Futó Iván: Természetes nyelvek

Jegyzet

A 70es évek elején voltak kísérletek gépi nyelvmodellekre. Az alapvető probléma hosszú ideig az volt, hogy a jelentésnek nem volt pontos, egységes, átfogó és jól gépesíthető formális kezelése.

90es évek végén megtalálható alkalmazások:

- Nyelvhelyesség ellenőrzők
- Automatikus elválasztók
- Beszédfelismerők

Az NLP (Natural Language Processing) aktuális feladata a régi MI módszerek újjáélesztése a jelenlegi hardverviszonyok közepette. Oka a megnövekedett sebesség és a megnövekedett tárolókapacitás.

Nyelvtechnológiai alkalmazások

- Alkotóelemei:
  1. formalizált nyelvtan :
     a. lexikális rész
     b. szabályrendszer
  2. ezt kezelő program
- Beszélhetünk:
  1. mondattan (szintaxis) : nagyobb szövegegységek létrehozása az alacsonyabb szintű formális elemekből
  2. jelentéstan (szemantika) : a szavakhoz rendelt atomi jelentések interpretálása a jelentéskombináció szabályok segítségével, hogy megkapjuk a mondat jelentését.

Szövegmegértés:
Nem beszélhetünk emberi szinten történő szövegmegértésről.
Feladat: szövegek aktuális nyelvi szintnek megfelelő gépi reprezentációja.

Morfológiai elemzés:

A minimális nyelvi egységek által hordozott információ a lexikon vagy szótár. Probléma: a szavak egy mondatban nem a szótári alakjukban fordulnak elő. Megoldások:

1. minden lehetséges szóalak megadása (80as évek közepéig)
2. szótő + lehetséges toldalékok, képzők

Problémák:
Tövek és toldalékok kombinálásának helyes kezelése; pl: jönni, jöhet, jövő vagy vésés, vesés.

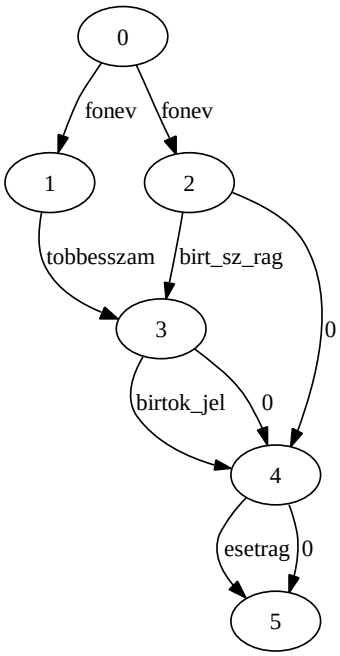
Általános módszer véges állapotú technológiák segítségével:
Koskenniemi - féle kétszintes morfológia (1983):

- a formalizmus lexikonból és szabályokból áll.
- Lexikon = szótő + toldalék.
- I. Szint : nyelvi elemek lexikális reprezentációi
- II. Szint : szóalakok felszíni reprezentációi
- A szabályok a két szint közti átmenetet definiálják.

Előnyei:

- Hatékony implementálhatóság
- Kétirányúság = elemzés + generálás

Ereje kisebb a generatív fonológiában használatos újraíró szabályokéhoz képest. A gond megoldható további lexikonok bevezetésével. Alapelve: a szóalaktani szabályok valójában reguláris kifejezések → véges állapotú automatáknak feleltethetők meg.
Nyelvészeti alkalmazások : gépi helyesírás-ellenőrzés → hatalmas szólisták.
A módszer továbbfejlesztései: Karttunen (1992), RELEX rendszer (Silberztein 1993), HUMOR formalizmus (Prószéky - Pál - Tihanyi 1994).
Egy példa a véges állapotú automatára:



Más módszerek:
lemmatizáló modulok segítségével:

- Átmenetháló éleit nyelvtani szimbólumokból álló párokkal helyettesítik
- Háló = lexikális alak + átmenethálók speciális kompozíciója.
- Használható elemzésre és generálásra is, ám mindkét irányban nem-determinisztikus

Egyértelműsítés:
Oka : a szavaknak többféle felbontása lehetséges.
Módszerei:

- Szabályalapú: nem minden esetben használhatóak, de ha igen kevés hibát követnek el.
- Valószínűségi: minden esetben tudnak dönteni, de gyakrabban tévednek. Pl. HMM.

Példa:

$$\begin{aligned}(nevelo)\ (igeto + TT)\ (fonev) \rightarrow TT &= bef - mn - igenev \\ (fonev)\ (igeto + TT) \rightarrow TT &= mult\ ideju\ ige\end{aligned}$$

Felszíni elemzés:
Több olyan módszer van, amely a mondat nagy részének elemzését végzi anélkül, hogy ismerné a teljes mondatszerkezetet. Az elemzés eredménye címkézett zárójelezés.

$$_{mondat}(La'ttam_{ta'rgy}(_{fn-csop}(egy_{igei_csop}(_{ta'rgy}(_{fn_csop})hordo'\ tokaji)t)hord)o'\ tokaji)t).$$

Mondatelemzés:
Módszerei a nyelvészek absztrakt konstrukcióinak a formális leírását tűzik ki célul.
Problémák:

- Bonyolult szerkezetű mondatok: veszők, gondolatjelek, stb.
- Ismeretlen szavakat vagy szerkezetet tartalmazó mondatok.
- Mondatnál nagyobb környezet alapján megfejthető szintaktikai-szemantikai egyértelműsítés.

Népszerű modell Noam Chomsky generatív grammatikája (lásd. Formális nyelvek):

- Természetes nyelvek végtelen sok mondata leírható egy véges szótár és egy szabályrendszer segítségével.
- A nyelv adott időpillanatbeli állóképét rögzíti.
- A generatív nyelvleírás valódi tárgya a kompetencia.
- Egyetlen ismeretlen szó nem jelenti az információfeldolgozás végét. Rá lehet kérdezni az ismeretlen szóra.

Probléma:
Metaforikus szóhasználat, ismeretlen szavak.
Megoldás a szótár nyitottá tétele, ez ellentmond a Chomsky-féle definíciónak. Ámde csak bizonyos szófajú szavak kerülhetnek később a szótárba.
Ezek alapján létezik egy minimál-nyelvtan, ezt nem lehet nyitott osztállyal definiálni, és van a lexikonnak egy nyitott része.
Korpusznyelvészet kialakulása lehetővé tette, hogy a szöveg teljes egészében előforduló ismeretlen szavakról információkat gyűjtsünk a további előfordulásuk segítségével.

Elemzés
A nyelvi szerkezetek elemzését leggyakrabban mondatfával adják meg. Forgatókönyv egy korábban megtanult, olyan hierarchikus ismerethalmaz, melynek az adott eset feldolgozásában lényeges szerepe van. Forgatókönyvek szótárszerű tárolása hatékonyíthatja a nyelvfeldolgozó modellünket.
Nyitott szótárak előnyei: képesek tetszőleges új képződmények, nyelvi fordulatok, szerkezetek felismerésére.

