Hunlex - morfológiai szótárkezelő rendszer

Trón Viktor*

Kivonat Cikkünkben[1] a HunLex szótárkezelő és morfológiai erőforrásgeneráló keretrenszert mutatjuk be. A HunLex lehetővé teszi, hogy egy könnyen fenntartható, átlátható de gazdagon specifikálható központi nyelvi adatbázisból kiindulva szószintű elemzőalkalmazások erőforrásait állítsuk elő. A HunLex prototípusa a Szószablya fejlesztés keretében megvalósított HunTools szóelemző eszköztár moduljai számára készít optimalizált nyelvspecifikus erőforrásokat, de elméletileg kész más rendszereket is kiszolgálni. A kimeneti erőforrások számos paraméter mentén igény szerint konfigurálhatók.

1. Bevezetés

A Szószablya projekt [4] legközvetlenebb célja egy nyílt magyar nyelvű morfológiai elemző kifejlesztése volt. Az ehhez szükséges nyelvi erőforrások – magyar morfológiai szótár és szabályrendszer – előállítását és továbbfejlesztését nagyban képes segíteni a HunLex előfeldolgozó komponens. A Szószablya szóelemző technológia [9,8] felépítését a 1. ábra szemlélteti.

A HunLex bemenete egy szakértői munkával létrehozott és fenntartott központi nyelvi adatbázis, kimenete pedig a valósidejű alkalmazások által közvetlenül értelmezhető erőforrás. Látható, hogy akárcsak a MorphBase elemző függvénykonyvtár rutinjai, úgy a HunLex is nyelvfüggetlen rendszer, amely két nyelvspecifikus morfológiai adatbázis közötti konverziót hivatott elvégezni.

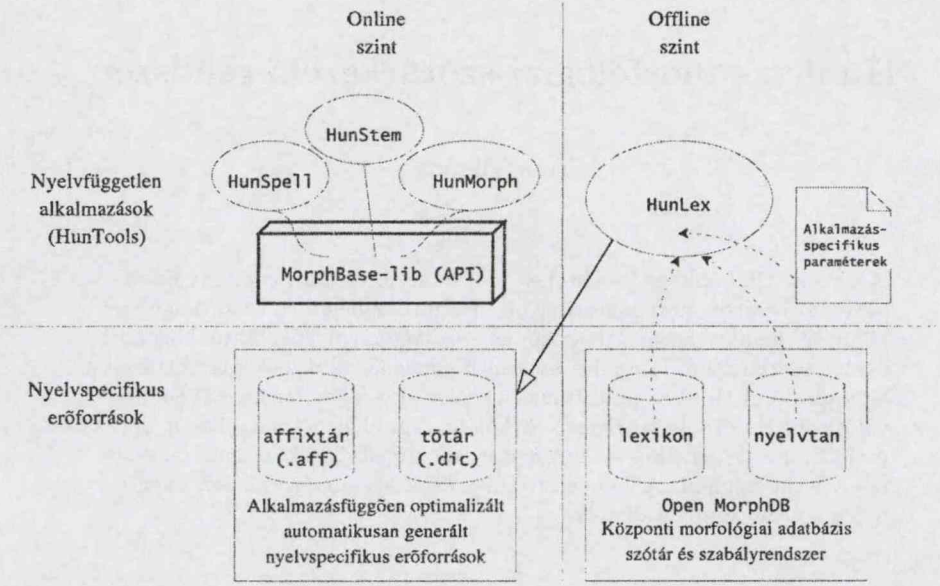
A cikk további részében ismertetjük a HunLex keretrendszert. Elsőként a HunLex elkészítésének motivációját tárgyaljuk (§2), majd röviden bemutatjuk a jelenlegi renszer fontosabb jellemzőit (§3). Végül a HunLex rendszer lehetséges további felhasználási lehetőségeit és a modul kiterjesztésére irányuló terveinket ismertetjük (§4).

2. Motiváció

Kényelmes bővíthetőség és fenntarthatóság. Alapvető elvárás, hogy egy valósidejű elemzőalkalmazás (helyesírásellenőrző, morfológiai elemző) nyelvfüggetlen legyen és az elemzéshez szükséges nyelvspecifikus tudást erőforrások formájában

* International Graduate College, Saarland University and University of Edinburgh, v.tron@ed.ac.uk

[1] Ezúton szeretnék köszönet mondani Halácsy Péternek, Konrai Andrásnak, Németh Lászlónak, Rung Andrásnak, Rebrus Péternek és Anne Benoitnak.



