DeepDict: Konteksta reta vortaro de vera lingvouzo

Eckhard Bick
Suddana Universitato & GrammarSoft

eckhard.bick@mail.dk

Resumo: *DeepDict estas sistemo por ĉerpi vortkontekstojn (precipe dependencajn 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 domajnojn 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 ofertas konkordancojn de ekzemplofrazoj. Aktuale DeepDict traktas 9 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, introspektatdon 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 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 kooperativaj 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, kombineblaj serĉoj (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-receptiva 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-mal-plia 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 konteksto elekti kiun tradukalternativon, kiel fleksii la cellingvan vorton ktp. Tial aktivaj vortaroj bezonas pli da (a) gramatikaj informoj, (b) ekzemplojn de sintaksa-semantika kompletigo, kaj (c) senco-distingajn kriteriojn (distingilojn). Temas do pri en-konteksta vortouzo, kaj multe helpas povi montri al la uzanto la plej tipan uzokonteksto 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 valenco kaj kombineblecoj (b), sed ankaŭ ofte - nerekte - sencaj distingoj (c), ĉar la kombinvortoj 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 al uzo de artikolo aŭ adjektivo.

Por fidinde 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 formo de tradicia literaturo, kaj la plej ofte uzata materialo estas gazetaro aŭ politikaj tekstoj (ekz. Europarl, Koehn 2005). Vikipedio ankaŭ estas interesa fonto, kaj pro ĝia liberfonteco, kaj pro ĝia 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. Gravas 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 (bazformigon), kio permesas rekoni fleksiitajn formojn de la sama vorto, tiel plibonigante la statistikan findindecon 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 balancigo de tekstotipoj, tre gravas por vortaristo kiom kaj kiel la strukturaj informoj en korpuso estas alireblaj. En neprilaborita, kruda tekstkorpuso, 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 ties 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-nomata parsilo, 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. dependencaj 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)

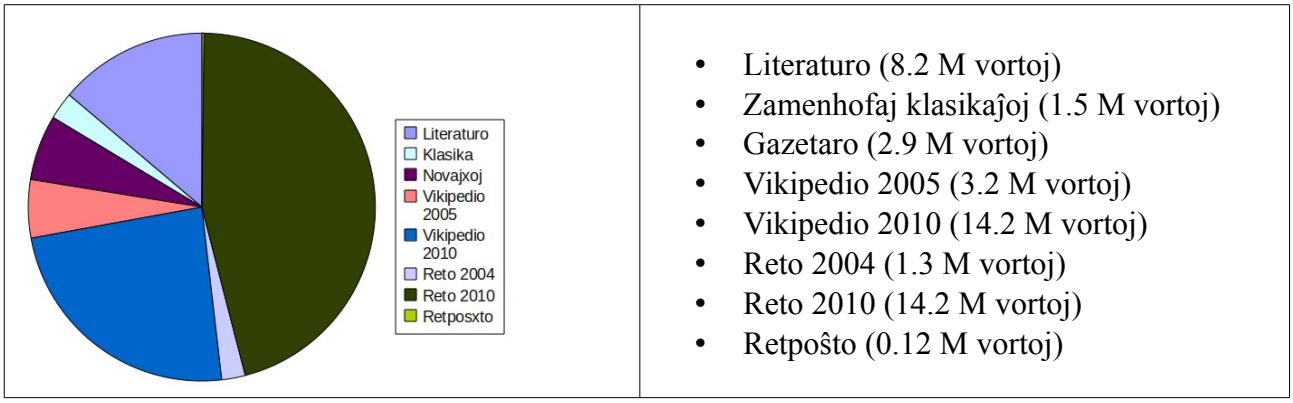

Fig. 1: Korpusaj fontoj

El la etikedizitaj korpusoj eblas facile ĉerpi ekzemplojn pri certaj strukturoj, ekz. statistikon pri 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 ofertas ekz. relativan frekvencordigon de la trovitaj substantivoj.

3 La kerna ideo de DeepDict

DeepDict volas esti konteksta vortaro, kiu klarigas vortojn kaj ties 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 dependenc-markita 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 dependencparojn 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 kion oni povas manĝi, kaj kiu povas esti manĝanto.

