SCIENCA REVUO de Internacia Scienca Asocio Esperantista BEOGRAD, Jugoslavio	El Vol. 23 n-ro 1-2(93-94) 25.4.1972.

PRI STRATEGIO DE AŬTOMATA LINGVO-TRADUKADO

(H. D. Maas, Universitato de Saarlando, FRG)

En la jaro 1967. en la kadro de la germanistika instituto de Universitato de Saarlando /FRG/ fondiĝis laborgrupo, okupiĝanta pri la eblecoj de aŭtomata tradukado de sciencaj rusaj tekstoj en la germanan lingvon. Komence oni planis modifi jam transprenitan el Usono tradukalgoritmon por la rusa-angla tradukado, sed tiu ĉi sistemo montriĝis tiel netaŭga dum testado, ke oni rezignis tute pri ĝi kaj anstataŭe evoluigis novan metodon.

1. ĜENERALA SUPERRIGARDO

Ni planas efektivigi la maŝinan tradukon en tri fazoj, nome

a/ analizo
b/ transfero
c/ sintezo

La unua paŝo analizo havas la taskon kolekti kiel eble plejmulte da informoj pri ĉiu unuopa tradukota frazo. La analizo devas liveri indikojn pri la strukturo de la frazo, pri la rilatoj inter unuopaj vortoj kaj vortogrupoj, kaj ĝi devas trovi la aktualan sencon de plursignifaj vortoj. Ĉiuj ĉi tiuj informoj estos sisteme kolektitaj por la sekvanta transfera paŝo.

Estas eble difini la nocion analizo per matematikaj rimedoj. Tiucele ni uzas la sekvantajn jam ĝenerale konatajn difinojn pri formala generativa lingvo:

Dif.1.1: Formala lingvo estas kvaropo /O,T,S,P/, kie
O estas finita ensemblo /signaro, alfabeto/
O' estas la ensemblo de ĉiuj sinsekvoj de elementoj el O /simbolĉenoj, signoĉenoj/
T estas nemalplena subensemblo de O /alfabeto de finaj simboloj/
S estas simbolo el O, kiu ne troviĝas en T / komenca simbolo/
P estas finita ensemblo de paroj de simbolĉenoj el O'.

La elementoj de P estas nomataj produkcioj. Se a estas produkcio, oni povas skribi ĝin kiel a = /p,q/ aŭ a = p→q.

Tiu lasta skribmaniero indikas la funkcion de la produkcioj: Ili servas por anstataŭado de simbolvicoj. La nomita produkcio a ekz-e permesas, ke anstataŭ p oni skribu q. Tial oni legas la regulon a ankaŭ kiel "skribu q anstataŭ p". La sistemo P do konsistas el difinita kvanto da tiaj substituaj reguloj.

Dif. 1.2: Estu x kaj y sinsekvoj de elementoj el O, do x = upv kaj y= uqv /kie u,v,p kaj q ankaŭ estas tiaj simbolĉenoj/ . Se ekzistas produkcio p→q, ni diras, ke y estas senpere derivita el x /pere de la produkcio p→q/ .

Dif. 1.3: Estu f = /x_o, x_1, x_2, . . . , x_n/ sinsekvo de simbolĉenoj el O', kie x_1 estas senpere derivita el x_o, x_2 senpere derivita el x_1 ktp. Tiukaze ni diras, ke x_n estas derivita el x_o /pere de la produkcia sistemo P/. Plue: x_o nomiĝas la origino de f /formule: x_o = O/f/ / kaj x_n estas la celo de f /formule: x_n = C/f/ /. f nomiĝas derivaĵo /de x_n/.

Ni vidas, ke ni ricevos derivaĵon, se ni sinsekve substituas helpe de la reguloj de iu donita produkciaro. Se oni derivas tiamaniere ke oni substituas ĉiam tiun anstataŭeblan simbolĉenon, kiu troviĝas plej maldekstre, oni ricevas specialan derivaĵon, nome kanonikan.

Dif. 1.4: La simbolĉeno z, konsistanta nur el elementoj el T, estas frazo de la lingvo /O,T,S,P/, se ekzistas derivaĵo f, kies origino estas la komenca simbolo S kaj kies celo estas z.

For nala lingvo nomiĝas generativa, se ĝia produkciaro plenumas la sekvantajn postulojn:

a/ Oni povas substitui nur unuopajn elementojn /ne-finajn/, senkondiĉe de la kunteksto.
b/ Ne ekzistas regulo kiu permesus elimini iun elementon.