Szemantika:
Mondatjelentés leírásához szükség van atomi jelentésekre, illetve a jelentések kombinálási szabályaira.
Hosszú ideig az elsőrendű predikátumkalkulus látszott az egyetlen alkalmas formalizmusnak.
1973ban Montague létrehozott egy jelentésreprezentáló rendszert. Fő szerepe a kompozicionalitásnak van. A modellelmélet első jelentős természetes-nyelvi alkalmazása. Lehetővé teszik a szemantika formális kezelését. Alkalmazása a következőképpen történhet:

- Egy részük csak az elmélet alapelveit használja föl.
- Másik részük a teljes Montague formalizmust használja, illetve egy erre épülő számítógépes eljárást.

A nyelvek leírásának lépései Montague nyelvtanok esetén:

1. a mondatszintaxis egy töredékének megadása elemzési fájával
2. intenzionális logika szintaxisának megadása a formulahalmazok induktív definíciójával
3. egy intenzionális logikai modell és a formulák intenziójának és extenziójának megadása erre a modellre
4. elemzési fákat formulákba fordító szabályok megadása
5. modellek halmazára megszorított jelentésposztulátumok megadása

Más szemantikus elemzésre alkalmas modellek:

- szituációs szemantika (Barwise 1983)
- frissítő szemantika
- dinamikus szemantika
- diskurzus-reprezentációs elmélet

Szöveggenerálás

Számítógépben tárolt ismeretek természetes nyelven történő megfogalmazása.

Nehézség: a hosszabb koherens szövegek generálása, a létrehozás tervezési lépéseinek a kidolgozása.

Egyik lehetséges mód a sémák kitöltése, ám ez nem valódi nyelvészeti rendszer.

Nehézségek:

- lexikonbeli elemek helyes kiválasztása (szinonímák)
- mondatok összefűzése, úgy hogy ne legyen köztük törés → mondattervezés

Diskurzus és párbeszéd

A dialógusok egyedi szerkezettel rendelkeznek:

- a beszélő és a hallgató két különböző személy
- a szerepek állandóan cserélődnek

Ha a beszélő nem változik diskurzusról beszélünk.

Diskurzus-reprezentációs elmélet

Kamp elmélete (1981):

- minden D szöveghez tartozik egy diskurzus-reprezentáló szerkezet, amely D-t kvantormentes klóz-alakban ábrázolja
- szöveg-reprezentációs szerkezet alakja: DRS=<REF, FELT>, ahol REF a DRS szöveg-referenseinek, Felt pedig az egyedekre vonatkozó feltételeinek halmaza
- a mondat rendszerbeli reprezentációja valamilyen DRS-ken operáló függvény lesz

- Heim állományváltoztató szemantikája segítségével az elmélet számítógéppel is ábrázolható, a DRS egy állomány, míg a diskurzusreprezentáció egy kártya.

Nyelvfeldolgozási módszerek
Nyelvészeti indíttatású módszerek:
Unifikációs nyelvtan = elméletcsalád, mely ma egyértelműen meghatározó
Unifikációs formalizmusok:

- Fejnyelvtan

- Lexikális funkcionális nyelvtan

- Fabővítő nyelvtan

- Kategoriális unifikációs nyelvtan

Ezek a szavakat, szószerkezeteket, mondatokat attribútum-érték párok halmazaként reprezentálják = jegy-együttesek. Ezek egymásba ágyazhatóak.

Alulspecifikáltság = egy adott jegy jelen van, de értéke nem vagy csak részben meghatározott.

Változókat is használhatunk, pl. alany és állítmány számának egyeztetésére.

Unifikáció = nyelvtani információk összeegyeztethetőségét vizsgálja, monoton művelet.

Statisztikai feldolgozás
Nyelvfeldolgozás = információátvitel zajos csatornán
A módszer alapelemei:

- Átviteli modell = felismert kimenet valószínűsége

- Nyelvmodell = egyes üzenetrészek adott környezetben való előfordulási valószínűségei.

Legnépszerűbb alkalmazott modell a rejtett Markov-modell (HMM)

Számítógépes elemzési technikák
Nyelvosztályok:

- Chomsky-féle osztályozás

- Osztályokra jellemző az elemzési sebesség

Új elmélet, jegylogika → jegyszerkezetek
A jelentés ábrázolása valamely logika keretében történik.
Elemzési technikák:

- Szabályalapú rendszerek

- Unifikációs formalizmusok

- Véges állapotú automaták

- Valószínűségi módszerek

Szintaktikai elemzés: LR elemzők (bizonyos módosításokkal alkalmas környezetfüggő nyelvek esetén is)

Jegyszerkezetre építő unifikációs formalizmusok egy alapvázra épülnek, osztályuk Turing ekvivalens, polinomiális időben nem elemezhető.

Kategoriális rendszerekben az elemzés logikai levezetésnek tekinthető → dedukció

Gépi nyelvészet által kifejlesztett speciális módszerek: neurális hálók.

Szövegkorpuszok:

Szövegkorpusz = gépi nyelvfeldolgozás számára összegyűjtött szövegek együttese.

Az egyes szavak különböző helyzetben való előfordulásainak tanulmányozására használják.

Párhuzamos korpuszok = eredeti szöveg és a fordítása.

Korpusznyelvészet módszereire elsősorban valószínűségi és statisztikai módszerek jellemzőek. Pl. Olyan szerkezetekre mint: erős légy.

Lexikonok és szótárak:

Lexikális tudás = a nyelv szavainak, kifejezéseinek ismerete.

Szótár = lexikális elemek listája + morfoszintaktikai, szemantikai, fonológiai viselkedésüket leíró jegyek összessége → szükség van egy jegyleíró formalizmusra.

Fontos a reprezentációs nyelv.

A reprezentációs nyelv szabványosítása az SGML (Standard Generalized Markup Language), szótárak leírásához pedig a TEI (Text Encoding Initiative) → formától függetlenül lekérdezhetővé válnak az egyes mezők és kombinációik.

Terminológiai adatbázisok:

Terminus = szakiránytól függő, akár teljesen más jelentéssel bír, állandóan születőben van.

Terminológiai adatbázisok dinamikusak.

Jellemzőek a soknyelvű adatbázisok.

A fogalmak egy fogalmi hálózat megfelelő relációkkal elérhető csomópontjaként jelennek meg. Jellemzésük tezaurusz-deszkriptorokkal, szinonimákkal, rövidítésekkel, definíciókkal, képekkel, relációkkal stb. történik.

A dokumentumkezelés műveletei:

- Létrehozás
- Keresés
- Kivonatolás

Szöveglétrehozás:

Szerzői eszközök: helyesírás ellenőrző, elválasztó, nyelvtani ellenőrző, szinonima szótárak.

Hibák:

- Billentyűzeten való melléütésből származó (környező betűk elhelyezkedése szerint)
- Magyar-angol billentyűzeten való y-z eltérés
- Magyar ékezetes betűk szabványos vagy nem szabványos elhelyezése

- Beszéd írásra való hatása "azt írjuk, amit mondunk"

Automatikus elválasztás:

Egy szó elválasztásához, annak minden lehetséges elemzését ismernünk kell. Az se jó ha nem ismeri mindet, de az se jó ha túl jól ismeri ezeket. Pl. Legelőre

Az elválasztó úgy működjön, hogy jelenléte alig észrevehető legyen → nem interaktív. A kézi elválasztás lehetősége biztosított kell legyen.

Rossz elválasztás hiba, ha nincs elválasztás az csupán az esztétikán látszik meg.

Keresés:

Egy szó minden alakjának felismerése.

Probléma: a szavak nincsenek szótári alakban → a mechanikus rendszerek gyakran tévednek.

