WikiTrans: La angla Vikipedio en Esperanto

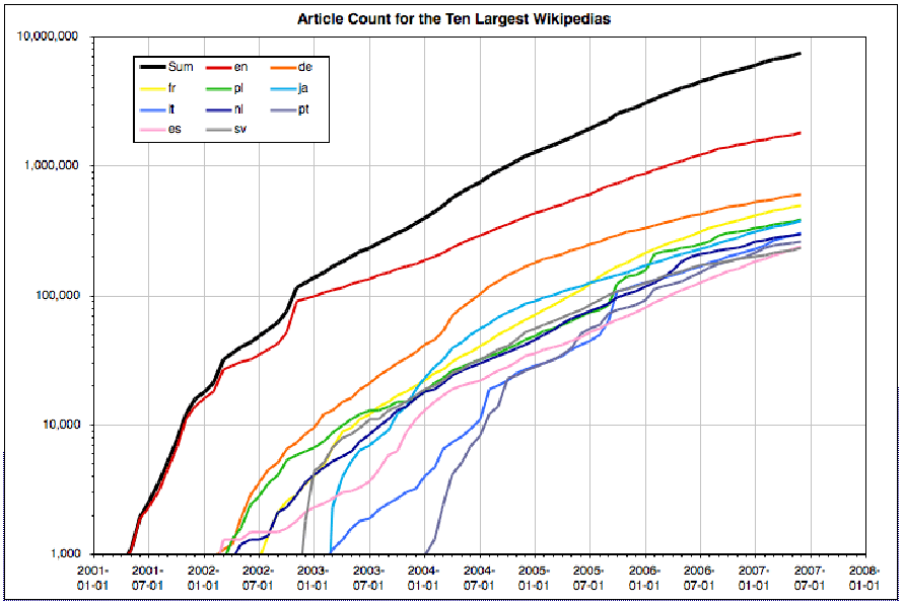
Eckhard Bick
GrammarSoft ApS & Suddana Universitato
eckhard.bick@mail.dk

Resumo: *WikiTrans estas tradukprojekto kaj retejo por tradukita(j) Vikipedio(j). Uzante la regulbazitan GramTrans-teknologion de la dana firmao GrammarSoft, la projekto kreis altkvalitan angla-esperantan MT-sistemon, kaj sukcesis aŭtomate traduki la kompletan anglan Vikipedion (ĉ. 3.000.000 artikoloj), kun rapideco de 17.000 artikoloj tage. La tradukitaj artikoloj estas plene serĉeblaj, kaj loke (www.wikitrans.net), kaj paralele en la originala esperanta Vikipedio, kun ebleco redakti (”originaligi”) tradukitajn artikolojn. Ĉi tie ni klarigas la koncepton kaj defiojn de la projekto, kaj montras kiamaniere ĝiaj tradukreguloj eluzas la gramatikajn analizojn de CG-parsilo (Celkonteksta Gramatiko).*

1 Motivigo

Praktike, Vikipedio estas nuntempe la ĉefa enciklopedia informfonto en la mondo, kaj laŭkvante kaj laŭuze, kaj kvankam la kvalito de la individua artikolo povas varii, sistemo de reciproka kontrolado, fontigo-devo kaj disputomarkoj funkcias kiel efika kvalitomarkilo kaj ebligas al konscia uzanto mem juĝi la fidindecon de ofertitaj informoj. Sed kvankam demokrata kaj egalema el verkista vidpunkto, Vikipedio neniel atingis lingvan egalecon, kaj ĝia informriĉeco estas multe pli granda en la angla kaj kelkaj aliaj ĉefaj kulturlingvoj ol en malgrandaj lingvoj (Fig. 1). La diferenco ne nur videblas en la kvanto de kovritaj kapvortoj, sed ankaŭ en la longeco kaj kvalito de la individua artikolo. Lingvaj baroj do malhelpas la ĉefan celon de Vikipedio - igi la scion de la mondo alirebla al ĉiuj.

Fig. 1: Kronologia lingvostatistiko en Vikipedio



La esperanta Vikipedio, kvankam impona relative al sia uzantobazo, kaj samgranda kun ekz. la dana, havas nur ĉirkaŭ 140.000 artikolojn, dum la angla havas pli ol 3.4 milionojn da artikoloj (= 2.345.000.000 words*), do estas 24-oble pli granda. Krome malsamas la meza longeco de artikolo[1], kun ĉ. 3.600 literoj (ĉ 600 vortoj) en la angla kaj germana, kaj iom pli ol 1500 literoj (ĉ 250 vortoj) en Esperanto, do la diferencofaktoro inter esperanta kaj angla vikipedioj, el enhava vidpunkto, kreskas al 57. Tio praktike signifas, ke pli ol 98% de la anglaj informoj ne habeblas en Esperanto. Oni povus argumenti ke la esperantaj artikoloj kovras ĝuste la gravajn kaj oftajn temojn, sed ĝuste ĉe tiaj artikoloj ofte sentiĝas la diferenco en profundeco, aŭ la kvanto de enaj ligoj.

