LiLe projekt:
Adatbázis mint „dinamikus korpusz”

Bódis Zoltán, Kleiber Judit, Szilágyi Éva, Viszket Anita

Pécsi Tudományegyetem, Bölcsészettudományi Kar, Nyelvtudományi Tanszék
7624 Pécs, Ifjúság útja 6.
lile@btk.pte.hu
http://lingua.btk.pte.hu/lile.asp

Absztrakt: A LiLe-projekt fő célkitűzése egy nyelvészeti lexikon MS-SQL-adatbázis-formában való felépítése. A kidolgozott adatbázis-struktúránk legfontosabb tulajdonsága, hogy a grammatika különböző szintjein létező szabályokat a morfémákkal azonos módon tárolja, így valósítva meg a legteljesebb lexikalizmust. A projektünk a grammatikus szóalakok és mondatok elemzésére és generálására vállalkozik. Ezeknek az eszközöknek a segítségével hozunk létre egy olyan korpuszt, amely az összes lehetséges szóalak és mondat generálására alkalmas. A generálási algoritmus kiválasztott adatbázisbeli elemeken (morfémákon és szabályokon) hajtható végre, így kívánjuk támogatni mind a nyelvészeti kutatásokat, mind a magyar nyelv oktatását. Ezt a speciális célokat megvalósító és sajátos felépítésű nyelvészeti lexikont nevezzük dinamikus korpusznak.

1. Célkitűzéseink

Kutatócsoportunk egy nyelvészeti adatbázis kifejlesztésére és feltöltésének koordinálására vállalkozott. A készülő adatbázis egy nyelvészeti lexikon, sajátos tulajdonságokkal. Szabad és kötött morfémák (tövek és toldalékok) egyaránt elemei a szótárnak, ezekből és a szintén adatbázisban tárolt szabályokból lehet (toldalékolt) szavakat előállítani. Mivel az alapok kidolgozásánál egy erősen lexikalista elméletre (GASG, Generatív/Általánosított Argumentumstruktúra Nyelvtan, Alberti 1998, 1999) támaszkodtunk, mindent a lexikai egységek leírásába kódoltunk bele, ami így egyszerre hordoz információt minden nyelvi szintről (fonológia, morfológia, szintaxis, szemantika).

Legfontosabb célunk egy leíró nyelvtan megalkotása, amiben (összhangban a lexikalista szemlélettel) nem a hagyományos utat követjük szabályok és kivételek felvételével. Ehelyett az egyedi eseteket rögzítjük – az egyes lexikai egységek tulajdonságait, viselkedését –, amelyekből statisztikai alapon a szabályok és kivételek „generálhatók”: a nagy elemkészleten működő eljárásokat lehet szabályként megfogalmazni, míg a kisebb halmazokon működők lesznek a kivételek. Ezzel a típusú leíró nyelvtannal kapcsolatos elképzeléseinket ismertettük a 2003-as MSZNY konferencián (Bódis és tsai 2003a).

Jelen előadásunkban az adatbázisunk egy újabb felhasználhatóságát mutatjuk be: egy „dinamikus korpusz” kidolgozását. Azért dinamikus, mert nem a *létező* (valaha létezett) alakokat tartalmazza, hanem az azokból visszaszármaztatott elemeket és szabályokat, így az adott nyelvállapot *lehetséges* szavai, kifejezései vagy akár mondatai bármikor generálhatók – a kompetenciánkat modellálja tehát.

Mire lehet alkalmas egy ilyen korpusz? Többek között kutatások támogatására, ahogy a „hagyományos” korpuszok is (Svartvik 1992). Amiben az általunk dinamikusnak nevezett

korpusz többet nyújt, az az, hogy tulajdonságra is kereshetünk benne. Kérhetjük például a programot arra, hogy generáljon főnévi igenevet tartalmazó mondatokat, vagy olyan szavakat, amelyekben több fonológiai váltakozás is van Így egyszerűen lehet majd egy-egy elmélethez példákat, illetve ellenpéldákat találni, vagy akár statisztikákat készíteni, hogy egy bizonyos tulajdonság (jegy) mennyire gyakori a nyelv lexikai egységeinek körében.

