A szupermorféma

Nyelvtechnológia és szöveg

Kis Ádám
SZAK Kiadó Kft.
adam.kis@szak.hu

Kis Balázs
MorphoLogic Kft.
kis@morphologic.hu

Kivonat. Az előadás a szöveg, illetve meghatározott részeinek, szintjeinek körülhatárolásával, ennek számítógépes alkalmazásaival, illetve szintjeivel foglalkozik. Nem a számítógépes nyelvészetben hagyományosnak tekinthető nyelvtani, illetve tartalomelemzési szempontokat használja fel, hanem a szöveg mint komplex egység tartalmát vizsgálja. A három alkalmazási példa – a hivatkozási példa, a keresés és a fordítás – a szövegek tartalmi vizsgálatának problémáját a szövegek gépi összehasonlításának, illetve a mintaillesztésnek a korlátaira vezeti vissza: fő kérdése, hogy ezek a korlátok meghaladhatók-e, s lehet-e szövegeket tartalmuk alapján összehasonlítani, illetve keresni.

1. Bevezetés

A tudományos szövegmeghatározások a szöveg dimenzionális meghatározását lényegében teljesen szemantikai alapokon végzik. De Beugrande és Dressler meghatározásrendszere (de Beugrande-Dressler, 1923) 7 ismérvet sorol fel, amelyeknek teljesülniük kell minden „szövegszerűségen" ahhoz, hogy érvényes legyen rá a kommunikatív jelző, és ez elengedhetetlen feltétel, hogy ezt a nyelvi jelenséget szövegnek tekintsék. Mint írják, „...a nem kommunikatív szöveget nem tekintjük szövegnek".

A szövegtan rövid történelmében kialakult szövegmodellek (Van Dijk, Petőfi) mind olyan ismérvek alapján közelítik meg a szövegfogalmat, amelyek algoritmizálása minden, csak nem triviális (Tolcsvai, 2000).

A nyelvészeti szövegfogalom nagyon erősen kötődik a jelentéshez. „...a szöveg elsődleges rendeltetése valamilyen értelemreprezentáció" – írja Tolcsvai (2000).

A de Beugrande-Dressler-féle meghatározás sutasága egyenesen elvezet a számítógépes szövegkezelés problémájához: az nyilvánvaló, hogy a számítógépen a szöveg – karaktersorozat, de nem minden (nyelvi szempontból értelmezhető) karaktersorozat tekinthető szövegnek. Vajon léteznek-e számítógépes módszerek az emlegetett kétféle karaktersorozat megkülönböztetésére?

A szöveg a számítógépen mindenképpen egy karaktersorozat, amelyet nyelvi kódolással hoznak létre. Szöveggé valójában akkor válik, ha bár dekódolják. Egy karaktersorozat a számítógépen gyakorlatilag háromféleképpen határolható körül (lokalizálható, kereshető, jeleníthető meg):

1. Fájlként.
   Ez azt jelenti, hogy a sorozat valamennyi karakterét közös névtérben helyezik el, itt hozzárendelik egy egyértelmű és megismételhetetlen azonosítóhoz (ez a sorozatot tároló számítógépre vonatkozó korlátozás, ami viszont, az adott számítógép egyedi azonosítása folytán végeredményben kiterjeszti a teljes digitális térre).
2. A karaktersorozat valamely pontjának megadásával.
   - Az elterjedt szövegszerkesztők *egy* általános konvenciót tartalmaznak ebben a tekintetben: a **szó** határolását. Közismert, hogy a számítógépen a szó két **szóköz** közötti karaktersorozat. Ez a konvenció igencsak alkalmas például a grammatikai összefüggések vizsgálatára.
   - Ugyancsak megvannak a **mondat** elhatárolását végző kitüntetett karakterek (a mondatzáró írásjelek és a bekezdésjel), azonban ez figyelemmel a bekezdésjel tipográfiai funkcióira koránt sem egyértelmű (előfordul, hogy – tipográfiai meggondolás miatt – egy mondat akár 3 bekezdést is alkothat).
   - A szöveg létrehozója a szöveg tetszés szerinti pontját felszerelheti olyan jelöléssel, amely a későbbiekben megtalálható, és így azonosíthatóvá teszi a szövegrészt. Ilyen eszköz például a könyvjelző (bookmark).
   - Az előző pont kiterjesztésével szövegintervallumok kijelölésére is mód van, mégpedig oly módon, hogy a könyvjelző nem egyetlen szövegpontra hivatkozik, hanem egy összefüggő karaktersorozatra (melyet pl. a Word szövegszerkesztőben *kijelöléssel* fogunk egybe).
3. Minta segítségével. A karaktersorozat egy részét a rendszer úgy lokalizálja, hogy vizsgálja, megegyezik-e egy adott mintával.