1. ábra. A Szószablya szóelemzési technológia felépítése

lehessen megadni. Az elemzés minőségét az erőforrásban megadott morfológiai és lexikai információ lefedettsége és pontosságát határozza meg. Emiatt nagyon fontos, hogy ezt az erőforrást könnyű legyen bővíteni és javítani. Az elemző hatékonyságának biztosítása érdekében azonban a nyelvi erőforrások formátuma gyakran nem alkalmas emberi feldolgozásra. Például a HunTools moduljainak erőforrásai bár szöveges állományok, igen redundánsak és nehezen áttekinthetőek, közvetlen szerkesztésük majdhogynem lehetetlen. Egy bonyolult morfológiájú, agglutinatív nyelv esetén az MorphBase affixumállománya számos toldalékmorf kombinációjából előálló affixumcsoportokat tartalmaz. Ha egy affixum viselkedését szeretnénk megváltoztatni, akkor az azt tartalmazó összes kombinációt figyelembe kell vennünk. Ez a feladat csak egy olyan keretrendszer segítségével végezhető szisztematikusan, amely lehetővé teszi a morfológiai szabályok és a lexikai elemek toldalákolási információinak rugalmas és következetes javítását.

Mindebből következik, hogy az elemzőalkalmazások erőforrásait érdemes offline automatikusan előállítani miközben az adatbázisok javítása és fenntartása átlátható központi formátum használatát igényli [1]. A kétféle szintű erőforrás között egy konfigurálható előfeldolgozó rendszer közvetít, egy ilyen keretrendszer mára a legtöbb elemzőtechnológiának része, így például a magyar nyelv elemzésére leginkább használt Humor rendszernek is [7].

Futásidejű elemzés hatékonysága. Mivel az elfogadás szempontjából nem fontos, hogy mit tekintünk tőnek illetve affixumnak, a helyesírás-ellenőrzőben az affixumok és tövek pontos meghatározása csak hatékonysági kérdésként merül fel. Egyes nyelvészetileg összetett alakok (általában a kivételes vagy improduktívan

toldalékolt alakok) felsorolással lehetnek kezelve, valamint a tő (itt: szótárban felsorolt sztring) fogalma nem azonos a lemma, vagy tőallomorf nyelvészetileg releváns fogalmával (például a *számat* „töve" *szám*, a *sarkam* „töve" pedig a *sarkak* a Magyar Ispell szótár eredeti állományaiban).

Hasonlóan, a kimeneti annotáció megvalósításához mind a tövek, mind az affixumcsoportok morfológiai annotációját meg kell adni. Gyakran előfordul, hogy a futásidejű elemzéskor használt tő–affixum felbontás nem feleltethető meg a kategóriák azonosítását szolgáló (és általában a morfológiai leírásként szolgáló) komponensekre-bontásnak.

Egyrészt számos imporduktív és kivételes alak a szótárban van felsorolva (pl. hatékonysági megfontolásokból), amelyeknak a morfológiai elemzését a lexikon kell, hogy kódolja. Másrészt egy affixumcsoport is potenciálisan tetszőleges számú morfológiailag releváns morf kombinációja lehet, ezért ezek „elemzését" is előre kódolnunk kell.

Az ilyen praktikus megfontolások azonban nem szabad, hogy befolyásolják a morfológiai elemzés kimenetét, vagyis az elemzés kimenete és az elemzés futásidejű implementációja ideális esetben függetlenítendő. Ugyanakkor a morfológiai adatbázis formátumát lehetetlen az egyes elemzési technológiák igényeihez optimálizni.

Algoritmusfüggő erőforrásoptimalizálás. Bár a helyesírás-ellenőrző tekinthető mint a morfológiai elemző egy leegyszerűsítése: ha a bemeneti szóalakhoz sikerül elemzést rendelni, akkor a szó helyes –, a kétféle elemzést hatékonyabb más módszerrel megoldani. Ugyanez igaz az információ-visszakereső (information retrieval) rendszerekben gyakran alkalmazott szótövező viszonyában, hiszen a tövek visszaadása során ugyan kezelni kell a tövek többértelműségét de például az egy kategorián belüli affixumtöbbértelműséget nem (irreleváns, hogy a *fürdik* alak 3SG-INDEF vagy 3PL-DEF). Egyértelmű tehát, hogy különböző elemzőrutinokhoz más és más erőforrás az optimális, előállításukat azonban érdemes egy központi adatbázisból automatikusan végezni.

Rugalmas alkalmazásfüggő erőforrásgenerálás. Az erőforrások alkalmazásfüggőségére további példa lehet, hogy egy morfológiai elemzőtől nagyobb rugalmasságot várunk el az akadémiai helyesírási szabályzat követésében, mint egy helyesírás-ellenőrzőtől (például hasznos, ha elemzi a gyakori **izület*, **lőjjünk*, vagy **adatbáziskezelő* szóalakokat is). Hasonlóan egy indexelésre használt szótövezőnél nem feltétlen hasznos, ha a szófaj-, illetve jelentős értelemváltozással járó képzések tövét adja vissza (például a *Sorstalanságról* töveként a *sors*-ot), ugyanakkor más feladatokhoz ez a tőinformáció hasznos lehet. Fontos szempont tehát, hogy egy központi adatbázisból szigorú, illetve engedékeny elemzők is előállíthatók legyenek, vagyis az erőforrásgenerálásnál lehetőséget kell adni az alul-, ill. túlgenerálásra.