Nyelvhelyesség ellenőrző:

Egyenlőre csak szóellenőrzőkről beszélhetünk.

Szöveg-visszakeresés:

Fontos a szinonimák illetve különböző nyelvekre történő fordítások közti keresés is, a szemantikát is figyelembe kell vennünk. Pl. Kutya - Kosárlabda EB

Automatikus szövegkivonatolás:

50es években már megfogalmazódott a gondolata.

90es évekre jelent meg a valódi igény.

Nincs külön elmélete, így mindenféle heurisztikák születhetnek.

Cél a szöveg tartalmának kevesebb mondatokkal való kifejezése.

Reális cél a szöveg releváns mondatainak kiemelése, és koherens szöveggé alakítása. A kiválasztás statisztikai alapon vagy kulcsszavak alapján történik.

Fordítás:

A számítógép keres egy hasonló szerkezetet a korábbi fordítások között.

Szabályok alapján történő fordítás nem célszerű.

Nyelvazonosítás:

Feladat: a fordításnak a gépi, illetve géppel támogatott létrehozása, továbbá a forrás- és a célszövegek szinkronizálása a későbbi feldolgozás számára.

Nyelvazonosítás elsősorban statisztikai alapon történik:

- Nyelvek legrövidebb szavainak eloszlását figyelik
- Egyes szó és karaktersorozatok gyakorisága
- Nyelvre jellemző speciális karakter és karakterkombinációk megfigyelése. Legelterjedtebbek a trigram-modellek, egymást követő betűhármasok gyakoriságainak megfigyelése.

Számítógépes fordítás:

Gépi fordításhoz használt számítógépes eszközök csoportosítása:

- Teljesen automatizált gépi fordítás (TAGF)

- Ember támogatta gépi fordítás (ETGF)
- Gép támogatta emberi fordítás (GTEF)

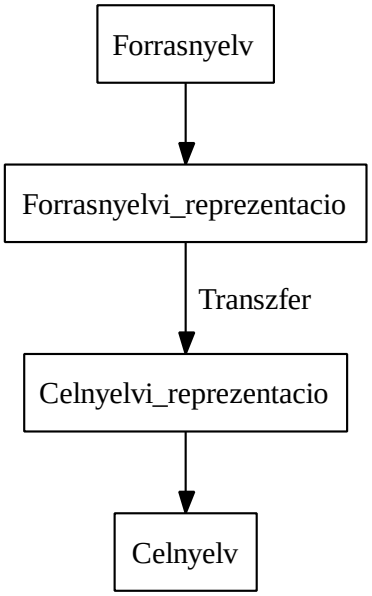
TAGF:
Közvetlen emberi beavatkozás nélkül működő rendszerek gyűjtőneve.
Legfeljebb technikai szövegek felszínes fordítására alkalmas.

ETGF:
A gép a felhasználó segítségével ad választ a többértelműségekre és bizonytalanságokra.

GTEF:
Hagyományos emberi fordítást jelent.
A fordító segédeszközei egy írógép és szótár funkcióját betöltő hatékony számítógépes rendszer.
Gépi fordítás csoportosítása az alapvető működési technikái alapján:

- Produktív vagy mintaalapú ha a fordítás célnyelvén a mondatokat a program maga szintetizálja, vagy csak kikeresi a forrásnyelv mondatai közül a leghasonlóbbat és annak "konzerv"-fordítását adja meg.
- Produktív fordítás lehet közvetlen vagy közvetett, ha a forrásnyelv analízise és a célnyelv szintézise függő vagy független
- Közvetett fordítás lehet interlingvális vagy transzfer fordítás, ha a jelentés-reprezentáció független-e a fordításban szereplő nyelvektől vagy sem.
- Minta alapú fordítás elsősorban fordítómemóriákat és a velük társítható fejlesztéseket jelenti.
- A produktív fordítás technikai előkészítését a kontrollált nyelvi eszközök végzik.

Produktív fordítás technikái:
Kezdetben a közvetlen fordítás volt elterjedve, azaz a forrásnyelvből a célnyelvre fordítottak kihasználva e két nyelv specifikus tulajdonságait. A fordító nyelv szótára és szintaxisa csupán a többértelműség feloldására szolgált. A meghatározó a célnyelv szórendje illetve a szavak meghatározása volt. Szemantika alig volt, csupán néhány jegy szerepelt a formalizált mondatokban.
60as évektől jelentek meg a közvetítő nyelvre épülő rendszerek. Egymástól függetlenek voltak az analizáló és szintetizáló komponensek, illetve külön fordító- illetve célnyelv szótárakra épült. Előnye a meglévő stratégiák módosítása nélkül kapcsolhatók a rendszerbe új nyelvek. A közvetítő nyelv elsősorban szintaktikai szerkezetet jelentett, szemantikai primitívekből szintén kevés volt. Hátrányai: bármely szinten végrehajtott rossz alternatíva-választás kihatott az összes további szintre; szintaktikai többértelműség miatt túl sok szerkezetet állítottak elő (szemantika hiánya miatt).
Ezen nehézségek kiküszöbölése miatt született meg az ún. Transzfer módszer:



90es évek TAGF rendszere kliens-szerver architektúrájú. A fordítás egy távoli gépen történik. Pl. AltaVista, Systran.

Nem teljesen automatikus gépi fordítás irányzatai:
Lényege a felhasználó aktív bevonása a fordítási folyamatba. Ide tartozik minden olyan gépi fordítást támogató rendszer, melynek célja a már meglévő fordítások hatékony felhasználása. Fontos olyan rendszerek léte, amelyek hatalmas nyelvi szövegek konzisztens fordítását garantálják.

Fordítói munkaállomások: kétnyelvű szótárak, szaknyelvi terminológiai adatbázisok, fordítómemóriák, és szerencsés esetben valódi gépi fordító rendszerek elérését is lehetővé teszik.

Fordítómemória feladata, hogy a fordításra váró szövegrészletekhez hasonló, korábban már lefordított anyagokat találjon. Legtöbbször neuronhálós technikákat használnak, azaz a nyelvészeti elemző módszerek feltételes felhasználásával történik a fordítás. A felhasználó kiválaszthatja a hasonló mondatok közül a legmegfelelőbbet, módosíthatja azt, illetve egy teljesen másat is megadhat. Nagy mennyiségű sémákra van szükség. Szövegszinkronizáló (aligner) programok segítségével a fordítómemóriába helyezhetünk már korábban elkészült fordítások anyagait a program kezdeti használatakor. A szövegszinkronizálás statisztikai alapon történik.

Az intelligens szótárak a szavakat akkor is megtalálják, ha a keresőkérdésben nem a szótári alakjukban állnak.

Szótár természetes közlési egysége a szócikk. Egy szócikkben a felhasználónak minden olyan információt meg kell találnia, ami az adott szótártípusra jellemző. Számítógépes szótárak esetén a cél, hogy egy szócikkhez kapcsolódó minden információ elérhető legyen, függetlenül attól, hogy benne van-e a szócikkben vagy sem.

Kiegészítő szótárnak minősül minden szakszótár.

Alapszónak számít minden olyan tőszó, melynek akár továbbképzett származékai, akár összetételbeli vagy kifejezésbeli előfordulásai is vannak.

Gépi szótár csak akkor ér valamit, ha képes támogatni a folyó szöveg elemzését, ellenőrzését is. Viszont egyes szavak származékalakjait szükséges önálló címszóként is megadnunk.