A szövegnek tekintett karaktersorozatok körülhatárolása a számítógép alkalmazásának lényeges funkcióihoz tartozik hozzá. Ezek közül hármat elemzünk: a hivatkozást, a keresést és a fordítást.

2. A hivatkozás

A szövegek hivatkozásokkal való összekapcsolása egyáltalán nem a modern kor terméke. Ez az eljárás lényegében a lineáris olvasás kötöttségeinek feloldására szolgál. Voltaképpen az írás, illetve az írott szöveg közelítése a gondolkodás struktúrájához, amely sokkal inkább bonyolult hálóra emlékeztet, semmint szekvenciára. Ősi, nevezetes példa erre maga a Biblia, amely tulajdonképpen szabályos indexszekvenciális struktúrát alkalmaz a nemlineáris olvasás megadásához.

A Károli-Biblia ekképp kezdődik (Biblia, 1948)[4]:

> *A világ teremtése*
> *(v.ö. Zsolt 104)*
> *1. Kezdetben teremté Isten az eget és a földet.*
> ** rész 2.4.5. Zsolt 33, 6.8,9, 12, 135.5*

[4] Szent Biblia, az Istennek Ó és új Testamentumában foglaltatott Szent Írás.

Mint látjuk, a kétségtelenül alapvető nyitómondathoz egy sor helyhivatkozás tartozik, amely módot ad az olvasónak arra, hogy a szöveg szerkezet által sugallt sorrendtől eltérjen.

Így, a tartalomjegyzék segítségével ellapoz a Zsoltárok könyvéhez (az adott kiadás 493. oldalára, ott megkeresi a 33. részt (ebben segít, hogy az élőfejben megadják az adott oldalon levő rész számát), ott megkeresi a 6. szakaszt, ahol ezt olvassa:

6. Az úr szavára lettek az egek és szájának leheletére minden seregök.
** 1. Móz., 1,6.7*

Máris teljes a kereszthivatkozás, azonban ez technikai értelemben roppant nehézkes. Igazából csak azért alakulhatott ki, és azért használták, mert csak egyetlen (néhány) könyv volt az emberek birtokában, melyet igazából nem is olvastak, hanem „használtak". A Biblia hivatkozásai a pontos és következetes struktúrán, illetve jelölésrendszeren alapulnak. Ez a jelölésrendszer lényegében *metanyelvnek* tekinthető, hiszen következetesen áttevődik a jó fordításokba is, biztosítva azt, hogy a hivatkozásrendszer a különböző nyelvi változatokban egyaránt érvényesüljön.

Tekintsünk egy másik példát az irodalmi hivatkozásokra! A következő szöveg Umberto Eco egy tanulmányából vett facsimile részlet [Eco:1994]

jelzések gyakran igencsak félreérthetők. Carlo Collodi *Pinokkió*-ja így indul:

Kezdődik a mese: – Volt egyszer egy...
– Király! – szóltok közbe tüstént, kis olvasóim.
*Csakhogy, barátocskáim, ezúttal tévedtek. Nem királyról szól a mese. Hol volt, hol nem volt, volt egyszer egy darab fa.**

Roppant csavaros kezdés. Collodi először mintha azt jelezné, hogy mesébe fog kezdeni. S

