PRÓSZÉKY GÁBOR[1,2] – INDIG BALÁZS[1]

[1] Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
[2] MTA–PPKE Magyar Nyelvtechnológiai Kutatócsoport

{proszeky.gabor, indig.balazs}@itk.ppke.hu

Magyar szövegek pszicholingvisztikai indíttatású elemzése számítógéppel

In this paper, we present the theoretical background and architectural overview of a novel framework for parsing Hungarian (short) texts. Our model aims to be psycholinguistically motivated, following our knowledge about the human language processing faculty. It is performance-based, it is designed to treat multiple sentences as text units, processes text strictly incrementally left-to-right, employs concurrent threads representing different knowledge sources and generates a non-tree representation.

1. Bevezetés

Írásukban egy az eddigi megközelítésektől több ponton is eltérő nyelvelemző rendszert ismertetünk, mely a következő alapelvek szem előtt tartásával készül.

a) *Pszicholingvisztikai indíttatású,* ami azt jelenti, hogy amennyire csak lehetséges, az emberi nyelvfeldolgozás (Pléh & Lukács, 2014) ismert mintáit követi.
b) *Performancia alapú* rendszerként minden olyan nyelvi megnyilatkozást megpróbál feldolgozni, ami (leírt szövegekben) előfordul (Prószéky et al, 2015), nem helyezve különös hangsúlyt az elméletileg létező, de a gyakorlatban meglehetősen ritka jelenségek kezelésére. Ugyanakkor bármilyen – rosszul formált, agrammatikus – szöveget igyekszik nyelvi megnyilvánulásnak tekinteni és értelmezni.
c) Szigorúan *balról jobbra* működve, szavanként dolgozza fel a szöveget. A még be nem olvasott, illetve el nem hangzott elemeket teljes mértékben ismeretlennek tekinti, rájuk semmilyen módon nem hivatkozik. Ha egy döntéshez nem elég az aktuálisan rendelkezésre álló információ, a rendszer továbblép és csak később dönt.
d) Az elemző architektúrája eredendően *párhuzamos*. A hagyományos megközelítésekkel szemben, ahol az elemzések általában egy láncot alkotó modulsor végén alakulnak ki, itt az éppen elemzendő szót folyamatosan, párhuzamosan jelen lévő szálak (morfológiai elemző, különböző grammatikai jelenségeket azonosító szálak, korpuszgyakorisági szálak, anaforafeloldó szálak, fókuszazonosító szál stb.) egyszerre vizsgálják és együttesen, egymással kommunikálva, olykor egymás esetleges hibáit javítva határozzák meg az elemzést.

e) Nem a mondatot, hanem a „rövid szöveget”, azaz az akár több mondatból álló *megnyilvánulást* tekinti reprezentálandó alapegységnek, lehetővé téve a mondaton belüli és mondatok közötti anaforikus viszonyok egységes kezelését.
f) Ennek megfelelően, illetve a különböző jelenségek egyidejű kezelése miatt a *reprezentáció* nem feltétlenül fa, hanem egy akár különböző típusú éleket tartalmazó összefüggő gráf.

Az elvi megalapozást követően az elemző alapvető megvalósítási lépéseit is bemutatjuk. Az elkészült mintaprogramra épülő példáink az alapelveket szemléltetik.

2. A performancia-alapú közelítés

A gépi nyelvészet kutatói hamar észrevették, hogy a nyelvészet utolsó évtizedeiben egyeduralkodónak mondható generatív modellek informatikai szempontból nem igazán nyújtottak hatékony megoldást a valóságban előforduló, azaz a nem feltétlenül tökéletesen szerkesztett szövegek elemzésére. Ennek az egyik – korán azonosított – oka, hogy a Chomsky (1957) által bevezetett és az ezt követő generatív technikákban a transzformációk nem invertálhatók. Azonban nem ez volt a fő ok, hiszen azóta már szép számmal léteznek a Chomskyétól eltérő, transzformációmentes generatív modellek is (GPSG, LFG, HPSG, TAG stb.). Ám a generatív közelítés elvei alapján ott nem játszhat szerepet a „hatékony elemezhetőség”, mert az nem a preferált kompetencia, hanem a performancia érdeklődési körébe tartozik.

A **performancia-alapúság** tehát számunkra elsősorban azt jelenti, hogy minden nyelvi megnyilatkozás feldolgozandó, ami előfordul. Ezzel szemben ami elvben ugyan lehetne, de valójában nem fordul elő egy megfelelően nagy és jó nyelvi fedést adó szövegkorpuszban, az valamilyen értelemben számunkra kevésbé lényeges. Az emberi nyelvfeldolgozás a nyelvi megnyilatkozással egy időben – ha tetszik: balról jobbra – halad, és igyekszik minden olyan információt felhasználni, mely a megnyilatkozás értelmezéséhez szükséges, még akkor is, ha az – a hagyományos grammatikai értelemben – nem feltétlen tökéletesen szerkesztett. Tehát rendszerünkben nincs mód a megnyilatkozások még el nem hangzott, vagy le nem írt részére hivatkozni, azaz legfeljebb feltételezni, valószínűsíteni lehet bizonyos még meg nem jelent összetevőket a már elhangzottak, leírtak alapján, egészen addig, míg a megnyilatkozás be nem fejeződik. Ez nem jelenti azt, hogy nem léteznek olyan megnyilvánulások, amelyek a legvalószínűbbnek tűnő elemzési megoldást „kijátszva”, olykor visszalépéses működésre kényszerítik az emberi elemzőt is, ám ezeket úgy tűnik, hogy a hétköznapi kommunikációban a grice-i maximák (Grice, 1975) betartásából következően a kommunikációban kerüljük, és inkább csak viccek, vagy szándékos félrevezetés alkalmával fordulnak elő. Ennek a bizonyítására nagyméretű szövegkorpuszokat használtunk (Prószéky et al, 2015), illetve az

interneten található magyar nyelvű tartalom összegyűjtésére magunk is elkezdtünk egy az immár tízéves Webkorpusznak megfelelő (Halácsy et al, 2004) mai szövegkorpuszt építeni. A korpuszfeldolgozásra irányuló kutatásunk egy másik fontos célja a modern grammatikaelméletek által sokat vizsgált, sokszor igen bonyolult – de a hétköznapi életben meglehetősen ritka – nyelvi szerkezetek előfordulási gyakoriságainak vizsgálata (Endrédy & Novák, 2013, Ligeti-Nagy, 2015).