3. Mit tud a hunlex?

Mindezen kívánalmak figyelembevételével terveztük meg a HunLex rendszert. A hunlex egy központi (gazdag információtartalmú) morfológiai adatbázisból dolgozik, de hogy pontosan milyen kimeneti erőforrást (a HunTools esetében ún. dic, illetve aff állományokat) kompilálunk, az számos szempont szerint változtatható.

Bemeneti források. A Hunlex konkrétan kétféle forrásból dolgozik: (i) a bázislexikon és nyelvtan a nyelv lexikonát és morfológiáját írja le; (ii) a többi állomány a kimeneti erőforrások kompilálását szabályozza.

A nyelv morfológiáját leíró hunlex lexikon és nyelvtan egyszerűen és átláthatóan specifikálható, így a folyamatos szótárbővítés és a morfológiai szabályok finomítása kényelmesen végezhető. A nyelvtanírást és a lexikon karbantartását segítik az egyszerűen definiálható makrók, amelyek reguláris kifejezésekhez is használhatók toldalékolási szabályok alkalmazási feltételeinek megadásához. Mivel lehetőség van a teljes nyelvtan és lexikon által generált nyelv előállítására, ezért a rendszerszerű tesztelés és a morfológiai leírás korábbi állapotaival való összevetés könnyen elvégezhető.[2]

Az erőforrásgenerálást vezérlő opciók beállításával a kimenet számos paraméter mentén konfigurálható.

- Állítható, hogy a kimenet helyesírás-ellenőrzés, tövezés, illetve morfológiai elemzés számára optimalizált dic illetve aff állományokat állítson elő.
- Kiválasztható, hogy mely toldalékolási szabályokat alkalmazza az elemző. Ezen belül megválasztható, hogy az elemző mely morfológiai szabályokat fogja alkalmazza futásidőben. Egyes morfológiai szabályok kompiláláskor alkalmazódnak a bázislexikon elemeire, így egyes morfológiailag komplex alakok is bekerülhetnek az elemző tőtárába. A hunstem tövező a tőtárból kikeresett tőinformációt adja vissza az elemzéskor, így ezzel az opcióval különböző mélységű tövezőket lehet kapni.
- Az futásidőben elemzendő toldalékmorfémák másik morfémákkal kombinálódhatnak és az eredményül kapott ún. affixumcsoportokat az elemző egy toldalékként (egy lépésben levágva) elemzi. Hogy mely toldalékok alkossanak csoportokat, azt a nyelvtantól függetlenül konfigurálható ún. szintek segítségével.
- Az egyes morfémaváltozatokat szabályozó morfofonológiai jegyek közül melyeket vegye figyelembe a rendszer. Bizonyos jegyek (részleges) kizárásával robusztus túlelemző nyelvtanok állíthatók elő.
- Korlátozható továbbá a rekurzív szabályalkalmazás mélysége.
- A morfológiai szabályok és a tövek különböző regiszter, ill. stílusjegyekkel lehetnek ellátva, amelyeket a kompilálás során figyelembe vesz a rendszer. Így például a helyesírásellenőrző számára szigorú normatív, egy robosztus elemző számára pedig hiperengedékeny forrás generálható.

[2] A hunlex morfológiai nyelvtant leíró formalizmusról és a specifikáció technikai részleteiről lásd a `http://www.szoszablya.hu` weboldalt.

- A kimeneti annotáció (tövező és morfológiai elemző számára) számos paraméter mentén konfigurálható. Többek között a hunlex képes beépített jegy-érték struktúrák kezelésére és unifikálására, ami igen rugalmassá képes tenni mind a kimeneti annotáció alakítását, mind a morfoszintaktikai kategóriák lexikai specifikációját [5].

A hunlex rendszer alkalmazása különösen hasznos olyan nyelvek leírására, amelyekhez szóelemző technológia nem áll rendelkezésre. Mivel a hunlex képes előállítani a megfelelő optimalizált erőforrásokat a nyílt licenszű HunTools csomag elemzőalgoritmusai számára, egyetlen egységes hunlex alapállomány segítségével akár ipari alkalmazásokba is beépíthető ellenőrző-, tövező- és morfológiai elemzőmodulok nyerhetők az adott nyelvre.

