

STROKOVNI ČLANKI

Peter Tancig

Računalniško prevajanje Machine Translation

The paper deals with many facets of machine (aided) translation (M(A)T). the topic of M(A)T is put into a broader framework, extending from simple word processing techniques, involving editing and formatting, across desk top publishing to a fuzzy area of natural language components such as spelling checkers, style and grammar checkers, and copy editing aids.

The less ambitious level of MAT consists of MAHT (machine aided human translation), HAMT (Human Assisted Machine Translation) and TD (Terminology Data banks).

MT proper is discussed from its enthusiastic beginnings, which dealt an almost fatal blow to the field of MT.

Today's setting of automated office and electronic networks is a natural 'habitat' for more ambitious MT systems, which are thus a natural extension of the information technology.

The paper continues with a more thorough discussion of 3 approaches to the design of MT systems and concludes with a taxonomy of MT generations and development periods.

More concrete details of operational MT systems and an assessment of the state-of-the-art in Yugoslavia are left for a subsequent presentation.

UVOD

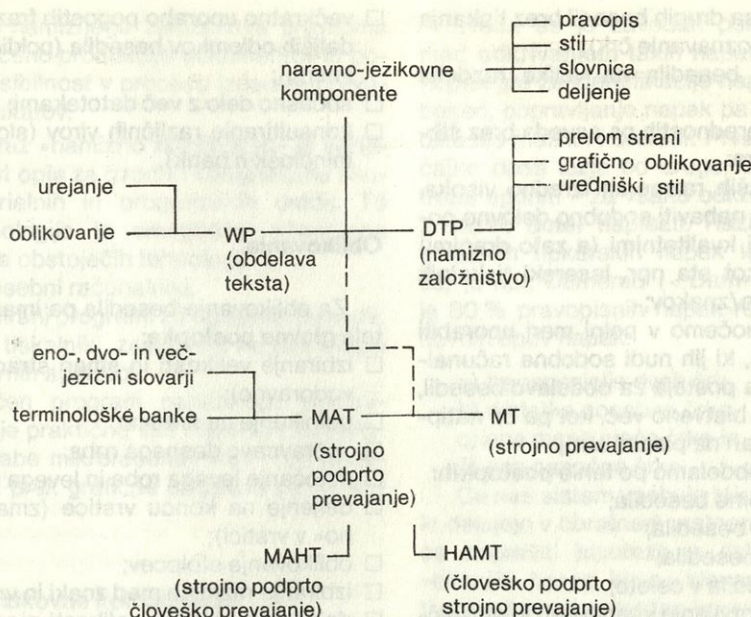
Jezik in računalnik

Jezik je osnovno komunikacijsko sredstvo med ljudmi, ki v vedno večji meri uporabljajo računalnike pri svojih dejavnostih. Spekter aplikacij računalniške obdelave jezika je zato izredno širok – naštejmo jih le nekaj:

- informacijski sistemi nad tekstualnimi bazami podatkov,
- leksikografija in leksikologija,
- komunikacija človek-računalnik,
- raziskovanje jezika in literature,

- pisanje besedil,
- prevajanje,
- dokumentalistika,
- založništvo,
- poučevanje.

Računalniško (podprto) prevajanje je eno od 7 aplikativnih področij računalniške obdelave naravnega jezika, za katere predvidevajo (<Joh-85>) za leto 1995 samo v ZDA prodajo v vrednosti 1,5 milijarde US\$. Teh 7 aplikativnih področij le naštejmo, nekaj več o njih pa je v (<TanVit-88>):



Slika 1 Širše okolje računalniškega in računalniško podprtega prevajanja

1. vmesniki za dostop v naravnem jeziku k velikim računalnikom;
2. vmesniki za dostop v naravnem jeziku k malim računalnikom;
3. vmesniki za dialog v naravnem jeziku;
4. pregledovanje vsebine besedil;
5. inteligentna obdelava tekstov;
6. strojno (podprto) prevajanje;
7. govorno-pisalni stroj.

Prevajanje in računalnik

Tudi če se omejimo na profesionalno **prevajanje s pomočjo računalnika**, moramo seči bolj v širino (slika 1) – tako proti različnim obdelavam teksta kot proti različno zahtevnim načinom računalniškega in računalniško podprtega prevajanja, ker je večkrat težko potegniti ločnico med posameznimi področji uporabe računalnika pri prevajanju.

OBDELAVA TEKSTA

Obdelava besedil (angleška kratica je **WP** – **World Processing**) danes ni več smotrna brez uporabe računalnikov, predvsem osebnih. Poglavitne **prednosti** takega načina dela so:

- ☐ veliko lažje vnašanje besedila – in to samo enkrat;
- ☐ izredne možnosti popravljanja in preurejanja besedila;
- ☐ možnost, da se spopademo z zelo dolgimi besedili – nekaj sto strani dolgo delo je mogoče urediti v nekaj dneh;
- ☐ velika izbira sredstev za grafično oblikovanje črk;
- ☐ vse na istem mestu – od ideje do rezultata na papirju nam ni treba zapustiti delovnega mesta;
- ☐ enostavno vključevanje slik v besedilo;

- ☐ možnost vnosa drugih besedil brez tipkanja (optično prepoznavanje črk);
 - ☐ hiter prenos besedila na velike razdalje brez papirja.
- Pri vseh teh prednostih pa seveda brez **šibkih točk** ne gre:

- ☐ **cena** je v naših razmerah izredno visoka, če si hočemo nabaviti sodobno delovno postajo z vsemi kvalitetnimi (a zato dragimi) pripomočki, kot sta npr. laserski tiskalnik in bralnik slike/znakov;
- ☐ **znanje**¹: če hočemo v polni meri uporabiti vse možnosti, ki jih nudi sodobna računalniška delovna postaja za obdelavo besedil, je treba znati bistveno več, kot pa če natipkamo eno stran na pisalni stroj.

