

## A gépi fordítás időszerű kérdései

A gépi fordítás gépi szótározásként indult. Az alapot ehhez a jelenleg elcsodálkoztató teljesítményekre képes elektónikus számológépeknek mai szemmel primitívnek nevezhető első szerkezetei adták, az ötletet elsőnek egy számológépes és egy desifrázó szakember közösen fogalmazta meg. A. D. BOOTH és W. WEAVER tudatában lévén az elektronikus számológép óriási lehetőségeinek, a szótározáson túlmenően komoly lehetőségnek tartották a használható fordítás megvalósítását is. A gondolat rövid kifejezését problémafelvetésként tartalmazó memorandumukra eltérő válaszok érkeztek. Többen voltak olyanok, akik a gépi fordítást elérhetőnek tartották.

### 1. A szavankénti fordítás

A gépi fordítás gyakorlati megvalósítása jellegzetes nyelvészeti feladat. Az első próbálkozások még a gépi szótározás elgondolására épültek és a szavankénti fordítás útján keresték a megoldást. Bár már kezdetben sikerült megvalósítani az egy szón belüli morfológiai elemzés bizonyos formáit, a szavak elszigetelt fordítása már előre is megjósolt súlyos szemantikai jellegű kérdéseket vetett föl. Egy-egy szó v. morféma a nyelvi elemek természetes többértelműsége révén a fordításban egy mondaton belül általában több szót ad, ami nemcsak zavarja az olvasót, hanem a mondat megértését is lehetetlenné teszi. A fordítás minőségét úgy lehetett javítani, hogy igyekeztek megszüntetni vagy legalábbis erősen lecsökkenteni a többértelműségek számát. A szavankénti fordítás szintjén a többértelműség határozottan szétválasztható grammatikai és tisztán szemantikai jellegű többértelműségeire annak alapján, hogy az elszigetelt morfémát grammatikai vagy lexikai elemnek tekintjük. A grammatikai többértelműség a közvetlen környezet vizsgálatával, a szemantikai többértelműség az egész szöveg jellegének szem előtt tartásával erősen leszűkíthető. A szavankénti fordítás javított módszerei ezt veszik alapul, amikor a megadható jelentést a vizsgált morféma előtti és mögötti elemek grammatikai tulajdonságainak figyelembe vételével, ill. a „mikroszótárban” található értelmezés felhasználásával tüntetik fel. (A „mikroszótár” az erősen leszűkített szakterületre érvényes jelentéseket tartalmazza csak).

A szavankénti fordítás használhatóságának, ill. alkalmazhatóságának az is erősen határt szab, hogy csak két olyan nyelv közti fordításra lehet használni, melyek szórendje nem tér el jelentősen egymástól. Abban az esetben, ha a fordítás célnyelvének szórendje a fordítás forrásnyelvének a szórendjétől erősen eltér, és az előbbit csak a szavak mondattani szerepének meghatározása alapján tudjuk megállapítani, akkor a szavankénti fordítás módszere nem használható. Ilyenkor (pl. az oroszról magyarra történő fordítás esetében is) a forrás-

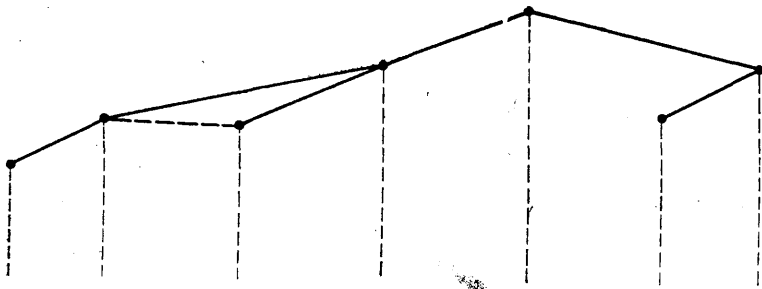
nyelvi szöveg szavainak vagy szókönyezetének elemzése helyett csak a mondattani elemzés vezethet célhoz.