4. Lehetséges kiterjesztések

További erőforrás-formátumok. Bár a hunlex elsődlegesen a MorphBase szóelemző eszközkönytrár algoritmusainak kiszolgálására készült, egy intelligens szótárkezelőtől elvárható hogy további futásidejű elemzőprogramok bemeneti erőforrásait is képes legyen előállítani. Ilyen például a véges állapotú technológiát használó SFST, illetve XSFT. Jelenleg is folyik annak a vizsgálata, hogy a hunlex formalizmusban leírt morfológiai nyelvtanok hogyan kompilálhatók a fenti programok által használt erőforrások formátumára. Amennyiben a formalizmusok ereje kompatibilisnek bizonyul, várható, hogy a jövőben a hunlex ezeket az nyelvi erőforrásokat is képes lesz előállítani, illetve a különböző nyelvtanformalizmusok közötti konverziót elvégezni. Ezzel egyrészt biztosítható, hogy a hunlex nyelvtanokkal leírt nyelvek más elemzőkkel is használhatók legyenek. Másrészt, így a véges állapotú modellel leírt nyelvtanokat az affixumlevágással dolgozó MorphBase algoritmusai is megértik, és az adott nyelvekre rögtön helyesírásellenőrző- és tövező-alkalmazásokat is kapunk.

Szintén tervezés alatt van a hunlex lexikonoknak szabványos XML kódolásra való átalakítása. Ezzel a lexikai adatbázis portabilitása biztosítható, ami elősegíti a szótári információ szélesebb körben való használhatóságát. Erre felkészülve a hunlex alapszótárban már jelenleg is lehetséges az elemzőrutinok által nem használt információ felvétele tetszőleges attribútumok bevezetésével.

Nyílt magyar morfológiai adatbázis. A BME Médai Oktató és Kutató az MTA Nyelvtudományi Intézetének munkatársaival közösen egy nyílt magyar morfológiai szótári adatbázis fejlesztésén dolgozik. A leírás keretéül a hunlex szolgál. A hunlex lehetővé teszi, hogy az nagy lefedettségű és naprakész Magyar Ispell szótárat [6] összevessük az Akadémiai Nagyszótárral (pontosabban az Értelmező Kéziszótárban önálló címszóval szereplő szókinccsel, amely Papp Ferenc Debreceni Tezauruszán keresztül digitális formában szabadon elérhetővé vált [3]), valamint a Magyar Ragozási Szótárral [2]. Ezeknek az adatbázisoknak a kritikus összefésülésével az eddigi legteljesebb magyar morfológiai nyelvtan és szótári adatbázis készülhet el és válhat szabadon elérhetővé. A HunLex keretrenszer

biztosíték arra, hogy a szótári adatbázis nagy lefedettségét és pontosságát a HunTools programcsomag szóelemző moduljai kihasználhassák és így leíró célja mellett az adatbázis közvetlenül a magyar nyelvtechnológia hasznára lehessen.

Hivatkozások

1. I. Aldezabal, O. Ansa, B. Arrieta, X. Artola, A. Ezeiza, G. Hernández, and M. Lersundi. Edbl: a general lexical basis for the automatic processing of basque. In *IRCS Workshop on linguistic databases. Philadelphia*, pages 1–10, 2001.
2. László Elekfi. *Magyar ragozási szótár*. MTA Nyelvtudományi Intézet, Budapest, 1994.
3. Mihály Füredi, András Kornai, and Gábor Prószéky. A szótár adatbázis. Kézirat, 2004.
4. Péter Halácsy, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón. A szószablya projekt. In *Proceedings of the 1st Hungarian Computational Linguistics Conference*. Szegedi Tudományegyetem, 2003.
5. András Kornai, Péter Rebrus, Péter Vajda, Péter Halácsy, András Rung, and Viktor Trón. Általános célú morfológiai elemző kimeneti formalizmusa. In *II. Magyar Számítógépes Nyelvészeti Konferencia*, 2004.
6. Németh László. Magyar Ispell – válasz a Helyes-e?-re. In *IV. GNU/Linux szakmai konferencia*, pages 99–107. Linux-felhasználók Magyarországi Egyesülete, 2002.
7. Attila Novák. Milyen a jó humor? In Zoltán Alexin and Dóra Csendes, editors, *Magyar Számítógépes Nyelvészeti Konferencia*, pages 138–144. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szegedi Tudományegyetem, 2003.
8. László Németh, Péter Halácsy, András Kornai, and Viktor Trón. Nyílt forráskódú morfológiai elemző. In *II. Magyar Számítógépes Nyelvészeti Konferencia*, 2004.
9. László Németh, Viktor Trón, Péter Halácsy, András Kornai, András Rung, and István Szakadát. Leveraging the open-source ispell codebase for minority language analysis. In *Proceedings of SALTMIL 2004*. European Language Resources Association, 2004.