Nova vojo al aŭtomata tradukado

17h00-18h00 mar IKU8 (Witkam), Lapenna

Toon Witkam

Naskiĝis 1944. Studis aeronaŭtikon ĉe la Teĥnika Universitato en Delft (Nederlando). Sekvis kariero en la softvar-industrio, de kiu li dediĉis la ĉefparton al perkomputila lingvistiko. Ĉe la Nederlanda firmao BSO en 1979 li iniciatis la esplorprojekton DLT (Distribuita Lingvo-Tradukado) por aŭtomata tradukado kun Esperanto kiel interlingvo. En 1990 li laboris kiel gast-esploristo ĉe ATR Interpreting Telephony Research Laboratories en Kioto. En 1991 li konsilis la Eŭropan Komisionon pri aŭtomata tradukado, kaj de 1992 ĝis 1996 li estis profesoro je informadiko kaj kogna ergonomio, denove en Delft. Ekde 2002, li okupiĝas pri vortstatistika analizo de Esperantaj tekstoj.

Resumo

Post kiam fervoraj esplorprojektoj pri permaŝina tradukado dufoje ekfloris en la pasinta jarcento, la ambicia celo de altkvalitaj aŭtomataj tradukoj ĝisnun ne realiĝis. Hodiaŭ, ĉiu interret-uzanto povas tuj ekhavi iun perkomputilan tradukon el pluraj lingvoj, foje eĉ senpagan, sed neniam seneraran. Fakte, la nuna komerce disponebla traduktehxnologio ankoraŭ radikiĝas en la frua evoluigperiodo de la 1960-aj kaj 1970-aj jaroj.

Tamen, tio ne signifas ke nuntempe la esplor-klopodoj haltis. La faktereno de aŭtomata tradukado ankoraŭ vigle vivas, kaj dum la lastaj ĉ. 15 jaroj disvolviĝis firma nova paradigmo: Perstatistika Aŭtomata Tradukado (PAT). Kontraŭe al antaxua esplorbazo, la lingvistiko, kiu kulminis en la 1980-aj jaroj, PAT surprize bazi?as sur la metio de la homa tradukisto mem. Alie ol la lingvisto, la tradukisto prilaboras tekstojn multe pli praktike, kutime el la vidpunkto de interkultura komunikado kaj ne laŭ la abstraktaĵoj de lingvistika teorio.

Ĝuste pro tio, longatempe la tradukista sperto ŝajnis malpli 'formaligebla' ol la lingvista scio. Sed finfine la konceptoj de la esploristaro ŝanĝis: la lingvista scio ne liveris sufiĉajn rezultojn, kaj la intertempe tre progresinta memorkapableco de komputiloj ebligas uzi grandegajn korpusojn de homaj tradukoj kiel ekzemplo-bazon.

Same kiel la metiistoj en tradukburooj daŭre lernas de la antaŭa propra laboro kaj tiu de pli spertaj kolegoj, ankaŭ komputiloj povus fari tion. Necesas ja ke la enkomputiligita 'tradukmemoro' estu fidinda, vastega, kaj daŭre aktualigita. Tehxnologie, la defio ekzistas en aŭtomate traduki per analogio tiujn tradukerojn aŭ parolturnojn kiuj ne laŭlitere troveblas en la ekzemplo-bazo. Kelkcent PAT-esploristoj tra la mondo koncentriĝas nun al tiu celo, kaj daŭros ankoraŭ kelkajn jarojn ĝis la tehxnologio maturiĝos.

Menciinda pri Esperanto estas ke verko de Victor Sadler (1989, kadre de DLT-projekto) agnoskitas en la faktereno kiel unu el la unuaj PAT-esploroj. Tamen, por ke Esperantio povu vere partopreni en tiu ĉi nova vojo al aŭtomata tradukado, nepre necesas amaso da aktualajhomaj tradukoj en nian lingvon.

A new road to Machine Translation

After repeated research efforts on machine translation in the 20th century, the ambitious goal of fully automatic high-quality translation has not been realized. Today, internet users can get computerized translations from various languages instantly, even for free, but far from flawless. In fact, the translation technology now on the market is still rooted in the pioneering developments of the 1960s and 1970s.

However, research efforts have not come to a halt. Machine Translation is still an active field of R&D, and over the last 15 years a new paradigm evolved: Statistical Machine Translation (SMT). In contrast to research based on linguistics, which culminated in the 1980s, SMT is surprisingly based on the craft of human translation itself. Unlike the linguist, the translator treats a text in a much more practical way, usually with communication-between-two-cultures in mind, rather than the abstractions of theoretical linguistics.

For this very reason, translator's skill always appeared much more difficult to 'formalize' than linguist's knowledge. But finally, the research community took a turn: linguistics did not deliver sufficient results, and the meanwhile dramatically increased computer capacity enabled extremely large corpuses of human-translation examples.

Just like humans acquire skill by learning from previous work or experienced colleagues, computers might act similarly. Of course, the computer's `translation memory' must be reliable, inexhaustible and constantly updated. Technologically, the challenge exists in automatic translation-by-analogy of all phrasings not literally found in the example database. Worldwide, several hundred SMT researchers now concentrate on this and it will take some years before the technology ripens.

As to Esperanto, the work of Victor Sadler (1989, as part of the DLT project) is being recognized as one of the earliest SMT researches.

However, in order to really profit from this new prospect for machine translation, the Esperanto community urgently needs to produce a vast amount of up-to-date human translations into its language.

Une nouvelle approche pour la traduction automatique.

Les recherches faites au 20 ème siècle sur la traduction automatique n'ont pas atteint les résultats escomptés. Aujourd'hui les internautes peuvent se procurer instantanément une traduction automatique à partir de diverses langues, parfois même gratuitement, mais jamais sans erreurs. En fait les traductions disponibles sur le marché sont toujours faites selon les premiers développements technologiques en ce domaine qui datent des années 60 et 70.

Cependant la recherche ne s'est pas arrêtée là. Elle se poursuit sur le terrain, et au cours des quinze dernières années s'est développé un nouveau paradigme : La Traduction Automatique Statistique (TAS)**)**. Contrairement à la recherche qui s'appuyait sur la linguistique et qui a connu son heure de gloire dans les années 80, la *TAS* se base sur l'art du traducteur. Contrairement au linguiste, le traducteur aborde les textes de manière plus pragmatique , privilégiant l'aspect « communication entre deux cultures », plutôt que la théorie linguistique.

C'est précisément pour cette raison que l'art du traducteur a longtemps paru plus difficile à formaliser que la théorie linguistique. Mais finalement la recherche a changé de cap. La linguistique ne donnait pas de résultats satisfaisants, et grâce à l'augmentation phénoménale de la capacité de mémoire des ordinateurs, on a pu utiliser d'énormes corpus d'exemples de traduction manuelle comme base de données.

Tout comme les traducteurs dans leur bureau progressent grâce à leur travail antérieur et à celui de collègues plus expérimentés, il en va de même pour les ordinateurs. Il est nécessaire bien sûr que les exemples traduits entrés dans l'ordinateur soient fiables, abondants et constamment actualisés. Sur le plan technologique, le défi consiste à traduire automatiquement par analogie toutes les expressions ou tours de phrase qui ne figurent pas dans la base de données. Des centaines de chercheurs de par le monde se concentrent sur ce problème et il faudra attendre encore quelques années pour que la technologie soit au point.

En ce qui concerne l'espéranto, Victor Sadler (1989 , dans le cadre de son travail pour le projet *DLT*) est considéré comme l'un des pionniers de la *TAS*. Cependant, pour pouvoir profiter vraiment de cette nouvelle avancée dans le domaine de la traduction automatique, la communauté espérantophone se doit de constituer un vaste corpus de traductions humaines dans cette langue.