1 http://stats.wikimedia.org/EN/TablesArticlesBytesPerArticle.htm

La evidenta solvo por tiu problemo, el la vidpunkto de la aŭtoro, estas tradukado de la angla Vikipedio al Esperanto, tiel ebligante al E-parolantoj aliron al la angla "ĉef-Vikipedio" - kaj/aŭ aliaj grandaj lingvoj, io kio permesus pli kritike kompari informojn kaj sintenojn. Per homfortoj, kun tradukrapideco de 500 vortoj hore, tia projekto por la angla-esperanta lingvoparo kostus 4.690.000 homajn laborhorojn aŭ. En Danio tio egalus al 3.000 laborjaroj, aŭ - kun 0.25 EUR/vorto - ĉ. 600 milionoj da eŭroj. Neimageble granda sumo, preter la subvenciaj fantazioj de eĉ la plej favora finvenkisto. Kaj eĉ se eblus iel trovi monon kaj fortojn, ne eblus facile ĝisdatigi la tradukitan Vikipedion, ĝi estus eterne ekstertakta kun la originalo, kaj kiam jam tradukita, rigida kaj malfacile adaptebla.

2 La solvo

La plej logika solvo al la skizita dilemo estas uzo de maŝintraduka sistemo por ŝpari homfortojn, kun libervola laŭokaza homa postredaktado, ekzemple por gravaj artikoloj aŭ simple laŭ la profesiaj /hobiaj emoj de redaktemaj uzantoj. Maŝintradukado (MT) kaj solvus la kvanto-problemon kaj la aktualeco-problemon, ĉar relative facile eblus retraduki, tuj traduki novajn artikolojn kaj eĉ eble sekvi ŝanĝojn en ekzistantaj artikoloj individue kaj tuje. Problemo tamen estas, ke ne temas pri simplaj tekstoj, ke la kovrota leksiko estas giganta, kaj ke uzanto postulus fluan kaj alireblan tradukon sen tro da eraroj kaj postrestantaj fontlingvaĵoj. Por la plej multaj lingvoj simple ne ekzistas MT de sufiĉa kvalito, kaj Esperanto krome tute mankas en la ofertogamoj de la komercaj MT-sistemoj, ĉu Google, Systran aŭ alia.

Teknike ekzistas du baze malsamaj aliroj al la tasko de maŝintradukado - unuflanke regulbazita, aliflanke statistika. Ambaŭ havas avantaĝojn kaj malavantaĝojn. La tradicia aliro estas la regulbazita, kiu bone enplektiĝas kun la analiza-struktura lingvistika tradicio. Tamen bona regulbazita MT estas tre laboriga, kaj tro dependas de lingvista specialscio por logi komercajn firmaojn se temas pri eta lingvo sen vera merkato kiel Esperanto. Statistika maŝintradukado ne bezonas lingvistojn kaj verkistojn, sed nur ties datumojn, kaj kun dulingva tekstokolekto kaj prefere iom da lingvistike markita tekstaro, eblas ”trejni” tradukmodelon malmultekoste por nova lingvo aŭ nova domajno. La problemo estas, ke la kvalito proporcias al la kvanto de trejndatumoj, kaj ke bona statistika MT bezonas gigantajn jam-tradukitajn, t.n. paralelajn korpusojn. Google ekzemple havas tion en la formo de dulingvaj retpaĝoj, sed ne en sufiĉa kvanto por malgrandaj lingvoj.

GramTrans (Bick 2007) estas nova aliro al MT, kiu ja estas regulbazita, sed eluzas la robustecon de CG-gramatiko (**C**elkonteksta **G**ramatiko, angle *Constraint Grammar,* Karlsson 1990) por analizi frazon/tekston, kaj tial povas oferti pli altkvalitan lingvistikan bazon ol kutime, kiun povas eluzi la traduka programaro mem (Fig. 2). Ekzemple, la tradukmodulo povas eluzi dependencajn ligojn inter vortoj, same kiel ties funkciojn (ekz. 'subjekto', 'predikativo') kaj semantikajn klasojn (ekz. 'ilo', 'vehiklo', 'manĝaĵo'), por formuli kondiĉojn por selektado de tradukalternativoj en ambiguaj kazoj. Reguloj en Celkonteksta Gramatiko havas kiel celon forigi, aldoni aŭ ŝanĝi lingvistikajn etikedojn de vortoj (ekz. vortklaso, fleksio, funkcio), kaj ĉiu regulo konsistas el listo de kontekstaj kondiĉoj (ekz.. agordoj, najbaraj vortoj, ĉeesto de homa aganto ktp.), kiuj devas esti plenumitaj por permesi la celitan operacion.

3 La WikiTrans-projekto

GramTrans estas la motoro en la tradukteknologio de la dana firmao GrammarSoft, kiu uzas ĝin, kunlabore kun la norvega firmao Kaldera, por traduki inter la skandinavaj lingvoj, kaj inter tiuj unuflanke kaj la angla (kaj iom la germana) aliflanke. GrammarSoft kunlaboras kun Suddana Universitato, kaj sciencaj esploroj ludas relative grandan rolon en ĝiaj projektoj. Do la firmao pretis, esplorcele kaj kiel konceptopruvon, lanĉi projekton sen komerca potencialo, WikiTrans, kun la intenco alirebligi grandlingvajn Vikipediojn traduke al malgrandlingvaj uzantoj, kun la angla-espearnta lingvoparo kiel la unua. Krom la aŭtoro, komputika lingvisto, kunlaboras Tino Didriksen,

la programisto de GrammarSoft.