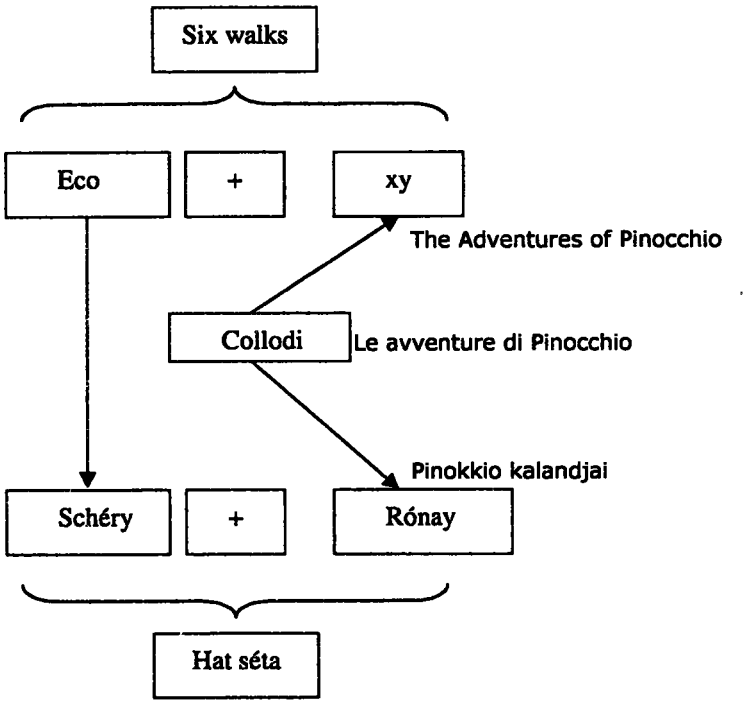
Mint látni fogjuk, ez a Bibilához képest igencsak bonyolult hivatkozásrendszer. A bonyodalmak azzal kezdődnek, hogy a beillesztett idézet – szemben azzal, hogy a bibliai szövegek végső soron ugyanannak a műnek különböző pontjaira hivatkoznak – egy egészen más műben őshonos. Míg a Bibliát kezelhetjük egyetlen strukturált szövegnek, addig Eco műve és Collodi *Pinokkiója* csak olyan alapon lehetnek egybefoglalható, ahogy minden ember rokon Ádámtól és Évától. A bonyodalmakat fokozza, hogy sem Eco tanulmánya, sem a Pinocchio, bár strukturált szövegek, nincsenek ellátva olyan könyvjelzőkkel, mint a Biblia. Mindehhez hozzájárul egy sajátos körülmény, mégpedig a szövegek nyelve.

A Biblia esetében egy-egy kötet rendszerint azonos nyelven szokott megjelenni, de mint említettük, a hivatkozásrendszer lehetővé teszi a különböző (korrekt) fordítások egységes kezelését. Esetünkben azonban – bár ez az első pillanatban nem érzékelhető, – a nyelvváltozatok figyelembevételével 5 szöveg is részt vesz a példában. Tételesen:

1. Eco angol nyelven írt szövege.
2. Collodi olasz nyelvű szövege.

3. Collodi szövegének angol fordítása[5].
4. Az Eco-szöveg magyar nyelvű fordítása.
5. Ezen belül a Collodi szöveg fordítása.

Nevezzük el a szövegeket a létrehozóikról (a fordításokat a fordítókról[6]), és vizsgáljuk meg a struktúrát!



Ezt a szövegkombinációt a mű valójában nem hivatkozásokkal oldja meg, hanem a megfelelő szövegek beemelésével. Az olvasó szempontjából ez tulajdonképpen kielégítő megoldás, hiszen a szövegek eredetiségi jellemzői inkább csak filológiai tekintetben érdekesek.

Van azonban a megoldásnak egy lényeges hátránya: az átemelt szöveg kikerül a szövegkörnyezetéből, és ezzel tulajdonképpen teljesen meg is változik. A Pinokkió néhány sora eredeti szövegkörnyezetében betöltött egyfajta funkciót, majd átkerült Eco szövegébe, és ott másfajta funkciót rendeltek hozzá. Figyeljük meg azonban, hogy ez a két funkció nem független egymástól. A Collodi-szövegrészletnek van egy olyan

[5] Nem volt alkalmunk meggyőződni arról, hogy vajon Eco a Pinocchio eredeti szövegét idézte, vagy annak angol fordítását, és ha az utóbbi történt, azt ki fordította angolra, Eco vagy más. Feltételezem azonban, hogy angol nyelvű volt az idézet, mert ellenkező esetben a hazai kiadói szokásnak megfelelően a fordításban is eredeti nyelven közölték volna, lábjegyzetben megadva a fordítást.