A kutatások előbb említett támogatása mellett egy konkrét számítógépes nyelvészeti programmal is együttműködünk: a szintén GASG-nyelvtanra épülő GeLexi-projekt (Alberti és tsai 2002a, b) számára biztosítunk dinamikusan bővíthető lexikont. Ennek a nyelvtan formalizálhatóságát bizonyító szoftvernek az adatbázisa jelenleg a Prolog programnyelven íródott elemző-kód része, ezáltal benne bármiféle változtatást végezni nagyon nehézkes. Ezért szükséges, hogy egy korszerűbb adatbázist építve legyen biztosítható az említett implementációhoz a megfelelő lexikon. Az általunk kínált adatbázis a lexikai egységeket az összes tulajdonságukkal együtt tartalmazza, így egy egyszerű unifikációs eljárással (amely az egyetlen művelet a GASG-ben) az elemek szavakká, illetve mondatokká tudnak épülni. A program jelenleg is működik, de ha elkészül hozzá ez a lexikon, sokkal nagyobb elemszámú korpuszon lesz képes ellenőrizni a szavak jólformáltságát (helyesírást), a mondatok grammatikalitását, és tud majd szemantikai reprezentációt társítani a beírt mondatokhoz.

Az épülő adatbázist teljes elkészültéig oktatási célokra kívánjuk hasznosítani. Két nagyobb alkalmazási területet különíthetünk el.

Az egyik a köz- és felsőoktatásban tanuló (magyar anyanyelvű) diákok nyelvtantanulásának segítése. Köztudott, hogy napjainkban a magyar nyelvtan tanítása igencsak elavultnak tekinthető, továbbá hiányzik belőle minden játékosság, problémaorientáltság, semmi nem motiválja a gyerekeket arra, hogy meg akarják ismerni anyanyelvük szabályszerűségeit (Takács 2000). Úgy gondoljuk, egy olyan programmal, amivel gyakoroltatni lehet az egyes szabályokat, megvizsgálni működésüket azáltal, hogy „ki-be kapcsolgatjuk” őket, érdekesen lehetne megtanítani a diákoknak, hogyan is működik a magyar nyelv. A program természetesen nem (csupán) a szabályok megtanítására és szemléltetésére lenne alkalmas, hanem a nyelvi tudatosság fejlesztésére is, ami nem csak a közoktatásban tanuló diákoknak fontos, hanem talán még inkább a felsőoktatásban tanuló, később magyartanárként elhelyezkedő fiatalok számára.

A másik fontos alkalmazási terület pedig a magyar nyelv tanítása idegen anyanyelvűeknek (hungarológia). Számukra talán még fontosabb, hogy lássák, hogyan, milyen szabályok szerint működik a magyar nyelv, illetve mely szavak vagy kifejezések viselkednek másképp. Emellett természetesen használhatják arra a célra is a programot, amire a magyar anyanyelvűeknek nincs szükségük: megnézhetik, hogy az általuk „összerakott” szóalak vagy mondat helyes-e, és abban az esetben, ha nem helyes, azt is láthatják, hogy mit rontottak el benne, melyik szabályt nem alkalmazták, vagy használták rosszul. Ha pedig idegen nyelvű lexikai egységek is szerepelni fognak az adatbázisban, használhatják azt (n-nyelvű) szótárként. Ebben az esetben természetesen nem csupán magyartanítás, hanem bármilyen más nyelv tanításának támogatására is alkalmas lesz a LiLe projekt által fejlesztett nyelvészeti lexikon.