Nova vojo al Aŭtomata Tradukado

Toon Witkam
Utreĥto, Nederlando
[toon.witkam@planet.nl]

En Hajdarabado staras granda centralo de aŭtomata tradukado. Tie, baterio da komputiloj tag-nokte tralegas amason da ĵurnaloj, gazetoj, libroj, raportoj, el la tuta mondo. Plejmulte per interreto, okaze per skanado de paperaj arĥivoj. Tiuj nelacigeblaj komputiloj ne faras tion pro plezuro. Iliaj homaj mastroj dresis ilin por tiu senĉesa legado. Fakte, la dreso jam daŭris jardekojn, kaj perfektiĝis tiamezure, ke la bravaj maŝinoj ĉiutage nur subtenas kaj aktualigas sian traduktalenton per multa legado kaj interkomparo de diverslingvaj tekstfontoj, pritraktantaj la saman temon. La revo de ambiciega tradukisto!

Antaŭ dudek jaroj, la avoj de ĉi tiuj hajdarabadaj maŝinoj estis ankoraŭ komencantoj pri la tradukarto, kiam ĉiu tago de kompara legado signifis modestan plibonigon sur larĝa skalo de traduksperto. Sed la elektronikaj filoj kaj fine la nepoj, ree pli rapidaj kaj memorpovaj, akiris kvalitnivelon pli altan ol tiun de la meza homa tradukisto.

La hajdarabada centro do servas al subteno de tiu kvalitnivelo - nivelo kiu rapide malsupreniros, se la tradukmaŝinaro ne adaptiĝas al la kontinuaj ŝanĝoj en la mondo: novaj terminoj, novaj idiomoj, novaj akronimoj, kaj pro politikaj kialoj novaj geografiaĵoj, kaj eĉ revivantaj lingvoj. Ĝuste por tio, internacia teamo de profesiuloj - poliglotoj, komunikil-specialistoj, sciencistoj - gardas siajn maŝinajn samklerulojn. Ili kontrolas ĉu la aparatoj ne ignoras iujn fontojn, nek trouzas aliajn; ĉu ili ne estas subtilmaniere infektitaj, ĉu ilia tradukado restas fidinda, ktp.

Krom en Hindio, estas similaj grandaj centraloj de tradukmaŝinoj en Usono (San Diego), Eŭropo (Nancy) kaj Koreo (Pjongjango). Ĉiuj kvar uzas diferencajn maŝinojn kaj softvaron, sekve la kvar funkcias sendepende unu de la alia. Samkiel en la pasinteco spertaj homaj tradukistoj ne liveris ekzakte identan tradukon de la sama fontteksto, ankaŭ la kvar centraloj ne faras tion.
Ĉi tiun feliĉan plurecon lerte ekspluatas interretaj tradukservoj. Ili disponas pri softvaro kiu tre precize kontrolas la samsencecon de la kvar tradukoj. Per tiu, ili povas garantii al sia klientaro des pli fidindan altkvalitan tradukprodukton.

Jen projekcio de la stato de permaŝina tradukado en la jaro 2055. Ĉu scienc-fikcio? Ne vere. La fikcia elemento kuŝas pli en la organiza aŭ entreprena flanko de la skizitaj centraloj ol en la teĥnologio mem. Kiu pretos larĝe investi en la afero? Komercaj entreprenoj rigardas lud-robotetojn kiel pli allogajn por la merkato. Ĉu registaroj, multnaciaj institucioj…?

Teĥnologio sur nia telero

En la hodiaŭa mondo de 2005, fulmrapida aŭtomata tradukado (AT [16]) jam disponeblas: per interreto, kaj ofte eĉ senpage. La fakto, ke tiaj tradukoj ne estas altkvalitaj, apenaŭ gravas por plimulto de ĝiaj uzantoj. Sur la reto oni intertempe kutimiĝis al neakurata, neglektema lingvo-uzo ĝenerale, do des malpli ŝoka aperas iomete fuŝa traduko.

Aliflanke, kie temas pri publikigendaj tekstoj, ĉu komunikaĵoj de publika servo, ĉu manlibroj en industrio, zorgema verkado de la originalo - kiom eble laŭ certa normo aŭ modelo koncerne vortprovizon kaj frazkonstruon - povas krei la ĝustajn kondiĉojn por apliki AT-teĥnologion altkvalite. Ekzemploj de tio estas la kanada traduksistemo Météo por veterprognozoj, la franca kaj usona plurlingvaj sistemoj Titus kaj Caterpillar por la teksoindustrio kaj eksporto de agrikulturaj maŝinoj, specialaj sistemoj por multlingva dokumentado de softvaro, ktp. Ekzistas ankaŭ konsilist-firmaoj, kiuj adaptas ekzistantan AT-sistemon kiel Systran al la bezonoj de internaciaj entreprenoj. En ĉiuj ĉi kazoj, temas pri altkvalita traduko de lingvofrakcio[17].

Escepte de tielaj por specifa apliko kaj klientaro adaptitaj sistemoj, la AT-kvalito de publike alireblaj tradukservoj sur la reto estas ankoraŭ tre modesta. Ne nur sintakse kaj stile, ankaŭ rilate la ĝustan vortelekton. Oni eĉ ne povas garantii, ke la hodiaŭaj maŝintradukoj fidinde transdonas la sencon de la originalo. Post lastatempa sondado [Hutchins 2003], rekonata specialisto pri AT-historio konkludis, ke la tradukkvalito[18] apenaŭ kreskis dum la lastaj du jardekoj.

Notu, ke *la nova vojo*, pri kiu tiu ĉi referaĵo temas, estas ankoraŭ esplor-ŝoseo! Ĝia ekesto ĝis nun ne atingis la AT-teĥnologion uzpretan sur la merkato. Tiu estas fakte fondita sur esplor- kaj evoluig-laboro el la sesa ĝis naŭa jardekoj de la pasinta jarcento.

La manko de imponaj rezultoj, post jardekoj da esploro kaj disvolvo, per si mem estis sufiĉa defio por novgeneraciaj sciencistoj. Novgeneraciaj - ne nur laŭ aĝo, sed prefere laŭ disciplino kaj metodaro. La profesiuloj kiuj dum la epoko 1950-1990 okupis sin pri AT estis komputor-sciencistoj, poliglotoj, gramatikistoj, vortaristoj, semantikistoj, logikistoj, praktikantoj de la formala lingvistiko, sed apenaŭ profesiaj tradukistoj. La plej granda projekto iam, EUROTRA, fine fiaskis pro troo da teoriaj lingvistoj [Hacken 2001].

Lernivaj maŝinoj anstataŭ lingvistoj

La nova generacio de AT-esploristoj ekestis ĉirkaŭ 1990. Ĝi kreskis dum la lastaj 15 jaroj kaj laŭgrade anstataŭis la antaŭan. Sed de ĝia naskiĝo [Brown 1988], ĝi eniris tutan novan vojon. Rezignante iun ajn lingvistikan konon, la novnaskitaj esploristoj[19] turnis sin al la modelo de profesia *homa tradukisto*! Jen preskaŭ revolucio, certe ŝanĝo de paradigmo. Ĉu oni nun esplordemandu spertajn tradukistojn, akurate observu tiujn dumlabore, divenu la procezojn en iliaj cerboj?

Pli simple: pristudu nur la rezultojn de ilia laboro, la tradukitajn tekstojn, kompreneble kun la originaloj apude. Aŭ pli oportune: lasu komputilojn fari tion. Bonega ideo! Komputiloj abundas, iĝas ĉiujare pli rapidaj kaj malpli kostaj, havas egan memoron, kaj ne plendas pri

[16] Tra la teksto, ni uzos la mallongigon AT (Aŭtomata Tradukado), konforme al MT (*Machine Translation*) en anglalingvaj faktekstoj.
[17] anglalingve *'controlled language'* aŭ *'sublanguage'*.
[18] Hutchins testis AT-sistemojn kiel 'Systran Personal', 'Personal Translator' ktp. kaj retajn AT-servojn kiel Bablefish, Lycos, Reverso, Promt, FreeTranslation, InterTran.
[19] anstataŭ lingvistoj ĉefe matematikistoj kaj komputikistoj.

longa labortempo. Plie, ekzistis ja la koncepto de ‘lerniva maŝino’[20], kies historio reiras ĝis la 1950-aj jaroj, do pli-malpli same malproksimen kiel tiu de permaŝina tradukado, kvankam la du terenoj ne estis interfruktigintaj unu-la-alian.