Tamen, por la plej multaj serĉvortoj, tiaj rilato-listoj fariĝus neuzeble grandaj se oni simple kompilus 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 interdependecon 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 kunan oftecon per la produkto de la unuopaj oftecoj (Church 1990). La formulo de DeepDict aplikas tiun metodon en logaritma skalo, kaj pezigas la kunan oftecon uzante ties kvadratnombron:

Korelacioforto: K = C * log(p(a->b) ^2/ (p(a) * p(b)))

Ni rigardas kiel tipajn tiujn vortligojn kiuj havas pozitivan korelacioforton super difinita sojlo. Krome, DeepDict markas la relevantecon de ligita vortparo per logaritma frekvencklaso. Frekvencklaso 0 tiel signifas ke la korpuso enhavis nur unu ekzemplon de la koncerna rilato, frekvencklaso 4 signigas 9-16 ekzemplojn (16 = 2 je la 4a potenco). Raraj ekzemploj estas danĝeraj - unue ili riskas esti tajperaroj aŭ analizeraroj, 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 substingoj 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. sekso (li - ŝi - ĝi), kvanteco (iom, tiom) aŭ informon pri adverbialaj komplementoj (tie, tien). Pro tio ni konservas pronomojn individue, sed rangigas ilin aparte, ekster la listo de aliklasaj vortligoj. Finfine, eksperimente, DeepDict ofertas ĝeneraligon per semantikaj prototipoj - do ĉe la vorto manĝi ĝi anstataŭ (aŭ kune kun) la listo de manĝaĵoobjektoj montras simple la kategoriojn <food> (manĝaĵo), <fruit> (frukto), <Aich> (fiŝo) ktp. Entute ekzistas ĉirkaŭ 200 tiaj prototipoj por substantivoj, kiujn liveras la etikedizo 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).

Angla-Esperanta *QuickDict* • 168.000 leksemoj (vortoj) • 192.074 tradukoj • 182.000 distingitaj signifoj • el tio 42.000 nomoj • (+ fiksaj esprimoj)	**give** give *(n)* — elasteco ; cedemo give *(v)* — doni 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

Fig. 2: QuickDick

DeepDict estas plurjara projekto de la dana firmao GrammarSoft, kunlabore kun Suddana Universitato, kaj kovras nun 9 lingvojn - krom Esperanto ankaŭ la grandajn romanidajn lingvojn, la skandinavajn lingvojn kaj la anglan kaj germanan. 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 kontekstaj informoj ol ordinara vortaro ofertas.

DeepDict ne estas la sola projekto pri rilatiga leksikografio. La plej proksima ”parenco” estas *Sketch Engine* (Kilgariff et al. 2004), kaj la Leipzig Wortschatz-projekto (Biermann et al. 2004). Ankaŭ tiuj projektoj estas plurlingvaj kaj ofertas grafikajn retinterfacojn, sed DeepDict distingiĝas en sia uzo de profunda sintaksa analizo kaj kapablo utiligi eĉ longdistancajn dependenco-ligojn. El historia vidpunkto, DeepDict eblis nur surbaze de multaj antaŭaj jaroj da parsilo-verkado en pluraj lingvoj, kaj eluzas spertojn el la CorpusEye projekto (corp.hum.sdu.dk) kaj esplorojn faritajn 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 serĉfunkcio[1].