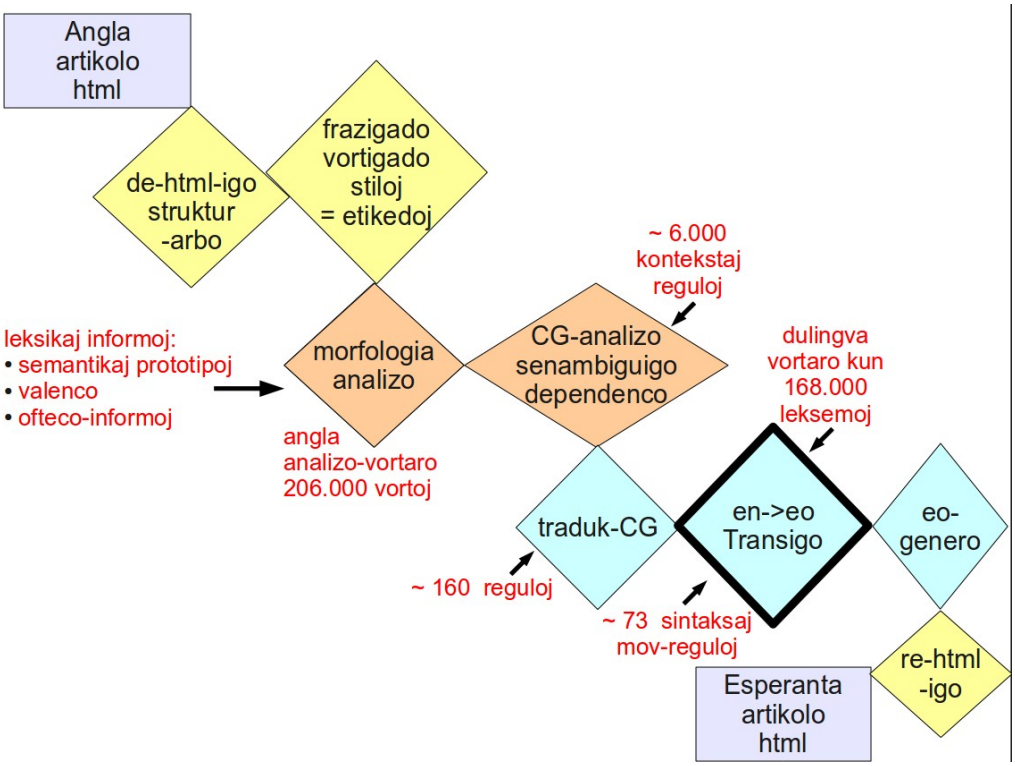


Fig. 2: Taskoskemo de la WikiTrans-traduko

La WikiTrans-projekto koncipiĝis fine de 2009 kaj trapaŝis la sekvajn fazojn:

- preparfazo: 2009 - februaro 2010: lingvistika kaj vortara laboro
- 1-a tradukfazo (feb/mar 2010): 100.000 plej oftaj artikoloj
- 2-a tradukfazo (mar-jun 2010): 500.000 plej longaj artikoloj, aŭ kun unuvortaj titoloj
- 3-a tradukfazo (jun-dec 2010): tri milionoj da artikoloj (ĉiuj)
- uzofazo: aktualigo, retradukoj, homa reviziado

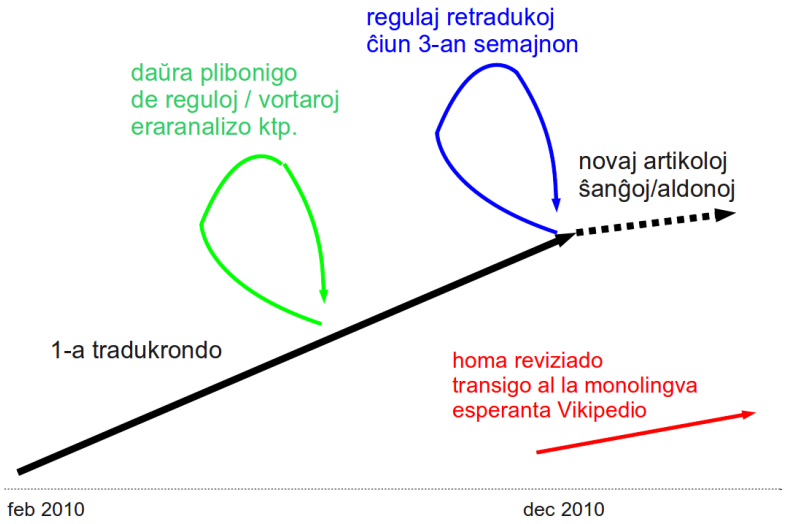


Fig. 3: Projektofazoj de WikiTrans

4 Kiel serĉi en WikiTrans

Grava kialo por traduki la tutan Vikipedion, kaj ne simple traduki unuopan artikolon, kiam uzanto petas tion, estas la ebleco aliri kaj sisteme serĉi informojn. Tradukante artikolon vive, necesus serĉi

anglalingve aŭ traduki serĉvorton al la angla, kaj poste elekti inter la anglaj artikoloj trovitaj antaŭ ol traduki. Tia servo vere nur utilus as dulingvanoj kiuj preferas legi ion en esperanto kion ili ankaŭ povus legi en la angla. Do, por serĉi en Esperanto, necesas ke la priserĉata tekstaro estu en Esperanto, despli se oni pripensas ke ofte serĉvorto(j) ne mem estas titoloj, sed jes aperas plurfoje ene de la tekstokorpo de iu aŭ alia artikolo.

