Nyelvtanfejlesztés, implementálás és korpuszépítés: A HunGram 2.0 és a HG-1 Treebank legfontosabb jellemzői

Laczkó Tibor, Rákosi György, Tóth Ágoston, Csernyi Gábor

Debreceni Egyetem, Angol-Amerikai Intézet
4032 Debrecen, Egyetem tér 1.
{laczko.tibor, rakosi.gyorgy, toth.agoston,
gabor.csernyi}@arts.unideb.hu

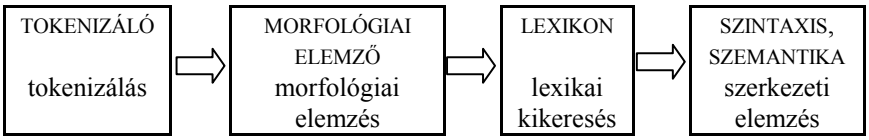
Kivonat: Cikkünkben beszámolunk kutatócsoportunk komplex elméleti nyelvészeti és implementációs vállalkozásának eddig elért eredményeiről. A kutatócsoport alapvető célja a magyar nyelv egy lehetséges generatív nyelvészeti modelljének a kidolgozása és ennek a modellnek az implementációja. Az elméleti keret a Lexikai-Funkcionális Grammatika, az implementációs platform a *Xerox Linguistic Environment*. Ebben a nagyobb ívű nyelvelméleti és nyelvtechnológiai kutatási folyamatban egy fajsúlyos részfeladatot is vállaltunk egy projekt keretében: egy jelentős méretű és sokféle felhasználási lehetőséget biztosító treebank létrehozását. A cikkben röviden jellemezzük az általános megközelítési keretünket, majd bemutatjuk azt, hogy ebbe hogyan ágyazódik bele a treebankes projekt: milyen nyelvelméleti és implementációs kihívásokkal kellett szembenéznünk, és milyen megoldásokat alkalmaztunk. Ezután részletesen tárgyaljuk és szemléltetjük a treebank legfőbb jellemzőit.

1 Bevezetés

Lexikai-Funkcionális Grammatikai Kutatócsoportunk (http://hungram.unideb.hu) 2008-ban egy OTKA projekt keretében kezdte el egy Lexikai-Funkcionális Grammatika (LFG) alapú magyar nyelvtan kidolgozását és – ezzel párhuzamosan – a nyelvtannak az implementálását (HunGram 1.0) az XLE (*Xerox Linguistic Environment*) platformon (a további részleteket l. a 3. szekcióban). Egy későbbi (de az előzővel párhuzamosan futó), TÁMOP projekt keretében egy másfél millió szavas magyar treebank összeállítását vállaltuk, amelynek a „rendező elve és motorja” a másik implementációs nyelvtannak egy olyan „átfejlesztése”, amely a treebank céljait közvetlenebbül és eredményesebben szolgálja. (A két projekt pályázati részleteit l. a *Köszönetnyilvánítás* szekcióban.) Cikkünkben egyrészt bemutatjuk ezt a treebankes nyelvtanváltozatot (a két nyelvtan közötti legfontosabb hasonlóságok és eltérések kiemelésével és illusztrálásával), másrészt beszámolunk az elvégzett korpuszfejlesztési munkálatokról (célok, eszközök, eredmények).

2 A HunGram nyelvtan automatikus elemzésre szánt változata

A HunGram 1.0 moduláris felépítése (az XLE-s nyelvtanok mintájára, vö. [1] és [8]), a következő.



1. ábra: a HunGram 1.0 fő komponensei.

A tokenizáló az adott szófüzért a magyar nyelv sajátosságainak megfelelő tokenekre bontja. Ezek szolgálnak bemenetül a morfológiai komponens számára, amely egy véges állapotú átalakító (*finite state transducer: fst*). A mi nyelvtanunk fst-je például a *játékot* főnevet és az *ették* igét a következő címkékkel (az angol eredeti alapján: *tag*ekkel) jellemzi.

(1) a. játék "+Noun" "+Sg" "+Acc"
b. eszik "+Verb" "+Past" "+Def" "+Pl" "+3P"

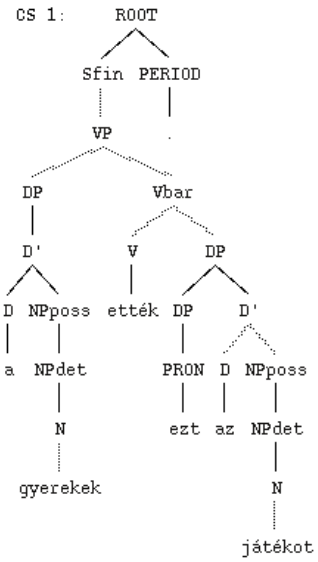
A lexikonban nemcsak szavaknak vannak megfelelő jellemzéssel ellátott tételei, hanem a különböző morfoszintaktikai jegyeket hordozó címkéknek is. Ezek lexikai tételeiben alapvetően funkcionális egyenlőségek révén rögzítjük az általuk kifejezett morfoszintaktikai jellegű információkat. Például a +Acc és a +Pl címkéknek a mi nyelvtanunk lexikonjában – egyebek között – a következő tételei találhatók.