## 2. Mondatelemzés a gépi fordításban

A mondattani elemzéstől a gépi fordítás szakemberei sokat vártak, és a kezdeti eredmények eléggé ígéretesek is voltak. A technika fejlődésével az elektronikus számológépek emlékezeti egységeinek kapacitása jelentősen bővült, és a gépi szótárba így az egyes szavak mellé grammatikai információkat is tárolni lehetett. Ugyanakkor a gépek működése is gyorsult, a program terjedelmesebb és részletesebb utasításokat tartalmazhatott.

A szintaktikai elemzés alkalmazása nem oldotta meg a gépi fordítás kérdéseit. Sőt, a nyelv szemantikai tulajdonságainak olyan jellegzetességeit fedte föl, amelyeket a korábbi elemzés nem hozott felszínre.

A beszélő és értő számára kétségtelen, hogy a mondat minden szava (a különleges közbevetésektől eltekintve) egy egységes kapcsolatrendszerben válik egységgé, melyben a kapcsolatrendszer csúcsán álló elemtől a már függő taggal nem rendelkező elemekig egy elemzésen belül csak egyetlen függőségi vonal vezet:



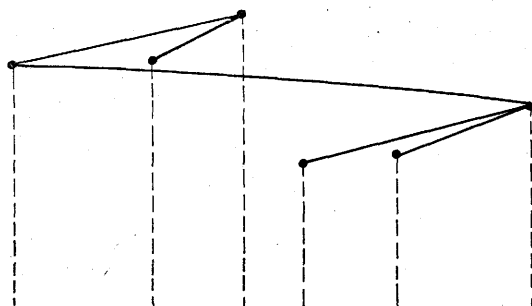
1. ábra

A mondat függőségi ágrajzán nincsenek zárt szemek: Zárt szem akkor keletkezne, ha a mondat egyik függő tagja nem egyetlen főrendelt szótól függne, hanem kettőtől pl.: ha a „függőségi” szó a „mondat” és „ágrajzán” szavaknak egyszerre függő tagja lenne.

A formális elemzés ezt a zárt rendszert a mondatok nagy számában sajnos nem tudja hibátlanul megalkotni, és a mondat szintaktikai elemzési módjai közül egyiknek sem sikerült még olyan utasításrendszert adnia, amely kielégítő módon teljes mondatelemzést nyújtott volna.

A formális elemzésben a mondat csúcsán álló független tagnak az alanyt az állítmányt vagy mindkét tagot egyszerre is lehetne venni. A függőségi grammatika az állítmányt veszi egyedüli független elemnek, mert így tudja az ágrajzra az egyébként más következtetések levonására szolgáló projektív függés szabályát érvényesíteni. E szabály szerint a függőségi rajz elágazási pontjaiból húzott függőlegesek nem metszik a függőséget jelző vonalakat. Ha a mondat főtagjául az alanyt vesszük, úgy ez a szabály sok mondatra nem lenne érvényes.

Pl.:



2. ábra

Nincsenek zárt szemek a függőségi rajzon

A mondaton belüli összefüggéseknek a függőségi grammatika által megrajzolt képe nem hagyja figyelmen kívül azokat az eseteket, amelyekben a mondat csúcsán álló elemtől egynél több „út” is vezethet a legelső elemekig. Az ilyen, grammatikailag többértelmű mondatokat a formális elemzéssel is természetesen többértelműnek kell értékelni. (A magyarban pl.: *Megsebezte a villám*; az angolban: *They are flying planes*.) A gépi fordítás számára mégsem az ilyen mondatok jelentik a nehézséget, hanem azok, amelyek a beszélő és értő számára egyértelműek, és amelyeknek az elemzése az ember számára nem okoz különösebb nehézséget. Az ilyen mondatokra a gép, — ha az elemzést az adható formális utasítások alapján mégis végrehajtja —, nem egyetlen, hanem több önmagában zárt és egységes szintaktikai szerkezetet fog adni, vagyis a mondatokat többértelműeknek tartja.