En WikiTrans, ni uzas la malfermfontan serĉprogramon *Lucene,* kiu permesas ankaŭ plurajn serĉvortojn samtempe, kaj mem ordigas la trafojn laŭ verŝajna relevanteco, uzante aritmetikon pri ofteco de (kun)okazo der serĉvorto(j) en la unuopa artikolo. *Lucene* markas tion per verŝajneco-indekso inter 0 kaj 1, kaj ni alprogramis kelkajn pliajn kriteriojn: Ekzemple, artikolo iras al la kapo de la listo, se la serĉvorto aperas en la titolo, aŭ - ĉe plurvorta serĉesprimo - se la vortoj aperas unu apud la alia. Kion la uzanto vidas, estas listo de la unuaj maks. 20 trovitaj artikoloj, kun kaj titolo kaj tekstoĉerpaĵeto (Fig. 4). Per klako la uzanto nun povas esplori unu aŭ plurajn el la trovitaj artikoloj, kiuj montriĝas - el aspektiga vidpunkto - ekzakte kiel la originalo, kun la samaj bildoj, tabelstrukturo ktp, sed tute en Esperanto.

El programista-teknika vidpunkto, la serĉoproblemo estas kompleksa ankaŭ pro la enorma volumo de la datenoj - pli ol 10 gigabajtoj da teksto (100 gigabajtoj kun gramatikaj etiketado). Por povi efike traserĉi tian datenspacon, necesas bona datenbaza sistemo, serĉmemoro ks. Male al ekz. korpuslingvisto kiu pretas atendi minutojn por efektiviki iun statistikon aŭ ekzemploĉerpadon el sia tekstobazo, la pacienco-horizonto de averaĝa uzanto de Vikipedio estas kelkaj sekundoj, prefere malpli ol unu sekundo. Alikaze multaj homoj reklakas la butonon, fakte eĉ malplifaciligante la situacion de la servilo, kiu nun dufoje devas fari la saman serĉon.

Fig. 4: De serĉesprimo al WikiTrans-artikolo

Fine, por permesi tradiciajn alfabetajn serĉojn, aŭ por havi superrigardon pri la gamo de artikoloj, ni ankaŭ permesas simple foliumi en WikiTrans, progresante de A-Z-listo por la unua litero en titolo, al A-Z-listo de la dua, tria ktp. ĝis individue klakebla unuekrana alfabeta sublisto de artikoloj.

5 Ligoj kaj bibliografio

Grava aspekto de elektronika enciklopedio, kaj unu el la ĉefaj avantaĝoj kompare kun papera, estas la internaj ligoj. Estas ili kiuj ebligas unuflanke havi legeble longan kaj fluan artikolon, aliflanke la profundecon de multe pli granda artikolo. Simplaj tien- kaj reen-klakoj permesas al ĉiuj legi la artikolon ekzakte je sia individua klereco- nivelo, uzante aŭ ne-uzante internajn ligojn por klarigi sciencajn fakvortojn, bildprezenti personojn aŭ esplori la fakan kontekston de iu aserto.

Teknike, la internaj ligoj estis unu el la gravaj solvendaĵoj dum la projekto. La problemoj estis pluraj: Unue, ne ekzistis garantio ke la ligita artikolo jam estas tradukita, kaj ni devis do aldoni funkcion de viva traduko dum la unua projektojaro, kaj zorgi ke estis sufiĉe da komputila forto je dispono por tio, faktoro kiun ni limigis uzante tradukmemoron. Due, ĉar temas pri inteligenta, kontekstdependa tradukado, la sama vortoĉeno povas ricevi malsamajn tradukojn en malsamaj lokoj, tiel ke la ligotraduko ne nepre identas al la traduko de la ligita artikoltitolo. Ni solvis tiun problemon konservante la originalan anglan terminon (aŭ ciferecan transformon de ĝi) en la <a href> marko mem, nevidebla al la uzanto. Post la traduka kaj datenbasiga fazoj do venis paŝo (ĉ. unu-seamjno da komputiltempo) trairante ĉiujn ligojn, kaj anstataŭante ilin kun tiu traduko kiun ricevis la koncerna angla titolo.

Eksteraj ligoj kaj referencoj estis teknike pli simplaj, sed ofte plenaj je nomoj, mallongigoj kaj ciferesprimoj malfacile tradukeblaj. Post unuaj provoj traduki la plejparton, ni nu aplikas pli singardan aliron, ne tradukante grandan parton de la nomoj, kaj diskutas tute ne tuŝi eĉ titolojn de verkoj ktp. Ĉar estas sufiĉe malfacile klasifiki kio estas verko, personnomo, eldonejo, urbo ktp en reerenco, la plej facila solvo estus tute ne tuŝi la bibliografian sekcion de Vikipedio-artikolo, konsciante ke ja ankaŭ la (ekstera) teksto al kiu oni ligas, ne estas en Esperanto, kaj iusence pli estas pruvilo por la ĝusteco de la artikolaj informoj, ol parto de la artikolo mem.

6 Integrigo kun la unulingva esperanta Vikipedio

La reagoj al la WikiTrans-projekto, kiujn ni ricevis el la esperanta komunumo, estis ĝenerale tre pozitivaj, kvankam por multaj ŝajne la movada aspekto (reklamo) pri gravas ol la faka aŭ simple informserva. Estas malfacile por nefakulo aprezi la malfacilecon de la tasko, aŭ propraokule kompari tradukkvaliton kun tiu de aliaj malgrandaj lingvoj en ekz. la traduksistemo de Google aŭ Babelfish, kaj la plej komprenebla kritiko estas tial, ke la tradukokvalito ne estas sufiĉa, kaj ke la projekto ”diluas” la kvaliton de la originala esperanta Vikipedio. Kaj prave, kvankam sufiĉe bonaj por glata legado, la tradukoj ne estas seneraraj, kaj tradukita artikolo ne estas nova originalo.