Kontrollált nyelvi alkalmazások Kontrollált nyelvi rendszerek garantálják, ha korlátozásaik segítségével egy szűkebben értelmezett emberi nyelvet használunk, akkor így készült szövegünk a rendszerhez csatolt fordítómodul segítségével fordítható lesz. Az így létrejött szövegek egymással konzisztensek, jól olvashatók és pontosan visszakereshetők. Felmerül a szegényesedés kérdése, az ilyen rendszerek nem ismerik teljesen az adott nyelvet.

Beviteli módszerek:

Klasszikus beviteli módszer a billentyűzet.

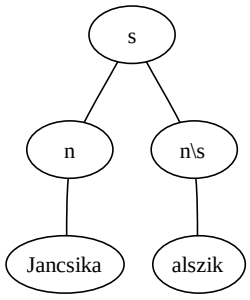
Újfajta beviteli módszerek: szkenner, mikrofon.

Írott és beszélt anyag bevitele zajos csatornán történik. Ezeknek a feldolgozásához statisztikai és valószínűségi módszereket használnak.

1. KIEFER FERENC
A nyelvészet és a számítástudomány

A nyelvészet és a számítástudomány kapcsolata az ötvenes években kezdődött, amikor amerikai kutatók a gépi fordítás lehetőségét felvetették. Az elképzelés egyszerű és logikus volt: mivel a számítógép mindenféle jelrendszer elemzésére képes, a természetes nyelvi jelekből álló rendszerek elemzése is megoldható a számítógép segítségével. Nem kell tehát mást tennünk, mint a szóban forgó nyelv nyelvtanát és szókészletét betáplálni a számítógépbe. Ahhoz, hogy ezt megtehessük, a nyelvtan szabályait formalizálva, a matematika szabályaihoz hasonlóan kell megadnunk, és természetesen a nyelv szavainak tulajdonságait megfelelő kódokkal kell ellátnunk. Egyszerűbb esetekben ez nem jelenthetett problémát, mivel a mondatok szerkezetét könnyen ki lehetett fejezni szimbólumok segítségével. Ha például n-nel jelöljük a főnév, és s-sel a mondat kategóriáját, és $n \setminus s$-sel azt a kategóriát, amely n-nel kombinálva s-t eredményez (magyarán: a mondat az alanyi főnévi szerkezetből és az állítmány szerepét játszó igei szerkezetből áll össze), akkor a "Jancsika alszik" mondat szerkezete:

1. ábra

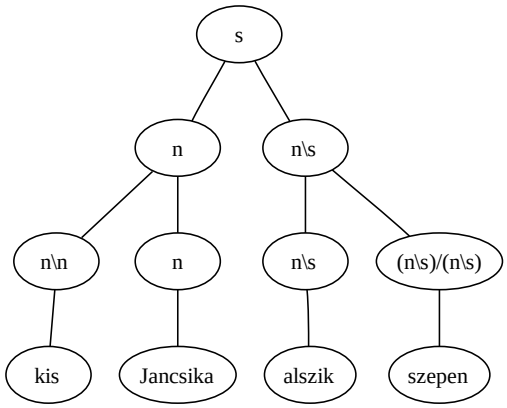


Az $n \setminus s$-ben szereplő ferde vonal azt jelzi, hogy a szóban forgó kategória balról veszi magához az n kategóriájú elemet. Ennek megfelelően az s/n azt jelenti, hogy az n kategóriájú elem jobb oldalon szerepel. Ezzel a jelölésmóddal az összetevők szórendi helye is kifejezhető.

Ez a szerkezeti felépítés kiterjeszthető a Kis Jancsika szépen alszik mondat esetére is: a kis melléknévi jelző kategóriája n/n, mivel egy n kategóriájú elemből újból egy n kategóriájú

elemet állít elő (a kategóriák kombinációja a törtekkel való műveletekre hasonlít), a szépen módhatározóé pedig $(n\backslash s)/(n\backslash s)$, mivel egy igei szerkezetből (jobbról) ismét egy igei szerkezetet állít elő. Tehát:

2. ábra



A vázolt elemzési módszer kategoriális grammatikaként vált ismertté (Lambek, Bar-Hillel). A szótár felépítéséről azt kell tudnunk, hogy n/n kategóriát kap az n kategóriájú főnevet balról módosító melléknév és $(n \setminus s)/(n \setminus s)$ kategóriát az $n \setminus s$ kategóriájú állítmányt balról módosító határozószó. A kategoriális grammatika ugyan nem váltotta be a hozzá főzött reményeket, de elindítója volt egy ma is virágzó kutatási irányzatnak, amelynek fő célja a természetes nyelvek grammatikájának formalizálása.

Ha egy A nyelvről (a forrásnyelvről) kívánunk egy B nyelvre (a célnyelvre) fordítani, akkor a megfelelő kétnyelvű szótáron kívül az A nyelvre egy elemző, a B nyelvre egy szintetizáló (generáló) rendszert kell kidolgoznunk. Az ötvenes évek végére a gépi fordítás lehetőségével kapcsolatban ugyan komoly kételyek merültek fel, az elemző, illetve szintetizáló rendszerek kutatása azonban tovább folyt. Ezek a kutatások a nyelvtudomány további fejlődésére is hatással voltak. A hatvanas évek elejétől a "számítógépes nyelvészet" (computational linguistics) elfogadott terminussá vált. A számítógépes nyelvészet a természetes nyelvek számítógépes feldolgozásával (natural language processing) foglalkozik. A megszokott klasszikus nyelvészeti területeken (hangtan, alaktan, mondattan, jelentéstan) kívül a fordítást, az automatikus kivonatolást, az információs és dokumentációs nyelvek kérdését, az automatikus indexelést, az automatikus kivonatolást, a mesterséges intelligenciakutatást, a párbeszédes rendszerek vizsgálatát is bizonyos mértékig nyelvészeti problémának kell tekintenünk. A számítógép és a nyelvészet szerteágazó kapcsolatairól tehát a jelen áttekintés nem adhat számot, meg kell elégednünk néhány jellemző példa bemutatásával.

Az alaktani elemzés, illetve szintézis

A szóalakok belső szerkezetének megállapítása, különösen a magyar és a magyarhoz hasonló nyelvek mind elméleti, mind pedig gyakorlati szempontból alapvető feladat. A számítógépes elemzés segítségével a korábbinál sokkal pontosabb képet alkothatunk a nyelv alaktani rendszeréről, ugyanakkor az alaktani elemzés előfeltétele mind a számítógépes szótárkészítésnek, mind pedig a számítógépes mondattani elemzésnek. A számítógépes alaktani elemzők ismertetése helyett néhány példán érzékeltetjük a problémát. Egy magyar főnévnek, mondjuk, a botnak van 18 esetragja: bot, bot+ot, bot+nak, bot+ban, bot+tal stb., többes száma: bot+ok, birtokos személyragos alakja: bot+om, bot+od, bot+ja stb., utóbbinak vannak több birtokra utaló alakjai: bot+jaim, bot+jaid, bot+jai stb., a főnév kaphat birtokjelet: bot+é, és a birtokjel

után is szerepelhet a többes szám jele: bot+é+i. A több birtokra utaló személyragos alakokat a következőképpen szoktuk szételemezni: bot+ja+i+m, bot+ja+i+d, bot+ja+i, ahol a ja az általános birtokviszonyjel, az i a birtokos személyragok társaságában megjelenő többesjel, az m, d birtokos személyragok; egyes szám harmadik személyben a birtokos személyrag zérus. Ha egyéb a névszókhoz járuló toldalékot is figyelembe vesszük, akkor összesen 842 paradigmatikus alakot kapunk. A toldalékok sorrendjét könnyű szabályba foglalni. A tő után bármilyen toldalék állhat, ha a többesjelet választjuk, akkor utána birtokjel, a birtokjel többes száma és esetrag következhet. Ha birtokviszonyjelet választunk, akkor ahhoz többes szám, birtokos személyrag, birtokjel, ennek többes száma és esetrag járulhat. A toldalékok sorrendjét legegyszerűbben egy grafikonnal ábrázolhatjuk (Kiefer 1999, 209). A 3. ábrában szereplő jelek magyarázata: Tsz = többes szám; Br = birtokos személyrag; Bj = birtokjel; Bv = birtokviszonyjel; Eset = esetrag.

