lile_MSZNY2004 előadás

1. dia

Sok szeretettel köszöntök mindenkit! [majd meglátjuk, az előttem előadó hogyan kezdi, ez a köszönés ahhoz lesz szabva] Viszket Anita vagyok, és egy négy fős csapatot képviselek, ahogy ez a diáról is kiderül. Az előadásunk címe: LiLe projekt, vagyis adatbázis mint dinamikus korpusz.

Az előadást Quntilianus kérdései alapján építettük fel. Feltűnhet, hogy a kérdések sorrendjén kicsit változtattunk, a klasszikus *ki*, *mit*, *hol* stb. kezdettel szemben a *mit* kérdésre mi egy kicsit később válaszolunk. Először tehát bemutatkozunk a *ki*, *hol*, *milyen segítséggel* kérdésekre válaszolva, majd ismertetjük a projektünket a *mit*, *miért*, *hogyan*, *mikor* kérdések segítségével.

klikk

2. dia

Az első kérdés tehát az, hogy *ki*? Négyen alkotjuk ezt a csapatot, akik a LiLe projekttel való foglalatoskodást szemelték ki egyik hobbijukul.

klikk

A LiLe kifejezés amúgy a Linguistic Lexicon rövidítése. Hogy mit értünk ez alatt, azt az előadás második részében részletesen elmagyarázzuk.

klikk

3. dia

A *hol* kérdésre azt válaszolhatjuk, hogy Pécsett. A projekt tagjai mindannyian a Pécsi Tudományegyetem Bölcsészettudományi Karának Nyelvtudományi Tanszékéhez kötődnek, ki oktatóként, ki hallgatóként.

klikk

4. dia

Milyen segítséggel. Három programozó segíti a munkánkat, mindannyian a pécsi Dexter Kft munkatársai. A rendszertervünk alapján ők készítették a feltöltő programot, és a folyamatábráink alapján az elemzőt.

klikk

5. dia

Immár sort keríthetünk az eddig elhanyagolt kérdésre: *mit* is csinál a LiLe projekt. Az előadás címében a dinamikus korpuszt emlegettük. Mivel ez egy általunk bevezetett terminus, illik ezen a ponton megmagyarázni, mire is gondoltunk.

Kezdjük a korpusszal. Egy hagyományos korpuszban morfémákat, szóalakokat, szószerkezeteket (egészen a mondatokig vagy több mondatot tartalmazó szövegrészleteig) találunk. Ráadásul mindezeket annotálva. Az annotálás történhetett kézzel vagy gépi módszerekkel, de az minden hagyományos értelemben vett korpusznak közös tulajdonsága, hogy a tárolt nyelvi egységek valamikor valahol elhangzott, leírt nyelvi egységek vagy azokból származó darabok. Így értjük azt, hogy „létező szövegekből" származik a hagyományos korpusz tartalma.

Hogyan definiáljuk a dinamikus korpuszt?

klikk

A dinamikus korpuszba a már létező leíró nyelvtanok alapján töltjük fel a zárt szóosztály elemei közé tartozó morfémákat (így értettük azt, hogy „elméleti alapokon"), kézi vagy gépi módszerrel rögzítjük a nyílt szóosztály elemeit, de a rögzítés mindig csak a disztribúciósan kimutatható legkisebb nyelvi elemre terjed ki, az összes többi elem generálás útján jön létre a korpuszban, a rögzített szabályok figyelembevételével. Vagyis a dinamikus korpuszban nem a valaha létező szavak és mondatok vannak, hanem a lehetséges szavak és mondatok, amelyek nincsenek is benne az adatbázisban, csak lehetséges őket abból generálni. Így a korpusz alkalmasabbá válik a kutatások támogatására. Természetesen ez nagyon gondos szabályrögzítést és rengeteg tesztelést igényel.

klikk

A dinamikus korpusz definiálásával tulajdonképpen a lilét definiáltuk, ami éppen ezekkel a tulajdonságokkal rendelkezik: a morfémákat és a tulajdonságaikat rögzítjük, és ezáltal tudunk szavakat vagy mondatokat generálni.

klikk

6. dia

Felmerülhet a kérdés, hogy *miért* pont ezt a célt tűztük ki magunk elé, ami semmiképp nem kecsegtethet gyors sikerrel, ha csak a fejlesztési és az adatrögzítési feladatokat tekintjük, már akkor sem. Tehát sok befektetést követelő és nagyon lassan megtérülő vállalkozásról van szó.

