

AMERIKA A GÉPI NYELVÉSZET SZEMSZÖGÉBŐL

Kiefer Ferenc

MTA Számítástechnikai Központ

1. Amerikai tanulmányutunk során a modern nyelvészet /matematikai nyelvészet/ elméleti és gyakorlati kérdéseit tanulmányoztuk. E beszámolóban a gyakorlati kérdésekre szorítok, vagyis arra, hogy a számítógépet milyen területen használják a nyelvtudományban és milyen eredménnyel. A matematikai nyelvészetnek ezt a gyakorlatibb részét számítógépes nyelvészetnek /computational linguistics/ is nevezik, de lehetne alkalmazott nyelvészetnek is /applied linguistics/, bár az angol terminus, - legalább is Amerikában - most jelent meg. /Elsősorban a nyelvtanítás problémáit öleli fel./

A gépi nyelvészet feladata, problémaköre éppúgy nem tisztázott még, mint a matematikai nyelvészeté. Minden kutatónak saját véleménye van e fogalmakról, ami a kutatás mai szintjén jogosult is. Általában azonban megegyezik a kutatók véleménye abban, hogy a számítógép felhasználása a nyelvtudományban hasznos, kifizetődő és majdnem elengedhetetlenül szükséges is. Abban azonban már eltérnek a vélemények, hogy a számítógépes nyelvészet önálló tudományágnak tekinthető-e, vagy pedig a számítógép alkalmazási lehetőségei közül egy terület, amelynek azonban nincsenek meg a külön, sajátos módszerei. Más szóval, a formális nyelvek vizsgálatára kidolgozott módszereknek alkalmazásáról van szó formális nyelvek helyett természetes nyelvekre. A módszer ugyanaz, csupán az eredmény lesz különböző. Természetes nyelvek esetén azonban nem várhatunk pontos eredményeket, hiszen az emberi nyelv nem formális rendszer. Abban szintén különböznek a vélemények, hogy a számítógépes nyelvészet érdekes-e a nyelvtudomány szempontjából, azaz, hogy vezet-e olyan eredményekre, amelyek eddig ismeretlen összefüggésekre világítanak rá, vagy pontosan eddig meg nem fogalmazott törvényszerűségekre vetnek fényt. Kétségtelen, hogy a számítógép sok adatot szolgáltat a nyelvészetnek, de az összefüggéseket a nyelvész állapítja meg és nem a gép. Másrészt a számítógépes nyelvészetnek - legalább is jelenleg - durva közelítésekkel kell dolgoznia, amennyiben gazdaságos kíván lenni. Ezek miatt a legtöbb amerikai nyelvész nem érdekli közvetlenül a számítógépes nyelvészet, de örömmel használja fel a gép munkáját, amikor csak teheti. Azt természetesen mindenki elismeri, hogy az ötvenes évek elején, amikor a számítógépes nyelvészet /elsősorban a gépi fordítás/ kezdetét vette és nyelvészek, matematikusok egyaránt

érdeklődéssel fordultak az új problémák felé, a matematikai nyelvészet elképzelhetetlen volt számítógépes nyelvészet nélkül. Pillanatnyilag tehát úgy látszik, hogy a tulzott bizalom a számológépben csak a mi országainkban él még, ami viszont valószínűleg elsősorban a megfelelő tapasztalat hiányával magyarázható. Tudvalevő ugyanis, hogy elgondolásainkat vagy egyáltalán nem valósítottuk meg eddig a gépeken, vagy pedig csak kísérletképpen, igen kis anyagon, ami még tulajdonképpen semmiféle következtetés levonására nem jogosít.

Az előbb említett kérdések tehát még tisztázatlanok és egyelőre mind a számológépeknek /elsősorban a perifériális berendezéseknek, mint pl. az automatikus olvasónak/, mind pedig a természetes nyelvek matematikai elméletének további fejlődésére van szükség ahhoz, hogy a lehetőségeket és korlátokat világosan lássuk.

Mivel valószínűleg a vélemények megegyeznek abban, hogy a számológépek hasznosak a nyelvtudomány számára és hozzátehetnénk sok más terület, így pl. a dokumentáció kutatás számára, a következőkben röviden áttekintjük a legfontosabb területeket, ahol eddig számottevő eredmények születtek.