Mint Prószéky (2000) utal rá, a nyelvi szerkezetek elemzés közbeni kiválasztása közben hozott döntéseink felül tudják bírálni a lexikont. A korábban kialakított nyelvi ismereteket összegző szótárakat és az eddig leírt szintaktikai mintázatokat adatbázisként használó szabályalapú elemzők és az egyes nyelvi konstrukciók korábbi gyakoriságára építő valószínűségi elemzők ezekre a „múltbéli" statisztikákra alapozva tudják meghozni döntésüket, az aktuális mondat „extra-lingvisztikai" (például adott esetben a elhangzás helyszínére vonatkozó) környezetét nem veszik figyelembe. (1. példa) Kiinduló hipotézisünk ugyanis az, hogy a nyelvhasználó fejében mindkét – a korábban megtanult szerkezetekre építő és az aktuális helyzet alapján döntéseket hozó – rendszer egyaránt él. Az utóbbi az elhangzó nyelvi elemek valós idejű feldolgozását akkor is képes megvalósítani, ha a „megtanult" szerkezetek az aktuális ismereteknek ellentmondó (például egymáshoz nem illeszkedő jegyszerkezeteket tartalmazó) nyelvtani információkat hordoznak. Ilyenre példa az (1) mondat, ahol háttértudásunk szerint a kutya harap, a postás pedig általában nem, vagy legalábbis nem a kutyát.

(1) *A postás megharapta a kutyát.*

- ha iskolai dolgozatban fordításként fordul elő, akkor valószínűtlen, és megpróbáljuk korrigálni úgy, hogy egy *A kutya harapta meg a postást* alakú mondat legyen belőle
- ha egy bulvárlapban olvassuk, akkor valószínűbbnek tekintjük, és megpróbáljuk szó szerint értelmezni

3. Pszicholingvisztikai és nyelvészeti motiváció

Több mint negyven éve jelentek meg azok a megértési stratégiák, melyeknek már lehetett, vagy inkább lehetett volna számítógépes implementációjuk. A generatív grammatika hatvanas évekbeli előretörése idején fontos volt Bever (1970) megállapítása, hogy a megértési folyamat nem a generatív levezetési folyamat egyszerű megfordítása. Ő stratégiákat adott meg, melyek olykor egymással versengve, és valószínűségi alapon működtek. Mi több, az elméleti nyelvészetben a mai napig uralkodó szemléletnek meglehetősen ellentmondóan felülbírálható, azaz nemmonoton folyamatokként vezette be ezeket. Már Bevernél megjelenik a hatékony elemezhetőség gyakoriságra való visszavezetése mellett a lassabb, de olykor megkerülhetetlen nyelvtani szabályrendszer használata. Ter-

mészetesen a nyelvtani szabályok is alapvetően a gyakorisági információk szimbolikus összegzéséből alakulnak ki. Még egy olyan „felsőbb döntéshozó" szerepét is bevezeti, aki a háttértudás és az aktuális nyelvi esetlegességek egymásnak való ellentmondása esetén a végső döntést meghozza. Kimball (1973) algoritmizálta az addigi ismereteket, és gyakorlatilag egy számítógépes elemzési stratégiát foglalt össze hét pszicholingvisztikai elv segítségével. Az ő közelítése természetesen nem volt független az akkor egyeduralkodó generatív grammatika gondolataitól és az emberi nyelv szinonimájaként használt angol nyelvtől. Frazier és Fodor (1978) aztán a nehezen elemezhető mondatok és a szerkezeti többértelműségek vizsgálata közben mondatelemzéshez olyan stratégiai közelítést javasolt, melyben az egyes nyelvi elemek mondatbeli szerepének meghatározásához nem mindig elegendő az aktuálisan rendelkezésre álló információ, ezért a végleges döntést olykor a bemenet továbbolvasásával, azaz bizonyos késleltetéssel lehet csak meghozni. A nyolcvanas évekre kialakuló interakciós megértései modellek (bemutatja: Pléh, 1998) hoznak az eddigi eredményekhez képest új gondolatokat is. Az egyik, hogy figyelnek a feldolgozás sebességére, másrészt észreveszik, hogy a világismeret olykor az elemzés alsó szintjeinél is jelen van, ha egy-egy döntést hatékonyan szeretnénk meghozni a megértés menetében tehát minden rendelkezésünkre álló információt azonnal felhasználunk, mihelyst szükségünk van rájuk, azaz az eddig hierarchikusnak gondolt nyelvi szintek egymással interakcióban működnek. Pléh (1998) a többféle megértési stratégia részleges összeegyeztetését ismertetve három folyamat együtteseként írja le a megértési folyamatot: ezek a **szófelismerés**, a **mondatmegértés** és a **szövegelemzés**. Szerinte a teljes megértést végző rendszernek a szóba jövő nyelvi szerkezetek leírására és ezek „kiszámítási sorrendjére" kell megoldást adnia, valamint azoknak a nyelvi jegyeknek az összegyűjtését kell megoldania, melyek alapján a fenti szerkezeteket a bemenet elemeit valós időben képes beazonosítani.

Vannak egyébként más olyan nyelvi jelenségek is, melyek feldolgozását az imént bemutatott szintaktikai szerepek kialakításával egyidejűleg végzi a rendszer. Ilyenek például az **aktuális mondattagolás** vagy a visszautalást tartalmazó, ún. **anaforikus elemek** kezelése. Az előbbi a mondatban közölt információnak ismertként és újként való megkülönböztetése a magyarban a szórend és bizonyos hangsúlysémák segítségével. A hangsúlyt az írott szövegekben nem jelöljük, így a nyomatékos szavakat tartalmazó mondatok felolvasásakor az ember az életében addig hallott hasonló mondatok hangsúlysémáját használja, viszont a pusztán csak a betűkkel dolgozó számítógépnek ilyen emléke nem lehet. Marad tehát a mondatban elrejtett olyan információmorzsák összeszedése, melyek segítségével bizonyos biztonsággal mégis ki tudjuk jelölni az ismert és az új információ határait. Ilyen például a mondat főigéje előtti szerkezet helye,

mely az esetek többségében a közölt információ legfontosabb elemét tartalmazza. Az ez előtt és az ige utáni pozíciók általában nem rendelkeznek hasonló kitüntetett információtöbblettel. A fókusz helyét azonban csak a főige megjelenésekor tudjuk kijelölni: ez lesz az ige előtti pozíciót kitöltő szerkezet. Ami ettől balra jelenik meg, az pedig nagy eséllyel a topik. (2. példa)