2. A megvalósítás módszerei

A korábbiakban ismertetett célkitűzéseink („dinamikus korpusz” létrehozása, a nyelvi intuíció támogatása és ezzel a nyelvelsajátítás vagy a nyelvi tudatosság fejlesztésének

segítése, a kutatások számára a tesztmondatokat generálni képes korpusz biztosítása) megvalósításához alapfeltétel a rugalmasan bővíthető lexikon és szabályrendszer. Az egyik legfontosabb célkitűzésünk, hogy a rendszerünk lehetőséget nyújtson a felhasználónak arra, hogy ne csak a lexikonban tárolt lexémákat tudja bővíteni, hanem ezekhez egy felhasználóbarát felületen szabályokat is tudjon rögzíteni, ami jelentheti akár teljesen új szabályok definiálását is. Továbbá lehetőséget akarunk adni a felhasználónak arra, hogy a szabályokat tesztelhesse: a különböző szabályok egymástól független ki-bekapcsolása alapján ellenőrizhesse, hogy grammatikus szóalakok, illetve mondatok generálódnak-e vagy sem, továbbá, hogy a felvett új szabályok nem generálnak-e túl.

A kiindulópontnak választott modell (GASG) olyan totálisan lexikalista grammatika, amelyben a fonológiai, morfológiai, szintaktikai, szemantikai szabályok egy szinten ábrázolódnak és időben egyszerre lépnek működésbe. Ezt a modellt vettük alapul a tervezett dinamikus korpusz megvalósításánál, és a célkitűzéseinknek megfelelően módosítottuk. Megtartottuk a GASG-nek azt az elképzelését, hogy a lexikonbeli egységek alapvetően morfémák, illetve néhol[1] allomorfok, amelyeket egy külön jeggyel kapcsolunk össze morfémákká. Továbbá megtartottuk a GASG modellből a szórendre vonatkozó szabályok megadásának módját: a LiLe-ben is megelőzési relációk, illetve ezeknek a rangparaméterezései fejezik ki a szükséges szórendi viszonyokat a mondatokban. A harmadik közös pont a mondatok szemantikai struktúrájának felépülése, amit szintén a GASG-ben definiált módon kívánunk megvalósítani.

A mi újításunk a GASG-hez képest az, hogy nem csak a különböző jegyek (feature) értékei, hanem maga a jegy-struktúra és a jegyekre épülő szabályok is adatbázisbeli elemek, vagyis nem a program (az elemző) részét képezik, hanem az adatbázisét. Így tetszőlegesen bővíthető a szabályok száma, és ez nem befolyásolja a program működését. Ezért valósítható meg az is, hogy a létező szabályok futását korlátozzuk vagy engedélyezzük, akár a lexémák tetszőleges halmazán. Ez természetesen némileg másképp definiált jegyeket és szabályokat tett szükségessé a GASG-hez képest. A definiálások során szem előtt tartottuk azt is, hogy az általunk létrehozott rendszer ne legyen nyelv-specifikus, vagyis tetszőleges nyelv szabályait (és ennek alapját képező jegy-struktúráját) kódolni lehessen a rendszerünkben. A legfontosabb különbség a GASG és a LiLe között az, hogy míg a GASG egy grammatikai modell (mint az LFG vagy HPSG), és így mind szemléleténél, mind felépítésénél fogva általánosításokat fogalmaz meg az univerzális grammatikával kapcsolatban is (pl. régens és vonzatok lehetséges egyeztetési módjai stb.), addig a LiLe nem tartalmaz az univerzális grammatikára nézve ilyen típusú megállapításokat, csupán arra törekszik, hogy minden elképzelhető kapcsolat rögzíthető és ellenőrizhető legyen.

A jegyek és szabályok alkalmazását egy példán keresztül mutatjuk be. Vizsgáljuk meg a *lovakat* szóalak elemzését a programban! (Elemzéseink kiindulópontja a Strukturális Magyar Nyelvtan Fonológia illetve Morfológia kötete – Kiefer 1994, 2000 – volt, ám ahol gyakorlati szempontból szükségesnek láttuk, apróbb változtatásokat végeztünk.)