Tekst lahko obdelamo po tehle **postopkih**:

- ☐ priprava vsebine besedila;
- ☐ pisanje delov besedila;
- ☐ popraviljanje besedila;
- ☐ ureditev besedila v celoto;
- ☐ ponovno popraviljanje vsebinskih in slovničnih napak;
- ☐ grafično oblikovanje besedila;
- ☐ prelom besedila na strani;
- ☐ končna ureditev (kazalo, indeks, opombe).

Včasih sta bili fazi **urejanja** in **oblikovanja besedila** povsem **ločeni** med seboj, kar se je odražalo tudi v posebnih programih (npr. interaktivni urejevalniki : paketni oblikovalniki). Danes to ni več potrebno, saj integrirani interaktivni paketi za obdelavo besedil vrste **WY-SIWYG (What You See Is What You Get)** vsebujejo tako urejevalne kot oblikovalne funkcije.

Urejanje

Za urejanje besedila imamo na voljo tele **osnovne postopke**, ki so tako ali drugače realizirani v urejevalnikih:

- ☐ pisanje besedila;
- ☐ premikanje po besedilu;
- ☐ vrivanje znakov, besed, vrstic, večjih kosov (blokov);
- ☐ brisanje znakov, besed, vrstic, blokov;
- ☐ iskanje po besedilu;
- ☐ zamenjavo nizov črk;
- ☐ prenos in kopiranje delov besedila;

- ☐ večkratno uporabo pogostih fraz (kratic) ali daljših odlomkov besedila (pokličemo jih po imenu);
- ☐ sočasno delo z več datotekami;
- ☐ konsultiranje različnih virov (slovarjev, terminoloških bank).

Oblikovanje

Za oblikovanje besedila pa imamo na voljo tele **glavne postopke**:

- ☐ izbiranje velikosti in smeri strani (pokonci, vodoravno);
- ☐ pomikanje na sredino;
- ☐ poravnavo desnega roba;
- ☐ določanje levega roba in levega umika;
- ☐ deljenje na koncu vrstice (zmanjša »belino« v vrstici);
- ☐ oblikovanje stolpcev;
- ☐ izbiranje razmika med znaki in vrsticami;
- ☐ določanje vrste in velikosti pisave ter nje-nega načina (kurzivno, polkrepko, podčrtano, stisnjeno, razprto, indeksi, potence);
- ☐ prelom besedila na strani;
- ☐ določanje položaja in velikosti slik.

Namizno založništvo

Malokatero področje je pritegnilo pozornost uporabnikov računalništva tako kot namizno založništvo (**DTP – Desk Top Publishing**). Povedano na kratko, namizno založništvo omogoča **popolno kontrolo nad tiskano besedo**. Omogoča produkcijo dokumentov s profesionalno podobo brez zunanjih tiskarskih in grafičnih storitev. S to tehnologijo je **mogoče**:

- ☐ pisati tekste,
 - ☐ urejati tekste,
 - ☐ risati ilustracije,
 - ☐ vgraditi kopije fotografije,
 - ☐ narediti načrt strani,
 - ☐ dokončno natisniti dokument,
- in vse to z relativno ne predragim računalnikom in laserskim tiskalnikom.

¹ Ki pa v nasprotju s ceno ni odvisno od tega, na kateri strani Alp živimo.

Sistemi namiznega založništva praviloma zmanjšajo ceno produkcije dokumentov in povečajo fleksibilnost v procesu izdelave pravočasnih rezultatov.

Toda izraz »namizno založništvo« je **varljivo preprost opis** za izredno **kompleksne** skupine materialnih in programskih orodij. To novo tehnologijo je omogočila edinstvena kombinacija obstoječih tehnologij:

- ☐ hitrejši osebni računalniki;
- ☐ sofinancirani programi za obdelavo besedil;
- ☐ laserski tiskalniki, zasnovani na tehnologiji fotokopirnih aparatov.

Povprečen program namiznega založništva vsebuje praktično vse elemente funkcionalne uporabe mikroročunalnikov – od obdelave teksta prek grafične obdelave pa do komunikacij.

Naravno-jezikovne komponente

Med klasično obdelavo besedil, v katero spadajo urejevaniki in oblikovalniki besedil, in naprednejšimi obdelavami naravnega jezika, kot sta npr. strojno prevajanje in komunikacija v naravnem jeziku z bazami podatkov, leži področje – z ne povsem definiranimi mejami – sistemov, ki z **vključevanjem znanja o jeziku** v t. i. naravno-jezikovne komponente omogoča bolj inteligentno obdelavo tekstov.

Preverjanje in popravljanje pravopisnih napak

Avtomatično odkrivanje in popravljanje **tipkarskih napak**, nastalih pri pisanju in urejanju tekstov je najbolj razvita tehnika vključevanja naravno-jezikovnih komponent v različne sisteme za obdelavo besedil. Sistemi identificirajo napačno zapisane besede na dva načina:

- ☐ glede na to, da take besede ni v slovarju;
- ☐ glede na to, da se takšno zaporedje črk nikoli ne uporablja v posameznem jeziku.

Oba načina imata svoje prednosti in slabosti; dober pregled te tehnologije je podan v (<Pet-80>).

Treba se je zavedati pomembne razlike med **odkrivanjem** takih napak: pri odkrivanju napak gre za identifikacijo napačno zapisanih besed, popravljanje napak pa vsebuje tudi indikacijo možnih korektur. Prva funkcija je običajno dosti lažja od druge, saj je pri slednji treba uganiti – za vsako odkrito napako – kaj je pisec **hotel** napisati. Raziskave o vrstah običajnih tipkarskih napak lahko pomagajo, saj je npr. Damerau (<Dam-64>) odkril, da je 80 % pravopisnih napak rezultat štirih osnovnih tipov napak:

- a) transpozicije dveh črk,
- b) dodatka posebne črke,
- c) ene manjkajoče črke in
- č) ene napačne črke.

Če naš sistem vsebuje taka pravila (a)–(č)), ki delujejo v obratnem vrstnem redu, je mogoče napraviti hipoteze o določenem številu »bližnjih« besed, ki so v slovarju, in jih ponuditi uporabniku kot možne korekture odkrite napake.