Do ekde ĉ. 1990, la historio de AT, sur ĝia tre nova vojo, certasence temas pri evoluiĝo de lerniva maŝino. Fakte, koncernas kaj evoluigon kaj evoluiĝon, ĉar la afero dependas ankaŭ de bonaj instruistoj! Jen la ĝeneralaj principoj de traduklernado fare de komputiloj:

- GRANDEGA KVANTO de lernmaterialo. La lernanta komputilo devas tralegi amason da ekzemploj de bona (homa) traduko: almenaŭ dekmilojn, prefere milionojn da frazoj el t.n. dulingvaj aŭ plurlingvaj korpusoj. Jen kelkaj:
  - Hansard (kanadaj parlamentaj debatoj):
    angla kaj franca, po 2.000.000 frazoj, po 40.000.000 vortoj
  - EuroParl (EU-parlamentaj debatoj, 1996-2001):
    11 lingvoj, po 740.000 frazoj, po 20.000.000 vortoj
  - CNS (Ĉina Novaĵ-Servo sur interreto):
    angla kaj ĉina, po 25.000 frazoj; 500.000 anglaj vortoj; frazparoj ekz.:

Mr. Luo said that 141,949 cases were handled by the Administrative Appeals Tribunals, between October 1990 and June 1995.	罗豪才说，从一九九O年十月至今年六月，全国各级人民法院共受理各类一审行政案件十四万一千九百四十九件。
Decision appealed came from 40 administrative areas including land, public security, urban construction, industry and commerce, environmental protection, prices, finance, customs, forestry, mining, taxation and technological supervision.	案件类型涉及土地、公安、城建、工商、环保、物价、金融、海关、林业、矿产、税务、技术监督等四十多个行政管理领域。

- SENĈESA NOMBRADO. Klarigenda ĉi tie estas la supre-menciita ago ‘tralegi’. Signifas ke la maŝino sisteme tralaboru ĉiun frazparon aŭ frazmultoblon, registrante kaj nombrante ĝiajn elementojn (vortojn, vortsinsekvojn, …) kaj certajn rilatojn inter ili, ĉio laŭ la precizaj instrukcioj de la instruisto. La nombrado nutras probablokalkulon, per kiu la maŝino poste (en la testfazo) mem provos traduki neviditajn frazojn.

- INSTRUADO FAR HOMOJ. La esploristoj estas ja la instruistoj. Eble ne plu necesaj en 2055, sed nun ankoraŭ la ĉefrolantoj. Ĉiu AT-esploristo, aŭ almenaŭ ĉiu esplorteamo, instruas la maŝinon siamaniere, eksperimentante kun sia propra metodo aŭ metod-variaĵo. Kelkaj esploristoj donas tre simplajn instruojn al la lernant-komputiloj, aliaj instruas ilin pli detale aŭ komplekse. Plejofte, la diversaj metodoj estas publikigitaj, kaj ekzistas internaciaj fakaj kunvenoj ĉiujare. Multe pri tio estas trovebla ankaŭ sur interreto.

Ĵus skizita estas la LERNFAZO. La en tiu ĉi fazo uzata korpuso - aŭ la por lernado uzata parto el ĝi - estas la LERNKORPUSO[21]. Tuj kiam la maŝino, kiel ĉiu lernanto, devas montri sian scipovon, komenciĝas la TESTFAZO, kiu iam finfine transiĝos en PRODUKTAD-FAZON.

En la testfazo, oni prezentas al la maŝino frazojn el la sama korpuso kiel la lernkorpuso. Tio estas justa, ĉar korpuso reprezentas certan teksttipon, kies lernadon la testo devas kontroli. Pro tio oni ĉiam dividas la uzatan korpuson en du partojn: la lernkorpuson, kutime la plej grandan parton, kaj malgrandan subaron de testfrazoj. Kompreneble, la lernanto ne vidu la testfrazojn dum la lernado!

[20] anglalingve: *‘machine learning’*.
[21] anglalingve: *‘training corpus’*.

La ripetfoja testado de lernivaj tradukmaŝinoj, karakterizaĵo de la nova vojo, urĝis la enkondukon de pli-malpli aŭtomata rezult-prijuĝo. La lastajn kvin jarojn, kelkaj vortstatistikaj helpiloj por mezuri la kvaliton de AT-produktoj estas inventitaj[22], sed ilia uzo estas ankoraŭ pridisputata. Cetere, krom la kvalitmezuro, ankaŭ la speco de la eraroj estas pristudenda.

La vere plej rapida fazo estas - kontraŭ-intuicie - la lernfazo, malgraŭ la grandega kvanto de la tralaborenda lernmaterialo. La lernfazo ja estas plenaŭtomata[23], kaj komputiloj iĝas ĉiujare pli rapidaj.
Tute alie, la malplej rapida kaj plej multekosta fazo estas la INSTRU-PREPAR-fazo: la hom-laboro de esploristoj, kiuj daŭre elpensas novajn instruarojn por siaj paciencaj maŝinaj adeptoj. Tiu ĉi fazo estas ankaŭ la plej kreiva parto de la ciklo.

Jen kelkaj ĉefakcentoj de la afero. Estas plue akcesoraj fazoj, ekz. la teda kaj temporaba antaŭa fazo por purigi kaj prepari[24] tekstkorpuson, por instali kaj agordi la diversajn softvarilojn, por frazliniigi la korpuson, ktp.

Sengramatikaj piratoj pioniras

La pioniroj de la nova vojo estis matematikistoj en esplorcentro de la usona firmao IBM[25]. Ilin inspiris antaŭaj progresoj en la teĥnologio de aŭtomata parolrekono - ĝis tiam tuta alia esplortereno ol teksttradukado. Rimarkante ke *"La problemo de lingvo-modeligo por AT esence egalas tiun por parolrekono"* [Brown 1993], ili ellaboris solidan matematikan bazon por pure Perstatistika Aŭtomata Tradukado (PAT [26]). Ilia fakrevua artikolo de 1993, kun 20 paĝoj plenaj de matematikaj formuloj, iĝis la plej referencita fonto en la AT-tereno. La en ĝi proklamita Fundamenta Ekvacio de Aŭtomata Tradukado:

$$\hat{s} = \operatorname{argmax}_s \Pr(s) \Pr(t \mid s)$$

resumas la tri defiojn de PAT: kompute taksi la probablecon de fontlingv-modelo Pr (s), kompute taksi la probablecon de tradukmodelo Pr (t | s) , kaj elpensi efikan, rendimentan serĉmanieron por trovi tiun fontlingvan vortĉenon, kiu maksimumigas la produkton de tiuj du probablecoj[27].

La lingvo-modelo temas nur pri la laŭlinia sinsekvo de vortoj en frazo, sen iu ajn uzo de sintaksa kono. En la tradukmodelo konata kiel IBM Modelo 1 vortsinsekvo eĉ tute mankas: anstataŭ vortoj demaldekstre-dekstren sur linio, oni kvazaŭ ĵetis ilin po frazlongeco en sakon. Vortpozicio tie ne ekzistas, kaj la fakto ke per lingvo-modelo de vorttrioj[28] (transprenita de parolrekona sistemo) 84 procentoj da testfrazoj[29] montriĝis aŭtomate rekonstrueblaj el iliaj vortsakoj, indikas la forton de simpla vortstatistiko.