A szintaktikai elemzés során nyert többértelműség más jellegű, mint az, amelyet a szavankénti fordítás esetében kapunk. Az első pillanatra grammatikai többértelműségnek tűnik, hiszen a mondat szerkezetre vonatkozik és a grammatikai elemzés során kapjuk. A grammatikai jelleg itt mégis teljesen más vonásokat mutat, mint a szavankénti fordításban. Itt ugyanis nem elszigetelt morféimakkal állunk szemben és nem azt kell eldöntenünk, hogy ezek a morféimák grammatikaiak-e, és ha igen, a többféle jelentés közül melyik az, amelyik a környezet alapján az egyetlennek tekinthető, hanem szóértékű morféimák egymás közötti kapcsolatáról van szó, ami egyrészt a kapcsolatban résztvevő elemek grammatikai tulajdonságai alapján jön létre, de eddigi formális módszereink alapján nem vezethetők le teljesen ezekből a tulajdonságokból. Ezek a kapcsolatok valójában szemantikai jellegűek is, és megoldásukra csak akkor számíthatunk, ha a gépi elemzésbe a szemantikát is sikerült belevonni.

A gépi elemzés számára legnagyobb akadályt jelentő mondat típusok:

$$N_{at} \quad V_{tr} \quad N_{at}$$

ahol:  $N_a$  — alanyesetű,  
 $N_t$  — tárgyesetű,  
 $N_{at}$  — olyan főnév, mely a formai ismérvek alapján alanyesetben és tárgyesetben állónak egyformán fölfogható;  
 $V_{tr}$  — tranzitív ige.

A magyarban ilyen mondatok kissé szokatlan formájúak és furcsán hatnak, de a szemléltetés kedvéért mégis adhatók:

*A bátyám olvassa a könyvem.  
A bátyám szórakoztatja a könyvem.  
Elismerésem váltja ki a könyved.*

A strukturális grammatika elgondolásai alapján az ilyen típusú mondatokat csak akkor lehet formális alapon eredményesen elemezni, ha a szóosztályokat megfelelő módon alosztályokra bontjuk, hogy azok segítségével megkülönböztethessük az alanyi és tárgyi viszonyban álló főneveket. A fenti első és második mondatra aránylag könnyen találhatunk főnévi alosztályokat, amelyek hozzásegítenek a megoldáshoz. Még a harmadik mondat két főnévre is találhatunk megkülönböztető alosztályokat, de nem mondható ugyanez az alábbi mondatokra:

*Elismerésem váltotta ki az ötleted.  
Problémám váltotta ki az ötleted.*

Ezekben a mondatokban az „ötleted” szó azonos ige mellett egyszer a mondat alanya, egyszer pedig a mondat tárgya.

A gépi fordítás számára problémát jelentő mondatok másik típusát inkább idegen nyelven lehet szemléltetni. A mondat típus képlete:

$$N_a V_z N_x Pr N_y$$

és itt  $V_z$  olyan ige, amely mellett nemcsak egy  $N_x$ , hanem egy  $Pr N_y$  főnév is állhat, de az utóbbi nem kötelező vonzatként. Ugyanakkor ez a  $Pr, N_y$  a közvetlenül előtte álló fánévnek is lehet jelzője. Pl.

*Я сижу в комнате со своими друзьями.  
Я сижу в комнате с большими окнами.*

Ezekben a mondatokban az alany és állítmány teljes azonossága mellett még egy formailag azonos szószerkezet van, mely a két mondatban teljesen eltérően viselkedik. A beszélő számára mindkét mondat egyértelmű, a gép számára ezek a mondatok kétértelműek.

Az ilyen határozói jelzők formális elmezésének nehézségére már BRANDWOOD rámutatott a gépi fordításnak egészen a kezdetén, de csak a legutóbbi időkben jelentek meg olyan írások, melyek a jelenség grammatikai elemzésének lehetőségeit kimerítően tárgyalják.

