kartuzia · 2:1 purrasa · 0.75:2 bronza · 1.54:1 kontrauxula · 0.29:2 cxela		PP postmodifiers: 0.78:1 el bronzo
oni povas ...	9.08:3 seli · 7.2:3 jungi · 8.2:2 sproni · 5.5:3 bredi · 6.19:2 bridu · 6.94:1 brosi · 6.84:1 elterigi · 6.05:1 rekapti · 4.87:2 vipi · 2.77:4 rajdi · 1.65:3 haltigi · 2.49:2 prunti · 1.16:2 peli · -0.73:3 frapi 7.13:3 desalti de · 3.46:5 rajdi sur · 5.47:1 parigxi kun	cxevalo(n)
cxevalo povas ...	10.37:3 heni · 5.94:1 ronki · 4.64:2 galopi · 1:4 tiri · -0.77:4 kuri	
cxevalo povas esti	4.84:2 jungi	...'ta

Fig. 5: substantivo

La grafika aspekto de serĉrezulto en DeepDict dependas de la vortklaso. Por substantivoj, montriĝas antaŭ- kaj postmodifiloj, resp. maldekstre kaj dekstre, por imiti la naturan legofluon de maldekstre al dekstre. Aliaj fakoj estas la subjekto-verbaj (xxx povas umi), verbo-objektaj (oni povas umi xxx) kaj pasivaj kombinoj. La ruĝaj ciferoj indikas unue la korelacioforton, kaj due la frekvencklason de la vortparo. Ambaŭ estas uzataj por ordigi la trovitajn rilatojn, kaj aparte "sekuraj" (do samtempe tipaj kaj oftaj) vortoj estas grasigitaj. La responda tabulo por verboj havas fakojn por subjekto (maldekstre) kaj objekto (dekstre), plus rubrikajn por adverboj kaj tipaj prepoziciaj komplementoj (t.e. prepoziciaj objektoj aŭ adverbialoj). Notu, ke la sistemo kaptas ankaŭ metaforajn uzojn (ekz. karesi + vento aŭ harbuklo).

karesi (verb)

total of 2390 relations

Hide Frequencies

Subjects: 6.55:4 PROP · 4.86:2 zefiro · 2.79:1 harbuklo · 2.64:1 vent · 2.09:1 venteto · 0.87:2 sorto · 1.5:1 nepino · 0.12:2 vento · 0.11:2 fingro · 1.08:1 sunradio · 0.16:1 regxido	Accusative objects: 4.14:3 PROP · 3.02:3 vango · 1.88:4 kapo · 2.78:3 hararo · 2.73:3 kato · 1.91:3 haro · 2.65:2 lipharo · 1.6:3 korpo · 1.93:2 barbo · 2.9:1 vilo · 0.64:3 vizagxo · 1.6:2 virgulino · 1.23:2 frunto · 0.13:3 koro · 1.95:1 herba · 0.84:2 ventro · 1.8:1 glano · 0.65:2 sxultro · 1.56:1 aspiro · 1.33:1 besteto · 1.21:1 buklo · 0.96:1 kapeto · 0.59:1 femuro · 0.41:1 nuko · 0.14:1 resto	
karesi ...	5.22:3 tenere · 2.53:2 milde · 3.08:1 ameme · 3.03:1 antaue · 1.55:2 ame · 2.37:1 mole · 2.3:1 bonhumore · 1.25:1 permane · 1.18:1 neordinare · 1.13:1 tremante	
karesi de ...	4.35:1 mangxmeto	
karesi de ...	0.19:2 mastro	
karesi per ...	1.13:1 maneto · 0.38:1 manplato · 0.25:1 beko	
karesi kiel ...	1.07:1 kompenso	

Fig. 6: verbo

Kaŝiĝas en DeepDict plia nivelo de profundeco: Per klako de ajna ligvorto oni malfermas konkordanco-fenestron (Fig. 8), kiu montras ekzemplofrazojn, fleksian statistikon kaj t.n. vortskizon (Fig. 7). En nia ekzemplo klako al 'korpo' donas la jenon:

korpon -> karesas	4	28.57%	<p style="text-align: right;"><i>word sketch:</i></p> <p>karesas sxian nematuran korpon { . ¶ / / . / dum } belajn virinajn</p>
korpon -> karesi	4	28.57%	
korpojn -> karesas	3	21.43%	
korpon -> karesis	2	14.29%	
korpojn -> karesis	1	7.14%	