(2) *Hétfőn csökkenéssel nyitottak a főbb európai értékpapírpiacok.*

| Hétfőn | csökkenéssel | nyitottak | a | főbb | európai | értékpapírpiacok. |

hétfő+n csökkenés+sel nyit+ottak a fő+bb európai értékpapírpiac+ok

V+Fin

Fókusz

Topik

A visszautalás kezelése sem a már kész szerkezeten, hanem a bemenet szintaktikai és mondattagolási elemzésével egy időben történik, akár egyszerű névmásfeloldásról, akár szinonim megjelölésről, vagy más visszautalási típusról van szó (3. példa)

(3) *Ismét nyert a Real Madrid: a királyi klub Granadában győzött 1-0-ra.*

| Ismét | nyert | a | Real Madrid | : | a | királyi klub | Granadában | győzött | 1-0-ra. |

4. Az pszicholingvisztikai motivációjú számítógépes elemzés alapjai

Az általunk megvalósított rendszer egy analitikus grammatikán alapul: innen a neve is: ANAGRAMMA. AZ elemző a Pléh (1999) által megfogalmazott elvek figyelembe vételével, értelemszerűen balról jobbra halad végig a nyelvi alapelemeken, amik a mi jelenlegi megvalósításunkban a **szavak**. Feldolgozza tehát a soron következő szót, tekintetbe véve az összes futó szál által szolgáltatott információt, majd (a) lezár, (b) elindít vagy (c) változatlanul hagy szükséges szálakat. Azt állítjuk, hogy a különböző nyelvi aspektusokat figyelő szálak együttműködésének mellékhatása a morfológiai egyértelműsítés és a kombinatorikus robbanások idejében történő megelőzése, mely utóbbi jelenség a szabály-alapú rendszereknél gyakran felmerül a hosszabb mondatok feldolgozása folyamán, akár még morfológiailag egyértelműsített tokenek esetén is. Ugyanezzel a problémával a statisztikai rendszerek a lokális optimalizálás segítségével szándékoznak megküzdeni, ami sokszor teljesen félre viheti a ritka szerkezetek elemzését.

A nyelvi bemenet feldolgozása diszkrét időpontok egymásutánjában történik. Modellünk az írott szöveg szavait – technikai szempontból betűközzel elválasztott egységeit – tekinti a feldolgozás alaplépésének. Más szavakkal azt is mondhatjuk, hogy egyfajta **órajelnek** tekinthetjük a bemenet szavainak egymásutánját.

Az első **feldolgozási szál**, ami minden órajel-lépés után elindul, a **morfológiai elemzés.** Ezt az egyszerűség kedvéért, illetve a szintaxisra való

koncentrálás miatt belső időfolyamatok nélkülinek, monolitikusnak képzeljük el, bár az emberi információfeldolgozásban ennek a modulnak a működése is – interakcióban a többiekkel – nyilvánvalóan több lépésben valósul meg. A morfológiai elemzés létrehozza azokat a jegyeket, amelyek segítségével a magasabb szintű elemzés folytatódni tud. A morfológiai elemzés létrehozza a szó lemmáját és megadja a további elemzési lépésekhez szükséges kiinduló morfoszintaktikai jegyeket. Ezek részben **kereslet** típusúak, azaz az ilyen szálak igényt jelenthetnek be bizonyos jegyekre, részben pedig a **kínálat** formájában megjelenő szálak a korábbi vagy továbbiakban megjelenő modulok keresletigényét elégítik ki.

A 4. példában látható mondat elemzési lépésein keresztül mutatjuk be az ANAGRAMMA-rendszer működésének alapjait.

(4) *Beüzemelték a Balaton vihar-előrejelző rendszerét.*

Beüzemelték	a	Balaton	vihar-előrejelző	rendszerét.
0	1	2	3	4 5
be|üzemel	a	Balaton	vihar-előrejelző	rendszer
Fin	Det+def	N	Adj	Acc
Past		Pers?		N
Nom?+Pl+3				Pers+Sg
Acc?+Def?				

Elsőként a *beüzemelték* szót olvassa be a rendszer. Ez egy finitum-alak, azaz a mondat főigéje. Ez az információ (FIN) mint kínálati szál jelenik meg, ám a legtöbb esetben ez nem is lesz alárendelve más csomópontnak. (hacsak nem ágyazódik egy másik tagmondatba). A jelen igealak töve a *beüzemel* igekötős ige, mely a vonzatkeret-szótárból a NOM? és az ACC? esetvégződésekre vonatkozó szálakat indítja el mint további keresletet, hiszen a szó tipikus használatakor egy alany és egy tárgy jelenik meg a megnyilatkozásban valahol az ige körül. Az alany specifikációjából annyi már világos, hogy az a finitum-végződésből kiolvasható többes szám harmadik személyű nominális szerkezet lesz. Ez a jelenleg szemantikusan meglehetősen üres információ azt fogalmazza meg, hogy várunk még egy nominatívuszi alakban álló többes szám harmadik személyű (NOM?+PL+3) főnévi csoportot. Ha találunk ilyet, akkor az alany kiegészül annak konkrét tartalmával (pl. *Beüzemelték a gyártók a …*), ha pedig nem – amint a jelen példamondatban is –, akkor az alany meghatározatlan marad a megnyilatkozás végéig, azaz általános alanyról beszélünk. A *beüzemel* tárgyát is keressük (ACC?+DEF?), ami határozott is kell legyen (ezt jelzi a DEF? kereslet), hiszen ellenkező esetben a *beüzemeltek* alak állt volna itt.

A következő órajelre az *a* névelő lexikális információját hozza be a morfológiai elemző: a DET szófajt mint kínálati szálat és a most megkezdett főnévi csoport határozottságára utaló DEF jegyet. A determináltságra vonatkozó információkat csak a főnévi csoport lezáró végződése (esetrag vagy névutó) felismerésekor kapcsoljuk majd a szerkezet fejéhez.

E pillanatban tehát két várakozó keresletünk van: a két igevonzat, melyek közül az alany már nem kötelező, hiszen az igevégződésből már beazonosítottuk, és a tárgy. Ezen kívül egy kínálat jelent meg, a DET, melyet majd a későbbiekben egy determinánsra vonatkozó kereslet fog a megfelelő csomóponthoz kötni.