A határozószói alakó jelzők elemzésére a formális elemzés a szókategóriák alosztályain kívül felhasználja az igei és főnévi vonzatok különböző fajait, a projektív függéssel adott szórendi szabályokat, állandó jellegű és egy meghatározott tárgykorú szövegben állandónak vehető szókapcsolatok jegyzékét és táblázatát, a mondat tényleges megformálásában föllelhető következetességeket, valamint a formális függés egyik végső eszközét, a disztribúciós elosztás statisztikai adatait is. A transzformációs analízis, mely a strukturális grammatikának egyik fontos formális eszköze, a gépi elemzésben nem használható.

Mivel a jelenleg a gépi fordítás rendelkezésére álló elemzési módokkal nem lehet minden mondatot hibátlanul elemezni, a jelzett módon többértelmű mondatok elemzésével foglalkozó cikkek nélkülözhetetlenek tartják, hogy

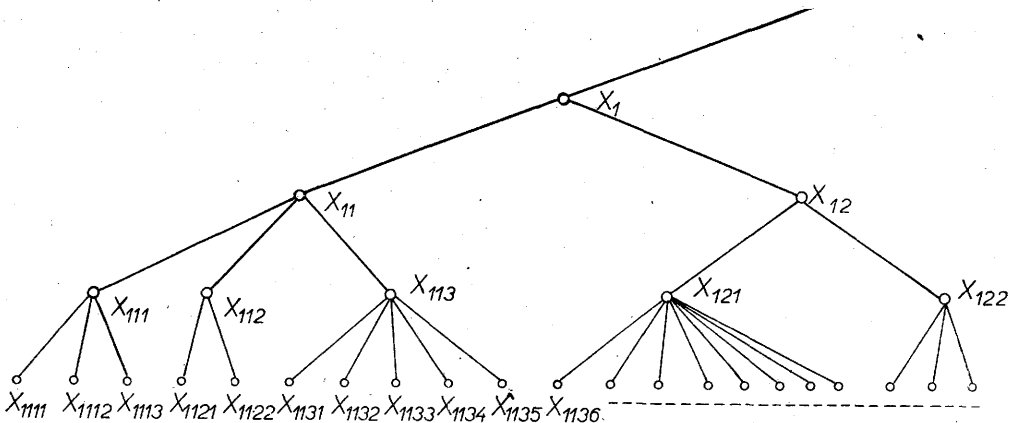
ennek a grammatikai kérdésnek megoldásához föl kell használni a szemantikai kapcsolatokat is.

### 3. Szemantikai rendszerek a gépi fordításban

Az utóbbi években többen próbálkoztak olyan szemantikai rendszert megalkotni, mely a gépi fordításban is eredményesen felhasználható. Ezek a kísérletek általában nem nyúltak vissza az általános nyelvészetben már meglévő szemantikaelméletekhez, hanem inkább a szótárírás gyakorlatára támaszkodva a hasonló jelentésű szavak csoportjaiból vagy az értelmező szótárak jelentésmeghatározásából indulnak ki. S. LAMB erősen grammatikai jellegű jelentéselméletétől eltekintve így két részletesebben kidolgozott gyakorlati rendszert találunk: a thezaurusznak nevezett rendszert és a szó jelentését jelentéselemek felsorolásával megalapozó elgondolást.

A *thezaurusz* elnevezés ROGET hasonló című szótárára nyúlik vissza, amelyben a szerző az azonos tárgyak, fogalmak jelölésére szolgáló szavakat rendszerezve gyűjti össze. A thezaurusz tulajdonképpen szinonima szótár, bár tartalmaz olyan szócsoportokat is, melyek egy tárgykör (pl. hajózás) körébe eső szavakat ad. A hasonló jelentésű szavak csoportjait ROGET egy eléggé önkényes fölosztás szerint rendezzi, s ez az, ami ezt a szótárt a betűrendes szinonímás szótáraktól megkülönbözteti, és a gépi fordításban való felhasználásra alkalmassá teszi.