Fig. 7: fleksistatistiko kaj vortskizo

monato-s37030	La maro per siaj varmetaj ondoj karesis iliajn korpojn kaj Marta kaj Aniko plezure naĝis en la mola mara ĉirkaŭbrako.
ttt-s76266	Lin multe plezurigas rigardi kaj karesi ŝian nematuran korpon , al kiu mankas ĉiu seksumindiko.
ttt-s46934	kaj iom post iom, dum forpasas la tagnoktoj, li ankaŭ karesas ŝian korpon
utt-233005	Multaj homoj fantazias pri tiu metio, kredante, ke oni karesas belajn virinajn korpojn dum la tuta tago, sed fakte la averaĝa aĝo kaj stato de la klientinoj estas iom seniluziiga.
monato-s42763	La zefiro karesas iliajn junecajn korpojn .

Fig. 8: konkordanco

La lasta granda vortklaso, adjektivo, pli ofte funkcias kiel dependeca ido ol kiel kapo/patrino, do plej plenas la fako kun la substantivoj, kiujn ĝi tipe modifas (Fig. 9)

peza (adjective)

total of 8201 relations
Hide Frequencies

Pre-modifiers: 3.98:7 pli · 4.07:6 malpli · 3.72:6 tro · 2.19:6 plej · 1.83:5 tre · 3.09:3 pli kaj pli · 1.02:4 iom · 0.39:3 tiom · 0.18:3 ege	Post-modifiers: 0.17:1 kiel plumbo	Premodifier of: 4.59:5 sxargxo · 3.91:5 industrio · 3.83:5 elemento · 4.25:4 metalo · 6.07:2 ciferpozicio · 2.84:5 objekto · 4.44:3 tanko · 4.4:3 mitralo · 3.24:4 fortikajxo · 4.17:3 bito · 1.95:5 laboro · 3.65:3 artilerio · 3.47:3 izotopo · 4.17:2 bitoko · 4.12:2 kvarko · 2.11:4 sxtono · 2.57:3 martelo · 1.55:4 tasko · 2.52:3 valizo · 3.52:2 masxinpafilo · 3.43:2 transirmetalono · 3.24:2 kavalerio · 2.24:3 imposto · 1.15:4 korpo · 3.99:1 pafvundo
----------------------------------------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Fig. 9: adjektivo

Notu ke la leksikogramo por 'peza' kaptas kaj la konkretan (*objekto, ŝtono, valizo*) kaj la abstraktan signifon (*laboro, tasko, pafvundo*) de tiu adjektivo, fenomeno kiun DeepDict permesas kompari kaj konstati por 12 lingvoj paralele.

6 Malkovri uzodiferencojn

Kvankam tradukeblaj en ĉiujn lingvojn, la adjektivoj *granda*, *alta* kaj *larĝa* malofte havas ekzaktajn ekvivalentojn, kaj por vere kompreni la signifon de tiaj vortoj, ne sufiĉas konsulti ordinaran vortaron, ĉar la fajnaj diferencoj montriĝas nur per uzo kaj kunteksto. DeepDict permesas eksplacitigi tiajn diferencojn per la plej tipaj korelaciaj vortoj. Tiel, en la kampo 'antaŭmodifikatoro de', ni ricevas la jenajn 3 listojn de tipaj kap-substantivoj:

alta	granda	larĝa
<ul style="list-style-type: none"> • nivelo • punkto • monto • temperaturo • pinto • grado • altitudo • protektanto • prezo 	<ul style="list-style-type: none"> • parto • kvanto • sukceso • nombro • urbocentro, urbo • buntpego • lago 	<ul style="list-style-type: none"> • strio, bendo • spektro • zono • flugilo (longa?) • senco • gamo • avenuo, strato • publiko
<p><degree> <height></p>	<p><size> <importance> <quant-mass></p>	<p><extension></p>

Fig. 10: adjektiva semantiko

Se oni forigas la kapvortojn kaj demandas al fluaj Esperanto-parolantoj, kiu listo apartenas al kiu kapvorto, la plej multaj verŝajne tuj elektus la ĝustajn parigojn. Tamen, sen la listoj, malmultaj sukcesus vortumi kio estas la semantika (aŭ uzo-) diferenco inter la tri adjektivoj. La avantaĝo de DeepDict estas ke ĝi per tipaj ekzemploj eksplacigas aliel kaŝitajn regulecojn. En nia ekzemplo la listoj montras ke 'alta' uziĝas por aferoj mezureblaj per unuoj, dum 'granda' uziĝas por graveco, kvanto kaj grandeco ne mezurebla per unuoj. Ĉe la abstrakta uzo de 'larĝa', fine, temas pri spektro da samnivelaĵoj. 'Granda' publiko, ekzemple, ne estas la sama kiel 'larĝa' publiko, kaj *larĝa lago* sonas pli strange ol *granda lago*, en la menso ĝi metamorfoziĝus al io longeca kiel rivero aŭ markolo, dum la mensa bildo de *granda lago* imageble aspektas pli prototipe ronda.