A következő okok motiváltak minket mégis arra, hogy belevágjunk:

klikk

Mind oktatóként, mind hallgatóként rendszeresen beleütköztünk abba a problémába, hogy milyen jó lenne az oktatásban valamilyen korszerű leíró magyar nyelvtan. A korszerű leíró nyelvtan akkor lenne számunkra a leghasználhatóbb, ha könnyen lehetne benne jelenségekre keresni. Továbbá az sem lenne baj, ha a saját észrevételeinket is tudnánk benne rögzíteni, netalántán még valahogyan ellenőrizni is. Azt is nagyon fontos szempontnak tartjuk, hogy egy ügyesen felépített, jó keresési lehetőségeket biztosító, esetleg saját szabályok rögzítésére alkalmas nyelvészeti adatbázis hatékonyan segíthetné a oktatást, nem csak felső-, hanem közoktatási szinten is.

klikk

Kutatóként is nagyon hiányzik egy olyan korpusz, amiben nem csak konkrét morfémákra, vagy esetleg a szűken értelmezett annotációs adatokra lehet keresni, hanem pl. tetszőleges szerkezetre is. Továbbá nagyon jó lenne egy olyan keret, amelyben tesztelhetnénk, mint kutatók, a nyelvészeti elképzeléseinket. Természetesen olyan keretet nem lehet építeni, amelyben minden grammatikai modellben felállított szabály egységesen ábrázolható. De abból a feltevésből indultunk ki, hogy minden grammatikai modellben megfogalmazhatóak ugyanazok az állítások, legfeljebb át kell fogalmazni őket egyik grammatikai modellről a másikra. Tehát ha tudnánk egy olyan könnyen kezelhető felületet létrehozni, ahol ugyan egy általunk létrehozott grammatikai modellben, de bármely kutató könnyen tudna szabályokat rögzíteni, akkor ez nagy segtség lehetne a nyelvészeti kutatásokban, nem utolsó sorban a saját kutatásainkban.

klikk

Egy harmadik szempont az volt, hogy az eddigi munkásságunk során mind a négyen foglalkoztunk adatbázisokkal, hosszabb-rövidebb ideig láttunk el mindenféle adatbázissal kapcsolatos feladatot, kezdve rendszertervezéstől, a rendszergazdai feladatokon át egészen az adatrögzítés irányításáig. Ezért motiváltak voltunk abban, hogy az általunk legjobban ismert MS-SQL adatbázist válasszuk formális kereteként a nyelvészeti lexikonunknak.

klikk

Összefoglalva az eddig elmondottakat: az elég nagyívű elképzeléseink szerint a LiLe egyszerre szolgál majd nyelvtani keretként működő adatbázisként, és az adatbázisban rögzített adatok alapján a magyar nyelv leíró szintű elemzéseként.

klikk

7. dia

Hogyan álltunk neki ennek a hatalmas tervnek? Először van az adatbázisunk, amely morfémákat tartalmaz, továbbá a nyelv különböző szintjeihez tartozó szabályokat, valamint természetesen a szabályok és a morfémk kapcsolatát, vagyis azt, hogy melyik morfémához milyen szabályok tartoznak.

klikk

A nyelvtani keretrendszerünk a GASG, a Generatív Argumentumszerkezet Grammatika, amely Alberti Gábor nevéhez fűződik. Ennek alapján feltételezzük, hogy mind a fonológiai, mind a morfológiai, mind a szintaktikai szabályok megadhatóak a morfémáknál rögzített, az adott morfémának a többi morfémával való kapcsolatáról állítást tevő szabályokkal.

klikk

Az adatbázison kívül van egy szoftver, amely egyszerre végzi el az elemzést és a generálást. Az elemző szabályok és a generáló szabályok nem különböznek lényegesen egymástól, a generáló szabályok generálta szóalakok közül valójában az elemző szabályok válogatják ki a megfelelőket. Maga az elemzés és a generálás nem más, mint az adatbázisból történő

lekérdezés, a szoftverben megadott sorrendben, továbbá az elemző meghívásakor kijelölt elemhalmazokon. A morfémákat jelenleg a nyelv alapján lehet szűrni, a szabályokat pedig az aktívságuk alapján.

Erről az aktívságról érdemes lenne még pár szót ejteni.