A thezauruszon belül egy-egy szó egy vagy több szinonimacsoportba tartozik, és így, ha ezeket a szinonima-csoportokat jelzéssel látjuk el, egy szó egy vagy több jelzőszámot kap. A thezaurusz rendszer a gyakorlati alkalmazásban arra épít, hogy egy összefüggő szövegen belül a szavak jelentése a szöveg egységes tartalmára vonatkozik, ami abban nyilvánul, hogy jelzőszámaik is egységesek lesznek. A gépi fordításban természetesen az egész folyamat fordítottja játszódik le. Ott a kérdés éppen az, hogy a fordításban a szavak sokféle idegennyelvi megfelelője közül, melyiket válasszuk meg. Azt, amelynek jelzőszáma a mondat többi szavainak jelzőszámával összhangban van. (Ez az elgondolás még azt is feltételezi, hogy a mondatban kell lennie egy-két olyan szónak, mely egyértelműen utal a szöveg tartalmára. Ez általában így is van, bár a tartalmat egyéb adottságokból is megkaphatjuk.) A thezaurusz általános fölépítése az alábbi képet mutatja:



3. ábra

A jelentést meghatározó szinonima-csoportokat különböző szinten hasonlíthatjuk össze. Az eljárás lényegében azt jelenti, hogy azonos, vagy egy magasabb kategória alá tartozó szavakat vagy csoportokat alkotunk a szó jelentésének meghatározása során.

A thezaurusz használhatósága ellen több kifogást emelnek. Az egyik az, hogy a thezaurusz csoportok önkényesen, valamilyen szubjektív magától érthetőség alapján vannak megalkotva és egyáltalán nem bizonyos, hogy a gyakorlatban is alkalmazhatók. Ez a kifogás nem mellékes jellegű, ha arra gondolunk, hogy a gépi fordításnak egzakt módon megfogalmazott szabályokra van szüksége, melyek nem egyéni véleményen, hanem a nyelvi anyag tulajdonságain alapulnak. A cambridgei csoport kutatásai is igazolták, hogy Roget rendszere változtatás nélkül nem alkalmazható. A thezaurusz megváltoztatása azonban nem egyszerű dolog. Ha a kiindulási elemeket újra akarjuk csoportosítani, úgy ezt nem tehetjük az elemeknek csak egy részével, hanem a dolog természeténél fogva az egész már megalkotott rendszert át kell alakítanunk. Újabban több olyan kísérlet van, mely a thezaurusz-csoportokat szövegek elemzése alapján, statisztikai adatok birtokában, objektív módon igyekszik megalkotni. Kész és használható rendszert a gépi fordítás irodalmában még nem lehet találni.

Az értelmező szótárak szerkesztési gyakorlatára épülő definíciós szemantikai rendszerek között kétfélet lehet megkülönböztetni. Az egyik a szójelentés meghatározásának a fogalom jegyeihez hasonló elemzésre épít, a másik a meghatározás szavaiból azokat használja föl, melyek további értelmezése nem jelent szóismétlést. Mindkét változat szerint a szavak jellemzőket kapnak. A szavak kapcsolódását ún. előíró, vagy elemző szabályokkal állapítják meg attól függően, hogy értelmes mondatok létrehozására vagy grammatikailag már elemzett mondatok lehetséges értelmezésére irányulnak-e.

A definíciós rendszer keretein belül a szavak többértelműsége is kitűnik, hiszen egy-egy szóhoz többféle értelmezés is fűzhető. Így más jelleggel ugyan, mint a thezauruszoknál, de lehetőség nyílik szinonimák csoportosítására azáltal, hogy a szavak kapcsolódását meghatározó szabályok az azonos lényeges jellemzőkkel rendelkező egységeket rendelik egymás mellé, illetve a fordításban az így kapott értelmezésnek megfelelő szavakat adják ki.