Tamen, al tiu argumento oni povas respondi, ke ankaŭ sen MT-sistemo, jam ĉiam, Vikipedio-verkistoj en malgrandaj lingvoj ĉerpas tradukante el grandlingvaj artikoloj. Fakte, la malfermfonta etoso de Vikipedio mem subtenas tiun fluon de tekst(er)oj de unu lingvo al alia. Do, ĉu ne estas pli bone povi fari tiun laboron pli efike kaj rapide kun la helpo de aŭtomata sistemo? Kio necesas, estas simple kontrolo de kio estas kio, kaj kie oni troviĝas en retumila klakĉeno - en la origina angla, origina esperanta aŭ tradukita esperanta Vikipedio. Mem ni proponas semaforan kolormarkigon - ruĝa makulo en angulo por pure MTa WikiTrans-artikolo, verda por plene reviziita kaj flava por parte reviziita. ”Verdajn” artikolojn oni do povus (kun reteno de la marko) movi al la vera Vikipedio, dum ruĝaj kaj flavaj estas serĉeblaj kaj tra WikiTrans kaj - kiel paralelserĉo kaj truoplenigilo - en la origina esperanta Vikipedio. Fig. 5 montras skemon de ebla integrigo de WikiTrans kun la originala Vikipedio.

Interinforme kun la administrantoj de Vikipedio, ni inter julio 2010 kaj februaro 2011 traktis la praktikajn aspektojn de tia integrigo, kiu nun funkcias tiel, ke lokaj javscript-programoj ĉe la uzanto interagas kun la programaro de GramTrans ĉe ties servilo. Prizorgis la uzantoflankajn programojn Marek Blahus (E@I), dum Tino Didriksen (GrammarSoft) realigis la necesan GramTrans interfacon kun traktado de la interna Wiki-sintakso, kaj grafikan redakciilon. Dum verkiĝas tiu-ĉi teksto, jam eblas individue alelekti la WikiTrans-redakto-sistemon por ĉiu registrita Vikipedio-uzanto, kaj simpla WikiTrans-kromserĉado (kaze de netrafoj en la originala Vikipedio) jam funkcias sen konscia alelekto.

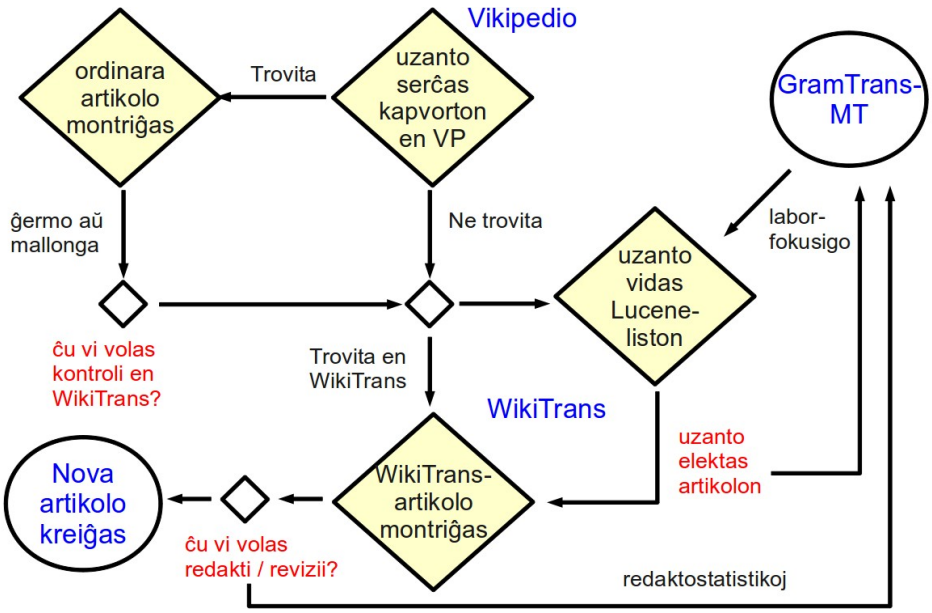


Fig. 5: Integrigo kun la originala Vikipedio

7 Lingvistikaj aspektoj de la traduksistemo

El klasifika vidpunkto, GramTrans estas nek surfaca nek interlingva traduksistemo (Fig. 6). Ĝi evitas, kompreneble, la problemojn de simpla vorton-post-vorto-tradukado, sed ne riskas abstrahi ĝis interlingva nivelo. La ”kostoj”, en formo de malrobusteco, por plena interlingva simboligo estas tre altaj, kaj eblas atingi kvazaŭ la samon per pli ”plata” transigo de font- al cellingvo, simple ĉar la plej multaj lingvoparoj havas, strukture kaj semantike, pli da komunaĵoj ol da diferencoj. Tio validas ankaŭ por la angla-esperanta paro - despli ĉar Esperanto kiel cellingvo estas ege fleksebla, kaj permesas esprimi trajtojn el multaj aliaj lingvoj sen ke la rezulto sentiĝas malnatura.