Praktikaj provoj montras, ke estas eble konstrui formalajn generativajn lingvojn, kies frazaro estas grandparte tiu de iu homa lingvo [1]. Do ni povas supozi, ke pli-malpli bone eblas alproksimiĝi al natura

lingvo pere de formala. Ties produkcia sistemo respegulas la gramatikon de la formale priskribita lingvo.

Dif. 1.5: Se por frazo z ekzistas n malsamaj kanonikaj derivaĵoj, oni diras, ke z estas n-signifa.

Plursignifaj frazoj laŭ tiu formala senco ankaŭ ĝenerale estas plursencaj por homa leganto. Tre ofte tamen la ĉirkaŭaj frazoj de la teksto unusencigas ĝin.

La aŭtomata analizo de frazo konsistas el la trovado de ĉiuj ĝiaj kanonikaj derivaĵoj. Ili ne estas konservitaj en la maŝino kiel sinsekvo de senperaj derivaĵoj /laŭ dif. 1.3/, sed en sistemo de informoj pri aparteno al sintaksaj kompleksoj kaj pri dependecoj.

La analizorezultoj estas pritraktataj de la transfera algoritmo, kiu konsideras la diversajn sintaksajn esprimeblecojn de la du lingvoj. Speciale la elekto de taŭga vortekvivalento por vorto el la origina teksto apartenas al la transfera fazo. Plej grandaj problemoj en la tradukado el la rusa lingvo estas la elekto de la artikolo kaj de la tempo. Entute la transfero estas kampo ankoraŭ preskaŭ ne esplorata.

Matematike oni povas priskribi transferon per tio, ke ĝi estas funkcio /mapo/, transformanta la derivaĵojn de la origina lingvo en tiajn de la cela lingvo.

La sintezo finfine estas prezentebla matematike kiel la celfunkcio de dif. 1.3. El donita derivaĵo ĝi produktas la koncernan frazon. La reguloj de sinteza sistemo povas multe simili al produkcioj de iu generativa formala lingvo. La diferenco kuŝas en tio, ke pere de normalaj produkcioj oni povas produkti kiun ajn frazon, elirante de la komenca frazsimbolo S. Tute alie dum sintezo la produkto-algoritmo estas katenita de la rezultoj liveritaj de la transfero kaj liveros nur unu frazon. Tiu ĉi produktado estas sendependa de la gramatiko de la fontolingvo.

Tial oni povas starigi la sintezan algoritmon sen scioj pri la fontolingvo, eĉ sen profunda kono de la analizo-strategio. De nia laborgrupo la sintezo estis ellaborita, antaŭ ol la analizo estis pritraktita. En [2] oni trovos pliajn informojn pri tiu ĉi temo.

2 PRI LA ANALIZO-STRATEGIO

La sintaksa analizo dividiĝas en du gravajn partojn. La unua paŝo "tekstopreparo" kontentigas teknikajn postulojn, la dua paŝo estas la analizo en ĝia propra senco.

2.1. TEKSTOPREPARO

Ni imagu, ke oni jam tajpis la tradukotan tekston sur trukartojn, tiel ke ĉiu karto estas plenskribita. Ĉar inter la unuopaj vortoj estas interspaco. estas relative facile ekkoni ilin per programo kaj aldoni al ĉiu kurantan numeron. Post tio eklaboros programo, kiu vicigas la vortojn alfabete, ne forgesante kuntreni ĉies lokonumeron. La celo de tiu ĉi procedo estas, ke nun eblas facile serĉi la vortojn sinsekve en sintaksa vortaro / sur magneta bendo aŭ diskaro/, kiu liveras necesajn informojn pri la vortoklaso / substantivo, verbo, pronomo,..../, pri nombro / singularo, pluralo/, pri tempo /prezenco, preterito, futuro/, pri la postulo de komplementoj /rekcio/ ktp. Pro la riĉeco de larusa fleksio, kiu kaŭzas, ke ĉiu vorto povas aperi en tre multaj formoj, oni ne klopodas listigi ĉiujn vortoformojn, sed nur radikojn. Tial necesas morfologia programo, kiu post detranĉo de la finaĵoj liveras gramatikajn informojn. Kiam ĉiu vortoformo de la origina teksto ricevis sian gramatikan maskon, oni reordigas aŭtomate la vortojn laŭ iliaj lokonumeroj, tiel ke post tiu operacio la teksto denove staras en sia origina sinsekvo. La vera analizo povas komenciĝi.