[6] A Collodi-szöveg angolra fordítóját XY-nak nevezzük (feltételezés szerint Nicolas Pewrella lehetett).

jelentése, amely lényegében szövegkörnyezettől függetlenül is létezik, értelmezhető. Ennek révén ez a szövegrészlet olyan viszonyba kerül az – otthagyott – környezetével, mint amilyen a mondatszintű nyelvstruktúrában a morfémák között szokásos. A szövegjelentés tekintetében morfémának tekinthetjük, és mivel szükségképpen a szó szoros értelmében vett morfémák halmaza, talán célszerű szupermorfémának nevezni.

Hozzá kell tenni, hogy a Pinokkió esetében ez a szövegrész azért válhatott szupermorfémává, mert maga a történet, és esetleg a szöveg is, közismert. Ugyanez fennáll a bibliai szövegeknél, ahol a hivatkozás mintegy figyelmezteti az olvasót (a szöveg „használóját"[7]), hogy az aktuális gondolatot kapcsolja össze egy másikkal, amelyet szintén ismer, illetve, ismerhet. A „szöveghasználatnak" más precedensei is vannak, pl. a jogszabályok olvasása. Nem véletlen, hogy a köznyelv az ilyen, mnemotechnikai jelentőségű kézikönyveket szívesen nevezi bibliának.

A számítógép a hivatkozások használatának relevánsan új lehetőségeit nyitja meg. Ha azonos szövegen belül akarunk hivatkozni, akkor módunkban áll a megfelelő hivatkozási pontról (melyet rendszerint a szövegekben eltérő színnel vagy aláhúzással, esetleg piktogrammal megjelölnek, de a folyamatos olvasás során jelöletlen is lehet, csak az egér ráhúzásával „bukkan elő") a szöveg egy másik pontjára ugrani. Ha nem egy szöveg (fájl) keretein belül vagyunk, akkor az úgynevezett hiperhivatkozások révén, az egész digitális kozmoszt átszelve juthatunk el vagy a hivatkozott fájlhoz (hogy ott más módszerrel keressük meg a szükséges helyet), vagy – az előző módszerrel kombinálva – annak egy pontjához.

Első pillanatra azt gondolhatnánk, hogy ez pusztán kényelmi szempont. Hiszen ugyanezt számítógép nélkül is el lehet érni: hogy megadjuk a hivatkozott helyet, oldalszámmal, illetve teljes bibliográfiával, és ebben az esetben a számítógép többlete „csak" annyi, hogy nem kell a megfelelő oldalra lapozni, illetve a könyvespolcon vagy könyvtárban keresgélni. A számítógép a hivatkozások mentén azonban többletet is nyújt: nemcsak a hivatkozott szöveget teszi elérhetővé, hanem annak környezetét is. Ez kegyetlen, de szükséges megoldás: megnehezíti a szövegek szabad értelmezésben való felhasználását, és megőrzi az eredetileg szándékolt jelentést.

3. Keresés

A számítógépes szövegkezelés másik kiemelkedő funkciója a keresés. Magát a funkciót itt nem ismertetjük, a Google-t vagy a Yahoo-t mindenki használja. Vegyük észre, hogy amikor szöveget keresünk a böngészők segítségével, lényegében ugyanaz történik, mint a hivatkozások esetében, csak éppen a tevékenységrendszer egy más pontján állunk. A hivatkozás során a hivatkozó szöveg szerzője kívánja a gondolatát valaki máséval kiegészíteni, és az olvasót ehhez a másik gondolathoz tereli. A keresés során magunk gondoljuk azt, hogy a gondolatunkat továbbviszi, illusztrálja vagy másféleképpen kiegészíti egy másik gondolat, és ezt próbáljuk megkeresni.