A következő bejövő szóalak a *Balaton*. Ez egy 0 végződésű főnév (N). Az elemzés jelen állapotában a *0*-ról nem lehet tudni, hogy a mondat alanya vagy egy esetleges birtokos szerkezet jelöletlen birtokosa, ám a mondat alanyáról azt már tudjuk, hogy többes számú, így a – tényleges elemzési lépéseket a példa kedvéért kicsit leegyszerűsítve – a *0* végződést elkönyvelhetjük a birtokosra utaló végződésnek, mely egy birtokos személyrag iránti keresletet jelent (PERS?).

A *vihar-előrejelző* szóalak töve ő maga, és szófaja melléknév (ADJ).

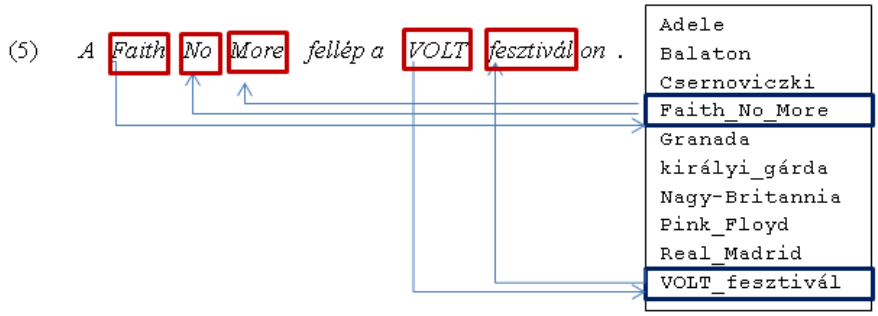
Az utolsó szó a *rendszerét*. Ennek töve a *rendszer*, szófaja főnév (N), ami a szótól balra előforduló esetleges jelzőket (ADJ) magához kapcsolja. Ez úgy történik, hogy bal felé egy opcionális „jelzőkereslet”-szálat indít bocsát ki (<ADJ?), és az ott található ADJ kategóriájú elemeket ADJ címkével magához kapcsolja. Így válik a *vihar-előrejelző* a *rendszer* jelzőjévé. A *rendszer* egyébként kínálatként az ACC jegyet ajánlja fel az elemzésnek, mely azonnal összekapcsolódik a mondat első szavának megjelenése óta ott várakozó ACC? kereslettel. Magyarul: a *rendszer* (egészen pontosan: a *rendszer* fejjel rendelkező csoport) lesz a *beüzemel* tárgya. ez a tárgy lehet határozott vagy sem, részben attól is függően, hogy áll-e határozott névelő a most lezárult szerkezet bal szélén. Az esetrag (most az ACC) elindít egy <DET? keresletszálat, ami sikeresen megtalálja az *a* névelőt, majd az ott talált DEF jegy hatására egy DEF kínálatszálat indít, amely azonnal összekapcsolódik a *beüzemel* tárgyára vonatkozó határozottnévelő-igénnyel. A tő és az esetrag között megjelenő birtokos személyrag (PERS) pedig mint kínálat azonnal összekapcsolódik a korábbi *Balaton* szó PERS? keresletével.

A szó végén szereplő pont nem része a lexikális alaknak (ahogy például a *stb*. szónak a pont része volna), így a mondat lezárul, és a *beüzemel* alanya most már végérvényesen az általános alany lesz, mivel más többes szám harmadik személyű alak nem állt a mondatban.

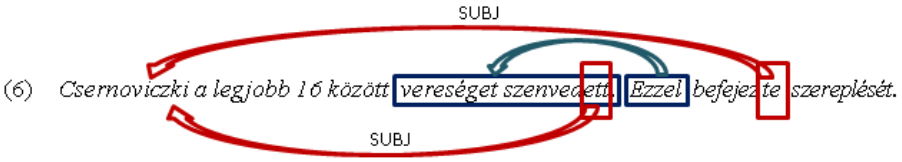
5. Az egyes nyelvi jelenségek kezelése az ANAGRAMMA-rendszerben

Elemzőnkben – mint a példából is látszik – egyidejűleg jelentkeznek keresletek és kínálatok, azaz a feldolgozás nem a hagyományos soros architektúra, hanem egy meglehetősen jól párhuzamosítható elképzelés mentén alakul ki. Az ANAGRAMMA rendszerben nemcsak a bemutatott „tisztán” morfológiai és szintaktikai relációkat leíró szálak, hanem akár statisztikai információk, korpuszgyakoriságok, vagy éppen ontológiák, világismeretek is párhuzamosan tudnak működni külön-külön**.**

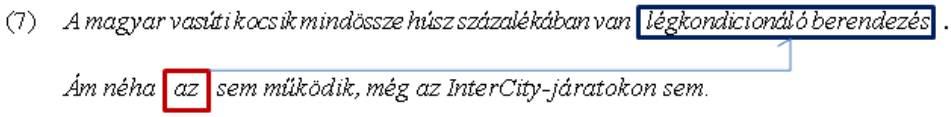
Az ANAGRAMMA elemző az emberi nyelvfeldolgozás hatékonyságából kiindulva igyekszik elkerülni a kombinatorikus robbanást, ezért használja az előismeretek összegzéseként kialakított **statisztikát**: a gyakori szerkezetek sokszor elemzés nélkül, kész belső szerkezettel jelennek meg a feldolgozásban. Informatikai szakszóval ezt gyorsítótárazásnak (angol szóhasználattal: cache-elésnek) mondanánk, ám a jelenség a pszicholingvisztikában (idegen szakkifejezéssel „Gestalt" néven) jól ismert (Pléh & Lukács, 2001). Az emberi nyelvértelmezés esetében ezt **egészleges feldolgozásnak** nevezik. Így tároljuk az akár több szóból álló tulajdonneveket, az idiómákat, de sokszor a nagyon gyakori tipikus nyelvi fordulatok szószerkezeteit is. Az ANAGRAMMA-rendszer tehát állandóan figyeli a világismeretből, nyelvismeretből adódó nagyobb egységek megjelenését, és ha ilyet észlel, megpróbál az órajel mentén továbbhaladni, hogy az a – kezdőszelete alapján valószínűsített – több szavas kifejezés teljes egészében megjelenik-e. Röviden azt mondhatjuk, hogy kár volna szavankénti részletes elemzést végeznünk addig, míg meg nem bizonyosodunk arról, hogy a – lépésenkénti belső elemzést nem igénylő – kifejezés teljes egészében jelen van a mondatban. Az 5. példa az ilyen egyszerű szerkezetű, de nem a nyelvi alapszótárba tartozó többszavas kifejezések feldolgozását mutatja.