2.2 PRI LA ANALIZO DE NOMINALAJ KOMPLEKSOJ

La "vera" analizo konsistas ĉefe el tri fazoj, nome el nominala analizo A_n, verbala analizo A_v kaj esploro de subfrazoj A_s. La nomitaj tri operatoroj pritraktas la frazon de maldekstre dekstren. Se P estas analizociklo, en kiu A_n, A_v kaj A_s estas uzataj unu post la alia, do formule $P = A_s A_v A_n$, tiam la analizo de frazo S_o reprezentiĝas kiel rekursiva sistemo

$$S_i = P/S_{i-1}/ \quad \text{por } i=1,2,3,...$$

La apliko de P al iu strukturo S_j signifas ties transformiĝon en S_{j+1} kaj la konservon de la tiel gajnitaj informoj. La operacio finiĝas, kiam aperas la fina simbolo S.

La plej grava kriterio por tio, en kiu programpaŝo vorto estas traktata, estas ĝia aparteno al unu el la tri vortkategorioj. Substantivoj,

atribuaj adjektivoj, adverboj, pluraj specoj de pronomoj, prepozicioj ktp. apartenas al la nominala kategorio /entute 15 klasoj/, finitivaj kaj infinitivaj verboformoj, predikataj adjektivoj ktp. al la verbala kategorio, dum la konjunkcioj /kunordigaj/ apartenas al la malampleksa kategorio konjunkcioj.

En la nominala analizo estas traktataj nur vicoj el nominalaj vortoj. La klasojn ni difinis laŭ kriterioj, havantaj decidan influon sur la sinsekvon de la unuopaj analizo-paŝoj.

La nominala analizo mem konsistas el kvar gravaj operacioj, nome

a/ ekkono de sinsekvo de nominalaj vortklasoj kiel subkomplekso,
ekz-e LA TRE DIKA, SED MALMULTEKOSTA LIBRO

b/ ekkono de kunordigitaj nominalaj kompleksoj,
ekz-e LA PATRO KAJ LA FILO

c/ ekkono de atribuitaj kompleksoj,
ekz-e LA LIBRO DE MIA FRATO

d/ ekkono de enkadrigitaj kompleksoj,
ekz-e LA LEGITA DE MIA FRATO LIBRO
aŭ LA DE MIA FRATO LEGITA LIBRO

2.2.1. Analizo de subkompleksoj

Seninterrompa vico de nominalaj vortklasoj estas rigardata kiel subkomplekso, se la sinsekvo de la klasoj estas sintakse allaseblakaj se la postuloj pri gramatika kongruo estas plenumataj. Aparta problemo aperas en la analizo de rusaj nominalaj grupiĝoj, se en la vico staras numeralo, kiu ekzercas specialan rekcion al sekvantaj nominalaj vortoj, kiuj devas montri difinitan kazon kaj nombron. Tiun malfacilaĵon oni majstras aldonante plurajn deklinaciajn maskojn al la numeraloj. Ekzemple ĉe DVA /=du/ oni havas unuan maskon pluralo nominativo /akuzativo maskulina/ neŭtra, kaj la duan genitivo singularo, ĉar substantivo post DVA devas esti en genitiva-singulata formo.

La sintaksa allaseblo de sinsekvo estas esplorata per la programo pri allaseblo. Tiu ĉi programo konstatas surbaze de la jam analizita parto de subkomplekso, kiuj dekstre sekvantaj vortoklasoj ankoraŭ povas aparteni al ĝi; la prognozo pri la allaseblaj vortoklsoj estas konservata en aparta memorloko. Antaŭ ol la kongruo de nova vorto kun la jam konstatita komplekso-parto estas kontrolata, la programo ekzamenas la allaseblon de la klaso de tiu vorto.

La kunordigado de adjektivoj, prepozicioj ktp., starantaj interne de subkomplekso, estas konstatata laŭ simila maniero en tiu ĉi unua programparto. Ĉi tie la allaseblo de supozita kunordigita vorto estas ekzamenata surbaze de la enhavo de memorilo pri kunordigebloj. Subkomplekso estas /provizore/ kompleta, kiam nova vorto estas ne-allasebla aŭ ne-kongrua. Per tiu ĉi nova vorto komenciĝas do

nova subkomplekso. Tiamaniere eblas malkovri la limojn inter tuŝiĝantaj subkompleksoj nominalaj.