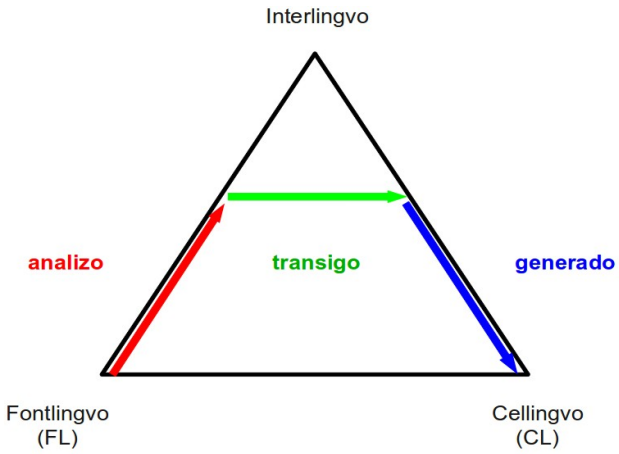


Fig. 6: La traduktriangulo

GramTrans baziĝas sur matura kaj robusta fontlingva analizo, atingita per la EngGram parsilo (http://visl.sdu.dk/visl2/constraint_grammar.html). EngGram estas CG-gramatiko kun pli ol 6000 reguloj, 200.000-vorta analizovortaro, kaj profunda dependenca analizo (Bick 2005), kaj GramTrans eluzas la kategoriojn kaj vortligojn de la fontlingva analizokonteksto por krei regulojn por la t.n. transigo-paŝo, kie solviĝas ambiguecoj kaj elektiĝas la ĝusta traduko el pluraj eblaj. La tri paŝo, generado, profitas krom el la flekseblecо de Esperanto ankaŭ el la fakto, ke kreado de esperanta morfologio (finaĵo aŭ afikso) preskaŭ egalas al koni la ĝustan lingvistikan kategorion (tempo, nombro, vortklaso ktp), do estas preskaŭ embarase simpla. Sola komplikaĵo estas la sintakso, ĉar kvankam oficiale lingvo kun libera vortordo, Esperanto tamen havas sufiĉe fortajn uzonormojn pri vortordo, kaj malatenti ilin malfaciligus fluan legadon.

7.1. Leksika transigo: La listo-mito

Laŭ naiva supozo, eblus traduki frazon vorton post vorto, 1-1, kaj eĉ ‘multaj-al-unu rilato (kie plura fontlingvaj vortoj havas la saman tradukon) ne ĝenus. La problemo estas, ke ofte temas pri unu-al-multaj rilato:

1. open - malfermi, malfermiĝi, malferma
2. goal - celo, golo, golejo
3. crack - fendi, kraki; fendeto, kraketo; drogo

En (1) temas pri vortklasa kaj transitiveca variaĵoj de la sama radiko, en (2) la cellingvo uzas kaj 2 malsamajn radikojn, kaj ankaŭ subdistingas inter loko (golejo) kaj poentoj (golo). (3) estas duoble kompleksa, kaj vortklasa (verbo/substantivo) kaj polisemia en ambaŭ vortklasoj. La plej simpla solvo por tiu problemo funkcias brute-ne-bele: Oni prenus tiajn dulingvajn vortolistojn, kiuj oni povas ricevi malfermfonte, kaj uzas simple la unuan tradukon, se estas pluraj. Pli bone estas almenaŭ ordigi la tradukalternativojn laŭ frekvenco, prototupeco aŭ multsignifeco. La lasta kategorio aparte bone funkcias se eblas trovi tradukon kiu havas la *saman* ambiguecogamon kiel la originalo. Sed eĉ tio ne solvas la bazan polisemion, kaj la cetero de la ĉapitro diskutos la solvojn kiujn aplikas GramTrans.

7.2. Leksika transigo: aldonante dimensiojn al la listo

La plej simpla solvo por la disambiguigo-tasko estas 1-dimensia, kaj bazigas leksikajn distingojn sur (1-2) vortklaso aŭ (3-4) fleksio, uzante nur la vorton mem. Konteksto en tiu metodo nur eniras ne-rekte, ĉar ĝin uzis la fontlingva analizilo (t.n. parsilo) por siaj decidoj:

1. type_N (substantivo) :tipo, :speco
2. type_V (verbo) :tajpi
3. force_NS (singularo) :forto
4. force_NP (pluralo) :armeo, :trupo

Pli ambicia alternativo estas 2-dimensia strategio, kun tradukmatrico anstataŭ listo, disigante vorton en plurajn signifojn (aŭ eĉ kontekste malsamuzajn sinonimojn). Sed dum la 1-a dimensio venas "senkoste" de la parsilo, la 2-a bezonas *senso-distingilojn* - specon da leksikaj mini-reguloj kiuj kombinas unu aŭ pluraj el la jenaj:

(a) kontekst-kondiĉoj (mallokaj distingiloj)
(b) afikso-kondiĉoj, difiniteco, funkcio, rolo (lokaj distingiloj kun foje malloka signifo)

Eĉ normale lokaj trajtoj povas foje esti ĉerpitaj de la konteksto, ekz. nominalagordo (inter substantivo kaj determinantoj), aŭ semantika projekciado de verbo al substantivo (ekz. ”homeco”-

trajto por la subjekto de kognitiva aŭ komunika verbo). La subaj ekzemploj montras ekzemplon de la uzo de sintaksa funkcio-markilo.

rather_ADV ... S=(@ADVL) :**prefere**; S=(@>A) :**sufiĉe**;
too_ADV ... S=(@ADVL) :**ankaŭ**; S=(@>A) P2?=(INFM)_por :**tro**; D=(@>A) :**tro**

[dependencaj rilatoj: S=Self/mem, D=Daughter/filino, M=Mother/patrino, B=Brother/frato GD=Granddaughter/nepino, GM=Grandmother/avino; dekstraj pozicioj: P1, P2 ... Pn; maldekstraj pozicioj: P-1, P-2 ... P-n]