Az egyes csomópontok kimeneti pontok is lehetnek. A Tsz és Bj közötti nyíl kétirányú, ami azt kívánja jelezni, hogy a többes számú alak felveheti a birtokjelet, de a birtokjel is kaphat többes számot (pl. bot+ok+é+i). A (3) véges állapotú automatával leírható nyelvtan, a formális nyelvtanok legegyszerűbbike.

A (3) alatti 'minigrammatika' generálásra és elemzésre egyaránt felhasználható. A generálásnál balról jobbra, az elemzésnél jobbról balra haladunk. Például a

```
bot + ja + i + tok + é + i + nak
```

szóalak esetében a (3) grammatikai modellnek megfelelően a bot tő a generálásnál először a ja, a ja után az i, az i után a tok stb. toldalékot kapja. Az elemzésnél a nak toldalékkal kezdjük, megállapítva, hogy a nak esetrag; a nak előtt áll az i, amely a birtokviszonyjel vagy a birtokjel társaságában a többes szám toldaléka, és így tovább.

A helyzet azonban nem mindig ennyire egyszerű. A generálásnál nyilvánvalóan nem elegendő a 3. ábrában megadott sorrendi információ. Így például az egyes szám harmadik személyben a birtokos személyrag lehet j-s vagy j nélküli: asztal+ a, de kalap+ja; a tő bizonyos toldalékok előtt megrövidülhet: víz - viz+et - víz+ben; rövid magánhangzóra végződő szavak esetében toldalékolásnál a szóvégi magánhangzó megnyúlhat: tábla - táblá+t. Vannak azután olyan tövek, amelyekben toldalékolásnál hangátvetés (pl. kehely - kehely+ben - kelyh+et), hangkivetés (pl. dolog - dolog+ért - dolg+ot), vagy hangbetoldás (pl. ló - ló+val - lov+at) következik be. Egyes esetekben külön problémát jelenthet a megfelelő toldalékok kiválasztása. A többesjel ötféle alakban jelenhet meg: -k, -ok, -ek, -ök, -ak, pl. ajtó+k, bolt+ok, kert+ek, fürt+ök, ház+ak. Ezek közül az első négy a magánhangzó-illeszkedéssel magyarázható, a ház+ak esetében azonban a helyes toldalék kiválasztásához szótári információra van szükségünk, ti. a szó hangalakja alapján nem jelezhető előre a toldalék alakja: ház+at, de gáz+t. Mindez azt mutatja, hogy a szóalakok generálásakor a helyes toldalékok kiválasztásához szótári információra is támaszkodnunk kell tudnunk. Elemzéskor elsősorban a tő azonosításához van szükségünk szótári információra, ha ugyanis nem szerepel a szótárban a bokor tő mellett a bokr is, akkor a bokr+ot toldalékolt alak elemzésekor a bokr tő nem azonosítható.

Hasonló meggondolások vonatkoznak a képzett és az összetett szavakra is. Az egyik legtermékenyebb igeképzőnk a -z, amely szintén többféle alakban jelenhet meg videó+z(ik), golf+oz(ik), internet+ez(ik), szörf+öz(ik), tehát -z, -oz, -ez, -öz.A képzők azonosítása és leválasztása a ragokéhoz hasonló módon történik. Az összetett szavak elemzésekor a szótári információ alapján a szótöveket kell tudnunk azonosítani. Összetett szónak számít minden olyan szó, amely egynél több szótövet tartalmaz. Tehát: miniszter+elnök, sötét+kék, beteg+ágy.Az összetételi tagok természetesen képzettek is lehetnek: autó+szerelő, levél+írás, írás+szakértő. Probléma

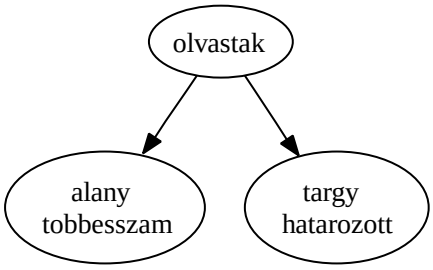
ezekben az esetekben sem adódik, hiszen a szavak összetétel voltát a helyesírás is jelzi: az összetételt alkotó szavakat általában egybeírjuk. Tudjuk azonban, hogy ez nem mindig van így. Íme néhány ellenpélda: önszabályozó rendszer, szilárd vázas állat, mágneses tér. Ebben az esetben a következő rendkívül bizonytalan és megbízhatatlan elv érvényesül: összetett szóról van szó, akár egybeírjuk az összetételi tagokat, akár nem, ha a szó egy fogalmat jelöl. Ez a kritérium természetesen a számítógépes elemzésben nem használható. Az egyetlen megoldás tehát az, ha a helyesírás által nem jelzett összetett szavakat is felsoroljuk a szótárban.

A vázolt kérdések számítógéppel jól kezelhetők. A magyar alaktan szintézise és elemzése nem okoz különösebb problémát (Prószéky, megjelenés alatt). Az említett morfológiai elemzés, illetve szintézis különböző módon implementálható. Ahhoz, hogy helyes eredményt kapjunk, mint láttuk, az elemző, illetve szintetizáló szabályokon kívül szükségünk van megfelelő információt tartalmazó tő- és toldaléktárra is. A morfológiai elemző fontos alkotóeleme a szintaktikai elemzőnek, mint látni fogjuk, de számos más számítógépes nyelvészeti alkalmazása is van. Nélküle nem készíthető számítógépes szótár, nem címkézhetők a szöveges adatbázisok szövegszavai, nem készíthető helyesírás-ellenőrző program.