Običajno trdijo, da pri 80–90 % napak obstoječi sistemi uspešno odkrivajo napake in predlagajo ustrezne korekture.

Preverjanje stila

To so dokaj **preprosti** programi, ki ne vsebujejo ambicioznih obdelav naravnega jezika – nimajo npr. sintaktičnega analizatorja ali realiziranih konceptov iz umetne inteligence, kot so npr. predstavljanje znanja in ekspertni sistemi.

Vsi tovrstni programi uporabljajo isto osnovno tehniko. Sistem uporablja tabelo »nepravilnih« zaporedij znakov in predlaganih nadomestil ter preprost postopek za iskanje zaporedij. Ti programi so osnovani na načelih **dobrega stila pisanja**, kot jih je najti v različnih »vodičih za stil«:

- ☐ izogibaj se obrabljenih, praznih ali često napačno uporabljenih fraz;
- ☐ raje uporabljaj specifičen, konkreten jezik namesto nejasnega in abstraktnega;
- ☐ ne uporabljaj besed, ki se redko pojavljajo v običajnem jeziku;

□ pravilno uporabljaj ločila in pazi na slovnično pravilnost.

Obstaja ogromno knjig s tovrstnimi pravili, vendar jih je težko naposredno pretvoriti v uporabna orodja, saj naše teoretično razumevanje jezika še ne omogoča ustreznih modelov za konstrukcijo delujočih popolnejših tovrstnih programov.

Preverjanje slovnice

Sistemi za preverjanje slovnice vsebujejo **slovar in sintaktični analizator** – spadajo torej med zahtevnejše obdelave naravnega jezika.

Tak sistem skuša najprej analizirati posamezne stavke, kot da bi bili slovnično popolnoma pravilni. Če ne uspe, poskusi znova z nekoliko bolj **ohlajnimi** pravili, npr. opusti preverjanje ujemanja posameznih kategorij (spola, sklona itd.) med posameznimi deli stavka. Če sedaj uspe, lahko iz narave (p)opuščenega pravila sklepa o naravi problema, ki je preprečil prvo (popolno) analizo.

Pri konstruiranju takih sistemov je treba poleg ustreznih formalnih modelov morfoloških in sintaktičnih kategorij konkretnega jezika imeti tudi dovolj npr. poslovne korespondence, iz katere je mogoče napraviti spisek običajnih napak, jih formalno modelirati in to znanje vgraditi v program.

Pomoč uredniku založbe

Dosedaj smo obravnavali aplikacije, namenjene predvsem avtorju, ki je sam svoj urednik. V založništvu pa obstaja močna distinkcija med avtorjem teksta in tistim, ki je odgovoren za to, da delo izide v založbi.

Tak **profesionalni urednik** često temeljito posega v tekst – bodisi v samo vsebino ali pa ga reorganizira na višji ravni. Kljub temu pa porabi pretežni del svojega časa za operacije, kot sta **urejanje** in **korigiranje**, tj. za kontrolo podrobnosti na nižjih nivojih. Založniki vztrajajo, da naj bodo končni izdelki v skladu s *hišnim stilom*, kar često opredeljujejo s podrobnimi specifikacijami, ki ne zadevajo samo uporabe

ločil, ampak za nepoučenega prave malenkosti, kot npr. format datumov, zapis števil, uporabo kratic, sklicevanje na literaturo, navedbe literature, citiranje besed v tujih jezikih itd.

Pri teh zadevah često ni mogoče govoriti o (ne)pravilnostih, ampak je pomembna **konsistentnost**. Zato vsebujejo taki sistemi splošen jezik, v katerem uporabnik z lahkoto specifičira nova pravila ali modificira obstoječa – precej podobno, kot se to dela v sodobnih lupinah ekspertnih sistemov.

STROJNO PODPRTO PREVAJANJE

Pri računalniških pripomočkih za prevajanje moramo upoštevati **dva vsebinska pola**: poznavanje **jezika** in obvladovanje **področja** (teme) prevajanja. Za uporabnike takih računalniških sistemov je značilen **kontinuum različnih stopenj kompetentnosti** na obeh področjih, ki ga lahko omejimo z **dveema ekstremoma**. En ekstrem je posebej prilagojen prevajalcem, ki dobro obvlada jezik (tako **izvirni**, iz katerega prevaja, kot **ciljni**, v katerega prevaja), le malo ali nič pa ne ve o specialni temi besedila. Drugi ekstrem je raziskovalec, inženir, programer, pravnik, ekonomist, ki ima veliko znanja o svojem strokovnem področju, a le pasivno znanje danega tujega jezika.

V prvem primeru mora računalnik, ki naj funkcioniira kot prevajalno orodje, zagotoviti ne samo potrebne jezikovne elemente – kot so semantične enote, izrazi, sintagme, fraze – posebnega predmetnega jezika, ampak tudi čimveč informacij o sami temi. V drugem primeru mora računalnik poskrbeti ne samo za pravopis, vključno z diakritičnimi znaki, ampak tudi za dele besedila, spol, uporabo predlogov ipd.

Le če je zadoščeno vsem tem **heterogenim zahtevam**, je lahko tak sistem za računalniško podprto prevajanje uspešen.

Sistemi za strojno podprto prevajanje (Machine / Computer Aided / Supported Translation) se delijo v dve skupini:

1. človeško podprto strojno prevajanje (Human Assisted Machine Translation),
2. strojno podprto človeško prevajanje

(Machine Assisted Human Translation), ki zaporedoma zasedata nižji mesti na lestvici ambicij računalniškega prevajanja.

HAMT se nanaša na sistem, kjer je računalnik odgovoren za izdelavo prevoda *per se*, toda pri tem lahko sodeluje s človekom – lahko ga npr. vpraša, kako razrešiti dvournost dela besedila, kako označiti mesto, kamor (sintaktično) sodi kaka fraza, katero besedo ali frazo naj izbere med številnimi možnimi, ki jih je odkril v slovarju sistema.