Ez a szó az elemzőnk szerint három morfémát tartalmaz: a *ló* szótövet, a főnévi többes számot kifejező morfémát, valamint a tárgyesetet kifejező morfémát. A szóalak grammatikus, mert a morfémák szófaja megegyezik, a morfémák a helyes sorrendben fordulnak elő, a *ló*-nak a megfelelő allomorfja szerepel benne, megfelel a hangrendi illeszkedés törvényének és a nyitásra vonatkozó hangtani szabályoknak is. Ahhoz, hogy ezeket megállapíthassuk, tárolni kell a szabályokat és az egyes morfémák és allomorfok tulajdonságait. A tulajdonságokat jegyekben tároljuk. Az egyes jegyeket el is neveztük, ami

[1] Ahol erre szükség mutatkozik, vagyis azokban az esetekben, amikor nem tudunk közös mögöttes hangtani reprezentációt rendelni az egyes allomorfokhoz.

nem a program működéséhez szükséges, hanem a szabályok megnevezéséhez, és így a felhasználó informálásához. A következő leírásban az egyes jegyek megnevezését adjuk meg, de a programban valójában csak a jegy azonosítójára hivatkozunk, a jegy nyelvészeti tartalma (szófajok, fonológiai tulajdonságok stb.) a program működése szempontjából irreleváns. Ezt azért tartjuk fontosnak megjegyezni, mert az egyes jegyek elnevezésénél nem törekedtünk arra, hogy valamely konkrét grammatikai (a példánkban: fonológiai) modell terminológiájához illeszkedjünk. A terminológia végső kialakítása az adatbázisunkat alapul vevő tananyagok része lehet.

A *ló* tulajdonságai közé tartozik, hogy főnévi tő, ami *ló* és *lov* alakokban fordulhat elő; továbbá mély hangrendű. A *lov* allomorf (a megjelölt toldalékok esetében) kötőhangot kíván meg; ún. nyitótő, ami befolyásolja az őt követő kötőhangot; valamint ún. v-vel bővülő tő. A többes szám jele (*-k*) tulajdonságai közé tartozik, hogy névszói toldalék, ami közepes erősséggel akar a szótőhöz közel kerülni[2]; kiváltja a v-vel bővülést; és a következő morfémára vonatkozóan ún. nyitótőként viselkedik. A lehetséges allomorfjai közül az *–ak* mély hangrendű és a nyitótövekhez társul. A tárgyeset ragjának a példánkban releváns tulajdonságai: névszói toldalék; kis erősséggel akar közel kerülni a tőhöz; az *–at* allomorfja pedig mély hangrendű és nyitó „tövekhez" (vagyis megelőző morfémákhoz) társul. A jegyek listájában megadjuk, hogy mely jegyek párosíthatóak, illetve azt is, hogy melyek zárják ki egymást, így egyfajta unifikációs módszerrel ellenőrizhető, hogy a szóalakban szereplő morfok jegyei illeszthetőek-e. Természetesen nem minden helyzet ilyen egyértelmű, ekkor szintén adatbázis-rekordként tárolt, bonyolultabb szabályok lépnek működésbe az egyszerű jegy-unifikáció helyett.

Az ismertetett lexikon- és jegy-struktúrát SQL-adatbázisban tároljuk, és az elemzéskor építünk az SQL-szerver gyors keresést biztosító működésére. A jelenlegi programverzió objektumorientált nyelven (Delphi) íródott, mivel ezzel tudjuk legkönnyebben biztosítani a felhasználóbarát felületet, de a későbbiekben terveink között szerepel a webes megvalósítás, aminek alapjául az szolgál, hogy az adataink akár online kinyerhetőek xml-formátumban.

Az adatbázis feltöltésének két fázisa van. Jelenleg az első fázisban tartunk, amikor a feltöltést kézi módszerrel végezzük, és eközben bővítjük a jegy-struktúrát és a szabályrendszert is. A feltöltés a magyar nyelv fonológiájának – morfológiájának oktatása közben történik, egyetemi hallgatók bevonásával. A második fázisban (egy éven belül) áttérünk az automatizált adatfeltöltésre: a programunk már jelenleg is alkalmas tetszőleges méretű szövegek szavakra bontására, és az ismert szavak elemzésére. Kidolgoztunk egy eljárást az ismeretlen szótövű alakok morfémáinak gépi jóslására és a rendelkezésre álló adatok alapján a jegyekkel való ellátására is. Természetesen ezeket a géppel létrehozott adatokat ellenőrizni kell, erre a legalkalmasabb a belőlük generált szóalakok és mondatok helyességének ellenőrzése: ekkor ugyanis az ellenőrzésbe nem csak nyelvészek vonhatóak be, hanem laikus magyar anyanyelvű felhasználók is.