(2) a. "+Acc" N_SFX XLE (↑ CASE)= acc
b. "+Pl" (i) N_SFX XLE (↑ NUM)= pl
(ii) V_SFX XLE (↑ SUBJ NUM)= pl

(2a) tanúsága szerint a +Acc címke főnévi tövekhez kapcsolódik, és azt a morfoszintaktikai információt kódolja, hogy a főnév (főnévi csoport) esetjegye akkuzatívusz. (2b) pedig azt mutatja, hogy egy címke különböző morfoszintaktikai információkat hordozhat attól függően, hogy milyen tőhöz járul. Egy főnévi tő esetében a főnév többességét jelöli, egy igei tő esetében viszont az alany többességét kódolja.

A szintaktikai elemzés alapját – az LFG felépítésének és elveinek megfelelően – egy olyan frázisstruktúrás szabályrendszer nyújtja, amelyben az egyes csomópontok szimbólumai megfelelő funkcionális annotációkkal is el vannak látva. Ilyen módon tud az LFG és (ebből következően) az XLE-ben implementált változata két parallel szintaktikai reprezentációt rendelni minden egyes jól megformált mondathoz: egy összetevős szerkezetet és egy funkcionális szerkezetet. Például a HunGram 1.0-ban a (3)-beli mondat egyik lehetséges összetevős szerkezetét a 2. ábra és az ennek megfelelő funkcionális szerkezetet a 3. ábra szemlélteti. Az összetevős szerkezet alapvetően a mondat kategoriális és szórendi jellemzőit ragadja meg, míg a funkcionális szerkezet a grammatikai viszonyokat ábrázolja (a grammatikai funkciókat és a releváns morfoszintaktikai információkat).

(3) *A gyerekekették ezt a játékot.*



2. ábra: *A gyerekek ették ezt a játékot* mondat összetevős szerkezete a HunGram 1.0-ban.

```
[PRED      'eszik<[2:gyerek], [95:játék]>'
 SUBJ      [PRED  'gyerek'
            NTYPE [NSEM [COMMON +]
                   NSYN common]
          2[CASE nom, DEF +, NUM pl, PERS 3]]
 OBJ       [PRED   'játék'
            NTYPE [NSEM [COMMON +]
                   NSYN common]
            SPEC 76[PRED 'pro'
                    CASE acc, DEIXIS proximal, NUM sg, PERS 3, PRON-TYPE demon]
          95[CASE acc, DEF +, NUM sg, PERS 3]]
 FOCUS     [2:gyerek]
 TNS-ASP   [MOOD indicative, TENSE past]
45[STMT-TYPE decl]
```

3. ábra: *A gyerekek ették ezt a játékot* mondat funkcionális szerkezete a HunGram 1.0-ban.

A HunGram 2.0-nak a HunGram 1.0-ból való kifejlesztése során a legfontosabb célunk az volt, hogy egy a magyar mondatokat automatikusan, a lehető legkevesebb többértelműséggel feldolgozni képes mondattani elemzőt hozzunk létre. Másrészt ugyanakkor alapvető kívánalom volt ezzel a nyelvtanváltozattal szemben is, hogy nyelvészeti szempontból is teljes és megbízható elemzéseket adjon. Ennek megfelelő-

en egy pusztán és szigorúan *csak nyelvészeti* megfontolások alapján szerkesztett nyelvtanváltozathoz képest sekélyebb, de a gépi feldolgozás kontextusában értelmezett valódi sekély nyelvtanoknál jóval gazdagabb nyelvtanváltozatot hoztunk létre. A nyelvtan hatékonyságát azáltal is igyekeztünk növelni, hogy az egyes köztes nyelvtanállapotok szerint leelemzett munkakorpuszból vett véletlenszerű elemzési minták helyességét több körben is manuálisan ellenőriztük, majd az észlelt hiányosságokat folyamatosan kiigazítottuk magában a nyelvtanban. A HunGram 2.0 így egy több szempontból is kiérlelt nyelvtanváltozatnak tekinthető.

A fenti követelmények jellege miatt a HunGram 2.0 elsősorban a pontosságra és kevésbé a lefedettségre törekszik. A következőkben röviden áttekintjük azokat a főbb tervezési jegyeket, amelyek ennek a közvetlen célnak a megvalósulását segítették elő. Az illusztrációként idézett mondatok forrása minden esetben maga a HG-1 Treebank.

A HunGram 2.0 szabályai nem generálnak pusztán nyelvészeti szempontból érdekes kétértelműségeket. Például az alábbi mondatban szereplő birtokos szerkezetben a határozott névelő elvileg tartozhat magához a birtokos főnévi csoporthoz (*a dzsungel*) vagy a teljes (félkövérrel szedett) birtokos szerkezethez:

(4) ***A dzsungel könyve*** *leszámol egy illúzióval*.

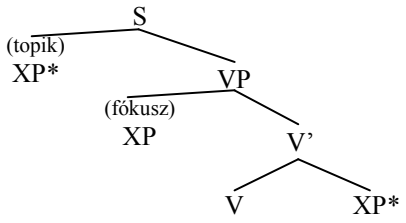
Ezt a fajta, inkább elméleti, mint gyakorlati jelentőségű kétértelműséget egyszerűen kizártuk azáltal, hogy a megfelelően megszorított nyelvtanunk csak a birtokos szerkezet egészéhez tudja hozzárendelni a határozott névelőt.