MAHT se nanaša na sistem, kjer je človek odgovoren za izdelavo prevoda *per se* (on-line), toda pri tem lahko komunicira s sistemom v nekaterih, vnapraj predpisanih, situacijah – npr. zahteva pomoč pri preiskovanju lokalnega slovarja ali tezavra, pri dostopu do oddaljene terminološke banke, pri iskanju zgledov za uporabo besede ali fraze ali pri obdelavi tekstov (npr. oblikovanju).

Terminološke banke podatkov (Terminology Data) so najmanj ambiciozni sistemi, kajti pogosto se ne uporabljajo med samim prevajanjem (ni nujno, da prevajalec dela on-line), ampak običajno pred človeškim prevodom. Banke podatkov so mogoče lahko dosegljive samo v obliki **iztiskanih glosarjev** za posamezna predmetna področja. TD omogoča dostop do tehnične terminologije, toda običajno ne do navadnih besed (le-te uporabnik itak že pozna). Glavna prednost TD ni to, da so avtomatizirane (celo ob on-line pristopu lahko najdemo besede v slovarju prav tako hitro), ampak, da so **ažur(ira)ne**; tehnična terminologija se nenehno spreminja in natisnjeni slovarji so običajno že zastareli, ko postanejo dostopni. Za TD je prav tako mogoče, da vsebujejo **več podatkov**, saj lahko računajo na večjo skupino aktivnih prispevkarjev – svojih uporabnikov.

STROJNO PREVAJANJE

Sistemi za strojno prevajanje (**SP**; **MT** ozir. **CT**: Machine / Computer Translation) so namenjeni za prevajanje brez človeške intervencije. To ne izloči poprejšnjega urejanja (pre-editing) niti poznejšega (post-editing),

vendar so sistemi SP popolnoma odgovorni za celoten proces prevajanja od vhoda v izvirnem jeziku pa do izhoda v ciljnem jeziku brez človeške pomoči. Pri tem ti sistemi uporabljajo:

- posebne programe,
- izčrpne slovarje,
- zbirke jezikoslovnih pravil.

Sistemi za SP so na vrhu lestvice ambicij računalniškega prevajanja.

Kratka zgodovina

Za zgodnja 50. leta tega stoletja je bilo značilno navdušenje računalniških krogov, saj se je zdelo, da bodo novi zmogljivi elektronski računalniki kmalu omogočili kakovostno avtomatično prevajanje različnih naravnih jezikov. V raziskovanja na tem področju so vložili veliko sredstev, toda rezultati so bili pičli, ker so hoteli v enem naskoku rešiti vse zapletene probleme, ki jih (je) postavlja(l) naravni jezik.

Šok iztreznitve, ki ga je povzročilo poročilo Ameriške akademije znanosti (<ALPAC-66>) o polomu avtomatičnega prevajanja, je pomenil začetek podpore temeljitejšemu proučevanju različnih vidikov naravnega jezika – **računalniško jezikoslovje** (računalniška obdelava jezika, računalniško razumevanje jezika, jezikoslovna informatika) je dobilo svoje osnovne obrise.

Zaradi spoznanja, da je prenašanje sporočil v naravnem jeziku zelo zapleten proces, ki ga ni mogoče zajeti le v abstraktnih nizih simbolov (črkah in besedah), so pričeli upoštevati teoretične jezikoslovne izsledke o naravnem jeziku, obenem pa je to spoznanje koristno uvedlo formalne modele in računalniške pristope v jezikoslovna raziskovanja.

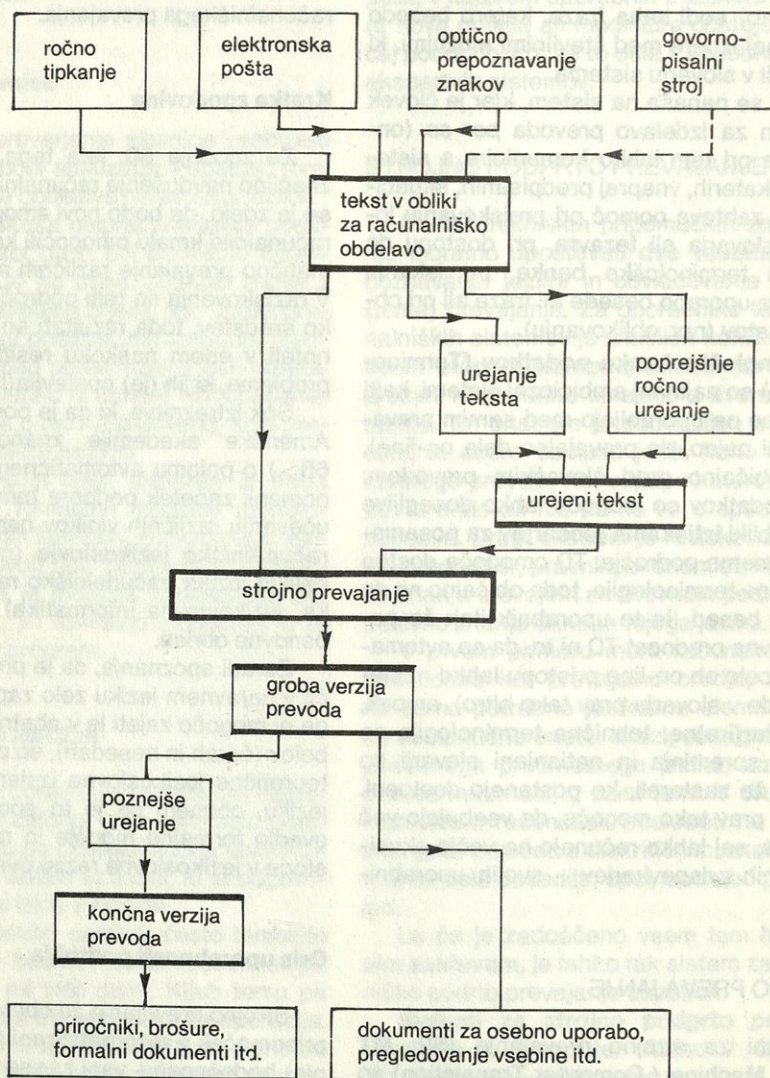
Oris uporabnosti področja

Strojno prevajanje že obeta postati bistven pripomoček v **avtomatizirani pisarni** (skorajšnje) bodočnosti – vsaj kar se tiče velikih uporabnikov (podjetij). Vsaka firma, ki bo namerala biti glavni dobavitelj opreme za elektronske pisarne, bo morala na neki način zagoto-

viti SP, bodisi z izdelavo lastnega produkta ali pa, bolj običajno, z integracijo produktov SP drugih dobaviteljev v lasten sistem.