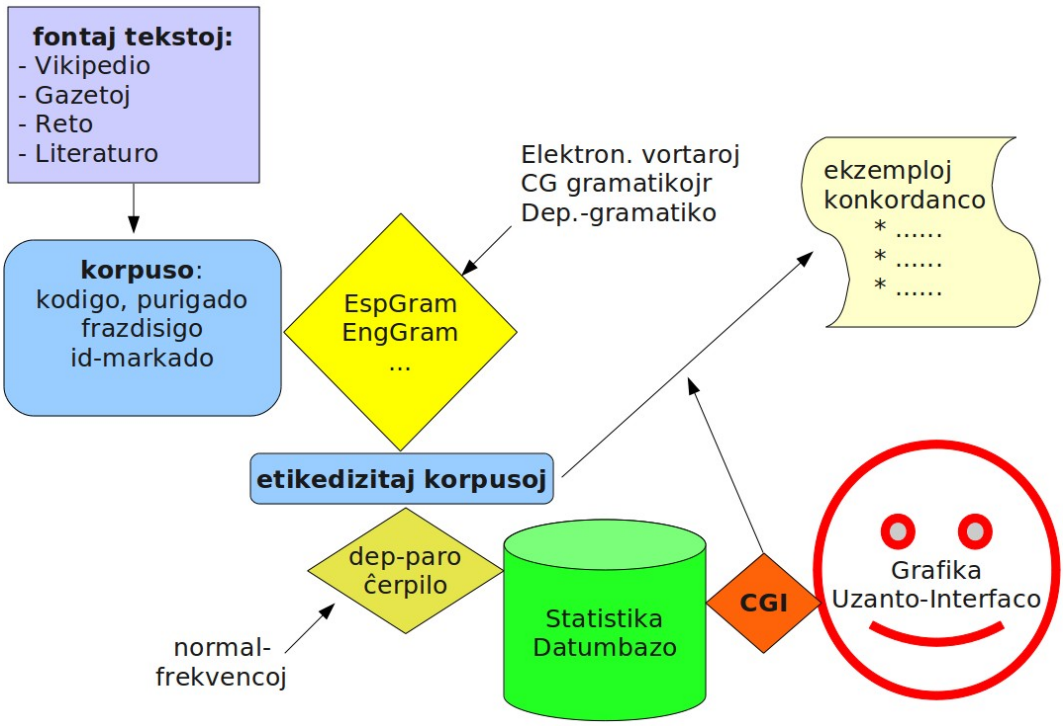

Fig. 3: Paŝoj por produkti (novan) DeepDict

En sia nuna, matura formo, DeepDict permesas aldoni novan lingvon relative facile - necesas nur dependencparkita korpuso en la konceerna 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 akademia kunlaborinterkonsento.

5 La DeepDict-interfaco

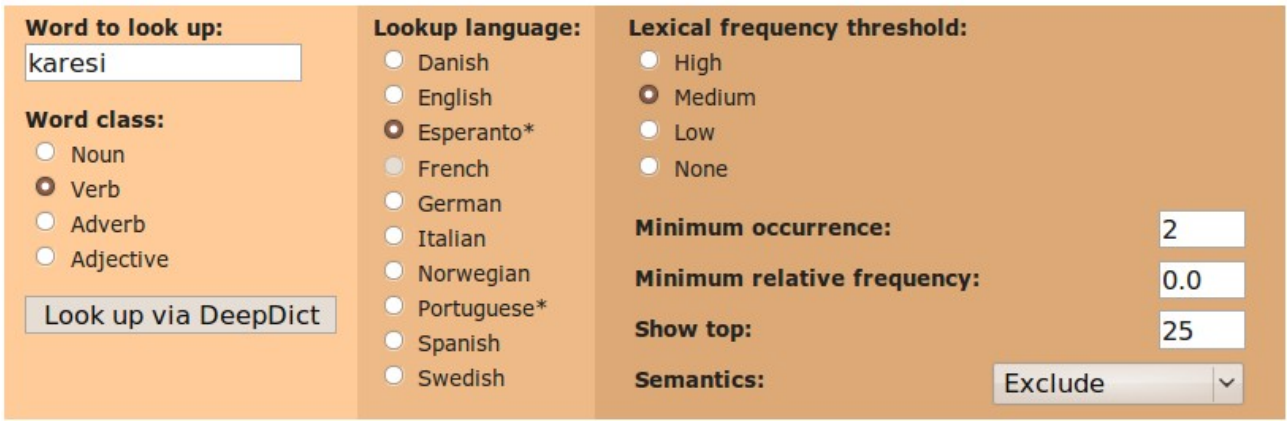

Fig. 3: DeepDict portalo

Krom administri entajpadon de vorto, kaj elekton de lingvo kaj vortklaso, la enira interfaco de DeepDict ankaŭ permesas man ŝanĝi diversajn agordojn, kiel la jam menciitajn frekvencsojlojn. La kvar niveloj de leksika frevenco (alta, meza, malalta kaj nenia) subtenas uzantojn de malsama

1 La teknikan optimigon de la datumbazo kaj serĉfunkcio prizorgis la programisto de GrammarSoft, Tino Didriksen.

lingvonivelo aŭ klereco. 'Alta' signifas ke nur oftaj vortoj montriĝas kiel *kontekstovortoj,* ĉar oni ekz. al lernantoj volas klarigi maloftan vorton per oftaj (do verŝajne konataj), 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 volas preferas vidi ĉion, dum lernantoj uzu ekz. '4', por nur vidi certajn ekzemplojn kaj forfiltri tajperarojn ktp.