Por ekhavi iom da intuicia kompreno de PAT, imagu vin en la rolo de la (lerniva) maŝino. Oni konfrontas vin kun centmiloj da frazparoj, en du al vi tute nekonataj lingvoj. Nek gramatikoj nek vortaroj estas disponeblaj. Krome, por fari la ekzercon pli amuza, oni antaŭpakis ĉiun

[22] i.a. BLEU *(BiLingual Evaluation Understudy)*, NIST *(National Institute of Standards and Technology)*, RED *(Ranker based on Edit Distances)*, ORANGE *(Oracle Ranking for Gisting Evaluation)*.
[23] escepte de kelkaj, i.a. la PAT-sistemo "Linear B", konstruata en Edinburgo [Callison-Burch 2004].
[24] anglalingve: *'preprocessing'*.
[25] IBM Research Laboratories, Yorktown Heights, NY, Usono.
[26] Tra la teksto, ni uzos la mallongigon PAT (Perstatistika Aŭtomata Tradukado), same kiel en anglalingvaj faktekstoj SMT egalas *'Statistical Machine Translation'*.
[27] la literoj s kaj t signifas vortĉenon en respektive la fontlingvo (*'source language'*) kaj la cellingvo (*'target language'*); ŝ signifas tiun fontlingvan vortĉenon, kiu maksimumigas la probablecon Pr(s|t), t.e. la probableco ke (inversa) traduko de ĉeno t produktos la ĉenon s.
[28] anglalingve *'word trigrams'* aŭ simple *'trigrams'*.
[29] temis pri frazoj kies longeco estas maksimume 10 vortoj [Brown 1990].

frazon en sakon, tiel ke ĝiaj vortoj[30] estas tute senordigitaj. Do vi nun havas antaŭ vi centmilojn da sakoparoj. Vi havos jarojn da tempo (kompare, por moderna komputmaŝino, milisekundo ja daŭras jaron), do vi kuraĝe ekas.
Vi malfermas la du sakojn de unua paro kaj blinde prenas vorton el ambaŭ. El la sako de Lingvo-1 aperas la vorto *'krhŝt'*, el tiu de Lingvo-2 la vorto *'uaaio'*. Vi rezonadas: "Interese! La frazo en unu sako estas la traduko de tiu en alia sako, do ekzistas iu ŝanco, ke la vorto *'uaaio'* estas traduko de la vorto *'krhŝt'* ".
Vi nun traserĉas ĉiujn sakoparojn pri enesto de tiuj du vortoj. Evidente vi nombras la fojojn ke ambaŭ vortoj aperas en la sama sakoparo, sed vi nombras ankaŭ la unuflankajn aperojn. Gravas la relativa frekvenco. Se *'krhŝt'* aperas en preskaŭ ĉiu frazo, ĝi verŝajne estas ofta funkcivorto (kiel ekz. 'kaj' en Esperanto). Se ankaŭ *'uaaio'* ĉeestas ĉie, eĉ tiam la ofta kunĉeesto de la du vortoj ne implicas, ke ili estas la traduko unu de la alia (*'uaaio'* povus esti alia ofta funkcivorto). La ideala kazo estas la konstato, ke *'krhŝt'* aperas nur en ekz. promilo da sakoparoj, ĉiam nur samtempe kun *'uaaio'*, kaj reciproke. Tio rezultigus: Pr (*'uaaio'* | *'krhŝt'*) = 1. Pli verŝajne, *'krhŝt'* kaj *'uaaio'* aperos ankaŭ kelkfoje unu sen la alia, aŭ eĉ plurfoje, se unu de la du vortoj estas de la plursenca tipo. Tiam ilia tradukprobableco kalkuliĝus je ekz. 0,95 aŭ nur 0,65.

Tiamaniere, en la unua eksperimento de la IBM-esplorgrupo [Brown 1990], la maŝino komputis por ĉiu kombino el 9000 anglaj kaj 9000 francaj vortoj[31] la probablecon, ke ĝi estas tradukparo, kio do rezultigis tabelon de 81.000.000 parametroj. Ties valoro estas: provizore indiki vortliniigon[32], gravan koncepton en PAT, trafe ilustritan per linioj inter du frazoj, interligantaj tiujn vortojn, kiuj estas traduko unu de la alia.

Tamen, la supre skizita procezo estis nur la eka ŝtupo en vico de pluaj: la IBM Modeloj 2-5 (inventitaj ĉiuj ĉirkaŭ 1990), trapasitaj iteracie, plipreciziga la antaŭe kalkulitajn probablec-parametrojn, surbaze de jenaj informoj: vortpozicio, fekundeco[33], kaj distordo[34].
La rolo de vortpozicio estas evidenta: tradukoj de vortoj el la komenca parto de la fontlingv--frazo aperos verŝajne ankaŭ en la komenca parto de la cellingv-frazo, ktp. Bela ekzemplo de vortliniigo tute konforma al vortpozicio estas[35]:

Among the many questions raised by the expanded membership of the European Union is the question of languages.

Inter la multaj demandoj levitaj de la plivastigita membreco de la Eŭropa Unio estas la demando pri lingvoj.

En tiu ĉi frazparo, la reguleco de laŭpozicia liniigo de vorttradukoj estas escepte bonŝanca! Plejofte, fraztraduko havas unu-du liniig-distordojn, ekzemple pro la inversigo de la adjektiv-substantiv-sinsekvo aŭ pro diferencoj en SVO (Subjekto-Verbo-Objekto)-vortordo (*'je le voit'* – 'mi vidas lin'), kiam oni tradukas de la angla (aŭ de Esperanto) al la franca. Sed inter ekz. la japana kaj la angla, la diferencoj en vortordo estas pli persistaj. Pro tio, se la vortpozici-informoj ne vere kontribuas, hodiaŭa esploristo uzas nur la IBM-modelon 1, ne la IBM-modelojn 2-5 [Ding 2003].

Ĉiea fenomeno estas la t.n. 'fekundeco'. Kiam fontlingv-vorto produktas du- anstataŭ unu-vortan tradukon en iu cellingv-frazo, ĝia fekundeco egalas 2 anstataŭ 1. Prominenta ekzemplo

[30] la nombro da vortoj en sako (frazo) povas varii inter proksimume 10 kaj 30.
[31] oni limigis la komputadon je la 9000 plej oftaj vortoj aperintaj en la korpuso.
[32] anglalingve *'word alignment'*.
[33] anglalingve *'feritility'*.
[34] anglalingve *'distortion'*.
[35] frazparo el la (esperanta kaj angla) informfolioj pri la Kvara Nitobe Simpozio (Vilno, 30 julio – 1 aŭgusto 2005).

estas la angla funkcivorteto *'not'* tradukita en la francan *'ne ... pas'* . Sed ankaŭ ĉe enhavvortoj, fekundecoj > 1 abundas. Rigardu jenajn frazojn[36]:

Tensions | between | the | two | powers | have increased | in | recent | months.

La streĉiteco |inter | la | du | grandaj regionaj potencoj | kreskis | dum| la lastaj | monatoj.

Se oni konsideras la esperantan frazon kiel fontlingvon, la vorto 'kreskis' havas fekundecon 2, ĉar ĝi produktas *'have increased'*. Alikaze, la anglaj vortoj *'tensions'* kaj *'recent'* havas tiun fekundecon, kaj la vorto *'powers'* havas eĉ fekundecon 3: 'grandaj regionaj potencoj'. Tre verŝajne, en la sama tekstkorpuso troviĝas ankaŭ frazparoj, en kiuj *'powers'* liniiĝas alimaniere, ekz. kun 'grandaj potencoj', simple kun 'potencoj' aŭ kun 'povo'. La esenco de PAT estas, ke ĝi kaptas en ĝiaj probablec-parametroj ĉiujn variaĵojn trovitajn en dulingva tekstkorpuso, do fakte la produktojn de la tradukistaj sperto kaj libereco - ne la regulojn de gramatiko aŭ la informojn de vortaro. Ĝuste pri tio PAT, la nova vojo, diferencas de la tradicia AT.

La sengramatikaj pioniroj de IBM, post perkorpusa[37] lernfazo de 40.000 frazparoj angla-francaj, kun entute proks. 1.600.000 tekstvortoj, atingis jenan rezulton [Brown 1990]: ilia lerniva maŝino kapablis bone traduki 48 procenton de 73 francaj testfrazoj. Modesta sukceso, kio estis tamen kuraĝiga kaj inspiriga. Certe impona estis ilia dua eksperimento [Brown 1993], en kiu la lerniva maŝino disponis pri 1.778.620 frazparoj, kalkulis la tradukprobablecojn de 2.437.020.096 vortkombinoj, kaj per pure statistika algoritmo komputis la ĝustan vortliniigon el ekzemple la 1.9×10^{25} teorie eblaj vortliniigoj en jena frazparo:

What is the anticipated cost of administering and collecting fees under the new proposal?