Prevajanje kar vabi k uporabi **informatične**

tehnologije. Obdelovalniki tekstov so očitno pomembni za tiste tekste, za katere predvidevamo, da bodo zahtevali preoblikovanje in urejanje. Uporabo obdelovalnikov besedila je



Slika 2 Alternativne poti za računalniško (strojno) prevajanje v avtomatizirani pisarni (elektronskem okolju)

oviralo pomanjkanje sistemov z večjezikovnimi nabori znakov, vendar se ta problem uspešno rešuje. Prevod je bil vedno odvisen od hitre komunikacije, kar danes čedalje bolj zagotavljajo podatkovne mreže in elektronska pošta. Poleg sprejemanja in oddajanja besedil in popravkov začenjajo prevajalci uporabljati on-line terminološke banke, kar jim omogoča spremljanje novosti v slovarjih in terminologiji.

Z večanjem deleža prevajanja, ki se generira in distribuira v **elektronskem okolju** – in često tudi neposredno elektronsko dosega – je smiselno in atraktivno dodati sisteme strojnega prevajanja kot naravni podaljšek sistemov za obdelavo tekstov (slika 2).

Strojno prevajanje je uporabno samo za **nekatero vrsto** prevajalskega dela. Uspešno se uporablja tam, kjer je izvorno besedilo iz omejenega, vnaprej določenega področja in kjer ciljno besedilo pustimo bodisi v grobi verziji, primerni za poznavalce tega področja, ali pa ga pozneje človeški prevajalec uredi v obliko, primerno za publiciranje.

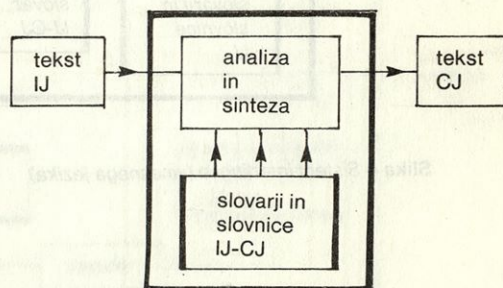
Očitni nadaljnji korak bo uporaba tehnologije **govorno-pisalnega stroja**, ki bo omogočala neposreden vhod izgovorjenih sporočil, ki jih bo najprej verificiral govorec, nakar jih bo sistem po potrebi parafraziral v obliko, ki bo omogočila zanesljiv avtomatični prevod. Sintetizator govora bo nato izgovoril končen prevod. Načeloma naj bi taka metoda omogočila **medjezikovno telefonsko komunikacijo**, ki jo že razvijajo nekatere japonske firme pod nazivom tehnologija prevajalnega telefona. Posamezne komponente takega sistema so bile individualno že demonstrirane z dokajšnjo »prepričljivostjo«, čeprav bodo praktični sistemi zahtevali višjo stopnjo delovanja vsake komponente, uspešno integracijo in zmožnost obravnavanja govora v obeh smereh v realnem času, za kar bo potrebno še kar nekaj let.

Načini načrtovanja sistemov za strojno prevajanje

Na splošno lahko ločimo tri načine načrtovanja (in realizacije) sistemov za SP.

Sistem direktnega prevoda

Pri tem načinu so sistemi zasnovani do vseh podrobnosti za **določen par jezikov** (slika 3).



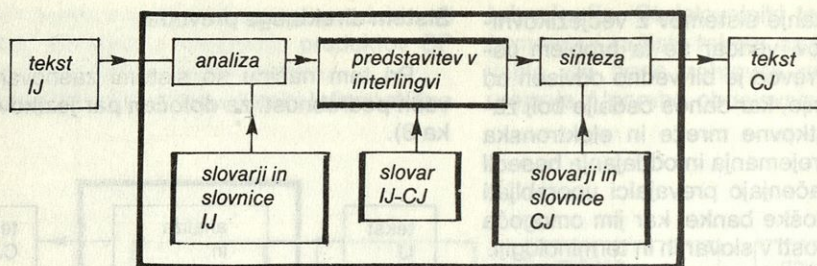
Legenda: IJ – izvorni jezik
CJ – ciljni jezik

Slika 3 Sistem direktnega prevoda

Osnovna predpostavka je, da **ni potrebno analizirati** slovarja in sintakse besedil v IJ bolj, kot je striktno potrebno za razrešitev dvoumnosti, pravilno identifikacijo ustreznih izrazov v CJ in specifikacijo besednega vrstnega reda v CJ. To npr. pomeni, če je zaporedje besed IJ dovolj blizu sprejemljivemu zaporedju besed CJ, ni potrebno identificirati sintaktičnih struktur besedila IJ. Večina sistemov za SP 50. in 60. let je bila zasnovana po tem načelu. Ti sistemi so se ločili po obsegu realizirane analize in/ali restrukturiranja izvirnega besedila v ciljno, kjer je bilo mogoče ločiti po stopnjah rastoče zahtevnosti:

- ☐ neposredni prevod slovarja;
- ☐ minimalno lokalno restrukturiranje v sistemih »beseda-za-besedo«;
- ☐ delno analizo strukture IJ;
- ☐ analizo celotnih stavkov IJ.