Az ismert kompetencia-alapú szintaktikai modellek a nem nyelvi információfeldolgozó alrendszerekkel „természetüknél fogva" semmilyen együttműködést nem feltételeznek. A performancia viszont nem választható el más kognitív folyamatoknak a nyelvre gyakorolt hatásától (v.ö. Frazier & Fodor, 1978 és Pléh, 1998), ezért az első elemzési lépéstől kezdve az ANAGRAMMA-módszer a nyelvi, és a modell kidolgozottságától függően bizonyos nyelven kívüli modulok (világismeret, hangulat stb.) párhuzamos kezelésére épít. Ráadásul, a szokásos megoldásoktól eltérően, nem egyes mondatokat, hanem teljes **megnyilvánulásokat** (egy gondolategységet átfogó, általában bekezdésnyi szövegeket) dolgozunk fel (6. példa), például a mondatokat összekötő egy-egy konjunktív elem jelenléte vagy hiánya nem okozhatja az azonos tartalom felszíni különbségek miatti radikálisan különböző feldolgozását, pusztán a mondathatárok különbözősége miatt.



A szerkezeti reprezentációk irányított élei – mint az illusztrációkból is látható – a **függőségi** nyelvtanokéra emlékeztetnek (Tesniére, 1957), a rendszer működése pedig azok inkrementális elemzéssel működtetett változataira (Menzel, 2013). Mivel a mondathatáron nem feltétlenül záródik le minden elemzési lépés, a részszerkezetek teljes összekapcsolása nem feltétlen egyetlen mondaton belül valósul meg. A **referenciális elemek** például az ugyanezen reprezentációban megjelenő, de a hagyományos generatív felfogás koindexálására emlékeztető éleket vezetnek be az ANAGRAMMA-reprezentációkba (7. példa). A szövegben előforduló események szereplőinek azonosítása, és koreferenciaviszonyaik meghatározása azért fontos, hogy a végső reprezentációban minél pontosabban lehessen látni, hogy mely szereplők azonosak a világban („ki kicsoda?"). Más szóval, szeretnénk helyesen kezelni, hogy mely szereplő „új" a szöveg egy adott pontján való megjelenésekor, és mely nyelvtani elem utal egy korábban már megjelent szereplőre, illetve van-e, és ha igen, milyen kapcsolata a korábbiakkal.



Az **elliptikus** jelenségek megfelelő kezelése egy másik ok, amiért megengedjük az elemzőnknek, hogy átlépjen a mondathatáron. Úgy véljük, hogy az elemzésnek nem szabad megállnia a mondatok végén, mert az egymagukban álló mondatokkal szemben a hosszabb megnyilatkozások az emberi kommunikáció természetes egységei. Az egymást követő mondatok témája sokszor azonos, ezért a természetes emberi kommunikáció során lehetséges – és többnyire meg is történik – az egyes elemek kihagyása (azaz: az ellipszis), ami a legtöbb hagyományos elemzőnél – akár egyetlen mondaton belül is (8. példa) – komoly problémákat okoz. A rendszerünk által feldolgozandónak szánt nyelvi egységek néhány mondatból álló összefüggő szövegek, a sok mondatból álló, nagyobb művek feldolgozását egyelőre nem szándékozzuk megcélozni.

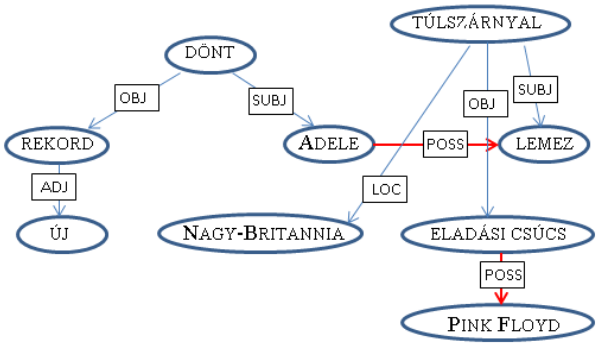
Az ellipszissel rokon jelenség még a **konjunkciós szerkezet,** mert az csak akkor azonosítható egyértelműen, ha egy konjunktív elem a bemeneten ténylegesen feltűnik. Ez lehet kötőszó (pl. *és*, *vagy),* vagy épp egy erre szolgáló

vessző, mert ezek vezetik be a konjunktív szerkezet következő tagját (9. példa). Ha a rendszer felismer egy ilyen elemet, akkor (de csak akkor!), módosítania kell az utolsóként feldolgozott elem reprezentációját a felismert szerkezetnek megfelelően, hiszen az előző elem volt ennek a konjunktív szerkezetnek az első tagja, amit az előző lépésben, annak feldolgozásakor még nem tudhattunk róla. A konjunkciót egyébként egyetlen egységként kezeljük, anélkül, hogy állást foglalnánk arról, hogy van-e az ilyen szerkezeteknek feje.

(9) *A spanyol mozdonyvezető telefonon beszélt és térképet nézegetett.*

A mondatok egyes részeinek referenciális alapon való összekötése (vonatkozó névmások, visszautalások kezelése stb.), az aktuális mondattagolás egységeinek felismerése és a szintaktikai függőségek megadása együtt egy sajátos összefüggő gráfot eredményez. Kimenetként tehát nem pusztán a szintaktikai, hanem más jellegű információkat is megkapunk, hiszen az elemző célja beazonosítani a nyelvi szituáció összes szereplőjét és a velük történt eseményeket, meghatározva a szükséges koreferencia-viszonyokat is. A rendszer végül is egy olyan, a mondatot, illetve a bekezdést reprezentáló **összefüggő irányított gráfot** hoz létre, amelynek segítségével válaszolni tud majd az olyan kérdésekre, hogy például ki, mit csinált, hol és mikor (10. példa).