---

[7] A szöveg „használata" adott esetben azt jelenti, hogy a kognitivitást az emlékezet pótolja, azaz az olvasás adott esetben nem új ismeret szerzését, hanem valamely korábbi felidézését célozza. Természetesen ez sem zárja ki a kogníciót, mivel az ismeretek közötti új kapcsolatok új következtetések levonására alkalmasak. A hivatkozásokkal a szöveg linearitásán tudunk túllépni, a várt kontinuitás helyett vadonatúj kontiguumokkal bővítve világismeretünkt.

Ennek során két eljárásmenet van: az egyik az, hogy sejtjük, hol kereskedjünk. Emlékeink, információink vagy feltételezésein vannak arról, hogy az adott témával ki foglalkozott, és ekkor többé-kevésbé pontos adatokkal megkeressük az illető művet. A „civil" életben katalóguskutatással, könyvek átpörgetésével stb. A számítógépes világban fájlkereséssel, feltételezve, hogy a nekünk szükséges információt tartalmazó fájl a szerző nevével és/vagy a mű címével megtalálható.

Ez az egyszerűbb eset. Bonyolultabb az, amikor nem tudjuk pontosan a szerzőt, illetve a mű címét, és úgy próbálunk rátalálni a keresett szövegre. Ebben az esetben a keresőprogramok általában kombinációs segítséget nyújtanak, azonban itt találkozunk azzal a nehézséggel, hogy amíg az emberi agy képes bizonyos asszociációs íveket létrehozni egymástól formájuk tekintetében távol álló szövegelemek között (pl. a szinonimákat képes felcserélni), erre a számítógépnek kevesebb lehetősége van. Igazán sikert csak egzakt kereséssel lehet elérni: a keresési szándékot valószínűsítő keresés jobbára még a jövő zenéje.

Léteznek olyan megközelítések, amelyek megpróbálkoznak mind a tárolt szöveg, mind pedig a keresőkérdés tartalmának egyfajta ábrázolásával. (Vö. Prószéky, 2003). Ez rendkívül komplex feladat, és a jelenleg rendelkezésre álló számítógépes kapacitás mellett csak rendkívüli mértékben leszűkített tárgykörön belül alkalmazható – ugyanis a rendszerben szemantikai keretek felhasználásával létre kell hozni egyfajta világmodellt.

Egy másik lehetséges megoldás a keresendő és a keresett szöveg közötti kapcsolat ábrázolására másfajta, „közelítő" jellegű modellt felhasználni. Ekkor az absztrakciós szint sem a keresett, sem a keresett szöveg ábrázolása (inkább: transzformációja) esetén sem éri el a szemantikai szintet, azonban lexikális kapcsolatok megjelennek. Lexikális kapcsolatokat tezauruszok, illetve felszíni „ontológiák" (pl. WordNet) hoznak létre (Mihaltz, 2003). Ezeket a kapcsolatokat kell megjeleníteni és felhasználni a például teljes szövegű keresőrendszerben, amely így nemcsak a keresett kifejezés kulcsszavait, hanem az azokkal kapcsolatban álló szavakat, kifejezéseket tartalmazó szövegeket is megtalálja. (Vö. Prószéky-Kis, 1999, pp. 176-202.)

Nézzünk egy példát! Ha egy szövegben idézetet akarok szerepeltetni, általában emlékezetből beírom, azután megpróbálom a számítógép segítségével ellenőrizni, hogy helyesen idéztem-e. Ez a feladat nem igazán nehéz, ha pontosan emlékszem a szerzőre, a mű címére. Elég felütni a kötetet, és megtalálni. A gyakorlat azonban azt mutatja, hogy sokan szívesen idéznek, de a memóriájuk nem tökéletes. Milyen kellemetlen például egy ilyen eset (fiktív szöveg):

> *„... szépen fejezi ki ezt Radnóti: Világunknak elme kell nagy fénybe, mely igazodni magára mutat!"*