Denaska angla-parolanto foje malcertas ĉu traduki "to know" per "scii" aŭ "koni", sed DeepDict povas malkaŝi ke la unuan oni uzas kun informoj kaj faktoj (scii ion), la alian por personoj (koni iun) kaj lokoj:

scii	koni
<ul style="list-style-type: none"> • malmulto, multo • lingvo 7, Esperanto 6 • <i>kialo</i> 5, <i>kaŭzo</i> 2 • nomo 6, vorto 5 • respondo, signifo • vero 4, sekreto 4 • detalo 4, dato 4 • <i>tion, kion, nenion</i> 	<ul style="list-style-type: none"> • lingvo 7, Esperanto 6 • historio • nomo 6, vorto 6 • homo • situacio, kulturo • vero 5, sekreto 5 • vojo, tero • persono, dio, -isto, viro • verko, fundamento • detalo 5, kialo 4 • <i>kiun > tion, lin > ĝin</i>
<p><quantity>, <information> <ion></p>	<p><person>, <place>, <system> <iun></p>

Fig. 11: verba semantiko (duoblaj vortoj kun frekvencklasoj)

7 Konkludoj kaj perspektivoj

DeepDict estas efika metodo por samtempe (sambilde) elmontri la ĉefajn korelaciojn kaj komplementojn de serĉvorto. Por konstrui sian leksikan sciobankon, DeepDict uzas bazon de gramatike analizitaj korpusoj kiuj permesas elĉerpon kaj statistikan traktadon de t.n. depgramoj (paraj rilatoj inter vorto kaj ĝia dependanto).

La eblaj uzoj de DeepDict estas multaj - krom la jam menciita utilo por uzanto de aktiva (fremdlingva) vortaro, ĝi ankaŭ povas liveri materialon por instruekzercoj. Ekzemple eblas facile krei vortkampojn - la serĉvorto *lingvo* tiel produktas liston de adjektivoj kiel *angla, franca, germana* ktp., kaj la kapvorto *voĉdoni* liveras vortkampon por eseeto pri balotoj. Fine, el komerca-eldonista vidpunkto, DeepDict estas valora ilo por profesia vortaristo, kiu povas en ĝi

- trovi kaj kontroli sintaksajn argumento-strukturojn
- trovi la plej tipan (ne simple la plej oftan!) ekzemplon
- trovi kandidatojn por plurvortaĵoj aŭ konstruverboj
- trovi kandidatojn por metafora uzo
- malkovri semantikajn distingojn kaj subsignifojn, kiuj ne estas memevidentaj, kaj kiuj normale ne estas kontrastigataj, ekz. la uzodiferencon inter *rigardi, spekti* kaj *vidi*
- kompari norman kaj veran uzojn, ekz. *komenci, eki* kaj *starti*

Interesa perspektivo, el lingvistika vidpunkto, estas ke la datenbazo de DeepDict pri dependecparoj invitas al multaj aliaj aplikoj. Unu tia ideo estus krei Esperantan valentvortaron, aŭ artefaritajn tipajn ekzemplofrazojn por instrua uzo. Tiel, simple ĉenigante la plej tipajn dependantojn de la verbo *aboni*, do kombinante la plej tipan subjekton kun la plej tipa objekto, adverbialo ktp., oni ricevas jenan iom strangan, sed tre tipan kaj "difinan" frazon:

esperantisto senpage abonas gazeton (al blogo, por jaro)

Eluzante la statistikajn informojn rekte, eblas kalkuli la probablecon ke subjekto aŭ objekto de iu verbo apartenas al certa semantika klaso. Frekvencetikedoj kun tia informo, ekz. <SUBJ/H:73> (73-procenta verŝajneco ke la subjekto de la koncerna verbo estas homa), jam uziĝas en kelkaj el niaj parsiloj, kaj eblus uzi ilin ankaŭ en Esperanta lingvoscienco, ekzemple por faciligi senambiguigon de funkciaj etiketoj en EspGram, aŭ por marki eblan metaforan uzon (kie komplemento *ne* obeas la statistikan regulon, do apartenas al neatendita semantika klaso).