La celitaj distingoj ne nepre devas sekvi tradiciajn vortarajn aŭ enciklopediajn distingojn. Unue, metaforoj aŭ genro-variado povas esti izomorfaj en ambaŭ lingvoj, do sen bezono eksplici tion en la cellingvo. Due, unu el la grandaj sekretoj de MT (kaj kialo ne iri ĝis la pinto en la MT-triangulo) estas la neceso *distingi,* pli ol *difini.* Alivorte, sufiĉas havi sufiĉe da konteksta kaj semantika scio en la sistemo por povi elekti unu aŭ alian tradukon, sed la fina *kompreno* ja okazas en la cerbo de la leganto, kiu havas multe pli da mondkonteksta scio ol la komputilo povas havi - do ne necesas por la sistemo ekspliciti ĉion je abstrakta, superlingva nivelo. Granda parto de la semantiko simple transportiĝas netuŝite de la font- al la cellingvo, sen iam vere disambiguiĝi. Ekzemple, metafora uzo de ujoj kiel kvantovortoj funkcias simile en ĉiuj lingvoj (2 *glasoj da* biero). Aliflanke, necesas foje ankaŭ devas distingi (precipe cellingvajn) *uzo-diferencojn* (sinonimoj, frekvenco), aldone al la signifo-diferencoj. Tiu problemo estas malpli akra en Esperanto ol en aliaj lingvoj, sed ĝi ekzistas.

7.3 Kontekstaj kondiĉoj - la skeleto de la sistemo

Kune, la diversaj eblecoj de disambiguigo permesas sufiĉe kompleksan leksikografian laboron, kaj precipe gravas la ebleco ligi verbajn informojn kun tiuj de la komplementoj. La suba ekzemplo montras kiel kontekstaj distingiloj permesas traduki la anglan verbon 'apply' al 9 malsamaj verboj en Esperanto. Se neniu alia kontekstokondiĉo aplikeblas, la unua traduko, 'uzi', esto elektita kiel la plej robusta.

apply_V :uzi;
D=("for")_pri :peti D=(<H> @SUBJ) D=("to" PRP)_por :kandidatiĝi
D=(@ACC) D=("to" PRP)_al :apliki
D!=(@ACC) D=("to" PRP)_por :validi
D=(<(conv|sem)> @SUBJ) D!=(@ACC) :validi
D=(<(cm.*|rem)> @ACC) :surŝmiri
D=("dressing" @ACC)_pansaĵo :surmeti
<¤vr> D=("to" PRP)_pri :koncentriĝi
D=("match")_alumeto :malestingi

[@SUBJ=subjekto, @ACC=rekta objekto, PRP=prepozicio, <H>=homa, <conv>=konvencio, regulo, <sem>=semantikaĵo, <cm>=konkreta masvorto, <rem>=medikamento, substanco, <¤vr>=refleksivo]

7.4 Plurvortaĵoj, tradukmemoro kaj nomoj

En kelkaj kazoj, ne havas sencon traduki vortvicon analitike, la signifo de la tuto estas pli ol la sumo de la partoj. GramTrans traktas tiaĵojn kiel "vortoj" kun spacoj. Povas temi pri kompleksaj substantivoj *(recovery_position - savpozicio),* kazo tre ofta en la angla, pri sintagmoj kun nelaŭparta traduko *(in_violation_of - malobee_al, every_inch_as - tute_same, all_year_round - tutjare)* aŭ pri aliel fiksitaj esprimoj *(see_also - vidu_ankaŭ).* Fine kunfando de kompleksaj prepozicioj aŭ konjunkcioj simpligas la markadon de sintaksaj funcioj kaj rilatoj: *each_other (unu_la_alian), instead_of (antataŭ), other_than (krom).*

Simile eblas trakti ”tradukmemorajn listojn” (TM), oftaj en MT-sistemoj, kaj taŭgaj por kovri specialvortojn kiuj tradukiĝu ĉiam same, do kiujn eblas enmeti en la tradukon sen uzo de kontekstaj reguloj. Aplikareo estas ekz. fakvortaj listoj (terminaroj), kiujn eblas ŝalti aŭ malŝalti depende de la tradukota domajno. Sed ankaŭ eblas per TM malhelpi oftajn erarojn de la sistemo, kie oni unufoje-por-ĉiam decidas fiksan tradukon. En la redaktointerfaco, kiun ni programis por WikiTrans-artikoloj, la sistemo tial memoras ĉiujn homfaritajn ŝanĝojn. Krom liveri superrigardon pri traktendaj problemoj, eblos ĉerpi el tiu ”korektobazo” materialon por la tradukmemoro, kaj eĉ oferti al posta redaktontoj proponmenuon ĉe ofte korektataj vortoj aŭ esprimoj.

Sendepende de la nomrekona kvalito de la parsilo (t.n. NER - *named entity recognition*), nomoj malfacile tradukiĝas, kaj grandparte mankas en vortaroj. Eĉ eblas fidi je majusklouzo, ĉar povas temi pri simpla emfazo aŭ frazkomenco, kaj tial estas malfacile centprocente distingi inter nomo kaj nekonata aŭ kunmetita vorto de alia klaso. El pure traduka vidpunkto la ĉefproblemo estas la decido ĉu entute traduki nomon, reteni la originan formon, aŭ transliterumi ĝin en esperanta fonetiko. Indas i.a. distingi inter du ĉefaj kazoj:

(a) institucioj kaj okazaĵoj, kiu tradukiĝu parton-post-parto
European Union - Eŭropa Unio, Olimpics - Olimpikoj, World War II - Dua Mondmilito
(b) personnomoj kaj produktnomoj, kiuj restu senŝanĝaj
*Geroge Bush - **Georgo Arbusto*

Aparte por Vikipedio-traduko, ni ankaŭ foje faras ambaŭ, do retenas la originalon, kaj aldonas tradukon inter krampoj, ekz. ĉe titoloj aŭ aliaj nomoj kiuj estas klare markitaj en la html-strukturo kiel tiaj.