Általában véve olyan mögöttes mondattant alkalmazunk a HunGram 2.0 nyelvtanváltozatban, amelyet kimerítően meghatároznak a morfológiai és a sorrendi információk. Így nincs például mondattanilag kódolt információs szerkezetünk (nincs pl. fókusz, topik vagy kontrasztív topik), mert ennek pontos beazonosításához a fenti információk *gyakran* nem elégségesek. Vegyük például azt a feladatot, hogy egy az igét megelőző főnévi csoportról el kell dönteni, fókusz-e vagy topik (a korábbi (3) mondatot (7)-ként idézzük újra).

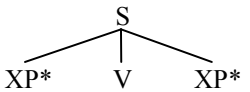
(5) ***Zoli*** *megette a radíromat*! **topik**
(6) *A sárkányok* ***csak a szüzeket*** *eszik meg, vagy rosszul tudom*? **fókusz**
(7) ***A gyerekek*** *ették ezt a játékot*. **topik/fókusz**
(8) *Az elmúlt század közepéig* ***a japánok*** *nem ettek marhahúst*. **topik/fókusz**

(5)-ben az igekötő ige előtti pozíciója kizárja a félkövér főnévi csoport fókuszos elemzését, (6)-ban viszont az igekötő posztverbális helyzete és a *csak* partikula jelenléte szükségszerűvé teszi azt. A HunGram 1.0 érzékeny is ezekre a szintaktikai információkra, és az ilyen szerkezetek esetén megbízhatóan el tudja dönteni, hogy lehet-e egy adott főnévi csoport a tagmondat fókusza vagy sem. (7) és (8) esetében viszont semmilyen morfoszintaktikai információ nem áll rendelkezésre, hogy el tudjuk dönteni, fókusz-e vagy topik a preverbális főnévi csoport. A HunGram 1.0 ilyenkor egy fókuszos és egy topikos elemzést is generál. Mivel az ilyen esetek igen gyakoriak, és az itt ismertetett megfontolásokhoz hasonlóak más esetekre is érvényesek, a HunGram 2.0-ban az ilyen kétértelműségek kiküszöbölése érdekében egy alapvetően lapos, az információszerkezetet nem grammatikalizáló mondatszerkezetet tételezünk föl. A HunGram 1.0-ban a számunkra most releváns, meghatározó mondatszerkezeti váz a (9a) mintázatát követi, míg a HunGram 2.0-ban a (9b) az alapvető rendező elv.

(9) a.



b.



A kategoriális vagy morfoszintaktikai jegyek miatti kétértelműségek egy részét igyekszünk gyakorisági megfontolások alapján már a lexikonban megszorítani. Tipikus példa az ilyen kétértelműségekre a melléknévként is lexikalizálódott melléknévi igenevek esete. A *borzasztó* szót például melléknévi lemmaként vettük fel a szótárban, letiltva egyúttal az (*el*)*borzaszt* ige folyamatos melléknévi igenévi használatát. Ennek következtében az alábbi mondatban a félkövérrel szedett szót csak melléknévként elemzi a nyelvtan, igenévként nem.

(10) *Hát* ***borzasztó*** *nevet választottál.*

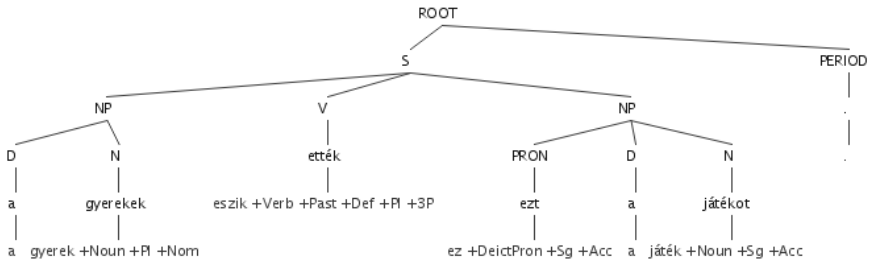
Mivel a szó valódi igenévi használata viszonylag ritka (vö. *a Jánost elborzasztó név*), többet nyerünk a potenciális kétértelműség kizárásával, mint a fenntartásával.

Végezetül egyes valós (vagyis akár jelentésbeli különbségekkel is járó) szintaktikai kétértelműségeket kizárunk a HunGram 2.0-ban. Például a HunGram 1.0-ban az általános LFG-s elméleti és XLE-s implementációs rendszerek felépítésének megfelelően az elemzőnk *összetevős szerkezeti* és *funkcionális szerkezeti* reprezentációt is ad. A funkcionális szerkezeti ábrázolásban igen gyakran van többszörös elemzés, és ennek az egyik rendszerszerű oka az, hogy bizonyos összetevőkhöz *oblikvuszi* és *adjunktumi* funkciót egyaránt képes a nyelvtan hozzárendelni. Vegyük például az alábbi mondatokat!