3. Amiben újat nyújtunk

Számos számítógépes nyelvészeti projekt dolgozik azon, hogy valamilyen alkalmazást, szótárt fejlesszen, elméleti (kutatási, pl. korpuszok, szótárak) vagy gyakorlati (pl.

[2] Vagyis megelőzi a ragokat, de követi a képzőket: ezeket a tulajdonságokat valójában rangparaméterekkel fejezzük ki.

helyesírás-ellenőrzés, fordítás) célokra. Ezekhez a projektekhez képest mi a célkitűzéseinkben, a felhasznált elméleti keretben, az alkalmazott technológiában és a működésben tudunk különböző szempontokból újat nyújtani.

A LiLe célja nemcsak a kutatások támogatása, illetve nem elsődleges célunk helyesírás-ellenőrző vagy fordító-programok kifejlesztése, ahogy a legtöbb számítógépes nyelvészeti projekt esetében, hanem az egyik speciális célkitűzésünk az eredmények taneszközként való felhasználása a nyelvoktatásban. A nem anyanyelvi beszélők hiányzó kompetenciája pótolható, kisegíthető azáltal, hogy egyes morfémasorok helyes vagy helytelen voltát el tudja dönteni programunk. Ennek alapja pedig nem karakteregyezés[3] (tárolt vagy generálással előállított elemekkel), hanem azok a szabályok, melyek az egyes elemeken működnek. Ebből kifolyólag a döntés alapjául szolgáló elveket is meg tudja nevezni, amely hasznos segítség lehet a nyelvoktatásban, illetve a nyelvi tudatosság fejlesztésében. Mivel szabályainkat – az elemekkel együtt – adatbázisban tároljuk, egymástól függetlenül kijelölhető az elemzendő elemek halmaza és az azokon működő szabályok köre is, amely az egyes nyelvi jelenségekkel kapcsolatos gyakorlásra, szemléltetésre ad módot.

A felhasznált elméleti keret kiválasztásánál sem az általánosan elterjedt frázisstruktúra-nyelvtan valamely modellje mellett döntöttünk. Mivel nemcsak a tárolt elemeket, hanem az azokon működő szabályokat is ugyanazon eszközrendszerrel kezeljük, legyen szó a nyelv bármely „szintjéről", egy totálisan lexikalista modellt választottunk elméleti háttérül, amely ezt biztosítani tudja. Az, hogy a korpuszunkban morfémákat tárolunk kész szavak helyett, nem példa nélküli (noha nem is általános) a számítógépes nyelvészeti alkalmazások körében. Továbbá az sem, hogy a nyelvtant nem szabályhalmazként képzeljük el, ebben ugyanis követjük az unifikációs modelleket, amelyekben a „szabályok" nem igazi szabályok önmagukban, hanem az egyes elemek tárolt tulajdonságainak (jegyhalmazainak) egyeztetésén alapulnak. A mi újításunk a megszokott unifikációs modelleknél (pl. HPSG) is szigorúbban lexikalista GASG kiválasztása, és annak a még erőteljesebb „lexikalizálása" az elemző-kód megszüntetésével, és minden szabálynak az adatbázisba való beépítésével. A morfémák tárolása a kész szavak helyett lehetőséget ad a speciális célunk (oktatás) támogatására, valamint a fő célkitűzésünk (dinamikus korpusz: összes lehetséges alak generálása) megvalósítására. A minél teljesebb lexikalizmus is ezt a célt szolgálja: az adatbázisunk szabad bővítésével vagy változtatásával a módosítások a „nyelvtan" működését is meghatározzák, így a felhasználó (akár tanuló, akár kutató) gyorsan és hatékonyan tudja tesztelni a nyelvi kompetenciáját vagy a nyelvészeti modelljének működését.