8 Struktura transigo

La lastaj paŝoj de traduko-transigo estas struktura generado kaj morfologia generado. Ekzemplo de struktura problemo estas, ke foje ne troveblas cellingva vorto kiu povas plenigi la sintaksan pozicion de la fontlingva vorto, aŭ ke necesas forigi aŭ aldoni strukturojn (angla negacio kun *don't,* angla demando kun *do,* esperanta demando kun *ĉu*). Krome necesas foje reordigi frazpartojn aŭ sintagmopartojn, kaj GramTrans enhavas regulojn ankaŭ por tio, ŝanĝante la sinsekvon de patrino-filino-grupoj en la dependenca arbo. Simpla kazo estas genitivoj *(Miachel's father - la patro de Michael).*

Kiel dirite, al dua genera tasko, morfologia generado, estas sufiĉe simpla en Esperanto, sed kie Esperanto havas pli da gramatika specifeco ol la angla, necesas konteksto - ekz. por eltrovi pri la plurala finaĵo de adjektivoj, aŭ la akuzativa -n. Ankaŭ, la du lingvoj ne uzas participojn kaj infinitivojn tute same (ekz. angla *have*-pasinteco), kaj foje malharmonias semantika kaj surfaca nombroj *(wages - salajro, stools - feko).*

9 Konkludoj kaj perspektivoj

WikiTrans (www.wikitrans.net), profesia lingvoteknologia projekto, sukcesis dum sia unua jaro krei angla-esperantan traduksistemon de sufiĉa kvalito por aŭtomate traduki enciklopediajn tekstojn, kaj en decembro 2010 finis tradukon de la ĉ. 3 milionoj da artikoloj en la angla Vikipedio, kun rapideco de ĉ. 17.000 artikolojn tage. La sistemo ne nur ofertas cellingvan serĉadon ene de artikoloj, sed ankaŭ bone integriĝas al la originala Vikipedio, permesante postredaktadon de tradukita artikolo en speciala interfaco. La perspektivo por 2011 estas ellaboro de sistemo kiu kapablas aŭtomate retraduki kaj aktualigi WikiTrans, kaj por tio ni planas instali ĉe Suddana Universitato, kun financa subteno de ESF (Esperanto Studies Foundation), apartan plursistemon *(cluster)* el 8 kvarkernaj komputiloj por permesi pli rapidan, kaj paralelan tradukadon. Depende de kiom bone la komunumo

akceptas kaj eluzas la redakteblecon por WikiTrans-artikoloj, ni antaŭvidas regulan traktadon de erarstatistikoj kaj oftaj korektoproponoj.

Lingvistika defio estas la terminologia laboro: Malgraŭ la fakto ke la WikiTrans-vortaro jam estas la plej granda angla-esperanta vortaro iam ajn, multaj fakaj terminoj daŭre tradukiĝas per heŭristikaj metodoj, t.e. analizaj aŭ partaj tradukoj, transliterigoj, latinismoj ktp. Necesas minimume permana validigo de tiuj proponoj (aŭ per la aŭtoro, aŭ per kunlabora retportalo), integrigo de ekzistantaj terminaroj (depende de kopirajto), kaj - se entute eblas por tioma kvanto da vortoj - terminologia diskuto en la esperanta lingvokomunumo. El kvanta vidpunkto, la dua WikiTrans-cellingvo, la dana, planita por 2011, similas al Esperanto, kun simile multaj artikoloj, kaj simile granda dulingva vortaro, - kaj ni atendas certan sinergion, ekz. ĉe la identigo de ”nekonataj” tradukendaj plurvortaĵoj, precipe kompleksaj anglaj substantivoj, aŭ ĉe la traktado (listigo, klasado) de nomesprimoj.

Alia logika evoluo estus aldoni pliajn fontlingvojn al la sama cellingvo - Esperanto, por permesi la plenigon de ”kulturaj truoj”, ebla problemo de ajna unulingva enciklopedio, aŭ por instigi al inter-kultura kompara de samtemaj artikoloj (ekz. politikaj, historiaj aŭ religiaj). GramTrans mem jam estas ellaboranta dan-esperantan sistemon, kaj ankaŭ eblus aldoni tradukojn el pliaj lingvoj per malfermfontaj sistemoj, ekz. Apertium (http://www.apertium.org/), se kaj kiam tia sistemo atingas la necesan kvaliton.

Bibliografio

Bick, Eckhard. 2005. ”Turning Constraint Grammar Data into Running Dependency Treebanks”. In: 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. 2007-1. ”Dan2eng: Wide-Coverage Danish-English Machine Translation”. In: Bente Maegaard (ed.), *Proceedings of Machine Translation Summit XI, 10-14. Sept. 2007, Copenhagen, Denmark*. pp. 37-43

Bick, Eckhard. 2007-2. ”Fra syntaks til semantik: Polysemiresolution igennem Dependensstrukturer i dansk-engelsk maskinoversættelse”. In: Henrik Jørgensen & Peter Widell (eds.), *Det bedre argument, Festschrift til Ole Togeby på 60-årsdagen* pp.35-52

Karlsson, Fred. 1990. Constraint Grammar as a Framework for Parsing Running Text. In: Karlgren, Hans (ed.), COLING-90 Helsinki: Proceedings of the 13th International Conference on Computational Linguistics, Vol. 3, pp.168-173