2. A különböző gépi nyelvészeti munkák középpontjában a lexicológiai munkák állnak, azaz különböző rendeltetésű egynyelvű szótárak gépi feldolgozása. Ilyenek az a tergo szótárak /a szavak fordított, "hátról előre" való megadása/, amelyek sok szempontból alapját képezik a szintaktikai jellegű kutatásoknak. A hagyományosabb nyelvész számára is sok érdekes adatot szolgáltatnak a végződésekre vonatkozóan. Ugyanazzal az anyaggal a gyakorisági szótár is elkészíthető, aminek ismét sok gyakorlati felhasználási lehetősége van. Így például elengedhetetlenül szükséges a gépi kivonatoláshoz, de fontos szempontokat szolgáltat a nyelvtanításhoz is. A morfémátárak /a szó helyett annak önálló alkotórészeit tartalmazó "tárak"/ is fontos és érdekes információt nyújthatnak a nyelvészeknek. Az értelmező szótárak gépi feldolgozása hozzásegíthet egy jelentéstani kategória-rendszer kidolgozásához. Műszaki vonatkozásban fontosak a szakszók glosszáriumok formájában történő feldolgozása. Sok helyen a különböző szakterületek szavait külön tárolják a gép memoriájában. Vannak azonban olyan kísérletek is, amelyek egyetlen szótárban egyesítik a különböző szakterületek szavait megfelelő kódokkal. Ilyen esetben az a probléma, hogy valamely szakterület esetén, megfelelő kódok alapján, a glosszáriumot a gép készítse el. Gépi módszerekkel könnyebbnek látszik a neologizmusok /új szavak/ felvétele a szótárba, mint emberi gyűjtőmunkával. Különböző időben kiadott szótárak összehasonlítása a szókészlet fejlődésére vethet fényt. Írói szótárak is nagy számban készülnek. Ami a gép számára könnyű feladat, ezekben az esetekben több ember valószínűleg hosszú évtizedeken át tartó munkáját venné igénybe. Statisztikai és gyakorlati /gépi nyelvészeti/ kutatások számára egyaránt fontosak az ún. konkordancia-szótárak. Hogy mennyire fontosak, azt mi sem bizonyítja jobban, mint az, hogy az egyik amerikai egyetem /Georgetown University, Washington/ jóformán kizárólag konkordancia-szótárak információira támaszkodva tud végezni sikeresnek mondható gépi fordítást.

3. A gépi nyelvészet egy másik nagy területe a statisztikai nyelvészet alkalmazási köréhez tartozik. Már az előző pontban említettük a gyakorisági szótárakat. A gyakorisági szótárakon kívül még a nyelvnek sok statisztikai jellegű tulajdonsága vizsgálható géppel. A szóstatisztika mellett fontosak lehetnek a morféma és fonéma /nagyjából: hang/ statisztikák és talán főleg statisztikai kutatások számára a fonémakombinációk /un. bi- és trigrammok, nagyjából: két, ill. három hang egymásutánja/ vizsgálata. A statisztikai jellegű vizsgálatokhoz tartoznak az un. redundancia vizsgálatok is, amelyek érdekes tényeket fedhetnek fel a nyelv, mint információk közlésének eszköze szempontjából. Szövegek különböző statisztikai paraméterek szerinti vizsgálata több területen hasznosítható. A statisztikai szemponton kívül itt megemlíthetnénk ismeretlen szerzőségű kéziratok szerzőségének valószínűsítését, továbbá ismeretlen, megfejtetlen írások megfejtésének könnyítését szolgáló adatok szolgáltatását. Mindezek a területek a nyelvtudomány számára nem egészen újak, hiszen különböző vonatkozású statisztikai jellegű vizsgálatokat már a múlt század vége óta végeznek. Ezek azonban többnyire kézi erővel készültek. A gép azonban nemcsak olcsóbb, hanem sokkal gyorsabb is. Másrészt a régebben kapott adatok sokszor megbízhatatlanként voltak /statisztikai szempontból/, mert a feldolgozott anyag nem volt elég nagy, vagy más szempontból nem tett eleget a statisztikai mintavétel követelményeinek. Újszerű lenne szintaktikai jellegű statisztikák végzése. Bár valószínűleg mindenki elismeri ezek fontosságát, egyelőre e feladat megvalósíthatatlan, megfelelő szintaktikai analízis hiányában.