A mondattani elemzés

A magyarhoz hasonló nyelvek esetében a nyelv mondatainak elemzése a morfológiai elemző által azonosított toldalékok mondatbeli funkcióit határozza meg. A mondat központi eleme, az ige a morfológiai elemző segítségével azonosítható. Például az olvastátok igealak esetében az elemzést jobbról kezdve megállapítható, hogy a k vagy az ok lehet névszó többes számú alakja, de mivel sem az olvastáto, sem pedig az olvastát nem található meg a tőtárban, a k, illetve az ok nem lehet ebben az esetben toldalék, a tok ugyan igei toldalék (pl. írtok), de az olvastá sem szerepel a tőtárban. A következő lépésben az átok leválasztását végezzük el, amely szintén szerepel a toldaléktárban, a maradék olvast, de ez alak sem található a tőtárban. A maradék utolsó betűjét, a t-t, nem számíthatjuk hozzá a toldalékhoz, mert tátok toldalék nincsen a toldaléktárban. Más szóval, átok a maximális toldalék, amely szóba jöhet. A toldalék kategóriáit (tárgyas ragozás, múlt idő, többes szám 2. személy) a toldaléktárból megkaphatjuk. Mivel az olvast nem szerepel a tőtárban, leválasztjuk a t-t; az olvas már megtalálható a tőtárban, a t pedig a múlt idő jelével azonosítható. Figyeljük meg, hogy ebben az esetben a -t tárgyrag már nem jöhet szóba, hiszen a tövet az átok toldalék alapján már igei tőként azonosítottuk. Az olvas tőről azt is tudjuk, hogy tárgyas ige töve, az átok toldalék alapján pedig arra következtethetünk, hogy az ige tárgyának határozottnak kell lennie. Tekintsük most azt az egyszerű esetet, amikor a mondatban egy határozott többes számú alanyi és határozott egyes számú tárgyi főnévi szerkezet található: A fiúk olvasták a hírt. Az olvasták igei alakról a vázolt módon meg tudjuk állapítani, hogy szerkezete olvas+t+ák, és hogy az ák toldalék a tárgyas ragozás, múlt idő, többes szám 3. személyű toldaléka. Ennek alapján az olvasták környezetében meg kell jelennie egy többes számú alanyesetben álló főnévnek (a fiúk) és egy határozott tárgyesetben álló főnévnek (a hírt). A fiúk esetében a k leválasztásával megkapjuk a fiú tövet, a hírt esetében a t tárgyrag leválasztásával a hír tövet.

5. ábra



Az 5. ábra mutatja, hogy az olvasták szóalakból milyen információ nyerhető a szintaktikai elemzés számára: a morfológiai elemző az olvasták szóalakot olyan igei alakként azonosítja, amely többes számú alanyt és határozott tárgyat kíván.

Az elemző bonyolultabb esetekben is a fent vázolt módon működik. A szótárban az igetövek jellemzése tartalmazza az ige vonzatkeretét, vagyis az igével jelölt cselekvés, folyamat kötelező és választható szereplőit (a predikátum-argumentum szerkezetet). A szótári információ része a vonzatok morfológiai toldaléka is, tehát például a megvesz vki vmit (vktől), ad vki vkinek vmit, tesz vki vmit vhová; zárójelbe szoktuk tenni a fakultatív, választható vonzatot. Láthatjuk tehát, hogy az ige azonosítása után a szótárból megtudhatjuk, hogy az ige környezetében milyen vonzatok várhatók. A Péter pénzt adott a fiúnak mondatban az ad igető azonosítása után tudni fogjuk, hogy az ad ige kötelezően három vonzatot vesz magához: egy alanyi, egy tárgyi és egy részeshatározói vonzatot. Az alanyi vonzatnak nincs morfológiai jele, a tárgyi vonzat t ragot, a részeshatározói vonzat pedig -nak/-nek ragot kap.

A nyelvekre általában jellemző, hogy minél gazdagabb az alaktanuk, annál szegényebb a mondattanuk, és fordítva, minél szegényebb az alaktanuk, annál gazdagabb a mondattanuk. A magyar nyelv a morfológiailag gazdag nyelvek közé tartozik, az angol nyelvnek ezzel szemben alig van morfológiája. Ebből következik, hogy a magyarban a morfológiai elemzéssel szintaktikai problémákat is meg tudunk oldani.

Ha egy mondatban több toldalékolt igei alakot azonosítunk, akkor összetett mondattal van dolgunk. Az összetett mondat pontosan annyi tagmondatból áll, ahány toldalékolt igei alakot találunk benne. Például A fiú látta, hogy Anna hazament, és hogy egy hatalmas bőröndöt vitt magával összetett mondat három tagmondatból áll. A tagmondatok elemzése az egyszerű mondatok mintájára történhet, az egyes tagmondatok összekapcsolásához természetesen a kötőszók funkcióját is ismernünk kell, amelyet ismét a szótárból nyerhetünk.

A szótár

A szótár a lelke, a legfontosabb komponense minden elemzésnek. Eddig csak arról beszéltünk, hogy a szótárban milyen szintaktikai és morfológiai információk találhatók (ezek minimálisan az igék esetében a vonzatkeret és minden szótő esetében a toldalékolásra vonatkozó információk), a jó szótár azonban arról is tájékoztat, hogy milyen gyakori a szó. A Collins-Cobuild angol szótár például a gyakoriság szerint öt csoportba osztja a szavakat. Az első csoportba tartoznak a leggyakoribb szavak (a, az, alá, fölé, mellé, mögé, már, mindig, beszél, válasz, terület, kar, fegyver, művészet), a másodikba a valamivel kevésbé gyakori szavak (híd, veszély, nyilvánvaló, vitatkozik, megérkezik, letartóztat).A két csoport szavai összesen a beszédben, írásban használt szövegszavaknak 75 százalékát teszik ki. A még mindig viszonylag gyakori szavak három további csoportjával együtt így összesen a szövegszavak 95 százalékát kapjuk. A hátralévő 5 százalék a ritka szavak csoportja. A gyakoriság megállapítása természetesen

számítógéppel történik. A modern számítógépek szinte határtalan tárolási kapacitása teszi lehetővé a nagy mennyiségű szövegek, ún. korpuszok alapján történő szótárkészítést. Az említett angol szótár (2. kiadása) 200 millió szövegszó alapján készült. A korpusz feldolgozásakor automatikusan megkapjuk a gyakoriságra vonatkozó információt, a morfológiai és szintaktikai információk pedig a korpusz alapján pontosíthatók és kiegészíthetők. A nyelv, tudjuk, állandóan változik; a kézi gyűjtéssel összeállított szótárak gyakran olyan információt is tartalmaznak, ami nem a mai nyelvhasználatot tükrözi. A szótárak azonban nemcsak ezért avulnak el viszonylag gyorsan, hanem azért is, mert állandóan keletkeznek új szavak. A nyelv változása leginkább a szókincs változásában érhető tetten. A mai írott és beszélt nyelvi korpuszok alapján készült szótárak ezt a problémát könnyen megoldják. Az elektronikus szótárt könnyen kiegészíthetjük új információval és új szavakkal. Annak érdekében, hogy a különböző szótárkészítő műhelyek adatbázisai egymással kompatibilisek legyenek, a nyolcvanas évek végétől egyre több helyen használják az SGML (=Standard Generalized Markup Language) reprezentációs nyelvet. Ez a nyelv független egy adott számítógépes rendszer adottságaitól, tehát könnyen adaptálható új rendszerek esetében. Ez azért is fontos, mert a szótárkészítő projekteket általában több évtizedre tervezik, és a számítógépes rendszerek ez idő alatt óriási fejlődésen és változáson mehetnek át.

A szótárkészítéskor igen nagy segítséget nyújtanak az ún. konkordanciák, amelyek a vizsgált szó környezetét mutatják. A konkordanciák lehetővé teszik a szó jelentésének meghatározását, több jelentésű szavak esetében az egyes jelentések szétválasztását, a vizsgált szóhoz kapcsolódó kifejezések, idiómák megállapítását. Konkordanciákat a szótárírók már régóta használnak, a mai szótárírónak azonban az a nagy előnye, hogy óriási korpuszból válogathat, a korszerű szoftverek a legkülönbözőbb szempontok szerinti keresést teszik lehetővé.