Ha a szerző biztos a dolgában, így hagyja, és az olvasók között biztosan lesz olyan, aki kapásból rájön, hogy ez nem Radnóti szövege, hanem József Attiláé. Ha ezt felismerjük, azért jó aggályosnak lenni, hátha így sem pontos az idézet. Beírjuk hát az egész szöveget a Google-ba. A válasz az, hogy ennek a keresésnek „egyetlen dokumentum sem felel meg." Ennek alapján arra kell következtetni, hogy ez a szöveg pontatlan, nem felel meg az eredetinek. A jelenlegi eszközrendszerben felhasználónak kell megfelelően „ravasznak" lenni ahhoz, hogy megfelelő választ kapjon. Olyan elemeit kell megkeresnie az idézetnek, amelyek egyrészről megfelelően szignifikánsak, azaz várhatóan elvezetnek az eredetihez, más részről kellően különösek ahhoz, hogy

belátható számú válasz keletkezzék. A nyelvérzék azt mondja, hogy ez a szócsoport: „elme kell nagy fénybe", megfelel ezeknek a kritériumoknak. A válasz meg is jön, az idézet helyesen *„társadalmunkba, elme kell, nagy fénybe', mely igazodni magára mutat*".

Mit kell ahhoz tenni, hogy ezt a ravaszkodást a gép elvégezze?

A számítógépnek képesnek kell lennie annak felismerésére, hogy két különböző szövegrész eléggé hasonlít egymáshoz. Mondhatjuk, hogy ez mindössze egy másik megfogalmazása annak a korábbi tételünknek, hogy a keresendő és a keresett szöveg közötti kapcsolatok ábrázolására van szükség. Megfelelő kapcsolat lehet például, ha a keresendő és a keresett szövegrész tartalmi (kulcs-) szavai – egy kis részük kivételével – megegyeznek vagy szinonimái egymásnak. A problémát azonban sokszor a szövegek matematikai jellegű összehasonlítására vezetik vissza, amely kizárja a szövegek nyelvi szerkezetének vagy tartalmának felhasználását, mégis sokszor alkalmas a hasonló szövegek közötti releváns kapcsolatok felfedezésére. Ez mellesleg bevett eszköz a fordítás számítógépes támogatásában.

4. Fordítás

Amikor a számítógéppel segített fordításról beszélünk, fontos hamar leszállni a fellegekről. A számítógépes fordítással kapcsolatban egy Alekszejev nevű professzor előadását hallgattam, aki elmondta, hogy a probléma 95%-ig meg van oldva, a fennmaradó 5% miatt azonban mindenképpen szükséges az emberi beavatkozás. Ez 1963-ban volt. Lehet, hogy Alekszejev túlzott a 95%-kal, azonban a megközelítés 40 év múltán sem jobb. Kétségtelen, hogy meghatározott igényszint kielégítésére alkalmas az ember nélküli fordítás, azonban a realitás mégis az, hogy a számítógép igazán hasznos két funkcióban lehet: vagy a szövegmegértés segítésében, minek során a cél igazából nem jól formált szöveg létrehozása, hanem az információ többé-kevésbé hiteles közvetítése nyelvi transzfer útján. A másik lehetőség az, hogy az egyszer megszületett emberi megoldások újbóli felhasználásának hozzáférhetővé tétele. A módszer lényege az, hogy amit egyszer lefordítottak, azt felesleges újra lefordítani, ha az eredeti szöveg valami miatt megismétlődik (és az a tapasztalat, hogy a szövegek mennyisége sokkal nagyobb, mint a szövegek fajtáinak száma, ami arra mutat, hogy az ismétlődés szükségszerű). Következésképp, ha egy forrásszöveget és egy fordítást együtt eltárolnak, akkor mód van arra, hogy egy újabb fordítási feladat során, megvizsgálva és megállapítva, hogy a forrásszöveg létezik a tárban, akkor annak újbóli fordítását nem kell elvégezni, elegendő elővenni a tárban hozzá asszociált tárgynyelvi szöveget. Ez a feladat voltaképpen nem különbözik a keresési feladattól. A haszna azonban mégis korlátozott, mert a forrásszöveg tekintetében gyakorlatilag 100%-os egyezés szükséges. Vajon lehet-e alkalmazni a módszert nem csak egyező, hanem hasonló szövegekre?