Az adatbázisunk felépítése, valamint az elemzőben lévő lekérdező szabályaink lehetővé teszik, hogy egy-egy szabályt ki-be kapcsolgathassunk. Ez a ki-be kapcsolgatás a lehető legkülönbözőbb szinteken mehet végbe: történhet magánál a szabályt tartalmazó rekordnál, vagy annak egy értékénél, de a morfémák egy halmazán vagy csak egy adott morfémánál is. Megvilágítom egy példával. A hangrendi illeszkedés alapján vannak szavak, amelyek mély, és vannak, amelyek magas hangrendű toldalékot várnak. Kikapcsolható ez a szabály általában, és ekkor minden hangrendi hibát tartalmazó szóalak (ha más hibát nem tartalmaz) grammatikus lesz a program számára. Kikapcsolható a mély hangrendű toldalékot váró szavak hangrendi illeszkedése a magas hangrendtől függetlenül (bár éppen ebben a példában egy ilyen kikapcsolásnak nem sok értelme van, hiszen a hangrendnél a viszony a morfémák között a legtöbb esetben szimmetrikus). De egyetlen morfémánál is kikapcsolható a szabály. A szabályoknak ez a dinamikus kezelése egyrészt a keretnyelvtanként való működést segíti, másrészt egy szintén fontos célkitűzésünket, a nyelvtantanulás motiválását, ugyanis egy igen érdekes, játékos eszközt adhat a kezünkbe az oktatásban.

klikk

Összefoglalóan: a Lile tehát egyszerre tud elemzőként és a generálással egy dinamikus korpuszként működni.

klikk

8. dia

A generálás egyelőre folyamatábra szintjén létezik a projektünkben, a megvalósítás az elkövetkező hónapok feladata. Az elemzésre azonban hoztunk egy példát a programunkból. A példa kiválasztásánál nem az motivált minket, hogy minél bonyolultabb szóalakot találjunk: valljuk be, morfológiai elemző sok van már az országban, nem okoz senkinek meglepetést valószínűleg, hogy az adott szóalakot elemekre tudjuk bontani. Az illusztrációval inkább azt akarjuk megmiutatni, hogy az elemzés és az eredménykiíratás különböző paraméterezéseivel milyen érdekes, kicsit mulatságos eredmények is jöhetnek elő.

Az elemzőben beállíthatjuk, hogy csak azokat az elemzéseket akarjuk-e látni, amelyek grammatikusnak bizonyultak, vagy az agrammatikusokat is. Most az állítottuk be, hogy csak a grammmatikus eredményeket akarjuk látni. Ilyenkor a felső pipálási lehetőség irreleváns.

klikk

Mint láthatjuk, az elemző kiírja, hogy az *ajtóban* szó az *ajtó* és a *ban/ben* morfémákból áll, valamint ez egy grammatikus szóalak eszerint az elemzés szerint.

klikk

Mi történik, ha azt jelöljük be, hogy az agrammatikus szóalak-értelmezéseket is látni szeretnénk. A felső pipa marad, ami azt jelenti, hogy szűrje ki az eredményből az elemző azokat a szóalakokat, amelyekben az elemzés nem egyező szófajú morfémákat mutatott ki.

klikk

Az eredmény az, hogy lehetséges az *ajtóban*-nak az az elemzése is, amikor a *ban* valójában két morféma, a *ba/be* és az *on/en/ön*. Azonban ez a szóalak nem lenne grammatikus, mivel a *ba/be* morfémát nem követheti az *on/en/ön*.

klikk

Az elemzés utolsó változata az, amikor az elemzővel kiíratjuk az olyan találatokat is, ahol a megtalált morfémák szófajilag ütköznek egymással.

klikk

Az eredménylista második sorából látható, hogy ekkor a *ban* sztring értelmezhető úgy is, hogy *ba/be* + a feltételes mód jele, ez azonban agrammatikus szóalakot eredményez, mivel a két morféma szófaja nem egyezik meg.

Az elemzés többi sorában az olvasható, hogy az *ajtó* mellett feltételezhető lenne még az *a* mint névelő is szótőként, utána a *j* mint a felszólító mód jele, a *t*-nek többféle értelmezése is lehetséges (tárgyrag, múlt idő jele, a befejezett melléknévi igenevet jelölő morféma), az *ó* mint a folyamatos melléknévi igenév jelölője, és végül a *ban* sztring felbontásának három lehetséges módja következik. A sok elemzési sor ezen lehetőségek összeszorzásából származik, ezért található összesen 9 elemzési lehetőség a névelővel kezdődő sorokban. Természetesen ezek a szófaji ütközéseken túl még morfémasorrendben is ütköznek, ezeket a problémákat mind kiírja az elemző az érintett morfémák megjelölésével, legfeljebb ez a hosszú sor miatt itt most nem olvasható.

Azért választottuk ezt a kicsit humoros elemzésű példát, hogy egyrészt illusztráljuk, hogy a főnévi és igei inflexiós morfológiát kezeli a program, másrészt hogy megmutassuk, hogy az automata elemző eljárás olyan eredményeket is adhat, amire az ember magától soha nem gondolna (hiszen annyira értelmetlenek nyelvtanilag), viszont a feltárásuk például az oktatásban érdekes és tanulságos lehet.

klikk

A következő dián a *mikor* kérdésre válaszolva azt összegezzük, hogy milyen eredményeket értünk el eddig, min dolgozunk most, és milyen munkákat tervezünk a jövőben.