4. Ami a fordítás gépesítését illeti, nagyjából következő a helyzet:

Gyakorlati fordítás gép segítségével több helyen készül, egy-egy szűk szakterületen /pl. szerveskémia, atomfizika, halmazelmélet/, a fordítások nyelvileg rosszak, de szakmailag érthetők. A fordítási algoritmusok egyrészt a már említett konkordancia-szótárakra támaszkodnak, másrészt elő- és utószerkesztőt is felhasználnak. Ilyen közelítő fordítás gazdaságilag is kifizetődő, elsősorban azért, mert nagy a hiány emberi fordítóban. Másrészt nem szabad elfelejteni azt sem, hogy a konkordancia-szótárak, továbbá a megfelelő algoritmusok elkészítése az ötvenes években nagy pénzt emésztett fel, és ma már csak telex /lyukasztási/ munkára és gépi időre van költség.

Több helyen fordítás helyett a gép csak szótár, azaz a kívánt szakterületnek megfelelően kiírja és lefordítja az adott szakszöveg szavait. Az emberi fordítónak így a szakszavak, a szövegben való előfordulásuk szerinti sorrendben, állnak rendelkezésére. Az ilyen, ember és gép együttműködésen alapuló fordítás elkészítése a tapasztalatok alapján legalább 50 %-os megtakarítást jelent, azonkívül pedig az előkészítő munkák is sokkal kevesebb anyagi megterhelést jelentenek.

Természetesen több helyen folynak kutatások, amelyek majd a jelenleg működő fordítói algoritmusok korrekciójához vezethetnek.

Ezeket a kutatásokat "gépi szintaktikai analízis" terminusszal szokás jelölni. Egyelőre minden kétséget kizáróan bebizonyosodott /elsősorban a matematikai nyelvészet elméleti eredményeinek alapján/, hogy az eddigi elemzési módszerek /és ide nemcsak az Egyesült Államokban használt módszerek tartoznak/ nem vezethetnek kívánt eredményekhez. Más szóval, az ezeken alapuló fordítás nem lehet lényegesen jobb az eddiginél. Kétségtelennek látszik, hogy az eddigi módszerek finomítása helyett minőségileg új módszerre van szükség, amit több kutató csoport a transzformációs grammatikában^{1/} vél megtalálni. E területen sok érdekes eredmény született, de még korai lenne megmondani, hogy a módszer beváltja-e a hozzáfűzött reményeket. Egyelőre mindenesetre úgy látszik, hogy hosszú évek kutatása kell ahhoz, hogy a kezdés eldönthető legyen.

5. A gépi nyelvészethez sok olyan terület is kapcsolódik, amelynek közvetlen haszna a dokumentáció kutatásban mérhető le. Jellemző, hogy sok olyan kutatócsoport, amely régebben kizárólag gépi fordítással foglalkozott, áttért a dokumentáció kutatásra, elsősorban talán azért, mert az előző területen már anyagi támogatást nem kapott. A dokumentáció kutatás szempontjából érdekes gépi nyelvészeti területek a következők:

a/ Gépi kivonatolás. E téren H.P.LUHN óta sok minden történt, a releváns mondatokat gépi úton mind meg tudják találni, de a kivonat még tulságosan bő, azaz a kivonat nagyon sok nem releváns mondatot is tartalmaz. A jelenlegi kutatások e fölösleges mondatok kizárására vonatkoznak. A feladat nem látszik könnyűnek, ezért a gépi kivonatolás esetén is csak kísérleti stádiumról beszélhetünk.

b/ Általános probléma, amely a nyelvtudományban is jelentkezik, az információk kicserélésének meggyorsítása. Gondolunk itt például a nyelvészeti bibliográfiák kiadására, amelyek általában kétévi késéssel látnak napvilágot. Bár e kérdéssel külön cikkben kívánunk majd foglalkozni, hadd jegyezzük meg itt is, hogy egyrészt az adatok /bibliográfiai kártyák/ egybegyűjtésének meggyorsításáról van szó, másrészt a nyomdai folyamatáról, ami lehetségessé válik a számológép által végzett kompozíció és sokszorosítás esetén. E témakörhöz kapcsolódik a decimális osztályozás nyelvtudományra vonatkozó részének felülvizsgálása és új osztályozási rendszer kidolgozása.