Erre a mai fordítástámogató programokban létezik megoldás, amely azonban tisztán matematikai, nincs nyelvészeti vonatkozása. A szöveget karakterkódok sorozatának tekinti, s ezeket a fuzzy logika elvei szerint hasonlítja össze. Jelenleg folynak kísérletek arra, hogy nyelvtani szerkezet, sőt, esetleg szemantikai tartalom alapján fedezzünk fel hasonlóságot (Kis-Lengyel, 2003; Gröbler-Hodász-Kis, 2004). A fordításban viszont nem ez jelenti az egyedüli problémát: a korábbi szövegek ugyanis fordításukkal együtt vannak tárolva, s amikor a forrásszöveghez „csak" hasonlót találunk, a

tárolt fordítás is legfeljebb hasonló lesz a kívánt fordításhoz. Így a gépi fordítástámogatásban az is kutatás tárgya, hogy a hasonlóság által reprezentált különbséget hogyan lehet csökkenteni, a tárolt fordítást hogyan lehet automatikusan átalakítani a tényleges forrásszövegnek megfelelően.

5. Összefoglalás

A szöveg számítógépes feldolgozása napjainkban nehéz feladatokat ró a mesterségesintelligencia-kutatókra. Neumann annak idején a számítógép lehetőségeit ekképp határozta meg

„...a tulajdonképpeni célkitűzést számokkal való műveletekkel kell először helyettesítenünk. Ez olyasmi, amit a gép maga nem tud elvégezni. Tehát előzetesen meg kell fontolni, hogy a kérdéses probléma hogyan fordítható le számokkal végzendő műveletekre." ([Neumann, 1954]:229).

A feladat világos: minden olyan szövegekkel kapcsolatos jelenséget, amelyet a szövegnyelvészet határozottan és mélyen szemantikai szinten határoz meg, le kell fordítani számokkal végzendő műveletekre. Ennyi. *Hic Rhodos, hic salta.*

Irodalom

BEUGRANDE, Robert De; DRESSLER, Wolfgang: Bevezetés a szövegnyelvészetbe. Corvina, É.n. [Beugrande-Dressler]

DIJK, Teun van: Some aspects of text grammmar. The Hauge: Mouton. in [Beugrande-Dressler] [Dijk]

ECO, Umberto (1994): Hat séta a fikció erdejében. Európa Könyvkiadó, Budapest. [Eco, 1994]

GRÓBLER, Tamás-HODÁSZ, Gábor-KIS, Balázs (2004): MetaMorpho TM: A Rule-based Translation Corpus. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004), Lisbon, Portugal.

KIS Balázs-LENGYEL István (2003): Új módszerek az emberi fordítás gépi támogatásá-ban. In: Az I. Magyar Számítógépes Nyelvészeti Konferencia gyűjteményes kötete, Szegedi Tudományegyetem, Szeged.

MIHÁLTZ Márton (2003): Magyar főnévi WordNet-ontológia létrehozása automatikus módszerekkel. MSZNY 2003, Szegedi Tudományegyetem, Szeged.

NEUMANN János (2003): A számítógép és az agy. In Neuman János Válogatott írások. TypoTeX, Budapest. [Neumann]

PETŐFI S. János (1990): Szöveg, szövegtan, műelemzés. Országos Pedagógiai Intézet, Budapest, 1990. [Petőfi, SZSZM]

PRÓSZÉKY Gábor (2003): NewsPro: automatikus információszerzés gazdasági rövidhírekből. MSZNY 2003, Szegedi Tudományegyetem, Szeged.

PRÓSZÉKY Gábor-Kis Balázs (1999): Számítógéppel emberi nyelven. SZAK Kiadó, Bicske.

Szent Biblia, az Istennek Ó és új Testamentumában foglaltatott Szent Írás. Brit és Külföldi Biblia Társulat és Magyar Biblia Társulat, Budapest, 1948. [Biblia, 1948]

TOLCSVAI Nagy Gábor (2001): A magyar nyelv szövegtana. Nemzeti Tankönyvkiadó, Budapest. [Tolcsvai, 2001]