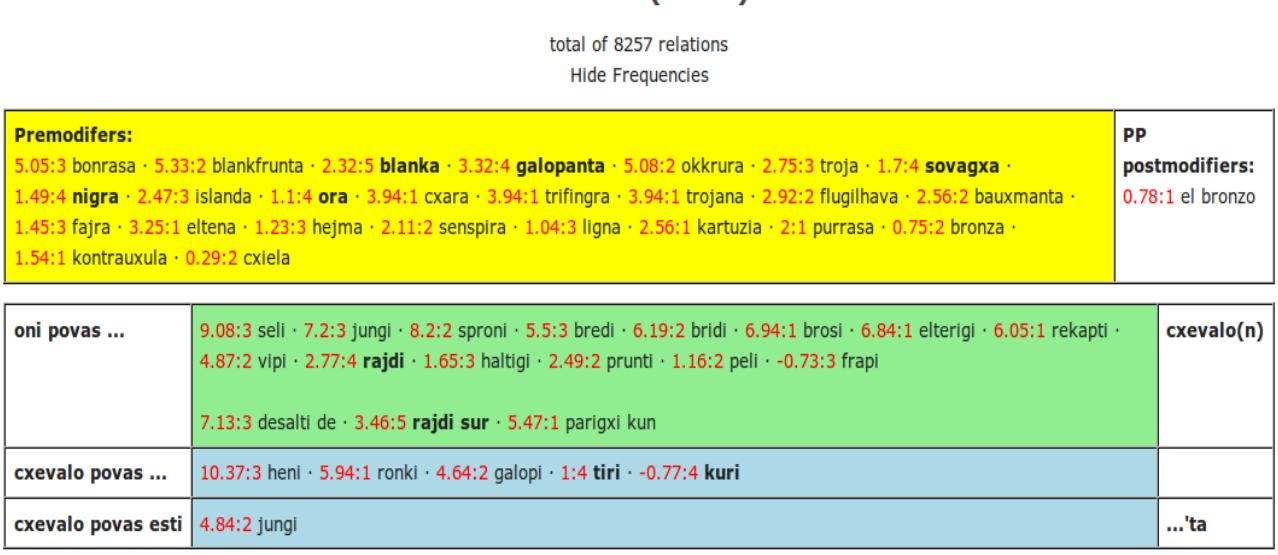

Fig. 4: substantivo

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 subjekt-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 smatempe tipaj kaj oftaj) vortoj estas grasigitaj.

La koresponda tabulo por verboj havas fakojn por subjekto (maldekstre) kaj objekto (dekstre), plus rubrikojn por adverboj kaj tipaj prepoziciaj komplementoj (prepoziciaj objektoj aŭ adverbialoj). Notu, ke la sistemo kaptas ankaŭ metaforajn uzojn (karesanta vento aŭ harbuklo).