Bibliografio

- Bick, Eckhard. 2007. "Tagging and Parsing an Artificial Language: An Annotated Web-Corpus of Esperanto", En: *Proceedings of Corpus Linguistics 2007, Birmingham, UK*. Elektronike publikigita ĉe (<http://ucrel.lancs.ac.uk/publications/CL2007/>, Nov. 2007)
- Bick, Eckhard. 2005-1. "Turning Constraint Grammar Data into Running Dependency Treebanks". En: Civit, Montserrat & Kübler, Sandra & Martí, Ma. Antònia (red.), *Proceedings of TLT 2005 (4th Workshop on Treebanks and Linguistic Theory, Barcelona, December 9th - 10th, 2005)*, pp.19-27
- Bick, Eckhard. 2005-2., "CorpusEye: Et brugervenligt web-interface for grammatisk opmærkede korpora". En: Peter Widell & Mette Kunøe (red.), *10. Møde om Udforskningen af Dansk Sprog 7.-8.okt.2004, Proceedings*. pp.46-57, Århus University
- Bick, Eckhard. 2011. WikiTrans, The English Wikipedia in Esperanto. En: *Constraint Grammar Applications, Workshop at NODALIDA 2011, Riga, Latvia*. (NEALT Proceedings Series, Vol. 14, pp. 8-16. Tartu: Tartu University Library. ISSN 1736-6305)

- Biemann, Chris & Stefan Bordag & Uwe Quasthoff & Christian Wolff. 2004. "Language-Independent Methods for Compiling Monolingual Lexical Data". En *Comp. Linguistics and Intelligent Text Processing*. Springer: Berlin, pp. 217-228
- Church, Ken and P. Hanks. 1991. Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics*, Vol.16:1, pp. 22-29.
- Karlsson, Fred et al. 1995: *Constraint Grammar - A Language-Independent System for Parsing Unrestricted Text*. Natural Language Processing, No 4. Berlin & New York: Mouton de Gruyter.
- Kilgarriff, Adam, Rychlý, P., Smrž, P. & Tugwell, D. 2004. "The Sketch Engine". Paper presented at EURALEX, Lorient, France, July 2004.
- Koehn, Philipp. 2005. Europarl: A Multilingual Corpus for the Evaluation of Machine Translation. MT Summit X, Sept.12-16, 2005. Phuket, Thailand.

Solvo de la tradukokvizo (sekcio 4):

advocacy group: aktivulgrupo, interesgrupo, lobio; **dropout rate:** ĉesprocento, rezignoprocento, ĉesofteco; **law-and-order:** nulkrima; jurprotekta; **whistle blower:** interninformanto, fajfulo, malkaŝisto; **tax cut:** impostredukto, impostmalpliigo; **write-off:** elkontigo; deprecado; rezigno; **spinoff:** kromprodukto; idofirmao; **sedan:** kabinaŭto, familio-aŭto, limuzino; **heartland:** kernregiono; kern- (ekz. heartland state); **forfeiture:** puncedaĵo; puncedo; **crawl the net:** troli la reton; **betting parlor:** vetsalono; **bandwagon:** festoĉaro, muzikkamiono; **to backspace:** reklavi; **bunkmate:** litnajbaro; **fiche:** minipaĝo; **SUV:** sportkamioneto, (altradulo?); **computer hacker:** komputilpirato, retpirato; **birdie:** birdeto; {golf} minustruo

Dr. phil. Eckhard Bick estas germana komputilingvisto vivanta en Danio. Li studis en Bonn, London kaj Århus kaj akiris diplomojn pri medicino, lingvistiko kaj Nordiaj lingvoj. Nuntempe li laboras por Suddana Universitato (Odense) kaj por sia propra firmao GrammarSoft, kreante porkomputilajn gramatikojn kaj leksikojn por la plej multaj ĝermanaj kaj romanidaj (latinidaj) lingvoj, uzatajn i.a. en maŝintradukado (gramtrans.com), perreta instruado (visl.sdu.dk) kaj korpusa lingvistiko (corp.hum.sdu.dk). Flankprofesie Eckhard Bick ankaŭ instruas akupunkturon kaj laboras pri Esperantaj projektoj, kiel Lingvohelpilo kaj WikiTrans. Li publikigis plurajn vortarojn, librojn pri ĉina medicino, Esperantajn tradukojn (Muminvalo) kaj eĉ Esperantan ludon (GEO-ludo), kaj li estas la ĉefaŭtoro de la Esperanta lernolibro "Tesi la Testudo".