Az alkalmazott technológia megválasztásánál a fő szempontunk az volt, hogy olyan technológiát válasszunk, melyben a sok különböző elem és összefüggéseik is egyféleképpen tárolhatóak, különbségeiket mégis megtartva. A relációs adatbázis ad erre lehetőséget. Továbbá az általunk használt MS-SQL-be beépített eljárások arra is módot adnak, hogy az általában tárolási célra használatos xml-t előállítsuk. Erre elsősorban a webes megjelenítésnél lesz szükség.

Mind a választott elméleti keret, mind az alkalmazott technológia lehetőséget ad a működésbeli újításra. A LiLe adatbázisában, azon túl, hogy az egyes morfémák külön-külön rekordként[4] szerepelnek, ugyanilyen formában tároljuk az egyes, jegyekkel definiált

[3] Például az oktató-gyakorló szoftverek közül az Interaktív Magyar Nyelvtan (http://www.sulinet.hu/kossuth/nyelvtan/), vagy a Magyar nyelvtan 1-2 (amelynek demo-verziója letölthető a http://www.oktatoszoftver.hu/ címről).

[4] Rekordon az egyes adatbázisbeli elemeket értjük.

tulajdonságokat is, amelyek vagy unifikációs eljárások[5] vagy „szabályok”[6] bemenetéül szolgálnak. A morfémák tárolásánál különböző változókat használunk az allomorfok elkülönítésére, kiszámítására. Ezek a változók is külön rekordok, melyek vagy az unifikációt szolgálják ki, vagy – szintén adatbázisban tárolt – eljárásokat hívnak meg az unifikációs módszerrel nem elemezhető jelenségek esetében. Ezek az eljárások műveleteket végeznek a tárolt elemeken. Mivel a változók a morfémák részei, azt a kört is pontosan definiáljuk, hogy hol jelennek meg a nyelvben az adott jelenségek. Ilyen értelemben a nyelvekben gyakran jelentkező „kivételek” kezelése sem jelent problémát, hiszen nem kell kivételeket tárolnunk – egy adott jelenség ilyen esetben egy kis elemszámú halmazon működik.

A felhasznált elméleti keret és annak általunk végrehajtott módosításai, valamint az alkalmazott technológia mind támogatják azt a már megfogalmazott célkitűzésünket, hogy egy olyan általános keretet tudjunk biztosítani, amely lehetőséget nyújt a grammatikai szabályok akár nyelvenként eltérő megfogalmazására. Vagyis nem célkitűzésünk egy univerzális grammatika létrehozása (szemben például a GASG törekvéseivel). A LiLe-ben az egyes nyelvek fonológiai/morfológiai/szintaktikai leírásában szereplő szabály-struktúrát az adott nyelv leírói határozzák meg, és az egyes nyelvtanok a közös struktúrájú szemantikai reprezentáción keresztül kapcsolódhatnak egymáshoz. Vagyis a lexikai egységek jellemzésében csak a szemantikai jegyek közösek (univerzálisak), a többi jegy szabadon alakítható. Így kívánunk bármely grammatikai modell számára hasznosítható lexikont biztosítani.

4. Eddigi eredmények

Kidolgoztuk azt a relációs adatbázis-struktúrát, amely minden további már megvalósult és jövőben megvalósuló elképzelésünknek az alapja. Ez az elképzelés merőben újszerű, célzott kutakodásaink ellenére sem találkoztunk egyetlen olyan nyelvi vagy nyelvvel kapcsolatos szoftverrel vagy kutatási projekttel sem, amely hasonló technológián alapult volna, és a magyar nyelv leírását, illetve feldolgozását tűzné ki célul.

A felépített struktúrának köszönhetően megvalósíthatóvá vált a totális lexikalizmus elve. Minden egyes lexikai elemet külön tételként, rekordként szerepeltetünk az adatbázisban, nem az összes előfordulási alakjában, hanem a tőmorfémákat és a toldalékokat külön, az összes releváns tulajdonságukkal együtt. A lexikai egységek mellett az adatbázisban kaptak helyet a lexikai egységek viselkedését befolyásoló szabályok is, tehát szabályaink is a lexikonunk részét képezik, ahogy ezt a korábbiakban már részletesen ismertettük.