A definíciós rendszer sem tudott mostanáig gyakorlatilag használható szabályokat adni a gépi fordításnak. Ennek okát abban kereshetjük, hogy bár igyekszik konkrét lenni, szabályainak megalkotásában mégsem olyan precíz, hogy kielégítő algoritmus megírására alkalmasak lennének. Így pl. nem tudni világosan, melyek azok a szemantikai jellemzők, amelyekkel a szavak jelentését meg lehet határozni. Az eddigi próbálkozások mind egyénileg eltérő jellemzőket használnak. Lényeges ugyanis, hogy a jellemzők minden szót elegendő módon meghatározzanak, ugyanakkor mégsem szabad minden szót egyénileg jellemezni, mert ez az egész rendszert alkalmatlanná tenné a gyakorlatra. Meg kellene tehát szabni az egyes szók jellemzésének határát és ez nehéz feladat.

Külön nehézséget jelent a szemantikai jellemzőkkel ellátott szók kapcsolási szabályainak megalkotása. Ez nemcsak a szó jellemzőinek sorát adó mátrixok matematikai egymáshoz rendelésének nehézségét jelenti (ez a kérdés sincs még megnyugtatóan megoldva), de főleg azt, milyen szabályokkal lehet a szókapcsolatokat megnyugtató módon megadni. Azok a tanulmányok, amelyek eddig ezekkel a kérdésekkel foglalkoztak, csak kiragadott példák szűk körén belül próbáltak a megvalósítás lehetőségére rámutatni.

#### 4. Gépi fordítás és általános szemantikaelmélet

De Saussure óta a szemantika nemcsak a jelentésváltozások különböző módjaival és a jelentésváltozások okaival foglalkozik, hanem keresi azokat a szempontokat is, amelyekkel a jelentéssel bíró nyelvi elemek sokaságát rendszerezheti. A szinkron nyelvleírásra jellemző szemantikai rendszereket Sp. Jones tárgyalja eléggé kimerítően. A jelentés lehetséges értelmezését két nagy csoportra osztja szét: az egyiket lingvisztikainak nevezi, a másikat, mivel a jelentést meghatározó nyelvi elemek rendszerezésében a denotátumnak is nagy szerep jut, csak igen erős fenntartással vagy egyáltalán nem tudja nyelvészetinek elfogadni. Jonest a jelentés „egzakt” meghatározása vezeti és ezért utasítja el pl. az asszociáción, incompatibilitáson, kollokáción, definíción, alapuló szemantikai osztályozási lehetőségeket. Kétségtelen, hogy elgondolását döntően befolyásolja az a szemléletmód, hogy a nyelvészet, különösen akkor, ha pontos szabályokat akar adni, csak a „tisztán nyelvi” jelenségekkel és azok leírásával foglalkozhat. Ilyen alapon számára a szinonimitás; a tulajdonképeni szemantikai osztályozási mód, és a szó jelentésének értelmezése is ebből adódik.

Azok a törekvések, amelyek a denotátum szerepét teljesen ki akarják küszöbölni a szemantikai osztályozásból, határozottan ellene vannak a hagyományos szemantikai osztályozási módoknak. De Saussure csoportjai, a paradigmaticus és szintagmatikus sorok, alapvetően számításba veszik a denotátumot. Ez nemcsak a szintagmatikus sorokra érvényes, hanem a paradigmaticusokra is, amelyekben a tisztán grammatikai sorokon kívül a tárgykör és a denotátum vezette asszociáció jutnak alapvető szerephez.