(11) *Képeket majd küldök jövő héten.*
(12) *Küldd a hírt a barátodnak!*

(12) datívuszos kifejezését szokás a *küld* ige oblikvuszi argumentumaként kezelni. Mivel azonban (11)-ben ez az összetevő nem szerepel, a (12)-ben található háromargumentumú *küld* tétel mellett mindenképpen fel kell venni a lexikonban egy kétargumentumú *küld* tételt is, amelynek csak alanyi és tárgyi argumentuma van. Emiatt viszont rögtön két elemzése is lesz (12)-nek, hiszen egyéb megszorítások híján elvileg egyaránt lehet benne akár a kétargumentumú vagy akár a háromargumentumú tétel is. Előbbi esetben a datívuszos kifejezés adjunktum, az utóbbiban oblikvuszi argumentum. A jelenlegi treebankünk szempontjából ez azért nem közvetlen gond, mert csak összetevős szerkezeti ábrázolást ad. Ugyanakkor a tervezett továbbfejlesztésre és a funkcionális szerkezeti ábrázolás beépítésére előre is gondolva a HunGram 2.0-ban az alkalmazható grammatikai funkciók közül kiiktattuk az oblikvuszt. Egy ilyen lépés *elméleti* motivációit részletesebben taglalja [5].

A HG-1 Treebankben tehát csak az összetevős szerkezeti elemzések jeleníthetők meg. (7) treebankbeli elemzését például az alábbi ábra mutatja (ugyanennek a mondatnak a HunGram 1.0 által generált összetevős szerkezetét a 2. ábrán láthattuk).



4. ábra: *A gyerekek ették ezt a játékot* mondat elemzése a HunGram 2.0-ban.

A szerkezetet generáló teljes nyelvtan bemutatására itt most nincs mód. Egy rövid átfogó ismertetés a treebank weboldalán érhető el (http://corpus.hungram.unideb.hu), a nyelvtant részletesebben ismertető publikációk jegyzéke pedig a kutatócsoportunk honlapján áll az érdeklődők rendelkezésére (http://hungram.unideb.hu).

Hadd álljon itt most csak egy rövid példa a nyelvtanfejlesztés során alkalmazott stratégiáink jellemzésére. A magyar nyelv egyik legnagyobb kihívást megtestesítő jelensége nyelvelméleti és implementációs szempontból egyaránt az igekötős igék viselkedése. A HunGram 1.0-ban [2] alapján kifejlesztettünk egy olyan megközelítést, amely az igekötős igék produktív és improduktív használatát is kielégítően és elvszerűen kezeli. Ennek részleteit [4] és [6] mutatja be. Tekintettel arra azonban, hogy ez a kezelésmód az improduktív használat esetében egyedi lexikai tételek bevitelét igényli, a produktív használat esetén pedig komplex frázisstruktúrás-annotációs mechanizmust, az automatizált elemzésre törekvő HunGram 2.0-ba ezt nem tudtuk átemelni. Ez utóbbi nyelvtanváltozatban lényegében – meglehetősen leegyszerűsített módon – közönséges határozói kategóriájúnak tekintünk minden igekötőt. Ezt annál is inkább könnyen megtehetjük, mert ez a nyelvtanváltozat egy olyan treebanket szolgál ki, amelyben nincs funkcionális szerkezeti reprezentáció. Márpedig az igazi kihívások ebben a dimenzióban jelentkeznek (grammatikai funkciók, argumentumszerkezet stb.). Mivel a jövőbeni implementációs nyelvtani munkálataink egyik legfontosabb célkitűzése az lesz, hogy a két nyelvtanváltozat azon vonásait, amelyek a másik változat számára is használhatók és hatékonyak lehetnek, igyekezzünk kölcsönösen figyelembe venni és minél teljesebb mértékben beépíteni, az igekötős igék kezelése esetében – a fentiek alapján értelemszerűen – azt fogjuk megvizsgálni, hogy van-e mód arra, hogy a HunGram 1.0 apparátusából átemeljünk olyan elemeket, amelyek a lehető legautomatizáltabb elemzést biztosítják úgy, hogy az igekötők természetének legfontosabb jellemzőit is megragadják.

3 Implementációs környezet, treebankfejlesztés

A magyar nyelvtan(ok) kidolgozása közvetlenül kapcsolódik a nemzetközi ParGram együttműködéshez (*Parallel Grammar*, lásd: [1]), amelyben több nyelvhez (angol,

német, urdu, francia, japán stb.) készül LFG nyelvtan, rendszeres egyeztetések mellett. A magyar grammatika minden elemét – a többi ParGram projekthez hasonlóan – az XLE munkafelületen [8] implementáltuk, mondattani elemzésre annak LFG-elemzőjét használtuk. Az XLE-ben található eszközöket, beleértve annak LFG-elemzőjét, nagyon hatékonynak találtuk, a rendszer robosztus és gyors, figyelembe véve azt is, hogy az LFG formalizmus szerinti mondattani elemzés inherens módon időigényes, NP-teljes probléma. A rendszer kifejlesztői több módon is igyekeztek kezelni ezt a problémát [9], többek között kiemelték a polinomiális időben megoldható feladatokat, és ezek feldolgozásával kezdik az elemzést, az exponenciális időben végrehajtható részfeladatok megoldása ezután következik. Az XLE elemzőjének egy másik tulajdonsága, hogy a szerkezeti többértelműségeket „csomagolt" módon, részfákra lokalizálva kezeli, ezzel segítve az eredmény gyors vizuális értelmezését. Ez ugyanakkor jól mutatja, hogy az XLE környezet elsősorban a humán feldolgozásra finomhangolt munkakörnyezet, miközben a további számítógépes feldolgozás szempontjából (a következőkben ismertetett treebankfejlesztési projekt esetében is) további munkafázist kellett ahhoz beiktatni, hogy a kimenet az általunk várt formát öltse.