A korszerű szótár azonban nemcsak hatalmas adatbázis alapján készül: a szótárírónak figyelemmel kell lennie a szójelentésre vonatkozó legújabb kutatások eredményeire is. A szótáríró régi problémája a szó különböző jelentéseinek egymástól való elkülönítése. A legtöbb szó több jelentésű (poliszém): a fest ige nem ugyanazt jelenti a tájképet fest, az ablakokat festi, zebrát fest az úttestre, festi a haját, festi az arcát kifejezésekben. A fest ige jelentéseinek elkülönítésekor érdemes a fordítás szempontjaira is gondolnunk. Ha valamely nyelvben a fest ige két jelentésének két különböző szó felel meg, akkor érdemes a szótárban a két jelentést megkülönböztetni. A jelentések megkülönböztetésekor további szempont lehet a jelentések kontextus (pl. az alany és a tárgy típusa, jelentése) alapján való előre jelezhetősége is.

A szokásos egynyelvű szótárokon kívül korpuszok alapján készül szótár a neologizmusokról (új szavakról, kifejezésekről, pl. globalizáció, internet), a kollokációkról (két szó együttes előfordulásáról, pl. egy melléknév és egy főnév együttes előfordulásáról: mély álom, gondos ápolás), az alapszókincsről, a frazeológiai egységekről (pl. iskolába jár, kukoricát tör, állást foglal, kifejezésre juttat),az összetételekről (szürkegazdaság, olajszőkítés),az igei vonzatokról (pl. rak vki vmit vmire, megrak vki vmit vmivel). Külön érdemes megemlítenünk a terminológiai szótárakat, amelyeknek korszerű változatai szintén számítógépes adatbázis alapján készülnek.

Az egynyelvű szótárokon kívül számítógépes adatbázis alapján készülnek a kétnyelvű szótárak is. Kétnyelvű szótár készítésekor a szótárkészítő gyakran ún. párhuzamos korpuszokra, vagyis hagyományos módszerrel lefordított szövegekre is támaszkodik. A párhuzamos korpuszokból információt nyerhetünk a fordításra vonatkozó ismeretekről (translation knowledge).

Szövegekből álló adatbázisok

Az SGML azonban nemcsak korszerű szótári adatbázisok reprezentációjára alkalmas, hanem az SGML szabályai szerint kódolt szövegek esetében gyors információkeresésre is. Minden egyes szöveg kódolt változata tartalmazza a szövegforrás legfontosabb bibliográfiai adatait és a tartalom azonosítását megkönnyítő kulcsszavakat. Szöveget tehát nemcsak szerző, kiadó, megjelenési hely, megjelenés éve szerint, hanem különféle tartalmi mutatók szerint is kereshetünk.

A szöveges adatbázisok a szöveg-koherencia és a jelentés vizsgálatában is új perspektívát jelentenek. A szövegek különböző célú számítógépes vizsgálata külön diszciplína, a korpusznyelvészet kialakulásához vezetett. A korpusznyelvészet is elsősorban lexikográfiai jellegű kérdéseket vizsgál, a hagyományos szótárkészítővel szemben azonban a korpuszokat nemcsak arra használja, hogy belőlük példákat merítsen, hanem rendszeres vizsgálatnak veti alá őket, vagyis gondosan szemügyre veszi a vizsgált szó összes előfordulását. A jelentések leírásakor nem vonatkoztat el a kontextustól, az utóbbit beépíti a jelentésleírásba. A korpusznyelvészetet nem egy elszigetelt nyelvi elem jelentése érdekli elsősorban, hanem a nyelvi elem és a kontextus közötti jelentésbeli viszony, illetve ezeknek a viszonyoknak az összessége. Sok esetben egy szónak nehéz megadni a jelentésdefinícióját, ilyen esetben a szó használatát tipikus példákkal érdemes illusztrálni. Ez különösen gyakori új szavak esetében, amikor a szó jelentése még nem állapodott meg teljesen (nem lexikalizálódott). Egy a korpusznyelvészetről szóló tanulmányból megtudhatjuk például, hogy a globalizáció szó a legkülönbözőbb jelentésekben használatos, amit a szerző különböző korpuszbeli mondatokkal illusztrál. A közeli szinonimák (a majdnem azonos jelentésű szavak) szétválasztása is csak korpusz alapján történhet. Az angol sorrow jelentése az angol-magyar szótár szerint 'szomorúság, bánat, bú, fájdalom', márpedig e négy szót nem használhatjuk egyformán tetszés szerinti kontextusban. A fájdalomnak van külső oka, a szomorúságnak nincs. A bánat tartós szomorúság, a bú pedig ma már inkább csak kifejezésekben fordul elő. A szótár nem említi, hogy a sorrow 'gyász'-t is jelenthet. A különféle jelentések csak gondos elemzéssel választhatók szét, ami megfelelő korpuszokat tételez fel. Mindebből az következik, hogy a korpusznyelvészeti megközelítés a kétnyelvű szótárak összeállításakor is nélkülözhetetlen.

Az írott nyelvi korpuszoknál talán még fontosabb a beszélt nyelvi korpuszok vizsgálata. Az ilyen korpuszok vagy spontán beszédet rögzítenek, vagy gondosan válogatott (szavakból, szószerkezetekből, mondatokból) szövegmintákból állnak, amelyeket több beszélő különböző akusztikai feltételek mellett mond ki. Az előbbi típusú korpuszok vizsgálata alapján megtudhatjuk például, hogy valójában hogyan is beszélünk (milyen szavakat, kifejezéseket használunk, beszédünknek milyen hangtani és alaktani sajátságai vannak, hogyan fest a beszélt nyelv mondattana). Az utóbbi típusú korpuszok alapján vizsgálják a beszéd akusztikai tulajdonságait, amelyeknek ismerete nélkülözhetetlen az automatikus beszédfelismerés szempontjából. Automatikus beszédfelismerés nélkül nem tudjuk a beszédet írássá alakítani, sem pedig a beszédet automatikusan egy másik nyelvre lefordítani. És természetesen a beszéd automatikus előállítása (beszédszintézis) szintén feltételezi a beszéd akusztikai tulajdonságainak ismeretét.

A fordítás

Térjünk most vissza a gépi fordítás kérdésére. A fordításhoz használt számítógépes eszközöket három kategóriába sorolhatjuk: (a) jó minőségű, teljesen automatizált gépi fordítás, (b) ember által támogatott gépi fordítás, és (c) gép által támogatott emberi fordítás. A teljesen automatizált gépi fordításról már az ötvenes évek végén kiderült, hogy nem megvalósítható. Az

ok egyszerű: a számítógép nem képes megérteni a fordítandó szöveget, márpedig ez a helyes fordítás legalapvetőbb feltétele. Hogy csak egyetlen példát említsünk: az első gépi fordítási kísérletekben oroszból angolra kívántak fordítani. Az oroszban nincsenek névelők, az angolban vannak. Kérdés: mikor kell az orosz szöveg angol megfelelőjében névelőt használni, és amikor szükség van névelőre, határozott vagy határozatlan névelőt kell-e használnunk? A kérdés eldöntéséhez nem elég az orosz mondatot megértenünk, a szövegösszefüggést is ismernünk kell. Hasonló jellegű problémát más nyelvpárok esetében is könnyen találhatunk. A jó minőségű, teljesen automatizált gépi fordítás tehát megvalósíthatatlan. Marad a (b) és a (c) lehetőség. Mivel azonban az emberi munka egyre drágább, inkább a (b) megoldást szokták választani, de azt is csak bizonyos korlátozásokkal. Egyrészt géppel csak műszaki-tudományos szakszövegeket fordítanak, másrészt az emberi beavatkozás a legfontosabb utószerkesztésre korlátozódik, amelynek feladata a durvább fordítási hibák és kétértelműségek kiküszöbölése. Ma több működő fordítási rendszert ismerünk, Az Európai Unióban a Systran-rendszert használják. Megjegyezzük, hogy EU támogatással nálunk is történnek előkészületek a Systran-rendszerbe való bekapcsolódáshoz.