En vertu des nouvelles propositions, quel est le côut prévu d' administration et de perception des droits?

Fine, la merito de la PAT-pioniroj de IBM komence de la 1990-aj jaroj ankaŭ estas, ke ili klare konsentis pri posta necesa aldono de morfologia kaj sintaksa komponantoj al PAT. La grando valoro de ilia laboro restas: ili taŭge enkondukis statistikajn metodojn en la AT-terenon, kaj konvinke montris ties forton.

Sintakso silente revenas

Sinkrone kun la evoluigo ĉe IBM komence de la 1990-aj jaroj en Usono, sed sendepende de ĝi, disvolvis en Japanio nova paradigmo. Ĝi estis proksima al PAT, sed konservis la sintakson: Ekzemplo-Bazita Aŭtomata Tradukado (EBAT [38]). Komuna trajto de PAT kaj EBAT estas la orientiĝo pri tekstoj el la tradukpraktiko, per uzo de bilingva korpuso aŭ datumbazo. Kiel la IBM-esploristoj, ankaŭ la japanaj samfakuloj estis parte inspiritaj de laboro pri perkomputila parolrekono.

La unuajn prototipojn de EBAT faris [Sato 1991]. Li komence eksperimentis per dulingva datumbazo de ekzemplaj frazeroj. Jen momentfoto de kalkultabelo (kun la frazer-vortoj re-ordigitaj laŭ VSO[39]) uzita interne de lia prototipa sistemo:

[36] fonto: "Ĉinio kontraŭ Ĉinujo", artikolo de Ignacio Ramonet en *Le Monde Diplomatique* de aprilo 2005, tradukita en la anglan de Ed Emery kaj esperantigita de Vilhelmo Lutermano; la angla kaj la esperanta tekstoj, kvankam ambaŭ estas tradukoj, povas mem esti uzataj kiel (parto de) angla-esperanta korpuso en PAT-esploro.
[37] la dulingva korpuso *Hansard*, arĥivo de la kanadaj parlamentaj debatoj.
[38] en anglalingvaj faktekstoj: EBMT *('Example-Based Machine Translation')*.
[39] VSO = Verbo-Subjekto-Objekto.

Source = (PLAY JAPANESE CARD)		*Weight-List* = (.211 .789)	
Rank	*Target*	*Distance*	*Most Similar Translation*
1	(する 日本人 トランプ)	1.25(4.34 .429)	(PLAY TARO TENNIS) -> (する 太郎 テニス)
2	(ひく 日本人 トランプ)	6.05(18.4 2.74)	(PLAY YOU VIOLIN) -> (ひく あなた バイオリン)
3	(ひく 日本人 カード)	6.72(18.4 3.58)	(PLAY YOU VIOLIN) -> (ひく あなた バイオリン)
4	(する 日本語 トランプ)	211.(999. 0.0)	(PLAY THEY CARD) -> (する 彼ら トランプ)
5	(する 日本語 カード)	212.(999. 1.45)	(PLAY THEY CARD) -> (する 彼ら トランプ)
5	(する 日本人 カード)	212.(999. 1.45)	(PLAY THEY CARD) -> (する 彼ら トランプ)
7	(ひく 日本語 トランプ)	213.(999. 2.74)	(PLAY I VIOLIN) -> (ひく 私 バイオリン)
8	(ひく 日本語 カード)	214.(999. 3 58)	(PLAY I VIOLIN) -> (ひく 私 バイオリン)
9	(演じる 日本人 トランプ)	792.(18.4 999.)	(PLAY YOU HAMLET) -> (演じる あなた ハムレット)
9	(演じる 日本人 カード)	792.(18.4 999.)	(PLAY YOU HAMLET) -> (演じる あなた ハムレット)
11	(演じる 日本語 トランプ)	999.(999. 999.)	(PLAY HE ROMEO) -> (演じる 彼 ロメオ)
11	(演じる 日本語 カード)	999.(999. 999.)	(PLAY HE ROMEO) -> (演じる 彼 ロメオ)

Traduki novan frazeron (*'japanese play card'*) el la angla en la japanan signifas kalkuli ĝian semantikan 'distancon' al ĉiu ekzempla frazero kun la sama verbo (*'play'*). La kalkulado funkcias per japanlingva tezaŭro[40], al kiu ankaŭ anglaj vortoj estis aldonitaj. Tiel la maŝino trovas la ekzemplon plej proksiman al la tradukenda frazero, kaj povas traduki tiun laŭe.

Makoto Nagao, la majstro de la japanaj AT-esploristoj, kiu lanĉis la ideon pri EBAT jam komence de la 1980-aj jaroj, bone eksplikas [Nagao 1992], ke ĝi superas la konvencian metodon, kiu dependas de teda laboro de lingvistoj. Kvazaŭ ili estas vortaristoj, ili devis permane meti semantikajn indikilojn ĉe ĉiu substantivo, ekzakte indiki verbvalentojn ktp. Tio estas malfacila, multekosta kaj longedaŭra afero. Aliflanke, provizo da ekzemploj sufiĉe ampleksa por traduki tutajn frazojn surbaze de ĝi estas simple neebla. Nagao kaj [Sato, 1990] gvidis esploristojn sur la novan vojon per priskribo de hibrida EBAT-kadro, kiu ebligas integri ekzemplajn frazerojn en la tutaĵon de la fraza sintaksa strukturo. Notinda en tiu proponita kadro estas la uzo de dependec-arboj anstataŭ la ĝis tiam kutimaj konstituant-arboj. Ankaŭ en sia dua, tutfraza prototipo Sato uzis dependec-arbojn.

Jardekon poste, [Yamamoto 2000] konfirmas la uzon de dependo-sintaksaj strukturoj por liniigo de frazeroj en PAT kaj implice en EBAT. Tio nun helpas solvi pli ĝeneralan problemon, kiun la usonaj IBM Modeloj 1-5 ne tuŝis: la liniigo de fontlingva vortsinsekvo al nur unu cellingva vorto. La klasika ekzemplo de tio estas la angla *'red herring'* kaj ĝia germana egalulo *'Finte'*, sed abundas tiaspecaj nekunmetivaj[41] tradukoj. Rilatigi (per nur unu konektlinio) tutan vortsinsekvon el la fontlingvo al tuta vortsinsekvo en la cellingvo, ankaŭ tion la IBM Modeloj ne kapablas, egale ĉu la vortnombroj en la du ĉenoj egalas aŭ diferencas. Oni nur pensu pri idiomoj kaj parolturnoj, ĝuste tiuj frazeroj, kies tradukon EBAT celas.

En la esplormondo, la hibrida tradukmaŝino, apogante kaj sur PAT kaj sur sintakso, nun laŭgrade gajnas terenon (PAT inklude EBAT, sintakso inklude morfologion). Estas tamen ankoraŭ fervoruloj, kiuj iom rezistas la revenon de sintakso. [Koehn 2003] interkomparis PAT-rezultojn ĉe du variantoj de liniigo: en unu metodo ĉiuj ajn trivortaj sinsekvoj[42] estis liniigitaj, en la alia nur sintaksaj vortgrupoj[43]. La aŭtoroj asertis, ke prefero por sintaksaj strukturoj malplibonigis la tradukon, kaj ili defiis la sintaksfavorajn samfakulojn.

