

PRESEK

List za mlade matematike, fizike, astronome in računalnikarje

ISSN 0351-6652

Letnik 2 (1974/1975)

Številka 1

Strani 28-31

Denis Poniž:

RAČUNALNIKI, JEZIKOSLOVJE IN LITERATURA

Ključne besede: matematika.

Elektronska verzija: <http://www.presek.si/2/2-1-Poniz.pdf>

© 1974 Društvo matematikov, fizikov in astronomov Slovenije

© 2010 DMFA - založništvo

Vse pravice pridržane. Razmnoževanje ali reproduciranje celote ali posameznih delov brez poprejšnjega dovoljenja založnika ni dovoljeno.

4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0

RAČUNALNIKI, JEZIKOSLOVJE IN LITERATURA

Kot obrabljena fraza zveni podatek, da danes ni več področja človekovega udejstvovanja, kjer ne bi bila dobrodošla uporaba računalnikov. Vendar to ne velja samo za t.im. eksaktne vede (fiziko, kemijo, meteorologijo, astronomijo itd.), marveč tudi za humanistične znanosti (sociologijo, zgodovino, literarno znanost itd.). Tudi na teh področjih so računalniki koristen in vsestransko uporaben pripomoček. Poskušali bomo na kratko opisati, kako

se je uporaba računalniških metod razvijala na področju jezikoslovja in znanosti o književnosti in kako lahko z računalniki raziskujemo in opisujemo tisti material, ki pomeni osnovo vsakega pisanega sporočila, jezik v vseh njegovih razsežnostih. Prav z uporabo računalniških metod so se zelo zblížala področja jezikoslovja, literarne teorije in literarne zgodovine. Vsa tri področja namreč potrebujejo podobne podatke, seveda pa jih vsako področje opisuje in uporablja na drugačen način.

Eno izmed področij, kjer so najprej začeli uporabljati računalnike, je t.im. *strojno prevajanje*. Ker vemo, koliko različnih knjig, člankov, poročil in časopisov izide na svetu vsak dan v množici različnih jezikov, je prenos sporočil, seveda pa tudi njihove vsebine, omejen na določeno jezikovno področje. Vsako širjenje sporočila preko teh meja je povezano s prevajanjem. Vendar pa ima prevajanje, ki ga opravljajo ljudje, več pomanjkljivosti: hitrost prevajanja je močno omejena, prevajalec mora poznati strokovne izraze za posamezna področja, pri prevajanju prevoda lahko nastanejo hude pomenske napake itd. Zato so menili, da bi delo prevajalcev lahko opravljali računski stroji - hitro, brez napak, kadarkoli in iz kateregakoli jezika v drug jezik. Vendar pa se je pri prvih poskusih pokazalo, da ni mogoče prevajati iz jezika v jezik, ne da bi naleteli na hude težave. Tu ne mislimo samo na težave, ki so nastale pri prevajanju enakozvočnic (homonimov) ali soznačnic (sinonimov), marveč tudi na vse pomanjkljivosti in napake, ki so nastale zaradi premajhnega poznavanja slovnične zgradbe, stavčne zgradbe in besednih značilnosti. Prevajanje *beseda za besedo*, t.j. računalnik skuša prirediti vsaki besedi iz jezika, iz katerega prevaja, ustrezno besedo v prevajanem jeziku, se je izkazalo za neuspešno. Zato je bilo treba začeti znova, vendar po drugi poti.

Pri teh poskusih se je izkazalo dvoje. Jezik je bolj zapleten mehanizem, kot so predvidevali. Poleg tega pa v vsakdanjem govoru in pri pisanju ne uporabljamo *idealnega* jezika, torej takšnega, ki bi bil zgrajen na podlagi *čistih* slovničnih zakonitosti, brez izjem ali posebnosti. Jezik, ki ga govorimo ali pišemo, je sestavljen v veliki večini iz izjem. Zato je nesmiselno pričakovati, da bi računalnik prevajal brez napak, če vemo, da je programiran tako, da upošteva predvsem slovnična pravila in šele nato tudi iz-

jeme. Ker pa so spomini danes uporabljanih računalnikov premajhni, da bi lahko vanje programirali vse izjeme, ki jih pozna neki jezik (n.pr. slovenščina), še ni mogoče govoriti o uporabnem računalniškem prevajanju.