Kutatócsoportunk a saját fejlesztésű nyelvtan és az XLE parser segítségével létrehozott egy 1,5 millió szavas treebanket (HG-1 Treebank). A treebank 1,5 millió szót (több mint 280 000 magyar mondatot) tartalmaz, morfológiai és mondattani annotációval ellátva. A HunGram 2.0 nyelvtanváltozat alapján végzett elemzést az NIIF szuperszámítógépes szolgáltatására támaszkodva, teljesen automatizálva állítottuk elő.

A korpuszt elsősorban az elméleti hátteret adó nyelvészeti kutatómunka kézzelfoghatóvá és felhasználhatóvá tételére, eredményeinek disszeminálására hoztuk létre és publikáltuk, ugyanakkor a fejlesztés felhasználható a nyelvoktatás, nyelvtanulás területén, a lexikográfiában, valamint elméleti nyelvészeti kutatásokban. A projekt website-ja (http://corpus.hungram.unideb.hu/) tartalmaz egy online lekérdezési felületet, valamint egy részletes leírást, útmutatót az elemzésekben található alaktani és mondattani jellemzők értelmezéséhez. Kiemeljük azt is, hogy a kutatócsoportunk nyelvtanírási projektjéhez a korpuszfejlesztési alprogram folyamatos tesztelési lehetőséget és visszajelzést biztosít.

Az elemzéseket tartalmazó korpusz kialakításához nyersanyagként a Magyar Webkorpuszt [3] használtuk, mely magyar weblapokról gyűjtött, elemzés nélküli szövegeket tartalmaz. A Magyar Webkorpusz (készítői által) szűrt változatából elemeztettünk annyi szöveget, hogy az elemzett mondatokban lévő szavak száma a 1,5 millió szót elérje.

Az elemzőnk által előállított kimenetet feldolgozva a mondattani fák tárolását a Tiger-XML leírónyelv segítségével oldottuk meg, amely kiváló eszköz fák reprezentálására [7]. Egy ágrajz kódolása a gyökérelem kijelölésével indul, utána a terminális szimbólumok felsorolása következik, melynek során a lexikai egységekhez kapcsolódóan a szófajt, a lemmatizált alakot és a morfológia által visszaadott összes jegyet tároljuk. Ezt követi az összes többi csomópont leírása legalább 1-1 kapcsolódó él meghatározásával.

Az adataink hozzáférhetőségének és kereshetőségének biztosítására létrehoztunk egy online lekérdezési felületet, amely lehetővé teszi a keresést szóra vagy lemmára, valamint a keresési találatok szűrését morfológiai jegyekre és a keresett szót tartalma-

zó összetevőre. A találatokról mondatelemzés-lista készül, ahonnan egy elemzést kiválasztva megkapjuk a megfelelő összetevős szerkezetet ágrajz formájában.

A korpuszfejlesztési projekt szoftver-infrastruktúrájának jelentős részét házon belül fejlesztettük ki. Az elvégzett programozási feladataink a következők voltak:

1. Mondatok elemeztetése a készülő nyelvtannal feltöltött XLE elemzővel, és a kimenet rögzítése (alternatív elemzésekkel).
2. Az összes lehetséges elemzés összetevős szerkezetének a kibontása és tárolása. A korpuszt ettől a ponttól XML dokumentumban tároljuk (TigerXML formátumban).
3. Alkorpuszok kezelése:
    a. korpuszfájlok darabolása és egyesítése,
    b. indexelés, statisztikák készítése (faszélesség, famélység, szavak és mondatok száma).
4. On-line lekérdezési felület létrehozása a következő főbb funkciókkal:
    a. keresés szóra vagy lemmára,
    b. keresés szűrése morfológiai jegyekre és a keresett szót tartalmazó összetevőre (szűrés beállítása űrlap segítségével),
    c. a találatok megjelenítése,
    d. a találati listából kiválasztott mondatelemzés ágrajzának megjelenítése.