(10) *Adele újabb rekordot döntött: lemeze Nagy-Britanniában túlszárnyalta a Pink Floyd eladási csúcsát.*



6. Az AnaGramma-elemzés kialakítását célzó kutatási alprojektek

6.1 Automatikus korpuszépítés

Az internetről származó szövegek begyűjtésekor számtalan probléma léphet fel. Az oldalakról ki kell vágni a fölösleges elemeket (boilerplate). Erre a feladatra készített eljárásunk a **GoldMiner-algoritmus**: (Endrédy, 2014). Ez a modul képes megkülönböztetni a különböző minőségű szövegeket is egymástól, tehát a

portálok lektorált szövegét a hozzászólók lektorálatlan anyagaitól. Ha nem keverednek össze a különféle módokon roncsolt szövegek, akkor az egyes szövegrészek konzisztensen rendelkeznek egy-egy hibával és könnyen javíthatóak, normalizálhatók maradnak. Például az ilyen hibák kiszűréséhez szükséges eldönteni, hogy az adott szöveg magyar nyelvű-e, fenn áll-e karakterkódolási probléma, illetve szükséges-e esetleg az ékezetek visszaállítása a szavakon.

6.2 Beavatkozások a normától való eltérés esetén

Az **ékezetesítés** problémájára a kutatócsoporton belül két független megoldás is létezik. Az egyik SMT-alapú (Novák & Siklósi, 2015), azaz tisztán statisztikai, míg a másik szabályalapú, mely morfológiai elemzés segítségével állítja vissza az ékezeteket (Endrédy-Novák, 2015).

Az elgépelések egy robusztus SMT-alapú **helyesírás-javító programon** átfuttatva érik el végleges állapotukat (Siklósi, Novák & Prószéky, 2015), melyben a forrásnyelv szerepét a tipikus módon elgépelt szövegek nyelve veszi át, míg a célnyelvi oldal a helyesírási szempontból javított szövegek nyelve lesz. A statisztikai gépi fordító tehát az előbbi nyelvről az utóbbira fordít. Ezzel olyan hibák is javíthatóvá válnak, amik a hagyományos szóalapú módszerekkel nem lennének javíthatóak (különírási hibák és a kontextusfüggő fals-pozitív elgépelések). Az elgépelést helyreállító és az ékezetesítő modulokkal kezelt szövegek már készen állnak az elemzésre. Szükség esetén ezek a modulok párhuzamosíthatóak is az elemző architektúrájának köszönhetően.

A szükséges beavatkozások elvégzése előtt az adott szöveg minőségére vonatkozó adatokat meg tudjuk adni a csoporton belül – eredetileg a fordítások minőségének meghatározására kifejlesztett – **minőségbecslő** szoftver segítségével. (Yang. Laki & Prószéky, 2015). Az elemző bemenetén várhatóan nem a korpuszokból ismert roncsolt szövegek jelennek meg, legalábbis ezt feltételezzük. Amennyiben a bemenet mégis eltér a normától, a feldolgozó rendszernek döntenie kell arról, hogy „megküzd-e" az eredeti szöveggel, vagy a benne valószínűsíthető devianciákat előbb kijavítja, és az elemzést csak a javított szövegen kezdi meg. Ezt a döntést készíti elő a minőségbecslő szoftver, melyet a bemenő szövegekre is lefuttatunk, és a kapott minőségértékek ismeretében döntünk a bemeneten történő esetleges beavatkozásokról.

6.3 Főnévi csoportok kinyerése az annotáció pontosításával és mondatséma-építés

A főnévi csoportok belső szerkezetének részletes leírását az InfoRádió-korpusz segítségével készítettük el (Ligeti-Nagy, 2014). Az annotációhoz használt morfológiai elemző címkéit (Prószéky & Kis, 1999) kellett pontosítani, hogy az új címkerendszer lehetővé tegye az NP-k nagy pontosságú kiemelését a szövegekből. A minimális NP-k nem tartalmaznak melléknévi igeneveket, konjunktív

elemeket és birtokos szerkezetet, míg a maximális NP-k a minimális NP-kből a most felsorolt három szerkezet kombinációjának segítségével állnak elő.

A szintaktikai mintázatok szövegkorpuszokban való vizsgálata (Endrédy & Novák, 2013) segítségével mondatvázakat nyertünk ki azáltal, hogy a maximális főnévi csoportokat egy elemmé zsugorítottuk, így kapva információt a mondatok valós szerkezetéről. A létrejövő mondatvázak „előelemzésével" a mondatfeldolgozási feladat tovább egyszerűsödik, mert a bemeneten jövő szavak sorára csak rá kell illeszteni a megfelelő mondatvázakat. Az illeszkedő mondatvázban az üres mezőket ki kell tölteni valódi, a bemenetről származó elemekkel, hogy előálljon a végleges reprezentáció.

6.4 Többszintű n-gram alapú statisztikák

Nyelvmodellünk kialakításához létrehoztunk egy faktoros nyelvmodellekre emlékeztető statisztikát, amelyben a gyakori egymást követő szavakból álló olykor különböző minőségben kötött – a kezdeti kísérletekben lemmából és szófajcímkéből álló – szerkezeteket egy **többszintű n-gram** modellel határozzuk meg. Így a – háromnál akár jóval hosszabb – n-gramok segítségével további gyakori szerkezeteket ismerhetünk fel (Indig & Laki elők.). A kapott minták a kitöltésnek gyakran csak egy-egy „faktorától" függenek. Például az *esik szó valamiről* kifejezés az *es-* szó *-SUB* alakú „vegyeshármasból" áll, azaz egy (tetszőleges igeidejű) szótőből, egy felszíni alakból és egy adott toldalékkategóriára végződő szóból. Szükségtelen tehát ilyenkor az ezeket lefedő redundáns esetek külön tárolása, ahogy ez a hagyományos módszerek esetében történni szokott. Továbbá, mivel csupán a gyakori szerkezeteket tekintjük, ezért az általánosan elfogadott trigram-szerkezeteknél hosszabb, egybefüggő és a szükséges mértékben kötött szerkezetek is feltérképezhetők modellalkotás céljából, a kombinatorikus robbanás elkerülése mellett. Ez a tulajdonság élesen elválasztja nyelvi modellünket a hagyományos faktoros elképzelésektől.

6.5 Prediktív szófaji egyértelműsítő

Alapelveinkből viszont például az is következik, hogy az elemzés folyamán nem használhatunk olyan tradicionális értelemben vett **szófaji egyértelműsítőt**, amely a mondatban megjelenő összes információ felhasználásával dönt egy-egy szó szófajáról. Az ANAGRAMMA-elemző nem építhet ugyanis az aktuális döntési helyzettől jobbra elhelyezkedő elemek tulajdonságaira, hiszen azok még nem hangzottak el, illetve nem kerültek beolvasásra, ezért egy csak a **balkörnyezetet leíró n-gram modellt** használunk, ami ugyan rendel valószínűségeket az aktuális szóhoz kapcsolható címkékhez, ám csupán az elhangzott, illetve leírt, az aktuális pozíciót megelőző szavak alapján. Ez a modul a PurePOS

egyértelműsítőnek egy az ANAGRAMMA-projekt számára módosított változata (Orosz, 2014).