Obenem z uporabo računalnikov na različnih področjih pa se je pričela razvijati posebna smer jezikoslovja, matematično jezikoslovje (matematična lingvistika). Matematična lingvistika, ki je posebej razvita v Sovjetski zvezi, na češkem, v Nemčiji in Združenih državah, poskuša v jezikih odkriti z matematičnimi metodami oblikovna in vsebinska pravila. Matematična lingvistika uporablja tudi statistične metode, saj je v jeziku mogoče meriti poprečne vrednosti mnogih količin: frekvenco posameznih črk in glasov, zlogov, besed, dolžine zlogov, besed in stavkov, frekvence posameznih besednih vrst (samostalnik, pridevnik, glagol, števnik, zaimek, veznik, medmet). Vse te količine govorijo o takšni ali drugačni naravi posameznih naravnih jezikov.

Vsi ti podatki pa so zanimivi tudi za tiste, ki preučujejo književnost. Stil pisatelja, pesnika ali esejista je v marsičem odvisen od teh izmerljivih in določljivih količin. Čimveč količin poznamo, lažje in zanesljiveje lahko določimo stil nekega pisca, njegovo izrabo jezikovnih sredstev. To je pomembno za ugotavljanje avtorstva nepodpisanih besedil: če poznamo podatke za znana dela nekega pisca, potem lahko te podatke primerjamo s tistimi, ki smo jih dobili pri obdelavi neznanega besedila. Če dobimo podobne vrednosti, je zelo verjetno, da je znani pisec tudi avtor nepodpisanega teksta. Določneje: če pesnik v vseh svojih pesmih uporablja besede s poprečno dolžino 4,95 črke na besedo, potem je jasno, da je majhna verjetnost, da bo napisal pesem, kjer bodo imele besede poprečno dolžino 5,5 črke na besedo. Že majhna odstopanja navzgor ali navzdol so zelo pomembna in jih ne smemo zanemariti.

Računalniki pa so za znanost o književnosti pomembni še iz drugih razlogov. Z računalniki lahko obdelamo velike količine podatkov, jih medsebojno primerjamo, jih na željo izpisujemo ali shranjujemo na magnetni trak. Če je bilo treba nekoč ročno opravljati zamudna izpisovanja določenih besednih vrst, danes to dela računalnik: precizneje, hitreje, podatki so uporabnejši, preglednejši in zato dostopnejši. Tako nastajajo banke podatkov tudi za

področje književnosti: posebne ustanove, inštituti za jezikoslovje ali književnost, hranijo dobljene in obdelane podatke na magnetnih trakovih, tako so vsak trenutek dostopni za nadaljne raziskave. Povejmo še, da smo nekatere podatke izračunali tudi za slovenski jezik! Tako n.pr. poznamo razvrstitev črk po njihovi pogostnosti. V slovenskem jeziku je najpogostejša črka *e*, sledijo ji *a*, *i*, *o* in *n*. Najredkejše črke v slovenskem jeziku so *h*, *š*, *c*, *ž* in *f*. Črka *f* se pojavi v slovenskem besedilu poprečno samo enkrat na tisoč črk! Najpogostejša črka (*e*) pa se na tisoč črk pojavi več kot devetdesetkrat! Torej je na približno deset črk vsaj ena *e*. To lahko vsakdo poskusi na poljubnem slovenskem besedilu.

Na koncu moramo omeniti še eno področje uporabe računalnikov. Povedali smo že, da so uspeli izračunati celo vrsto podatkov o naravi posameznih jezikov, obenem pa so uspeli izračunati tudi nekatere slovnične zakonitosti. Da bi jezik še bolje spoznali, poskušajo danes s pomočjo računalnikov pisati t.im. sintetične (umetne) tekste. Le-ti nastajajo tako, da uporabijo določeno, močno poenostavljeno slovnično zgradbo (n.pr. priredni stavek, ki ima osebek, povedek in predmet) in omejen besedni zaklad. Računalnik potem izpiše vse možne stavke. Na podlagi dobljenih podatkov potem sklepajo o naravi posameznih jezikovnih zgradb, o slovničnih zakonitostih, pogostosti in uporabnosti posameznih besed, pomen-skem bogastvu, ki ga je mogoče še ohraniti pri okrnjeni slovnični strukturi. Včasih so lahko ti izpisi podobni modernističnim pesmim. Odtod tudi vesti, da *računalniki pišejo poezijo*. To dvakrat ni res. Prvič računalnik ne more pisati poezije, saj računalnik ne misli in ne čustvuje. Drugič pa takšni poskusi nimajo namena ustvarjati pesmi, marveč rabijo raziskovanju jezikovne zgradbe in določanju obsežnosti besednega zaklada.

Na področju jezikoslovja in književnosti so računalniki postali nepogrešljiv pripomoček. Na tem mestu ni mogoče opisati vseh možnosti, ki jih bo uresničila bodočnost. Verjetno pa mednje sodijo računalniško prevajanje (predvsem strokovnih tekstov), učenje tujih jezikov in shranjevanje vseh mogočih podatkov o literarnih delih.