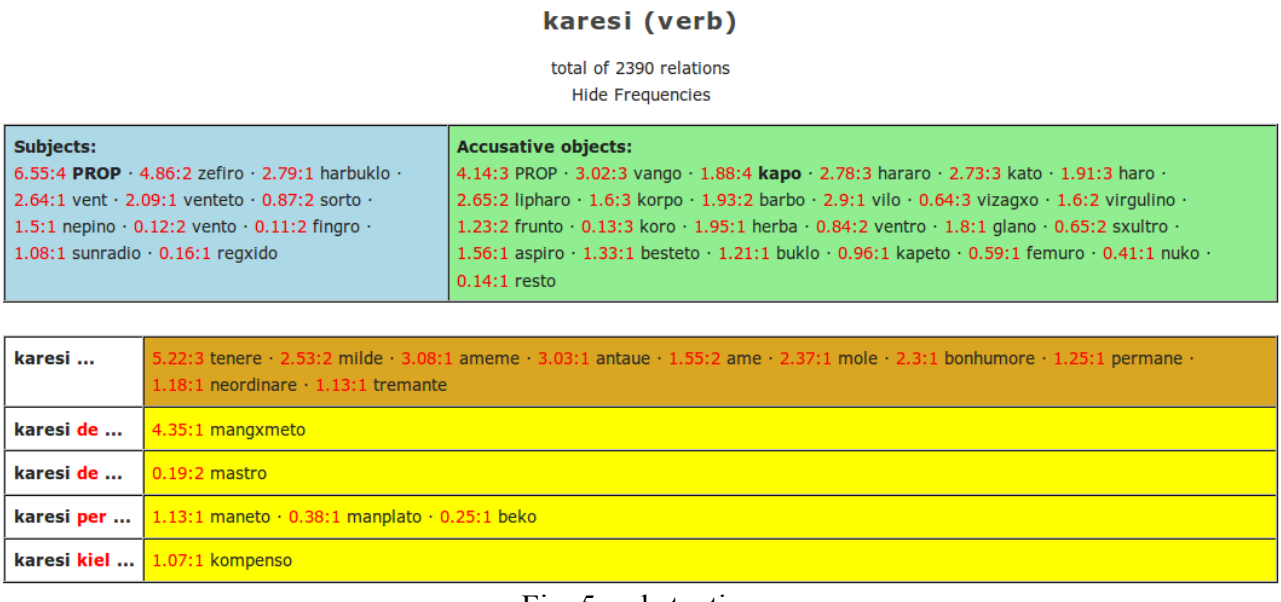

Fig. 5: substantivo

Kaŝiĝas en DeepDict plia nivelo de profundeco: Per klako de ajna ligvorto oni malfermas konkordanco-fenestron (Fig. 7), kiu montras ekzemplofrazojn, fleksian statistikon kaj t.n. vortskizon (Fig. 6). En nia ekzemplo klako al *'korpo'* donas la jenon:

korpon -> karesas	4	28.57%
korpon -> karesi	4	28.57%
korpojn -> karesas	3	21.43%
korpon -> karesis	2	14.29%
korpojn -> karesis	1	7.14%

word sketch:

karesas sxian nematuran belajn virinajn **korpon** {. ¶ / , / . / dum}

Fig. 6: fleksistatistiko kaj vortskizo

monato-s37030	La maro per siaj varmetaj ondoj **karesis** iliajn **korpojn** kaj Marta kaj Aniko plezure nagxis en la mola mara cxirkauxbrako .
ttt-s76266	Lin multe plezurigas rigardi kaj **karesi** sxian nematuran **korpon** , al kiu mankas cxiu seksumindiko
ttt-s46934	kaj iom post iom , dum forpasas la tagnoktoj , li ankaux **karesas** sxian **korpon**
uttt-233005	Multaj homoj fantazias pri tiu metio , kredante , ke oni **karesas** belajn virinajn **korpojn** dum la tuta tago , sed fakte la averagxa agxo kaj stato de la klientinoj estas iom seniluziiga
monato-s42763	La zefiro **karesas** iliajn junecajn **korpojn**

Fig. 7: konkordanco

La lasta granda vortklaso, adjektivo, pli ofte funkcias kiel dependenca filino ol kiel kapo/patrino, do plej plenas la fako kun la substantivoj, kiujn ĝi tipe modifas (Fig. 8)

peza (adjective)

total of 8201 relations

Hide Frequencies

Pre-modifers:	Post-modifiers:	Premodifier of:
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	0.17:1 kiel plumbo	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 transirmetalo · 3.24:2 kavalerio · 2.24:3 imposto · 1.15:4 **korpo** · 3.99:1 pafvundo

Fig. 8: 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 9 lingvoj paralele.

6 Malkovri uzodiferencojn

Kvankam tradukeblaj en ĉiujn lingvojn, la adjektivoj *granda, alta* kaj *larĝa,* malofte ekzistas ekzaktaj ekvivalentoj, kaj por vere kompreni la signifon de tiaj vortoj, ne sufiĉas konsulti ordinaran vortaron, ĉar la fajnaj diferencoj nur montriĝas per uzo kaj konteksto. DeepDict permesas ekspliciti 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**
• nivelo	• parto	• strio, bendo

• punkto • monto • temperaturo • pinto • grado • altitudo • protektanto • prezo	• kvanto • sukceso • nombro • urbocentro, urbo • urbo • buntpego • lago	• spektro • zono • flugilo (longa?) • senco • gamo • avenuo, strato • publiko
<degree> *<height>*	*<size>* *<importance>* *<quant-mass>*	*<extension>*

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 eksplicitas 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 metamorfozus al rivero aŭ markolo, dum norma lago estas pli simetrie ronda.

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 bazo de gramatike analizitaj korpusoj kiuj permesas elĉerpon kaj statistikan traktadon de t.n. depgramoj (paraj rilatoj inter vorto kaj ĝia dependento).

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 (la kompletecon de) 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 ka subsignifojn, kiuj ne estas memevidentaj, kaj kiuj normale ne estas konstrastitaj, ekz. la uzodiferencon inter *rigardi, spekti* kaj *vidi*
- kompari norman kaj veran uzon, ekz. *komenci, eki* kaj *starti*

Interesa perspektivo, el lingvistika vidpunkto, estas ke la datenbazo de DeepDict pri dependencparoj invitas al multaj aliaj aplikoj. Unu tia ideo estus krei esperantan valencvortaron, aŭ artefaritajn tipajn ekzemplofrazojn por instrua uzo. Tiel, simple ĉenigante la plej tipajn dependentojn 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 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 disambiguigon de funkciaj etikedoj 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 publikita 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

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.