A kidolgozott struktúra és a technológia lehetőségei együtt szinte korlátlan bővíthetőséget biztosítanak. Ennek köszönhető például az is, hogy a munka előrehaladtával, más célok megvalósításával dinamikusan fejlődhet, bővülhet nem csak a hagyományos értelemben vett lexikon, hanem a nyelvtan maga is. Vagyis gond nélkül felvehetünk újabb értékeket, tulajdonságokat mind a lexikai egységekhez, mind a tulajdonsághalmazunkhoz, sőt ez a lehetőség nyitva hagyja előttünk az n-nyelvűség megvalósíthatóságának kapuit. A munka jelenlegi szakaszában, a korábban definiált első fázisú adatfeltöltési módszerünkkel egyidőben változott és változik az adatbázis, méghozzá az alapötlet megsértése vagy változtatása nélkül.

---

[5] Egyszerűbb esetekben, lásd az idézett *lovakat* szóalak elemézése.

[6] Bonyolultabb esetek kezelésénél.

Már a munka kezdetekor célul tűztük ki magunk elé, hogy az adatbázist úgy tervezzük meg és hozzuk létre, hogy az ne csak a saját fejlesztésünk alapja lehessen, hanem más nyelvészeti elmélethez és gyakorlati megvalósuláshoz is hasznos segítséget nyújthasson, vagy azoknak is alapjául szolgálhasson. Az építkezés során mindig a szemünk előtt lebegett egy olyan elméletektől független, vagy legalábbis kevéssé befolyásolt lexikon képe, amely többféle próbálkozáshoz, elképzeléshez tud kapcsolódni. Ennek egyik fontos technológiai alapja lehet a rendelkezésünkre álló XML-adatszolgáltatás.

Az adatbázis mellett a projektünk eddigi eredményei közé tartozik egy szoftver[7] is, amely mostanra két egymástól élesen elkülönülő funkcióval rendelkezik.

Az egyik funkció az adatbázis feltöltése. Az adatrögzítést azzal segíti a szoftverünk, hogy a táblastruktúra részletes ismerete nélkül is bővíthető az adathalmaz. Ezt a feltöltő programot a jövőben weben elérhető formában, autentikációval kibővítve fogjuk a felhasználók rendelkezésére bocsátani. Az autentikáció előnye, hogy a nyelvtani rendszeren végrehajtott változtatások vagy a lexikon bővítései felhasználókhoz köthetőek.

A szoftver másik része egy szóelemző program, amely jelen állapotában a magyar főnévi és igei inflexiós morfológia és az írásban jelölt fonológiai szabályok alapján működik. Figyelembe veszi a szófaji egyezést, a helyes morfémasorrendet és a fonológiai szabályokat. Az egyedisége abban rejlik, hogy nemcsak azt tudja megállapítani, hogy a beírt szóalak helyes-e, hanem azt is megmondja a felhasználónak, hogy mely nyelvtani szabály helytelen működése vagy figyelmen kívül hagyása eredményezte a nem megfelelő szóalakot. Úgy gondoljuk, hogy ennek köszönhetően a programunk továbbfejlesztett változata igen hathatós segítséget nyújthat a külföldiek magyartanításában és a magyar anyanyelvűek nyelvi tudatosságának fejlesztésében.

A projekt már jelenlegi fázisában is az oktatás részét képezi, hiszen az elmúlt félévben a PTE BTK Nyelvtudományi Tanszékén egy kutatószeminárium keretében a hallgatóink megismerhették az adatbázis felépítését és működésének elveit. Miközben az adatbázis releváns adatokkal való feltöltésével a magyar morfofonológia tanulásához gyűjthettek gyakorlati tapasztalatokat, a mi munkánkat is segítették egy-egy remek ötlet felvetésével.

5. Jövőkép

Az adatbázis kidolgozásán túl az elsődleges célunk a lexikon feltöltése magyar nyelvi adatokkal, és ezáltal egy leíró magyar nyelvtan definiálása.