Osnovna značilnost zgodnjih sistemov direktnega prevoda je bila zabrisana meja med analizo IJ in sintezo CJ. Novejši direktni sistemi imajo bolj modularno strukturo.



Slika 4 Sistem interlingve (vmesnega jezika)

Sistem vmesnega jezika (interlingve)

Pri tem načinu predpostavljamo, da je mogoče besedila IJ pretvoriti v **semantično-sintaktične predstavitve**, ki so skupne več kot enemu jeziku. Iz takih interlingvalnih predstavitev naj bi se nato generirala besedila v drugih jezikih (slika 4).

Osnovni argument za tako prevajanje je **ekonomičnost v večjezičnih okoljih**. Če gre za prevajanje med vsemi pari n jezikov, bi sistem potreboval $n(n-1)$ binarnih programov direktnega prevoda in $2n$ interlingvalnih programov. Takoj ko število jezikov preseže 3, postane zato drugi način bolj ekonomičen.

Pri takih sistemih prevajanja iz IJ v CJ gre za **dve ločeni in neodvisni stopnji**:

- v prvi stopnji se opravi polna analiza besedil IJ v interlingvalno predstavitev;
- v drugi stopnji so interlingvalne predstavitve vir za izdelavo (sintezo) tekstov CJ.

Postopki za analizo IJ so specifični za IJ in niso zasnovani za kak poseben CJ v sistemu; podobno je tudi sinteza CJ specifična za posamezen CJ. Interlingvalni sistemi se ločijo po svojih **zasnovah interlingvalnega jezika**: »logični« umetni jeziki ali »naravni« pomožni jezik (kot je npr. esperanto), množica semantičnih primitivov, ki so skupni vsem jezikom, ali »univerzalni« slovar itd. Interlingvalni projekti SP so se razlikovali tudi v poudarku na leksi-

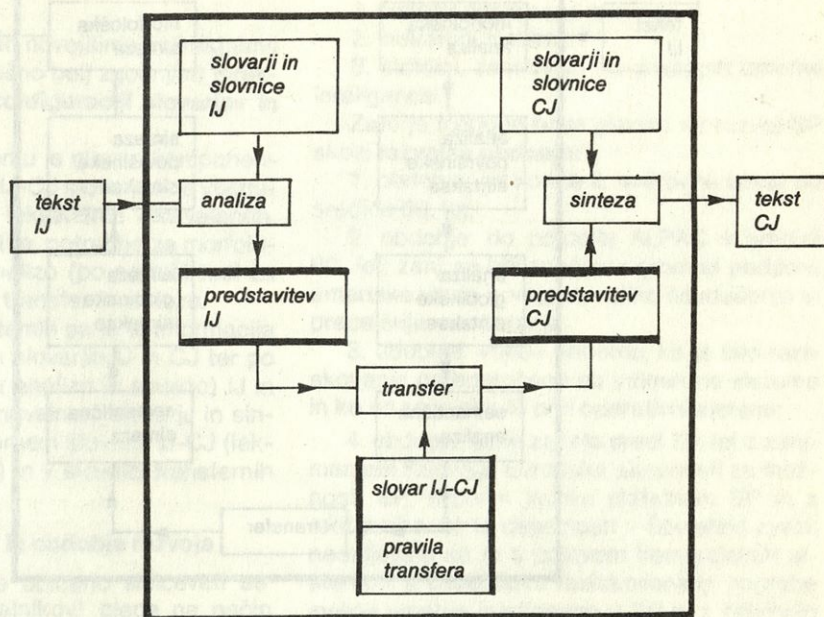
Sistem transfera

kahnih (semantičnih) vidikih in sintaktičnih vidikih – torej interlingvalna leksika proti interlingvalni sintaksi.

Namesto operacije v dveh stopnjah preko ene same interlingvalne predstavitve so pri transfernem sistemu **tri stopnje** z ustreznimi predstavitvami tako za besedila IJ kot za besedila CJ:

- prva stopnja pretvori besedila IJ v IJ transferno predstavitev, ki rabi kot vhod za
- drugo stopnjo, ki jo pretvori v CJ transferno predstavitev, ta pa rabi kot vhod za
- tretjo stopnjo, ki izdela končno obliko besedil CJ.

Interlingvalni način nujno zahteva popolno razrešitev vseh dvoumnosti in anomalij v besedilu IJ, kar omogoča prevod v katerikoli drugi jezik, transferni sistem pa se ukvarja edino z dvoumnostmi, ki so inherentne obravnavanemu jeziku. Razlike med jezikoma vrste *know-savoir/connaitre* se tako obravnavajo med transferom. Pri analizi angleščine se *know* obravnava kot nedvoumen in ni potrebno določiti, za katero vrsto »(po)zna(va)nja« gre. Interlingvalni način bi zahteval tako analizo, transferni pa je ne zahteva; problemi neujemanja med leksikalnimi dosegi IJ in CJ se razrešijo v komponenti transfera.