[Lin 2004], kiu jam en la 1990-aj jaroj estis esplorinta multpovajn parsilojn helpe de dependo-sintakso[44], respondis la defion. Dum Koehn et al. bazis sian sintaksan varianton sur konstituant-

[40] *"Word List by Semantic Principles"*, NLRI (National Language Research Institute), Syuei Syuppan, Japanio, 1964.
[41] anglalingve: *'non-compositional'*.
[42] anglalingve: *'clump', 'word trigram'* aŭ simple *'trigram'*.
[43] anglalingve: *'syntactic phrase'* aŭ simple *'phrase'*.
[44] vidu [Lin 1995].

arboj, la lerniva tradukmaŝino de Lin ekstraktas padojn[45] el fontlingvaj dependec-arboj de vortliniigita korpuso, kaj tradukas tiujn en fragmentojn de cellingva dependec-arbo. Samtempe, ne nur la dependec-rilatoj, ankaŭ la laŭlinia sinsekvo de vortoj estas enkodita. Tiel la korpus-bazita lernprocezo rezultigas aron de transiraj[46] reguloj kun certaj probablecoj. Poste, la traduko de nova frazo disvolviĝas jen: parsu la frazon por akiri ĝian dependec-arbon; ekstraktu el tiu ĉiujn padojn, kaj retrovu iliajn tradukojn; serĉu kombinaĵon de transir-reguloj, kiu pritraktas la fontlingv-arbon komplete kaj produktas cellingvan dependec-arbon senkonflikte; se pluraj tiaj kombinaĵoj estas trovataj, elektu tiun kun la plej alta probableco.
La sistemo de Lin trairis lernfazon de 116.889 frazparoj (angla-francaj, kun 3.4 miliono da vortoj entute), el kiuj 2.040.565 sintaksaj padoj estis ekstraktitaj. La testfazo enhavis 1775 frazojn de 5-15 vortojn longeco. Kvankam la tradukkvalito estis ankoraŭ modesta (BLEU-poentaro: 0.26), promesplena estas la lerta transira modelo, kies sintakso kapablas pritrakti deviojn kiel la angla- germanan paron *'there is' - 'es gibt'* kaj la angla-hispanan *'swim across' - 'cruzar nadando'* .

Francaj arboj revivas, usonaj sekiĝas

La nova vojo de AT havas ankaŭ jenan karakterizaĵon: kreskanta prefero por dependo-sintakso. Tio estas rimarkinda, ĉar dum jardekoj ĝia ega frato, la konstituant-sintakso, regis la AT-mondon preskaŭ sola. Ĉi tie, anstataŭ la teĥnikaĵojn mi volas substreki la preskaŭ kulturan diferencon inter la du.

Dependo-sintakso devenas de la franco Tesnière, meze de la 20-a jarcento, kaj akiris certan adeptaron inter eŭropaj lingvistoj. Sed kiam en Usono la AT-esploro disvolviĝis, la Chomsky-a transform-generiva gramatiko tre influis la. tieajn lingvistojn. Tiu modelo kun ĝiaj abstraktaĵoj kaj konstituant-sintakso fariĝis vera modo, kiu penetris ankaŭ la rondojn de AT-esploristoj en Eŭropo kaj Japanio. Juĝe al ties publikaĵoj, dependo-sintakso tute ne ekzistis. Jen la situacio ĝis ĉirkaŭ la fino de la 1980-aj jaroj.

Oni ne forgesu, ke en la mondo de AT kaj perkomputila lingvistiko ĝenerale, la angla lingvo havas elstaran pozicion. La plejmulto da esploroj, sistemoj, korpusoj, parsiloj, softvariloj ktp koncernas la anglan lingvon. Pri ĝi akumuladis la plej ampleksa scio kaj sperto. Do, estas kompreneble kaj certagrade pardoneble, ke tia prominenta lingvo pli-malpli speguliĝas en la elekto de metodoj kaj iloj. Cirkonstanco kiu kontribuis al tio, estas la fakto, ke multaj anglalingvaj AT-esploristoj, eĉ la (modernaj) lingvistoj inter ili, havas nur tre limigitan konon pri "fremdaj" lingvoj. La simpla uzo en anglalingvaj esplorpublikaĵoj de tiu epiteto, por indiki aliajn lingvojn, malkaŝas tion.

Dum la konstituant-gramatiko estas sufiĉe bona por la angla lingvo, kies sintaksa strukturo estas bazita ĉefe sur vortordo (konstituanto fakte estas vortsinsekvo), ĝi malpli utilas por lingvoj kun pli morfologi-bazita sintakso. Por diverseco de lingvoj, pli taŭga estas dependo-sintakso, ĉar tiu proksimiĝas al kontrastiva sintakso [Schubert 1986].

Tendenco al dependec-arboj estas nenegebla. Laŭ [Lopez 2002], la sukceso de lastatempaj parsmetodoj [Charniak 2000; Collins 1999; Ratnaparkhi 1999] estas dank'al ideoj esence propraj al dependo-sintakso. [Hwa 2002] konfirmas tion kaj lerte ekspluatis la antaŭecon de multpova parsilo por la angla lingvo, kiu konvertas frazojn en dependec-arbojn. Tia parsilo ankoraŭ ne ekzistis por la ĉina. Per vortliniigo inter anglaj kaj ĉinaj frazoj de korpuso[47], Hwa (aŭ pli precize: ŝia lerniva maŝino) prenis vortdependecojn el la angla flanko kaj projekciis ilin

[45] *'paths'* en la anglalingva terminologio.
[46] anglalingve *'transfer rules'*.
[47] 56.000 frazparoj de Hong Kong News.

sur la ĉinan flankon, tiel kreante dependo-arbojn tie. Per tiu eksperimento ŝi montris, ke vortdependecoj estas pli konvenaj por interlingva projekcio ol la vortsinsekvaj konstituantoj.

Konstituant-arboj ankoraŭ ne malaperis en AT, sed sur la nuna PAT-vojo ili iom post iom perdas sian forton. [Knight 2004] koncedas, ke liniig-distordo kiel en frazparo *'I had bought the car'* - *'Ich hatte das Auto gekauft'* ne estas pritraktebla sen dependo-sintakso, kaj [Koehn 2002] raportas pri neceso limigi je 6 vortoj la frazlongecon en eksperimento, kiu celis konstituant-sintaksan pliriĉigon de PAT.

Fine, ankaŭ kiel ponto al semantiko, dependo-sintakso funkcias pli bone ol konstituant-sintakso. [Hwa 2002] asertas: *"semantikaj dependecoj konstituas superaron bazitan sur sintaksaj dependecoj"*, kaj referencante al [Baker 1997]: "sur la tereno de leksika semantiko, esploroj pri la rilatoj inter sintaksaj elementoj unuflanke kaj supranivelaj konceptoj kiel *aganto, profitanto, temo,* aliflanke, koncentriĝis ĉefe sur sintaksaj dependecoj, ne sur konstituantoj".

DLT-rezultoj montriĝas daŭraj

Rigardante malantaŭen, kiagrade rilatas al PAT la tiama DLT-projekto[48]? Tiu projekto, kiu inkludis ambician esploron pri AT en kaj el Esperanto, fakte okazis antaŭ la paradigmo-ŝanĝo de ĉirkaŭ 1990, same kiel sia dekoble pli granda konkuranto EUROTRA[49].

Des pli notinde estas, ke DLT-ĉefgramatikisto Klaus Schubert jam meze de la 1980-jaroj saĝe kaj kuraĝe antaŭiris la supre-menciitan tendencon al dependo-sintakso. En periodo, kiam tiu metodo estis ankoraŭ ĝenerale ignorita en la AT-rondoj, li perceptis ĝin kiel plej taŭgan por diverslingva traduksistemo, kaj amplekse publikigis pri ĝi [Schubert 1986, 1987].
Kiel indikita supre, la celo de dependo-sintakso en AT estas plifaciligi la projekcion aŭ transiron de elementoj el fontlingva strukturo al tiu de la cellingvo, do kontrastivan sintakson aŭ *'metatakson'*, kiel Schubert nomis - honore al Tesnière - la procezon.
Krome, Schubert ne nur priskribis kaj motivis la principojn de metatakso, sed de 1986 ĝis 1989 ankaŭ kunorganizis verkadon de konkretaj dependo-sintaksoj por 10 lingvoj[50]. Ankaŭ ties rezultoj estis publikigitaj [Maxwell 1989].