A HG-1 Treebankben szereplő mondatok és elemzéseik adatbázisban történő rögzítését és kereshetővé tételét egy több fázisból álló feldolgozás előzte meg. Az előkészítés egy jelentős részét az XLE által előállított mondatelemzések összegyűjtése, egy köztes adatszerkezetre hozása (egy adott programnyelven), majd XML-formátumra alakítása képezte. A kiinduló pont az volt, hogy a keretrendszer az elemzéseket – amelyeket a korábban megírt mondatfeldolgozási és elemzési szabályok alapján állít elő – kimenetként Prolog programozási nyelven kódolt reprezentációban adja meg. A munkafolyamat ezen szakaszához tartozott tehát többek között a kimeneti Prolog fájlok (egy mondat a hozzá tartozó elemzésekkel = egy fájl) szerkezeti és tartalmi értelmezése, majd feldolgozása (vagyis a megfelelő adatszerkezet implementálása és az elemzések adatainak ebben az adatszerkezetben való rögzítése). Ezt követően kerülhetett sor a feldolgozott elemzések közül a (fentebb is említett) szabályok által létrehozható duplikált elemzések kizárólag egyszeres letárolására, a (gráfelméleti szempontból) köröket tartalmazó, és emiatt hibás összetevős szerkezetek kiszűrésére, valamint az ezáltal esetlegesen előálló elemzés nélkül maradt mondatok kizárására. További részfeladat volt a feldolgozás miatt a későbbiek során fontosnak bizonyuló információk, mint például az összetevős szerkezet (mint fa) mélységének és szélességének megállapítása, valamint az elemzések zárójelezett reprezentáció formájában történő elkészítése és letárolása olyan formában, amely az általunk is használt – az elemzések vizualizációját végző – phpSyntaxTree alkalmazás működéséhez szükséges.

A Prolog kódban tárolt elemzések egyedi, köztes adatszerkezetre való leképezése után lehetővé vált azoknak XML-alapú adatbázisba való építése. Az XML formátuma a már korábbi treebank alapú nyelvészeti alkalmazásokban is gyakran használt TigerXML lett, a fentebb is említett adatok (fa mélysége, szélessége; zárójelezett reprezentáció) tárolásához szükséges szerkezeti kibővítésekkel.

Az XML adatbázis létrehozását követően a további feladatokat az határozta meg, hogy egy weben használható, online lekérdezőfelületen keresztül – akár összetett feltételeket is tartalmazó, ugyanakkor viszonylag gyors – kereséseket lehessen végrehajtani a treebankben. Mivel a vállalt célok között szerepelt a keresési feltételek szóalakok és lemmák formájában történő megadása, továbbá a keresési lehetőségek részét képezte a szűrési feltételek morfológiai jegyekre és domináló szintaktikai kategóriákra szűkítése is, egy több táblából álló, SQL-alapú relációs adatbázis tervezése, valamint egy, a TigerXML forrásból adott SQL adatbázis formátumra alakító program készítése vált indokolttá. Így került sor egy megfelelő szerkezetű MySQL adatbázis kidolgozására, amelynek során fontos szempont volt, hogy külön táblában legyenek letárolva a mondatok, azok elemzései, az elemzésekben előforduló szóalakok morfológiai elemzésükkel, valamint azok lemmái szófaji kategóriájukkal. A kereshetőség felgyorsítása céljából a táblák indexekkel lettek ellátva.

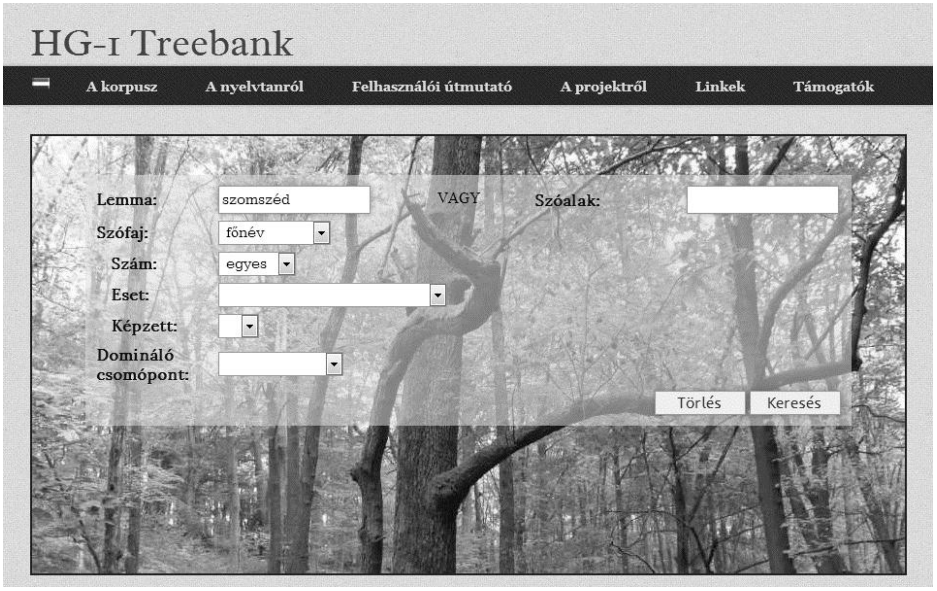
Az alkalmazás, amely a TigerXML forrás beolvasására, feldolgozására, és SQL-adatbázist töltő szkriptek írására lett kidolgozva, NIIF szuperszámítógépes környezetben került kipróbálásra és alkalmazásra. Ennek indokoltságát a relációs adatbázis jellegéből adódó, az integritás fenntartása érdekében végzett – a szóalakokra és lemmákra vonatkozó „szerepelt-e már”, „volt-e már” típusú – ellenőrzésekhez nélkülözhetetlen magas memóriaigény és hosszú, növekvő futási idő támasztotta alá.