2.2.2. Kunordigoj de kompleksoj

Tuj post la apero de du konstatitaj nominalaj kompleksoj, enfermantaj kunordigan konjunkcion, la maŝino esploras, ĉu ili povas formi kunordigon. Ĉe tio estas postulataj kongruo laŭ kazo kaj laŭ sintaksa funkcio /ekzistas normalaj, relativaj, subjunkciaj kaj demandaj kompleksoj/. Oni antaŭvidis, ke prepoziciaj kompleksoj povas esti interjungitaj kun adverboj. Se la kontrolaj operacioj ne alportas negativajn rezultojn, tiam ni rigardas la kunordigon kiel certa. La informoj de la dekstra komplekso estas translokitaj el la normala memorilo en la memorilon por dependaj grupiĝoj, kie estas konservataj la informoj pri atribuitaj kaj kunordigitaj kompleksoj. Ĉe la maldekstra komplekso oni konservas la memorilan adreson de la dekstra, kaj inverse.

Se tamen la dekstra grupo ne povas esti kunordigita kun la ĵus antaŭa, tiam la komplekso, reganta tiun maldekstran, devas esti esplorata koncerne kunordigeblecon ktp.

Se komplekso estas neniel kunordigebka, ĝi estas rigardata /provizore/ kiel memstara.

2.2.3. Atribuitaj kompleksoj

Atribueblecо de komplekso B estas ekzamenata, kiam ĝi estas konstatita kiel kompleta /vidu 2.2.1./ kaj se nominala komplekso A staras rekte antaŭ ĝi. Necesa kondiĉo por atribua rilato estas tio, ke rekcio-postulo de A estas kontentigita de B. Se la rekcio povas esti saturita, tiam oni supozas, ke la atribua rilato estas certa. La informoj de la dekstra /atribua/ grupo estas poste transportitaj el la nominala memorilo en la dependo-memorilon. Se la kontrolo liveris negativan respondon, tiam estos esplorita, ĉu ekzistas komplekso A', havanta liberan valenton kaj reganta la grupon A, tiel ke tiu ĉi A' estas ebla reganto de B... Genitiva grupo estas ĉiam konsiderata kiel atribuita al la ĵus antaŭa komplekso, se ties lasta elemento estas substantivo.

Se supozita atribuebla komplekso estas neniel atribuebla, ĝi estas rigardata kiel /provizore/ memstara.

2.2.4. Enkadrigitaj kompleksoj

Estu A nominala komplekso. entenanta rekciipovan adjektivon /aŭ participon/, kiun povas saturi la grupo B. Tiam A estas disigebla laŭ la formo $A = A_1 A_2$, kie la menciita adjektivo situas ĉe la dekstra rando de A_1, kaj $A_1 A_2$ povas pluformiĝi al $A' = A_1 B A_2$ /vidu ekzemplon /d/ de 2.2/. En la germana lingvo la reganta adjektivo devus esti maldekstre en A_2, dum en Esperanto ambaŭ pozicioj estas permesitaj. Tiun ĉi fenomenon ni nomas enkadrigo.

La aŭtomato enketas pri enkadrigo, kiam la nominala enkadrigo entenas almenaŭ tri sinsekvajn grupojn. Necesaj postuloj por la ekzisto de enkadrigo estas: La lasta vorto de A_1 devas esti rekciipova adjektivo; tiu ĉi rekcio devas esti saturita de B; post la elimino de B devas montriĝi, ke A_1 povas esti kompletigita /laŭ 2.2.1/. Se ĉiuj postuloj estas plenumitaj, oni rigardas la enkadrigon definitiva, translokas la informojn pri B en la enkadrigo-memorilon ktp.

2.3 PRI LA ANALIZO DE VERBALAJ KOMPLEKSOJ

Kiam aperas vorto el la verbala kategorio, ĝi estas sendita en la analizoparton A_v, kie unue la programo esploras, ĉu en la sama analizoĉeno / = parto de subfrazo/ jam ekzistas verbalo. Se jes, ekfunkcias subprogramo /"subrutino"/, kontrolanta la allaseblon de la lasta vorto. Se ĝi estas permesebla, la subrutino redonas karakterizan numeron por la tipo de la nova verbala grupiĝo, ekz-e finitiva verbo estas karakterizita per 1, infinitiva per 2, finitivo plus infinitivo per 10. Entute ni distingas ĉ. 30 malsamajn tipojn.

Kiam la maŝino konstatas la /provizoran/ finon de analizoĉeno, ĝi saltas en la programon por subfrazoj /A_s/. Tio okazas, se aperas komo aŭ kunordiga konjunkcio, kiuj ne estas partoj de kunordigita nominala grupiĝo.