Kétségtelen, hogy e denotátumok nem tartozhatnak bele a maguk valóságában a nyelvtudományba, hanem csak mint a jelölt valóságnak a jelhez való viszonyukban. A jel azonban megszűnik jel lenni és képtelenek vagyunk jeltermészetét valóban fölmérni, ha eltekintünk attól a kölcsönös viszonytól, amely a denotátummal összeköti és amely jelszerű viselkedését is befolyásolja. Ha a jel tulajdonságait a denotátumtól elszakítva vizsgáljuk, értékes adatokat kaphatunk, de nagyon sok esetben csak statisztikai valószínűséggel.

Az ember nem a statisztikai szabályok alapján alkalmazza a nyelvet, s a gépnek, ha valóban az emberhez hasonló minőséggel akar fordítani, többet kell alkalmaznia, mint csak statisztikai szabályokat.

A gépi fordítás tapasztalatai szükségszerűvé teszik a szemantika formalizálását, adatszerűvé tételét. Az eddigi eredmények, mint láttuk, kétféle megoldást nyújtanak. A kettő közül a thezaurusz rendszer tekinthető formai szempontból inkább kielégítőnek (bár, alapjában véve, a szinoníma-csoportok is a denotátumok helyesen értelmezett azonosságára épülnek), a gépi fordításban mégsem szolgálhat a legnehezebb kérdések, a homonim szintaktikai kapcsolatok megoldására. A szinoníma csoportok ugyanis nem szókapcsolatokat határoznak meg, hanem az elszigetelten szemlélt szavak összhangját adják. (L. a szavankénti fordítás szemantikai kérdéseit.)

A mondatelemzéses fordításban fevetődő kérdések megoldására a definíciós szemantikaelméletek megfelelő alapot nyújtanak. Ezeknek a rendszereknek legfőbb bajuk azonban éppen a szemantikai jegyek megállapításában van. A szó szemantikai jegyei ugyanis nem közvetlenül vannak adva, hanem más szavakkal való kapcsolódás, szembenállás révén tűnnek elő a megfelelő vonatkozásban. Az is lényeges, hogy az értelmezési meghatározással alapul vett

jegyek között nemcsak mennyiségi és minőségi jegyek vannak, hanem térbeli és időbeli megkötöttségek is, melyek a szókapcsolatokban következetesen érvényesülnek, de amelyeket jegyekben megadni túlzott gyakorlati követelménynek látszik. Így pl. a *Он читает доклад о машинном переводе в Венгрии*. mondatnak *csak* akkor lehet olyan értelmezést adni, hogy az előadó a magyarországi gépi fordításról beszél, ha tudjuk, hogy Magyarországon van gépi fordítás. (Ehhez hasonló mondat sok van.)

A szemantikai kérdések megoldásának szükségessége a gépi fordítás számára igen komoly akadályokat jelent. Az a körülmény, hogy a szemantikát adatszerűvé kell tenni, azonban nem jelenti okvetlenül azt, hogy az ember szemantikai jellegű gondolkodását kell formalizálni. Biztató eredményeket ígér az a megoldási mód, amely nem az emberi gondolkodás menetét, hanem tárgyköri tudását akarja adatszerűvé tenni, Különösen érdekes, hogy a nyelvi megkötöttség, a nyelvi forma, melynek az anyag rendszerezésében érvényesülnie kell, nagy mértékben lecsökkenti a különben igen nagynak ígérkező adathalmazt. Mindezek megoldása még a jövő feladata.

HELL GYÖRGY

### Irodalom

1. A. D. BOOTH, L. BRANDWOOD, J. P. CLEAVE, *Mechanical Resolution of Linguistic Problems*. London, 1958.
2. K. SPARCK JONES, *Synonymy and Semantic Classification*. Cambridge Language Research Unit. Cambridge, 1964.
3. M. ZARECHNAK, *Applied Radical Semantics; Mechanical Translation*. V. 8. (1965) 1965. N. 3-4.
4. F. KIEFER: *Some Questions of Semantic Theory: Comp. Linguistics IV*. Budapest, 1965.
5. GY. SIPÓCZY, *The Analysis of Prepositional Constructions*. . *Comp. Linguistics, IV.*, Budapest, 1965.