A TigerXML forrásból ilyen módon létrehozott (MySQL) adatbázisból történő lekérdezésekhez ezt követően egy PHP-alapokon működő, weben elérhető interfész lett kialakítva, amelyben mint online űrlapban van lehetőség keresési kritériumok megadására. A keresendő adat minden esetben egy szóalak vagy lemma lehet, és további szűrési paraméterek is beállíthatók domináló szintaktikai csomópontok és morfológiai tulajdonságok/jegyek formájában. Ez utóbbiakat választómezőkön keresztül, opcionálisan lehet specifikálni. A keresési feltételeknek megfelelő eredményhalmazban megjelennek a mondatok és releváns elemzéseik is, az elemzések (mint összetevős szerkezetek) ágrajzos ábrázolására pedig a phpSyntaxTree (v1.10) alkalmazás került beépítésre az online keretrendszerbe.

Az interfész segítségével keresendő kifejezés minden esetben csak egy lemma vagy egy szóalak lehet. Ezek egyikének megadása kötelező, a keresés fő feltétele ezen alapszik. További szűrési kritériumként domináló csomópont bármilyen kereséshez beállítható, valamint szófaji kategória is megadható mint keresési feltétel. Ez utóbbi kiválasztása után (az adott szófajtól függő) további feltétel(ek)ként az egyéb morfológiai tulajdonságok szolgálhatnak. A szűrési feltételek szófaji kategóriánként csoportosított listája a következő:

- *főnév* (benne a tulajdonnevekkel): szám, eset, képzett-e;
- *határozószó*: képzett-e;
- *ige*: szám, idő, mód (feltételes/felszólító/kijelentő), műveltető-e, határozott-e, képzett-e;
- *igenév*: típus, szám, eset (csak melléknévi igenevek esetén), képzett-e;
- *kötőszó*;
- *melléknév*: szám, fok, eset, képzett-e;
- *névmás*: típus, szám, eset;
- *névutó*;
- *számnév*: szám, eset.

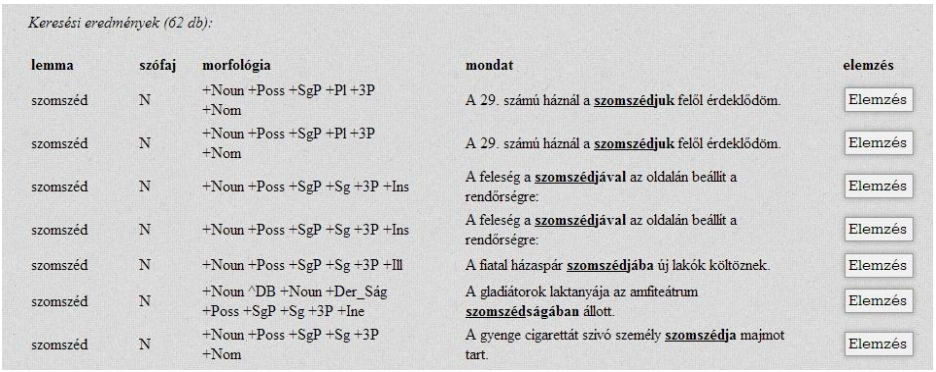
Az egy keresésre beállított szűrési opciók úgy működnek, hogy azok mindegyikének (egyszerre) teljesülnie kell a treebankben történő kereséskor. Az 5. ábra egy keresést szemléltet.



5. ábra: Keresés a *szomszéd* lemmára mint egyes számú főnévre.

A keresés eredményeit a lekérdezőfelület táblázatos formában jeleníti meg. Lemmára kereséskor listázásra kerül a lemma annak szófaji kategóriájával, azon szóalakoknak a morfológiai elemzése, amelyeknek az adott lemma tényleges lemmája, valamint a mondatok, amelyekben a lemma (bármilyen szóalakkal) előfordul. Abban az esetben, ha a fő szűrési paraméter szóalak, a táblázat értelemszerűen leszűkül a szófaji kategória oszloppal. A találatok számát a keresési mezők alatti szövegrész mutatja. A táblázatot képző eredménylistába a mondatok ábécésorrendben kerülnek.

A keresés találatainak áttekinthetőbbé tételét kiemelések is segítik (l. a 6. ábrát). Azon mondatokban, amelyekben egyszer fordul elő az adott keresési kifejezés (lemma vagy szóalak), lemmára kereséskor a mondat oszlopban aláhúzás jelöli a lemmát – amennyiben az morfo(fono)lógiai alternáció nélkül van jelen –, a szólak pedig félkövér betűstílussal jelenik meg, függetlenül attól, hogy lemmára vagy szóalakra szűrünk.