2.4. PRI LA ANALIZO DE SUBFRAZOJ

Ĉi tie oni esploros komence, ĉu ĝia verbala komplekso estis ĝuste analizita. Surbaze de la supozata tipo oni devas kontroli saturitecon de rekcioj kaj necesaj postuloj pri tempoj kaj aspektoj. Se la verbala grupo estas ankoraŭ nekompleta, ankaŭ la analizoĉeno estas ankoraŭ nekompleta subfrazo; sed se ĝi estas perfekta, oni povas kontroli, ĉu la sendependaj nominalaj kompleksoj saturas la rekciojn de la verbala grupo. En tiu ĉi paŝo oni unusencigas nominalajn grupojn, kiuj restis plursignifaj laŭ kazo aŭ nombro. Krome eblas elekti tiun ĉi varianton de verbo, kies rekcioj estas realigitaj, se la ekkono de tiu varianto necesas por la ĝusta traduko.

Same kiel ĉe nominalaj kompleksoj oni devas konsideri enkadrigon, ekz-e

LA LABORO, KIUN MI FARAS, ESTAS INTERESA /tri analizoĉenoj/

atribuitecon, ekz-e

MI TRINKAS, ĈAR MI SOIFAS /du ĉenoj/

kaj kunordigon, ekz-e

MI LEGAS, SED VI SKRIBAS /du ĉenoj/

Oni antaŭvidas ke tiuj ĉi fenomenoj povas kombiniĝi, ekz-e LA LABORO, KIUN MI FARIS, DUM VI DORMIS KAJ LA INFANOJ KRIIS, ESTIS INTERESA.

2.5. HOMOGRAFECO

En la analizo la plej grava problemo estas la homografeco. Ni nomas vortoformon x homografa, se ekzistas almenaŭ du vortoj /leksemoj, lemoj/ a kaj b, tiel ke x troviĝas en la provizo de vortformoj de a kaj de b. En Esperanto homografeco estas malofta, ĉar la formoj de la ĉefaj vortoklasoj ĉiam distingiĝas laŭ siaj finaĵoj. Sed ekz-e la vortoformo kiu estas homografa: ĝi apartenas al la leksemo kiu kiel relativa pronomo kaj al KIU kiel demanda pronomo /la lasta krome funkcias subjunkcie/. Tial la frazo

KIU SCIAS, KIU VIDAS

estas plursignifa laŭ nia difino 1.5. La subfrazo KIU VIDAS povas esti kunordigita aŭ "atribuita" al KIU SCIAS.

En germanaj tekstoj preskaŭ ĉiu dua, en rusaj ĉiu deka vorto estas homografa. En la malnova aŭtomata analizo por germanaj frazoj, ellaborita antaŭ kelkaj jaroj de la Saarbrückena laborgrupo, troviĝas paŝo solvo de homografeco antaŭ la vera analizofazo !3!. Sed tia procedo ne estas pravigebla lingvistike nek analizopraktike, ĉar homografeco ne estas produktiva lingvofenomeno - ĝi estas pli-malpli hazarda. En la difino de formala lingvo - kaj cele al analizo ni devas rigardi niajn lingvojn formale - oni ankaŭ ne parolas pri homografeco, ĉar ĝi ne starigas ian problemon en la produktado de frazoj.

Homografeco do estas nura analiza problemo.

Tial en la nova rusa analizo la homografeco ne plu estas speciala unua programfazo /kiu plue anticipas jam grandparte la rezultojn de la posta analizo/, sed ĝi prezentiĝas kiel operacio, kiun oni helpe alvokas dum ĉiu analizostadio. Tiu ĉi procedo postulas bone elpensitan memorigan teknikon, tiel ke la solvo de homografiaj problemoj similas al serĉado de eliro el labirinto. Verŝajne tio postulos pli da kalkula tempo ol se ni disponus pri specialaj solvoprogramoj por ĉiu homografo-klaso. Sed la vojo nun elektita ŝajnas plej praktikebla, konsiderante la limigitan laborforton de malgranda teamo kaj la deziron kiel eble plej baldaŭ atingi rezultojn.

3. MENCIITA LITERATURO

[1] H.D.Maas, Die Beschreibung der deutschen Sprache durch eine af-Sprache. En: Beiträge zur Linguistik und Informationsverarbeitung. kajero 12, 1967.

[2] H.D.Maas, Die Synthese deutscher Sätze im Zusammenhang mit maschineller Sprachübersetzung. En: Muttersprache, kajero 9/10, 1969.

[3] H.Eggers k.a., Elektronische Syntaxanalyse der deutschen Gegenwartssprache, Max Niemeyer Verlag Tübingen, 1969.