Magától értetődik, hogy a gépi fordításhoz szükség van (a) egy megbízható morfológiai elemzőre, (b) egy jól működő szintaktikai elemzőre és (c) egy megfelelően kódolt információkat tartalmazó szótárra.

Szövegmegértő rendszerek

Az információszerző rendszerek (ide tartoznak a különféle szövegmegértő és szövegkivonatoló rendszerek is) a hagyományos morfológiai és szintaktikai elemzésen túl szemantikai-fogalmi összefüggéseket reprezentáló formalizmusokon alapulnak.

A szövegmegértő rendszerek eredetileg a nyelvészettől független célok megvalósítására törekedtek, és nem is mindig támaszkodtak a szövegek grammatikai elemzésére kidolgozott nyelvészeti eszközökre. A szövegmegértésben nagy szerep hárult a tudásreprezentációra, mindennapi ismereteinknek a számítógép számára érthető ábrázolására. A nyelvészettel való találkozás azonban elkerülhetetlen volt: a szövegmegértés nemcsak számítógépes, hanem a megismerés és a nyelv kapcsolatát vizsgáló kognitív nyelvészeti és a nyelvhasználat szabályszerűségeit kutató pragmatikaelméleti probléma is. A keret (frame) sztereotip szituációkat jellemző ismeretrendszer, minden kerethez tartozik egy forgatókönyv (script), amely az adott kerethez tartozó esemény részeseményeinek a sorrendjét szabályozza (Minsky1975). A vendéglő, a telefonbeszélgetés, az iskolai tanóra mind egy-egy keretet hív be a hozzá kapcsolódó forgatókönyvvel együtt. Ezen az elgondoláson alapulnak azok a korai nyelvészeti munkák, amelyek keretszemantika (?frame semantics') néven váltak ismertté (pl. Fillmore 1976). A keretszemantika szerint a szavak jelentésének leírásához igen gyakran a tudásreprezentációban használt kerethez hasonló információkra van szükség. Például az adás-vétel eseményéhez kapcsolódó nyelvi keret magában foglalja a vesz, elad, fizet, költ, kerül igéket és a pénz, fizetés, kereskedő, vásárló főneveket. Ezeknek a szavaknak a jelentése egy szemantikai-fogalmi keret részét alkotja, e keret nélkül jelentésük nem érthető. A keretszemantika ma is élő kutatási irányzat (pl. Konerding1994).

A szövegmegértés fejlettebb modellje már az események szereplőinek céljait is mérlegeli, és figyelembe veszi a célok között fennálló viszonyokat. A célok lehetnek például szembenállóak (6) versengők (7), beágyazottak (8).

6.ábra

```
Tomi el akart menni moziba, de másnap vizsgáznia kellett fizikából.
Tomi elment moziba.
Tomi megbukott fizikából.
Kérdés: Miért bukott meg Tomi fizikából?
Válasz: Mert tanulás helyett moziba ment.
```

7.ábra

```
Tomi meg akarta nyerni az autóversenyt.
Henrik meg akarta nyerni az autóversenyt.
A verseny elott Tomi tönkretette Henrik kocsiját.
Kérdés: Miért tette tönkre Tomi Henrik kocsiját?
Válasz: Mert meg szerette volna nyerni a versenyt.
```

8.ábra

```
Tomi és Éva házasok voltak.
Tomi meghalt.
Évának munka után kellett néznie.
Kérdés: Miért kellett Évának munka után néznie.
Válasz: Mert férje meghalt és valamibol meg kellett élnie.
```

A rendszer nemcsak rövid szövegek megértésére alkalmas, hanem a szövegre vonatkozó kérdések megfogalmazására és megválaszolására is. Ezekben az esetekben is nyilvánvaló a nyelvészeti pragmatika, illetve a szövegnyelvészet szerepe. A pragmatikaelméletből ismert relevancia- elv miatt a fenti szövegeknek koherenseknek kell lenniük, a kérdésre adott válasz erre a koherenciára épít. Szövegkivonatoló rendszerek

Utolsóként említjük a szövegkivonatoló rendszerek problémáját, amelyek gyakorlati alkalmazásuk miatt különösen fontosak. A szövegkivonatoló rendszerek legalább hatféle műveletet tételeznek fel (Prószéky1989, 271).

1. szegmentálás: a szöveg elemi egységekre való bontása;

2. reprezentáció: az egységek indexelése, az ige vonzatkeretének meghatározása;

3. osztályozás: a szövegnek előre megadott kategóriák segítségével megfelelő osztályba történő besorolása;

4. módosítás: a szöveg átírása;

5. konvertálás: a szöveg tartalmi elemeinek a szövegkivonatoló rendszer által előírt formátumba való hozása;

6. differenciálás: egy adott specifikációhoz illeszkedő elemek megtalálása.

A felsorolásból is kitűnik - különösen az (a) és (b) lépések esetében - a szövegkivonatoló rendszereknek a nyelvészettel való szoros összefüggése. Mindkét lépés feltételezi a szöveg morfológiai és szintaktikai elemzését. Nem kimondottan nyelvészeti jellegű a (c) művelet: a (c)

alapján soroljuk be például a szöveget az orvosi szövegek közé, és az orvosi szövegeken belül például a májbetegségekről szóló szövegek közé. A (d) művelet viszont ismét kapcsolódik a nyelvészethez: a (d) bizonyos transzformációk elvégzését jelenti, pl. a hiányzó ige behelyettesítését (pl. skin no eruptions bőr nincs erupció" skin showed no eruption a bőrön nem látható erupció", a mellérendelő szerkezetek felbontása mellérendelő mondatokká (pl. jobb mellkasi és felkari fájdalom jobb mellkasi fájdalom és jobb felkari fájdalom), a vonatkozó névmás helyettesítése teljes értékű főnévvel (pl. a törés, amely patologikus lehet a törés, úgy hogy az a törés patologikus lehet).Az (e) esetében egy mondat elemeihez különböző mutatókat rendelünk. Például A 80 éves kaukázusi asszony rosszullétről, hányingerről panaszkodott mondat logikai reprezentációja többek között az alábbi elemeket tartalmazza: [80 éves: MELLÉKNÉV KOR], [kaukázusi: MELLÉKNÉV FAJ], [nő: FŐNÉV NEM] stb. Itt egyrészt szintaktikai (főnév, melléknév, ige stb.), másrészt a szemantikai kategóriákat (kor, faj, nem) találunk.

A szövegkivonatoló rendszerek szinte kizárólag tudományos és műszaki szövegek számítógépes reprezentálására készülnek.