Legközelebbi céljaink közül az első nyelvészeti indíttatású. Az adatbázis-struktúra morfológiai bővítésén dolgozunk, a morfológiai szabályrendszert kívánjuk kiterjeszteni a képzők körére is. Számítógépes nyelvészeti feladat az adatbázis automatizált feltöltésére való áttérés a már kidolgozott eljárási elveink alapján. Tisztán informatikai feladat az elkészült program webes környezetbe való átültetése, amely szélesebb körben teszi elérhetővé a megvalósuló eredményeket.

További céljaink között kívánjuk megvalósítani a nyelvészeti feladatok terén a magyar szintaxis kidolgozását a modellünkön belül. A számítógépes nyelvészeti munka területén annak a már bemutatott dinamikus korpusznak a megvalósítását tervezzük, amelyhez minél több elméleti és gyakorlati kutatás kapcsolódhat. Végül pedig ugyancsak egy-két éven belül tervezzük létrehozni azt az oktatástámogató programcsomagot, amely a hungarológia-oktatás és a közoktatás segítségére lehet.

[7] A programozási munkákért szeretnénk köszönetet mondani Lőcsei Gábornak.

A távlati célok közül az egyik legfontosabb az n-nyelvűség kérdésének gyakorlati megvalósítása, a másik pedig egy olyan lexikalista szemantika felépítése, amelyet adaptálni tudunk a már meglévő adatbázis-struktúránk keretei közé.

Hivatkozások

Alberti Gábor (1998): *GASG: Minimal Syntax, Maximal Lexicon and PROLOG, Paper read at ALLC/ACH '98, July 9;* In Hunyadi László szerk.: ALLC/ACH '98. KLTE, Debrecen; 81-83.

Alberti Gábor (1999): *GASG: The Grammar of Total Lexicalism;* in: Working Papers in the Theory of Grammar 6/1, Elméleti Nyelvészet Program, ELTE és MTA Nyelvtudományi Intézet.

Alberti Gábor – Balogh Kata – Kleiber Judit – Viszket Anita (2002a): *A totális lexikalizmus elve és a GASG nyelvtan-modell*; In Maleczki Márta szerk.: *A mai magyar nyelv leírásának újabb módszerei V.* Szegedi Tudományegyetem; 193-218.

Alberti Gábor – Balogh Kata – Kleiber Judit – Viszket Anita (2002b): *Towards a totally lexicalist morphology. Talk at 6th International Conference on the Structure of Hungarian (ICSH6),* Düsseldorf, Németország. Megjelenés előtt: Kenesei István – Piñón, Chris szerk.: *Approaches to Hungarian 9.*

Bódis Zoltán – Kleiber Judit – Szilágyi Éva – Viszket Anita (2003a): *Leíró nyelvtan – adatbázisból;* In: Magyar Számítógépes Nyelvészeti Konferencia MSZNY2003 Szeged, 2003 december 10-11. konferenciakötet, SZTE; 300-302.

Bódis Zoltán – Kleiber Judit – Szilágyi Éva – Viszket Anita (2003b): *Nyelvészeti lexikon – oktatási és kutatási adatbázis fejlesztése;* Konferencia-előadás: Multimédia az oktatásban, Pécs, 2003. október 10.

Kiefer Ferenc szerk. (1994): *Strukturális magyar nyelvtan 2. Fonológia;* Bp. Akadémiai Kiadó.

Kiefer Ferenc szerk. (2000): *Strukturális magyar nyelvtan 3. Morfológia;* Bp. Akadémiai Kiadó.

Mitkov, Ruslan szerk. (2003): *The Oxford Handbook of Computational Linguistics;* Oxford University Press.

Prószéky Gábor – Kis Balázs (1999): *Számítógéppel - emberi nyelven. Intelligens szövegkezelés számítógéppel;* Szak Kiadó, Bicske.

Prószéky Gábor – Olaszy Gábor – Váradi Tamás (2003): *Nyelvtechnológia* In: Kiefer Ferenc szerk.: *A magyar nyelv kézikönyve*, Akadémiai Kiadó; 567-589.

Svartvik, Jan szerk. (1992): Directions in Corpus Linguistics, Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1991; Mouton de Gruyter, Berlin – New York.

Takács Viola (2000): *Attitűdvizsgálat – strukturális elemzéssel*. Iskolakultúra 2000/6-7, 199-201.