La de Schubert elektita dependo-sintakso montriĝis solida bazo, sur kiu en la jaroj 1987-1989 lia kolego kaj ĉefsemantikisto ĉe DLT, Victor Sadler, konstruis avangardan metodon por ebligi specon de EBAT (Ekzemplo-Bazita Aŭtomata Tradukado). La aŭtoro mem titolis tiun '*per-analogia semantiko*'[51] kaj publikigis sian verkon en libro [Sadler 1989], kiu ofte estis referencita en japanaj faktekstoj komence de la 1990-jaroj. Dum la supre-menciitaj EBAT-prototipoj en Japanio uzis tezaŭron por kalkuli semantikajn 'distancojn' inter vortoj aŭ frazeroj, la metodo **de Sadler nur bezonas la tekstkorpuson mem, kiu en sia tutaĵo funkcias kiel ekzemplo-bazo kaj tezaŭro samtempe. Tiu arĥitekturo metis DLT-on sur la sojlo de la nova (PAT-)vojo. Ĉi-rilate, vidu ankaŭ la superrigardon en [Hutchins 1992].**

Semantika vortdistanco, aŭ *semantika proksimeco*[52], kiel Sadler nomis ĝin, estas la kerno de lia invento. Ne konfuzu ĝin kun la laŭlinia distanco inter du vortoj[53], kiu egalas la nombron de

[48] Distribuita Lingvo-Tradukado, esplorprojekto ĉe la tiama nederlanda softvarfirmao BSO, dum la jaroj 1982-1990 (vidu: http://ourworld.compuserve.com/homepages/profcon/e_dlt.htm).
[49] plej grandskala AT-esplorprojekto iam, en kiu ĉ. tricent universitatanoj tra Eŭropo partoprenis, kaj kiun financis Eŭropa Komunumo (1978-1993).
[50] la lingvoj kaj verkistoj estis: angla (Bieke van der Korst, Dan Maxwell), bengala (Probal Dasgupta), dana (Ingrid Schubert), Esperanto (Klaus Schubert), finna (Kalevi Tarwainen), franca (Luc Isaac, Dorine Tamis), germana (Henning Lobin), hungara (Gábor Prószeky, Ilona Koutny, Balázs Wacha), japana (Shigeru Sato) kaj pola (Marek Świdziński).
[51] anglalingve *'analogical semantics'*.
[52] anglalingve *'semantic proximity'*.
[53] anglalingve *'word co-occurrence'*.

interaj vortoj plus 1, kaj kiun uzas la teĥnologio de serĉmaŝinoj sur interreto, foje eĉ certaj traduksistemoj. Tamen, por akiri altkvalitan tradukadon, la celon de PAT, pli subtila ilaro necesas. Por bone kompreni la naturon de *semantika proksimeco*, imagu ke vi urĝe bezonas kompletan bildon de la sencdiferenco inter du vortoj, tiel kiel tiuj du vortoj estas uzataj en la praktiko, kiun spegulas granda tekstkorpuso. Nek vortaro nek aparta tezaŭro disponeblas, do vi petas pri konkordanco, listo de ĉiuj kuntekstoj, en kiuj vorto numero 1 aperas[54]. Se vi havas komputileskan memoron kaj rapidecon, vi tuj enkapigas tiun kuntekstaron. Poste, vi senprokraste enŝaltas konkordancon de vorto numero 2, kaj en la sekvaj mikrosekundoj vi sumigas la diferencojn inter la du kuntekstaroj kaj deduktas el tio la *semantikan proksimecon* de la du vortoj: ciferan valoron je du decimaloj[55]. Kelkaj ekzemploj:

registaro	estraro	0,89
registaro	federacio	0,78
registaro	konvencio	0,64
registaro	komunikado	0,35
registaro	principo	0,27
registaro	garbo	0,11

Ĉi tie, la noveco kuŝas en speciala difino de 'kunteksto': dependo-sintaksaj rilatoj[56] kun najbaraj vortoj, anstataŭ pure laŭliniaj proksimecoj, eĉ se la lastaj hazarde koincidas kun la unuaj. La formulo per kiu Sadler en 1989 instruas la lernivan maŝinon kalkuli la *semantikan proksimecon* do baziĝas sur dependo-rilatoj, samaj kiel tiuj[57] enkondukitaj de supre-menciita esploristo [Lin, 2004] pli ol jardekon poste. Pliriĉigo de korpuso per parsado, kiun la metodo de Sadler postulas, estas efektivigebla, ĉar pofrazera parsado - spuri unuopajn dependo-rilatojn - sufiĉas.

Ni konsciu pri tio, ke estas la korpuso sola, kiu kaŭzas la dudecimalajn semantik-ciferojn per la Sadler-a dependo-sintaksa mezurilo. Se hazarde la korpuso estus romano, en kiu homoj senĉese dirus, ke ili ne fidas sian edzon, nek sian registaron, ke kaj la edzo kaj la registaro forĵetadas monon, ke pro tio ili ĝojos pri ŝanĝo de edzo kaj pri ŝanĝo de registaro, tiam la semantika proksimeco inter la vortoj 'registaro' kaj 'edzo' povus eble atingi valoron 0,90.
La korpus-bazita aspekto tamen havas grandan avantaĝon kompare kun la uzo de tezaŭro, taksonomio aŭ ontologio[58]. Tiaj enciklopediaj strukturoj ne nur necesigas kontinuan aktualigon (tiun fakte ankaŭ korpuso bezonas), sed ilia prizorgado implicas: precizigi en hierarĥio la lokon de ĉiu nova aldono. Ĝuste tio estas - certe ĉe pli abstraktaj konceptoj - ofte riska kaj foje nedecidebla afero. Saĝe prizorgi kaj egalpeze plivastigi korpuson kiel ununuran sciobazon ne estas senzorga, sed almenaŭ farebla.

La *semantika proksimeco* ellaborita de Sadler, kaj ligita kun la dependo-sintakso provizita de Schubert, estas la postlasita trezoro de DLT. Ĝia valoro estas daŭra kaj aktuala, ĉar en 1989 ĝi sufiĉe anticipis la evoluojn en la faktereno, kreinte *"semantikajn dependecojn bazitajn sur sintaksaj dependecoj"* (kp. [Hwa 2002]). Oportune, la trezoro estas ankaŭ alirebla. Ĝi estas amplekse kaj detale dokumentita per publikaĵoj[59].

Konklude, vetlude?

[54] anglalingve konata kiel *'KWIC (KeyWord in Context)'*.
[55] maksimume 1,00 (kaze de ekzakta egaleco inter la kuntekstoj).
[56] subordigaj rilatoj, ekz. 'Verbo – Objekto', 'Substantivo – Adjektivo', 'Prepozicio – Substantiva grupo'.
[57] esploristo Lin nomis tiujn dependo-rilatojn 'padoj'.
[58] ekz. WordNet, EuroWordNet, kaj *"Word List by Semantic Principles"* (NLRI).
[59] [Schubert 1986, 1987] estas haveblaj ĉe la Libroservo de UEA; ĉe www.amazon.com aĉeteblaj estas [Schubert 1987], [Maxwell 1989] kaj [Sadler 1989].

Rigardante antaŭen, je kiaj monsumoj ni kuraĝas veti, ke altkvalita tradukado per maŝinoj pretos en 2020, aŭ en 2030…?

Certe la memorkapacito de la estontaj komputiloj ne estos problema, nek la rapideco de tiuj universalaj aparatoj. Jam nun ili sufiĉas por preskaŭ ĉiuj maŝintraduktaskoj. Ankaŭ la provizo de tekstkorpusoj (lernmaterialo por la lerniva tradukmaŝino) daŭre kreskas kaj aktualiĝas. Fakte la interreto mem pli kaj pli funkcios kiel grandega multlingva korpuso, kaj kreskanta nombro de esploristoj utiligas ĝin tiamaniere.

La nova bazo de PAT, en lerta hibrida aranĝo kun i.a. sintaksaj elementoj, aspektas sana kaj promesplena. Kompare kun la regul-bazita raciismo de la tradicia paradigmo, kiu tro emis al perfektigo de abstraktaj lingvo-modeloj, la nuna perstatistika kaj empirisma alpaŝo aperas pli taŭga por laŭgrada kaj daŭra plibonigo de tradukmaŝinoj.