A konferencia előtt is tisztelegve a két konferencia-alkalomra fűztük fel az eddigi teljesítéseinket.

A tavalyi tanévben hoztuk létre az adatbázist, és teremtettük meg a szoftverünk alapjait. A projektünk valójában már korábban indult, amikor olyan tanulmányokat végeztünk pár féléven át, amelyek szükségesnek látszottak ehhez a munkához (SQL, XML, objektumorientált programozás stb.), és amelyeket magyar szakosként korábban nem tanultunk. Tavaly készült el a névszói elemzések rendszere, amiről beszéltünk is egy évvel

ezelőtt. Közben tartottunk olyan kurzust a Pécsi Egyetemen, amelyen magyar és idegen nyelv szakos hallgatók jellemezték a rendszerünkben a magyar névszókat, és sok ötletet adtak az adatbázis módosításához is.

klikk

Az idei tanévben ugyanezzel a munkamódszerrel az igei fonológia-morfológia volt soron. Azoknak, akik bármilyen számítógépes megoldáson dolgoznak, felesleges is részleteznünk, hogy a gyakori módosítás mellett párszor az is előfordult, hogy az elmúlt egy évben újabb felismerések alapján teljesen elölről kellett építenünk a struktúrát és elölről kellett írnunk a programokat is. És természetesen az sincs kizárva, hogy a jövőben hasonlókra kényszerülünk. A célunk egy annyira gumi-struktúra kialakítása, amiben minden jelenség kezelése megoldható, de ugyanakkor a rendszer még működőképes, és a feltöltése (akár a grammatika kialakítása, akár a morfémák jellemzése) nem igényel programozói ismereteket, sőt az adatbázis-struktúra ismeretét sem.

klikk

Már kidolgoztuk a jelenlegi struktúránkban azoknak a szavaknak az elemzését, amelyekből a szótő nincsen meg az adatbázisunkban. Ezeket nevezzük a rendszer szempontjából nonszensz tőnek. Ugyancsak elkészült a generálás folyamatábrája, jelenleg ennek a két feladatnak a programozása folyamatban van.

klikk

A projektünk hallgatói szakdolgozat keretében dolgoznak a modellben a képzők kezelésén, mert a képzők természetesen egészen más módon kezelendőek nem feltétlenül az elemzés (bár a szófaj- és argumentumstruktúraváltozások miatt ott is), sokkal inkább a generálás során.

A szintaxis megoldásában is a GASG elveit fogjuk követni (valójában már alig várjuk, hogy azzal foglalkozhassunk!), természetesen az adatbázis-struktúránkra alkalmazva, és ezáltal némileg módosítva a GASG-ben mint elméleti keretben alkalmazott megoldásokat.

Folyamatban van az automatikus adatbázis-feltöltés kidolgozása. Lényegileg már kitaláltuk, de az előbb említett folyamatok megvalósítása még változtathat az elképzelésünkön.

klikk

Ez a felsorolás a távolabbi jövőbeli terveinket tartalmazza. Ezek közül kiemelném az oktatás-támogatást, ami nem feltétlenül nyelvészeti, sokkal inkább fejlesztési feladatot jelent. Vannak reményeink arra, hogy ilyen feladatban együtt működhetnénk felnőttoktatással foglalkozó szakemberekkel, de szeretnénk az ötleteinket pedagógus továbbképzésekben is hasznosítani.

A programunk webes felületre való átültetése a külsősök számára teszi elérhetőbbé a rendszert. A nyelvészeti nagyon távoli fejlesztésekben pedig szintén a GASG megoldásait fogjuk alapul venni.

Természetesen tisztában vagyunk azzal, hogy a magunk elé kitűzött cél nagyon merész és nagyívű, főleg egy négyfős, nonprofit, mondhatnánk hobbi-csapat számára. De azt

tapasztaljuk, hogy már maga a munka is nagyon sok új felfedezéssel jár, és kifejezetten élvezetessé teszi a magyar nyelvtan tanulását mind a magunk, mind a részt vevő hallgatók számára. Abban reménykedünk, hogy minimálisan ezen a területen, az oktatástámogatás területén fogunk tudni még ebben a kicsi projektben is eredményeket felmutatni. A nagyobb célkitűzésekre pedig talán előbb-utóbb találunk partenereket is.

klikk

Köszönöm a figyelmet!