6.6 Vonzatkeret-adatbázis létrehozása

Az elemző szabályait eleinte kézzel hoztuk létre, de a kutatás statisztikai moduljai segítségével ezt hatékonyabbá tettük, felhasználva a rendelkezésünkre álló nyelvtani adatbázisokat is

Elemzőprogramunk építésekor felhasználtuk a **MetaMorpho**-projektben (Prószéky & Tihanyi, 2002) kialakított, mintegy 34.000 magyar igei konstrukciót felvonultató, rendkívül átfogó szabályrendszert. A működés a következő: egy kínálat típusú szál alapvető szintaktikai-szemantikai jellemzőkkel (élő, ember, absztrakt stb.) annotálja az egyes argumentumokat a rendelkezésre álló mintegy 118.000 szót és többszavas kifejezést tartalmazó adatbázisból. Az elemző bemenetén megjelenő finitum ige pedig lehívja a hozzá tartozó vonzatkonstrukciókat az adatbázisból, és megteszi a szükséges felajánlásokat, amiket a munkamemóriában lekötetlenül álló argumentumigények tudnak kielégíteni. Egyszerre több szabálynak is teljesülhetnek a felajánlásai, ugyanis csak a tagmondat végén szükséges a meglévő elemzések összesítése, amikor is egy vagy több konstrukció minden argumentuma lekötésre került, míg a többi hamis elemzési ágnak bizonyult.

A **HuWordNet** (Prószéky & Miháltz, 2008) segítségével az argumentumokat az őket alkotó szavak hipernimái alapján további osztályokba tudjuk sorolni, amivel egyben megszorításokat is tudunk adni az egyes MetaMorpho-szabályoknak az adott szerkezethez igazodására, így az igei szerkezetek argumentumainak olyan szemantikus szelekciós megszorításai jelennek meg, mint pl. az alábbiak:

```
iszik  ACC folyadék
kigombol    ACC ruha
olvas  ACC könyv
ül          SUP ülőbútor
vádol  INS bűncselekmény
megold      ACC nehézség
```

6.7 Architektúraépítés

Az elemző rendszer architektúrájának kialakításakor sokat merítettünk a GIT verziókezelő rendszerből, ami – eredeti célját felülmúlva – az élet mind több területén jelenik meg például adatbázisrendszerként. A GIT elgondolása és az elemző működése a következő párhuzamokat mutatja. Egyrészt időben lineárisan előre – gyakorlatilag balról jobbra – halad, kizárólag a változások tárolásá-

val, memóriát spórolva, ahol egy ún. kommitot lehet a mi fogalmaink szerinti órajelnek tekinteni. Az egy kommit alatt bekövetkező változások a tárolásból fakadóan könnyen visszaállíthatók: a különböző állapotokon oda-vissza lehet lépegetni egyszerűen tetszés szerint, ha felülbírálásra kerül a sor. Az elemzés több ágon való folytatását a GIT rendszer – felépítéséből fakadóan – támogatja. Így az „elemzés" bármelyik pontján egyszerűen hozhatók létre párhuzamos ágak, amiknek szükség szerinti szinkronban tartásával az elemző rendszernek nem szükséges azonnal döntenie olyan helyzetekben, ahol egy átmeneti ideig több valószínű elemzési ág is lehetséges. Ezekből is látható, hogy a GIT kiválóan illeszkedik az elemző architektúrájához.

A GIT az objektumok tárolásából fakadóan rendkívül hatékony, bár a fájlok és mappák hierarchikus elrendeződése miatt jobban hasonlít a frázisstruktúra-nyelvtanok által épített mondatfákhoz. A függőségi reprezentációra emlékeztető, illetve az ezt a reprezentációt létrehozó elemzési mód miatt néhány helyen el kellett térni a GIT rendszer eredeti felépítésétől, hiszen a GIT természeténél fogva könyvtár-hierarchia, ami a közvetlen összetevős nyelvtanokra emlékeztet. Valójában nem befolyásolja a működését, hogy mi függőségi jellegű reprezentációk ábrázolására használjuk, hiszen bármilyen DAG-ra működik. A lényeg, hogy az objektumokat (az esemény egy állapotát) egyszer tárolja, és ezek relációit úgy frissíti időben előrehaladva, hogy különböző kitüntetett belépési pontokról indulva minden "időpillanat" bejárható, a rendszer lényegi átalakítása nélkül.

7. Összefoglalás

Kutatásunkkal egy pszicholingvisztikai motivációjú, performancia-alapú, párhuzamos feldolgozást végző nyelvi elemzőt céloztunk meg. A rendszer működéséhez szükséges ismeretek kidolgozása közben végigtanulmányoztuk a szükséges irodalmat, és azt találtuk, hogy az emberi nyelvfeldolgozás általunk vizsgált aspektusait egyetlen ma működő elemző sem elégíti ki teljes mértékben, így nem tudjuk rendszerünket a hasonló rendszerekkel összehasonlítani. Az egyetlen olyan algoritmikus közelítés, ahol az elemzési teljesség igénye, és a párhuzamosíthatóság is megjelent, a szószakértő elemző, a Word Expert Parser (Small, 1983) volt. A WEP-ben az egyes szavak mint az akkoriban népszerű szakértői rendszer egy-egy modulja működött. Ezeknek a kicsiny programoknak az interakciója némiképp emlékeztet az ANAGRAMMA-rendszer elemzési szálaira, ám a WEP működése elsősorban szemantikus, illetve fogalmi információkra épült, így a rendszerszerű morfológia és szintaxis teljességgel hiányzik belőle. Dolgozatunkban így igyekeztünk ismertetni egy –az ismert számítógépes nyelvi elemzőkhöz kevéssé hasonlító – új elképzelés, az ANAGRAMMA alapjait.

Irodalom

Bever, T. (1970) The Cognitive Basis for Linguistic Structures. In: Hayes, J. R. (ed.) *Cognition and the Development of Language*. New York: Wiley.

Chomsky, N. (1957) *Syntactic Structures*. The Hague: Mouton.