Intuicie ni atendu, ke statistika kerno faru sistemon pli fleksiĝema, kvazaŭ savoreto por ĉiuj tiuj neantaŭvideblaj kontraŭ-regulaj kazoj, inkluzive preserarojn, senmajusklajn proprajn nomojn, alilingvajn citaĵojn ktp. La lastatempa progreso sur la nova vojo jam montras, ke sintaksa analizo de eroj, sekve kunigendaj per la korpus-bazita statistiko, estas pli sukcesa strategio ol la eternaj klopodoj konstrui perfektan parsilon, kiu senriproĉe trovos la solan ĝustan analizon de ĉiu ajn tuta frazo[60]. La perstatistika operaciiado certasence entenas redundon: pluraj tradukoj povas rezulti, eĉ kun neglekteblaj probablec-diferencoj. Tio povas plifortikigi la tradukprocezon.

Aliflanke ni ne forgesu, ke la lernivaj maŝinoj en testfazo ĝis nun tradukas nur ĉ. 50 procenton de la prezentitaj frazoj bone. Unu rimedo por progresi estas pligrandigi la korpuson. Ju pli granda la datumbazo, des pli fidinda la statistiko. Alia rimedo estas plivastigo kaj plibonigo de la diversaj procezoj (korpus-preparado, liniigo, parsilo, transiro, tekststruktura analizo).

Sed la plej kriza faktoro, de kiu dependos trarompo de altkvalita tradukmaŝino estos organiza, ne teĥnologia! La esploristoj, disaj tra universitatoj, laŭnature emaj krei ĉiam novajn variantojn, rare devontigas sin al komuna plukonstruado de unu sistemo. La komerco trovas altkvalitan ĝeneralan traduksistemon ne sufiĉe alloga, kaj internacia registaro kiel tiu en Bruselo ne kuraĝas (denove) riski grandan elspezon por ĝi. Necesas premo kaj elstara organizado, por ke kompetentaj fervoruloj per kunigitaj fortoj efektivigu malfacilan multjaran kunlaboron. Kiel esprimis longtempa AT-esploristo [Carbonell 1992]: *"en Aŭtomata Tradukado, persisteco gravas"*.

Renaskiĝo de (P)AT-projekto en Esperantio, kvazaŭ posteulo de DLT, ĉu ne tio estas vetinda? Ĉu eble internacia reto aŭ mirakla loka grupiĝo de lingvo-konsciaj komputikistoj… kompetentuloj, por kiuj engaĝiĝinta kunlaborado provizore kompensos la mankon de laborkontrakto en Hajdarabada centralo?

Bibliografio

[Baker 1997] Mark C. Baker. *Thematic Roles and Syntactic Structure*. Kluwer. p. 73–137.

[Brown 1988] Peter F. Brown et al.: *A statistical approach to language translation.* Proceedings International Conference on Computational Linguistics (COLING-88). Budapeŝto. p. 71-76.

[60] Eĉ por la angla, pri kiu la ilaro estas plej avancita, kvar jardekojn da AT-historio ne disponeblis deca parsilo.

[Brown 1990] Peter F. Brown et al.: *A statistical approach to language translation.* Computational Linguistics, junio 1990, vol. 16, n-ro 2, p. 79-85.

[Brown 1993] Peter F. Brown et al.: *The mathematics of Statistical Machine Translation: Parameter Estimation.* Computational Linguistics, junio 1993, vol. 19, n-ro 2, p. 263-311.

[Callison-Burch 2004] Chris Callison-Burch, Colin Bannard, Josh Schroeder: *Improved Statistical Translation Through Editing.* School of Informatics, University of Edinburgh; Linear B Ltd., Edinburgh Technology Transfer Centre.

[Carbonell 1992] Jaime G. Carbonell, Teruko Mitamura, Eric H. Nyberg, 3rd: *The KANT Perspective: A Critique of Pure Transfer (and Pure Interlingua, Pure Statistics, ...).* Fourth International Conference on Theoretical and Methodological Issues in Machine Translation, Montréal.

[Charniak 2000] Eugene Charniak: *A maximum-entropy-inspired parser.* Proceedings 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL), Seattle.

[Collins 1999] Michael Collins: *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

[Ding 2003] Yuan Ding, Daniel Gildea, Martha Palmer: *An Algorithm for Word-Level Alignment of Parallel Dependency Trees*. Proceedings MT Summit IX, New Orleans.

[Hacken 2001] Pius ten Hacken: *Revolution in Computational Linguistics.* Language and Computers, decembro 2001, vol. 37, n-ro 1, p. 60-72(13).

[Hutchins 1992] W. John Hutchins, Harold L. Somers: *An Introduction to Machine Translation.* Academic Press.

[Hutchins 2003] John Hutchins: *Has machine translation improved? Some historical comparisons.* Proceedings MT Summit IX, New Orleans.

[Hwa 2002] Rebecca Hwa, Philip Resnik, AmyWeinberg: *Breaking the Resource Bottleneck for Multilingual Parsing.* Institute for Advanced Computer Studies and Department of Linguistics, University of Maryland.

[Knight 2004] Kevin Knight, Philipp Koehn: *What's New in Statistical Machine Translation*. Tutorial at Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), Boston.

[Koehn 2002] Philipp Koehn, Kevin Knight: *ChunkMT: Statistical Machine Translation with Richer Linguistic Knowledge.* Pere de: http://people.csail.mit.edu/people/koehn.

[Koehn 2003] Philipp Koehn, Franz Josef Och, Daniel Marcu: *Statistical Phrase-Based Translation.* Proceedings (Main Papers) Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), Edmonton. p. 48-54.

[Lin 1995] Dekang Lin: *A dependency-based method for evaluating broad-coverage parsers.* Proceedings International Joint Conference on Artificial Intelligence (IJCAI-95), Montréal. p. 1420–1425.

[Lin 2004] Dekang Lin: *A Path-based Transfer Model for Machine Translation.* Proceedings International Conference on Computational Linguistics (COLING-2004), Genevo.

[Lopez 2002] Adam Lopez, Michael Nossal, Rebecca Hwa, Philip Resnik: *Word-level Alignment for Multilingual Resource Acquisition*. Language and Media Processing Laboratory

(LAMP) Technical Report 085, Institute for Advanced Computer Studies, University of Maryland (UMIACS).

[Maxwell 1989] Dan Maxwell, Klaus Schubert (eds.): *Metataxis in Practice - Dependency syntax for multilingual machine translation.* Foris Publications.

[Nagao 1992] Makoto Nagao: *Some Rationales and Methodologies for Example-based Approach.* Proceedings, International Workshop on Fundamental Research for the Future Generation of Natural Language Processing (FGNLP). Sofia Ananiadou (ed.), Manchester.

[Ratnaparkhi 1999] Adwait Ratnaparkhi: *Learning to parse natural language with maximum entropy models.* Machine Learning, 34(1-3) p. 151–175.

[Sadler 1989] Victor Sadler: *Working with Analogical Semantics: Disambiguation Techniques in DLT.* Foris Publications.

[Sato 1990] Satoshi Sato, Makoto Nagao: *Towards Memory-based Translation.* Proceedings, International Conference on Computational Linguistics (COLING-90), Helsinki.

[Sato 1991] Satoshi Sato: *Example-Based Machine Translation.* Ph.D. thesis, septembro 1991, Universitato de Kioto.

[Schubert 1986] Klaus Schubert: *Syntactic Tree Structures in DLT.* BSO/Research, Utreĥto.

[Schubert 1987] Klaus Schubert: *Metataxis - Contrastive dependency syntax for machine translation.* Foris Publications.

[Yamamoto 2000] Kaoru Yamamoto, Yuki Matsumoto: *Acquisition of Phrase-level Bilingual Correspondence using Dependency Structure.* Proceedings, International Conference on Computational Linguistics (COLING-2000), Saarbrücken.