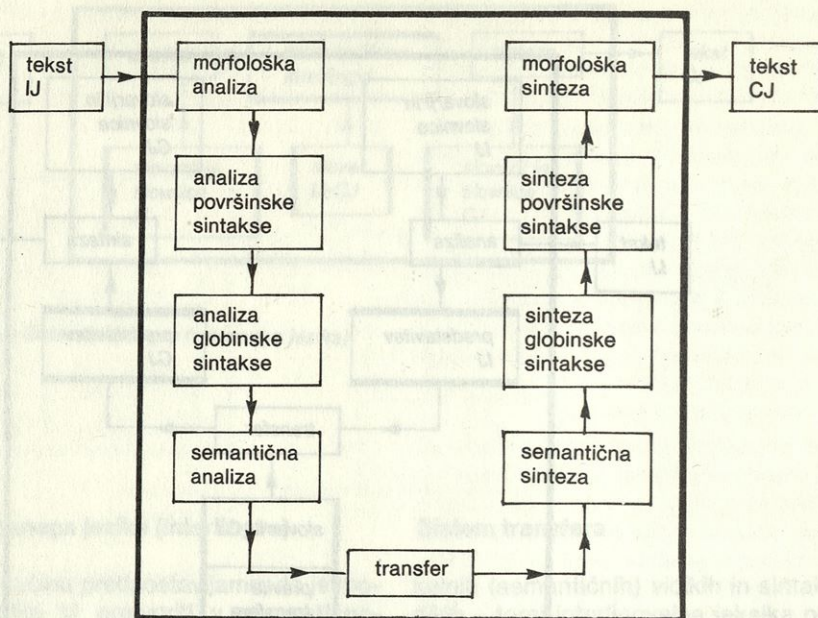


Slika 5 Sistem transfera

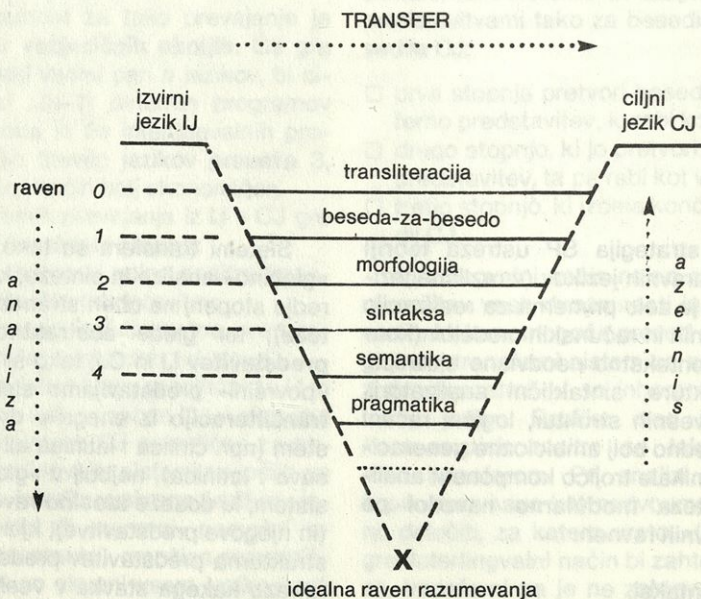
Transferna strategija SP ustreza **teoriji stratifikacije** naravnih jezikov (v različne jezikovne ravni) in je zelo primerna za realizacijo različnih formalnih in računskih modelov (končni avtomati, kontekstno neodvisne slovnice, drevesne strukture, sintaktični analizatorji, pretvorniki drevesnih struktur, logični računi itd.). Tako so vedno bolj ambiciozne generacije sistemov pomikale trojico komponent analiza-transfer-sinteza **modularno navzdol** po različnih jezikovnih ravneh:

- ☐ morfologiji,
- ☐ površinski sintaksi,
- ☐ globinski sintaksi,
- ☐ semantiki (slika 6).

Sistemi transfera se tako ločijo glede na »globino« analize in sinteze, kjer gre za zaporedje stopenj na obeh straneh (analize in sinteze), ter glede **abstraktnosti transfernih predstavitev** IJ in CJ; tako si lahko najbolj na »površini« predstavljamo sistem, ki opravlja **transliteracijo** iz enega v drug znakovni sistem (npr. cirilica : latinica ali pa japonska pisava : latinica), najbolj v »globini« pa idealni sistem, ki doseže **idealno raven razumevanja** (in njegove predstavitve), kjer bi ista osrednja strukturna predstavitev predstavljala vse parafraze kakega stavka v vseh jezikih (pri takem sistemu pa transferne komponente zato nenadoma ni več! – slika 7).



Slika 6 Stratifikacija jezikovnih ravni v sistemih transfera



Slika 7 Asimptotično približevanje idealni ravni razumevanja v sistemih transfernega prevoda

Razlike med direktnim načinom in indirektnima načinoma

Glavna razlika med novejšima indirektnima načinoma in (na splošno bolj zgodnjim) direktnim načinom je v **konfiguraciji slovarjev in slovnici**.

V direktnem sistemu je glavna komponenta en sam dvojezični IJ-CJ slovar, ki ne vsebuje le informacije o leksikalnih ekvivalentih, ampak tudi vse podatke, potrebne za morfološko in sintaktično analizo (po potrebi tudi za semantično analizo), transfer in sintezo.

Pri indirektnih sistemih pa je ta informacija razpršena po ločenih slovarjih IJ in CJ ter po ločenih slovnicih (za analizo in sintezo) IJ in CJ – bodisi po interlingvalnem slovarju in sintaksi ali pa po transfernem slovarju IJ-CJ (leksikalnih ekvivalentov) in v slovnici transfernih pravil struktur IJ-CJ.

Generacije sistemov in obdobja razvoja

V računalništvu je običajno sklicevati se na generacije računalnikov¹ glede na način realizacije **materialne opreme** računalnikov:

1. generacija: elektronke;
2. generacija: tranzistorji;
3. generacija: integrirana vezja;
4. generacija: integrirana vezja velike gostote.

Pri pričakovani »5. generaciji« računalnikov bodo **spremembe globlje** oz. vzdolž več dimenzij, saj bodo serijske arhitekture nadomestile paralelne, pojavljajo se nove organizacije računalniškega spomina, novi programski jeziki bodo vsebovali koncepte umetne inteligence, pojavljajo se dosti zahtevnejše aplikacije, med katerimi strojno prevajanje ni zadnja.

Podobno poimenovanje po generacijah velja tudi za metodologijo razvoja programske opreme in, ne presenetljivo, tudi za različne sisteme za SP, vendar tu **ni takega soglasja**, kaj je osnova za delitev. Nekateri takole razvrščajo generacije SP:

1. generacija: preprosti sistemi »beseda-za-besedo«;
2. generacija: sintaktična analiza;
3. generacija: vgraditev semantike določene vrste.

Nekateri pa takole:

1. sistemi direktnega prevoda;
2. indirektni sistemi;
3. sistemi, zasnovani na pristopih umetne inteligence.