c/ Az előbbi, elsősorban gyakorlatinak látszó munka két területen kapcsolódik /persze a gyakorlatban közvetlenül hasznosítható/ elméleti kutatásokhoz. Az egyik az osztályozás problémáival foglalkozik, a másik az információ tárolás és visszakeresés kérdése. Természetesen ez a két kérdés is szorosan összefügg. E rövid beszámoló keretei nem engedik meg, hogy ezekkel a témákkal részletesebben foglalkozzunk, de az érdekes eredmények miatt, mindegyiknek, - amelyeket láttunk - külön-külön cikket szánunk.

^{1/} Vö. KIEFER F.: Matematikai nyelvészeti tanulmányok. Bp. OMKEK, 1965. 179 p.

d/ Külön hangsúlyozni kell, hogy már a 2-4. pontban említett területek eredményei is hasznosíthatók a dokumentáció kutatásban. Maga a gépi fordítás is főleg a dokumentációs szakembereknek jelent majd az első időben hasznót.

6. Röviden megemlíthetjük a nyelvtanítás "gépesítését" is. Nemcsak arról van szó, hogy ember helyett gép végezze a nyelvtanulás kezdeti feladatait, hanem arról is, hogy a nyelvtanár "algoritmus" formájában, mintegy imitálva a gépet, tanítsanak idegen nyelvet. A kezdeti sikerek és az újszerűség vonzóereje azonban nem elégséges ahhoz, hogy végleges eredményekről beszélhessünk. Tisztán kell látnunk, hogy e téren is egyelőre csak kísérletekről beszélhetünk.

7. E beszámolóban nem törekedtünk teljességre abban az értelemben sem, hogy nem soroltuk fel a gépi nyelvészet minden területét. Beszélhettünk volna még a beszédfelismerési kísérletekről, sok érdekes pszicholingvisztikai kísérletről és sok minden egyébről. Mi a legfontosabb témákat kívántuk kiemelni és ezek helyzetét röviden vázolni. A gazdag /azt is mondhatjuk, minden túlzás nélkül, hogy e téren a leggazdagabb/ tapasztalatokkal rendelkező amerikai kutatók, kutatócsoportok véleményét, állásfoglalását érdemes figyelembe venni a mi lényegesen kisebb méretű kutatási programunkban is. Tapasztalatuk elsősorban arra tanít bennünket, hogy a számítógép szerepét a nyelvtudomány szempontjából, vagy a számítógépes nyelvészet szerepét a dokumentáció kutatás szempontjából helyesen tudjuk értékelni. Helytelen lenne, ha e szerepet lebecsülnénk, de az is helytelen és káros, ha azt tapasztalatlanságunkból fakadó optimizmusból túlbecsülnénk.

.oOo.

KIEFER, F.: Research in mechanical linguistics in the United States

The article reviews the research work conducted on mechanical linguistics in the United States. It covers the most important fields of linguistics: lexicology, statistical linguistics, mechanical translation, and linguistic problems connected with research in documentation. A brief survey and evaluation of the results achieved in these fields are given.

UUU

КИФЕР, Ф.: Исследование по механическому языкознанию в США.

Статья обсуждает научно-исследовательские работы по механическому языкознанию в США, охватывающие важнейшие отрасли языкознания, в том числе лексикологию, статистическое языкознание, механические переводы и лингвистические проблемы, связанные с исследованием в области документации. Даются сводка и оценка достигнутых в этих областях результатов.

□ □ □

KIEFER, F.: Amerika vom Gesichtspunkt der mechanischen Linguistik

Es wird eine Übersicht der amerikanischen Forschungen auf dem Gebiete der mechanischen Linguistik gegeben. Die wichtigsten Gebiete dieser neuen Disziplin werden der Reihe nach behandelt: die lexikologischen Forschungen, die Probleme der statistischen Linguistik, die Fragen der mechanischen Übersetzungen und die mit der Forschung auf dem Gebiete der Dokumentation verbundenen linguistische Probleme. Die erzielten Ergebnisse werden kurz geschildert und bewertet.

o^oo