Endrédy I., Novák A.: (2012) Egy hatékonyabb webes sablonszűrő algoritmus - avagy miként lehet a cumisüveg potenciális veszélyforrás Obamára nézve. In: Tanács Attila; Vincze Veronika (szerk.) *A IX. Magyar Számítógépes Nyelvészeti Konferencia előadásai*, 297-301. SZTE, Szeged

Endrédy, I. & Novák, A. (2013) More effective boilerplate removal – the GoldMiner algorithm. *POLIBITS* 48, 79-83.

Endrédy I., Novák A.: (2015): Szótövesítők összehasonlítása és alkalmazásaik. In: Navracsics J. (szerk.) *Alkalmazott Nyelvtudomány*, XV. 1-2., Veszprém (2015)

Endrédy I.: (2015) Corpus based evaluation of stemmers. Submitted to the *7th Language & Technology Conference, Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznań.

Endrédy I.: (2015) Improving chunker performance using a web-based semi-automatic training data analysis tool). Submitted to the *7th Language & Technology Conference, Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznań.

Endrédy I. & Indig, B. (2015)**:** HunTag3, a general-purpose, modular sequential tagger – chunking phrases in English and maximal NPs and NER for Hungarian. Submitted to the *7th Language & Technology Conference, Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznań.

Frazier, L. & Fodor, J.D. (1978) The sausage machine: A new two-stage parsing model. *Cognition* 6, 291-325.

Grice, H. P. (1975) Logic and Conversation. In Cole, P. & Morgan, J.P. (eds.) *Speech Acts*. New York: Academic Press, 41–58

Halácsy P., Kornai A., Németh L., Rung A., Szakadát I., Trón V. (2004) Creating open language resources for Hungarian In *Proceedings of the 4th international conference on Language Resources and Evaluation (LREC2004),*

Indig B., Laki L. (előkészületben) Mozaik nyelvmodell az AnaGramma-elemzőhöz.

Kimball (1973) Seven principles of surface structure parsing in natrual language. *Cognition* 2, 15-47.

Ligeti-Nagy N. (2014) Szövegkorpuszok pontosabb annotációja gépi elemzéshez. *MANYE-2014*, Kolozsvár.

Ligeti-Nagy N. & Endrédy I. . (előkészületben) Magyar mondatvázak elemzése módosított annotációval, In: Tanács Attila; Vincze Veronika (szerk.), *XII. Magyar Számítógépes Nyelvészeti Konferencia előadásai,* SZTE, Szeged

Menzel, W. (2013) Incremental and Predictive Dependency Parsing under Real-Time. Condition. In: Galia Angelova, Kalina Bontcheva, Ruslan Mitkov (eds.) *Proceedings of the international conference Recent Advances In Natural Language Processing RANLP 2013*. Hissar, Bulgaria.

Miháltz M., Indig B. & Prószéky G. (2015) Igei vonzatkeretek és tematikus szerepek felismerése nyelvi erőforrások összekapcsolásával egy kereslet-kínálat elvű mondatelemzőben. In Tanács Attila, Varga Viktor, Vincze Veronika (szerk.*): XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015)*, SZTE, Szeged,. 298-302.

Miháltz M. & Sass B.: (2014) Mit iszunk? A Magyar WordNet automatikus kiterjesztése szelekciós preferenciákat ábrázoló szófajközi relációkkal. In: Tanács Attila, Varga Viktor,

Vincze Veronika (szerk.): *X. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2014*), SZTE, Szeged, pp. 109-116.

Mihăltz, M., Sass, B. & Indig, B. (2013) What Do We Drink? Automatically Extending Hungarian WordNet With Selectional Preference Relations.. *Proceedings of the Joint Symposium on Semantic Processing*. *Textual Inference and Structures in Corpora*. Trento, Italy. 105–109.

Novák, A. & Siklósi, B. (2015): Automatic Diacritics Restoration for Hungarian. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing,* 2286–91. Lisbon, Portugal: Association for Computational Linguistics.

Orosz, Gy. & Novák, A. (2013): Purepos 2.0: a hybrid tool for morphological disambiguation. In: Galia Angelova, Kalina Bontcheva, Ruslan Mitkov (eds.) Proceedings of the international conference Recent Advances In Natural Language Processing RANLP 2013. Hissar, Bulgaria. 539–545

Pléh Cs. (1998) *A mondatmegértés a magyar nyelvben***.** Budapest, Osiris.

Pléh Cs. & Lukács Á. (2001) *A magyar morfológia pszicholingvisztikája* Budapest: Osiris.

Pléh, Cs. & Lukács, Á. (2014) *Pszicholingvisztika* 1-2. Budapest: Akadémiai.

Prószéky, G. (2000) Számítógépes morfológia. In: Kiefer Ferenc (szerk.): *Morfológia (Strukturális magyar nyelvtan III).* 1021–1064. Akadémiai, Budapest

Prószéky, G. & Mihăltz, M. (2008) Magyar WordNet: az első magyar lexikális szemantikai adatbázis. *Magyar Terminológia* 1(1), 43-58.

Prószéky G., Indig B., Mihăltz M. & Sass B. (2014) Egy pszicholingvisztikai indíttatású számítógépes nyelvfeldolgozási modell felé. In: Tanács Attila, Varga Viktor, Vincze Veronika (szerk.): *X. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2014),* SZTE, Szeged, 79-90

Prószéky, G. & Tihanyi, L. (2002). MetaMorpho: A Pattern-Based Machine Translation System. *Proceedings of the Translating and the Compter 24 Conference*. London: ASLIB.

Sass B. (2015). Egy kereslet-kínálat elvű elemző működése és a koordináció kezelésének módszere.. In: Tanács Attila, Varga Viktor, Vincze Veronika (szerk.) *MSZNY 2015*. *Magyar Számítógépes Nyelvészeti Konferencia,* Szeged: JATEPress, 309-311.

Siklósi, B., Novák, A. & Prószéky, G. (2015): Context-aware correction of spelling errors in Hungarian medical documents, *Computer Speech & Language*, vol.35, pp. 219-233

Simonyi, A.. Indig, B. & Mihăltz, M.. (előkészületben) Exploiting Linked Linguistic Resources for Semantic Role Labeling.

Small, S. (1983) Parsing as cooperative distributed inference: understanding through memory interactions. In: King, M. (ed.) *Parsing Natural Language*, 247-276.

Tesniére, L. (1959) *Éléments de syntaxe structural*e. Klincksieck, Paris.

Yang Z. Gy., Laki L. & Prószéky G. (2015) Gépi fordítás minőségének becslése referencia nélküli módszerrel. *MSZNY 2015*