Zato je mogoče bolje gledati na razvoj SP skozi **razvojna obdobja**:

1. obdobje: od konca II. svetovne vojne do sredine 50. let;
2. obdobje: do poročila ALPAC v sredini 60. let; zanj so bili značilni ogromna podpora ameriške vlade in vojske, veliko navdušenje in precejšnja razočaranja;
3. obdobje: »tiho« obdobje, ko je bilo raziskovanje osredotočeno na indirektno sisteme in ko so se uveljavili prvi operativni sistemi;
4. obdobje: se je začelo sredi 70. let z zanimanjem Komisije Evropske skupnosti za možnosti SP, s prvim javnim sistemom SP in z reorganizacijo te dejavnosti v Sovjetski zvezi; nadaljevalo se je s pojavom komercialnih sistemov, s povečanim raziskovanjem uporabe metod umetne inteligence v SP in z oživitvijo japonskega interesa;
5. obdobje: ker so bila ostala obdobja približno desetletja, bi radi verjeli, da se začenja novo obdobje, za katero so značilni prava eksplozija raziskovalnih dejavnosti, povezanih z japonskim projektom »pete generacije«, in renesansa raziskovanja SP v ZDA.

ZAKLJUČEK

Naj vstopa v novo obdobje ali ne, bodočnost SP je varna,² saj:

- ☐ zadovoljuje resnično nujne potrebe;
- ☐ je predmet raziskav in razvoja, razširjenih (in medsebojno povezanih!) po cellem svetu;
- ☐ postaja komercialni produkt kot drugi tehnični pripomočki in pisarniška oprema;
- ☐ uporaba računalnika v prevajanju je realnost, za mnoge je to že tako del življenja kot sam računalnik.

¹ Po svoje zanimiva je uporaba izraza »generacija« za opisovanje (bistvenih) konceptualnih sprememb, ki so se na tem področju dogodile v eni sami človeški generaciji.

² Vsaj v razvitih okoljih.

Neobravnavana problematika

Zaradi kompleksnosti in obsežnosti obravnavanega področja je obseg tega članka precej narasel in ga tako ni mogoče nadaljevati. Verjetno bi bilo zato smotno v **posebnem članku** prikazati še:

- bolj znane delujoče sisteme za strojno (podprto) prevajanje (<Gar-72>, <MTT-87>, <PEMT-82>, <Sch-88>, <TC-79>);

- »obnašanje« sistemov za prevajanje v realnih okoljih (prihranki, omejitve, motivacije uporabe, prednosti, slabosti);

- področje naprednih obdelav naravnega jezika (<Ing-88>, <Tan-86>);

- stanje računalniške obdelave jezikovnih podatkov pri nas (<ROJP-88>) in možnosti uporabe dosedanjih rezultatov oz. znanja (<VitTan-88>);

- nekaj specifičnosti slovenščine za računalniško obdelavo (<Tan-82>);

- možne smeri domačega usmerjenega razvoja strojnega (podprtega) prevajanja (<VitTan-88a>);

- slovarček pogostejših angleških izrazov in kratic iz področja obravnavane problematike.

Literatura

- | | |
|---|--|
| <p><ALPAC-66> Language and Machines, Computers in Translation and Linguistics. A Report by the Automatic Language processing Advisory Committee, National Academy of Sciences, National Research Council. Publication 1416. Washington, D.D., 1966</p> <p><Dam-64> Damerau F. J.: »A Technique for Computer Detection and Correction of Spelling Errors«; Communications of the ACM, 7, 1964</p> <p><Gar-72> Garvin P. L.: ON MACHINE TRANSLATION; Mouton, Teh Hague, 1972</p> <p><Ing-88> Ingria R. J. P.: »Natural Language Processing – Where It's Been and Where it Might Be Going«; v <ROJP-88></p> <p><Joh-85> Johnson T.: NATURAL LANGUAGE COMPUTING – THE COMMERCIAL APPLICATIONS; Ovum Ltd., London, 1985</p> <p><MTT-87> King M. (ed.): MACHINE TRANSLATION TODAY – THE STATE OF THE ART; Edinburgh University Press, 1987</p> <p><PEMT-82> Lawson V. (ed.): PRACTICAL EXPERIENCE OF MACHINE TRANSLATION; North-Holland, Amsterdam, 1982</p> <p><Pet-80> Peterson J. L.: COMPUTER PROGRAMS FOR SPELLING CORRECTION – AN EXPERIMENT IN PROGRAM DESIGN; Springer-Verlag, Berlin, 1980</p> | <p><ROJP-88> Tancig P. (ur.): Zbornik 4. Konference »RAČUNALNIŠKA OBDELAVA JEZIKOVNIH PODATKOV«, ROJP-4, Portorož, oktober 1988</p> <p><Sch-88> Schubert K.: »Modularity in Linguistic System Design – Experiences from the DLT Machine Translation Project«; v <ROJP-88></p> <p><Tan-82> Tancig P.: »SOVA – General Purpose Software Environment for NLU Systems«; v Trappl R. (ed.): CYBERNETICS AND SYSTEM RESEARCH; North-Holland, Amsterdam, 1982</p> <p><Tan-86> Tancig P.: »Računalniško razumevanje naravnega jezika – pregled, primerjava, ocena«; IJS Delovno poročilo, DP-4536, 1986</p> <p><TanVit-88> Tancig P., Vitas D.: »Računalniško razumevanje jezika – aplikacije in produktivnost«; posebna izdaja »Infomatika i produktivnost« revije Tehnika, 1988</p> <p><TC-79> B. M. Snell (ed.): TRANSLATING AND THE COMPUTER; North-Holland, Amsterdam, 1979</p> <p><VitTan-88a> Vitas D., Tancig P.: »Okvir za definisanje YUTRA – jugoslovenskog sistema za avtomatsko prevođenje«; XXXII konf. ETAN-a, knjiga XII, Sarajevo, 1988</p> <p><VitTan-88b> Vitas D., Tancig P.: »Skice za izgradnju integrisanog ambijenta za obradu tekstuelnih informacija«; v <ROJP-88></p> |
|---|--|