Keresési eredmények (62 db):

lemma	szófaj	morfológia	mondat	elemzés
szomszéd	N	+Noun +Poss +SgP +Pl +3P +Nom	A 29. számú háznál a **szomszédjuk** felől érdeklődöm.	Elemzés
szomszéd	N	+Noun +Poss +SgP +Pl +3P +Nom	A 29. számú háznál a **szomszédjuk** felől érdeklődöm.	Elemzés
szomszéd	N	+Noun +Poss +SgP +Sg +3P +Ins	A feleség a **szomszédjával** az oldalán beállít a rendőrségre:	Elemzés
szomszéd	N	+Noun +Poss +SgP +Sg +3P +Ins	A feleség a **szomszédjával** az oldalán beállít a rendőrségre:	Elemzés
szomszéd	N	+Noun +Poss +SgP +Sg +3P +Ill	A fiatal házaspár **szomszédjába** új lakók költöznek.	Elemzés
szomszéd	N	+Noun ^DB +Noun +Der_Ság +Poss +SgP +Sg +3P +Ine	A gladiátorok laktanyája az amfiteátrum **szomszédságában** állott.	Elemzés
szomszéd	N	+Noun +Poss +SgP +Sg +3P +Nom	A gyenge cigarettát szívó személy **szomszédja** majmot tart.	Elemzés

6. ábra: Találati lista kiemelésekkel a *szomszéd* lemmára mint egyes számú főnévre.

A találatok összetevős szerkezetének vizualizációja egy külön ablakban jelenik meg az adott mondat melletti „Elemzés” gombra kattintást követően. Egy ilyen példát mutatott be a 4. ábra. Az ágrajzban az alapértelmezés szerint külön színnel jelölt terminálisok szintjén a mondatban szereplő szóalakok morfológiai elemzései láthatók. (Ahol nem jelenik meg morfológiai elemzés, ott az a nyelvtanírás során már korábban felül lett írva LFG-formalizmuson alapuló lexikai tétellel.) Az összetevős szerkezetet szemléltető felület a rendszer sajátosságait kihasználva lehetőséget biztosít a megjelenítést kényelmesebbé tevő beállítások változatására is (pl. a betűméret változtatása, a terminálisok külön színnel való jelölése stb.), amely nagymértékben hozzájárul a Treebank komplex, integrált, ugyanakkor felhasználóbarát rendszeréhez.

A korpusz legfőbb célja és értéke a munkacsoport által kifejlesztett magyar LFG nyelvtan kézzelfoghatóvá és felhasználhatóvá tétele. Alkalmazható a nyelvoktatás és a nyelvtanulás területén – a korpuszalapú megoldások összes előnyével: motiváló autentikus élőnyelvi szövegekkel dolgozhatunk olyan módon, hogy a tanulás nyelvi felfedezéssé válik. Ugyancsak fontosak számunkra a lehetséges lexikográfiai alkalmazások, valamint a korpusz felhasználása elméleti nyelvészeti kutatásokban (melyre közvetlen példa saját nyelvtanírási projektünk is, amelyhez a korpusz folyamatos tesztelési lehetőséget és visszajelzést biztosított).

Köszönetnyilvánítás

A cikk elkészítését részben az OTKA K 72983 számú kutatási projekt, részben pedig a TÁMOP 4.2.1./B-09/1/KONV-2010-0007 számú projekt támogatta. A TÁMOP projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósult meg. Laczkó Tibor és Rákosi György kutatásait a Magyar Tudományos Akadémiának a Debreceni Egyetemen működő Elméleti Nyelvészeti Kutatócsoportja is támogatta.

Hivatkozások

1. Butt, M., King, T.H., Niño, N., Segond, F.: A grammar writer's cookbook. CSLI Publications, Stanford (1999)
2. Forst, M., King, T.H., Laczkó, T.: Particle verbs in computational LFGs: Issues from English, German, and Hungarian. In: Miriam, B., King, T.H. (eds.): Proceedings of the LFG'10 Conference. CSLI Publications, Stanford (2010) 228–248
3. Kornai, A., Halácsy, P., Nagy, V., Oravecz, Cs., Trón, V., Varga, D.: Web-based frequency dictionaries for medium density languages. In: Kilgarriff, A., Baroni, M. (eds.): Proceedings of the 2nd International Workshop on Web as Corpus ACL-06 (2006) 1–9
4. Laczkó, T., Rákosi, Gy.: On particularly predicative particles in Hungarian. In: Butt, M., King, T. H. (eds.): Proceedings of the LFG '11Conference. CSLI Publications, Stanford (2011) 299–319
   Online: http://cslipublications.stanford.edu/LFG/16/papers/lfg11laczkorakosi.pdf
5. Rákosi, Gy.: Non-core participant PPs are adjuncts. In: Butt, M., King, T. H. (eds.): Proceedings of the LFG '12Conference. CSLI Publications, Stanford (Megj.e.)
6. Rákosi, Gy., Laczkó, T.: Inflecting spatial particles and shadows of the past in Hungarian. In: Butt, M., King, T. H. (eds.): Proceedings of the LFG '11Conference. CSLI Publications, Stanford (2011) 440–460
   Online: http://cslipublications.stanford.edu/LFG/16/papers/lfg11rakosilaczko.pdf
7. The TIGER-XML treebank encoding format.
   http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/doc/html/TigerXML.html
8. XLE Documentation. http://www2.parc.com/isl/groups/nltt/xle/doc/xle_toc.html
9. XLE. http://www2.parc.com/isl/groups/nltt/xle/