Igei szerkezetek gyakorisági szótára

Egy automatikus lexikai kinyerő eljárás és alkalmazása

doktori (Ph.D.) disszertáció tézisei

Sass Bálint

témavezető:
Prószéky Gábor,
az MTA doktora

Pázmány Péter Katolikus Egyetem,
Információs Technológiai Kar,
Multidiszciplináris
Műszaki Tudományok
Doktori Iskola

Budapest, 2011.

Bevezetés

'*Részt vesz vmiben.*' '*Górcső alá vesz vmit.*' Bár a természetes nyelvfeldolgozás kutatási hagyományában és a szótárírási hagyományban is két elkülönült területet jelentenek az igei vonzatkeretek és a többszavas kifejezések, számos nyelvben előfordulnak az effajta bonyolultabb szerkezetek, melyek *egyszerre* igei vonzatkeretek és kollokációk. Ezeket a szerkezeteket (legalább) két tartalmas elem – alapesetben egy ige és egy (ragos/névutós/elöljárós) névszó – alkotja, és ezen kívül még egy (vagy több) vonzat is szerves része a szerkezetnek. A fenti magyar nyelvűek mellett valóban számos nyelven látunk ilyenre példát: '*get rid of*' (angol; megszabadul vmitől), '*få lov til*' (dán; engedélyt kap vmire), '*imati pravo na*' (szerb; joga van vmihez), '*houden rekening met*' (holland; számításba vesz vmit), '*zijn van toepassing op*' (holland; vonatkozik vmire), '*avoir effet sur*' (francia; hatása van vmire).

Az idézett példákban az ige mellett mindig két bővítményt találunk: az egyiket egy konkrét, kötött szó tölti ki, ez alkot jelentéshordozó kollokációt az igével, a másik bővítménynek viszont csak a helyét jelöli ki a szerkezet egy esetrag vagy elöljáró segítségével. Látható, hogy általában ugyanazokkal a nyelvi eszközökkel – ragokkal, névutókkal, elöljárókkal vagy szórendi megkötéssel – kapcsoljuk a bővítményeket az igéhez; függetlenül attól, hogy a kollokátumról, vagy pedig a vonzati helyen éppen megjelenő tartalmas szóról (ilyen lenne például a '*játék*' a '*részt vesz a játékban*' esetén) van szó. A '*részt vesz vmiben*' szerkezetben például a kollokátum jelenik meg tárgyként, a '*górcső alá vesz vmit*' szerkezetben pedig a vonzat. Ez a váltakozás persze egy-

azon ige szerkezetei között is előfordulhat, a '*pillantást vet vkire*' és a '*szemére vet vmit*' szerkezet egyaránt tárgyat és egy '*-ra/-re*' ragos bővítményt tartalmaz, de az elsőben a tárgy a kollokátum és a '*-ra/-re*' ragos bővítmény a vonzat, a másikban pedig éppen fordítva.

Az ilyen szerkezetek – sokszor anyanyelvi intuíciónkkal ellentétes módon – kifejezetten gyakoriak, nagyon fontos szeletét képviselik egy nyelv szerkezeteinek, nem kezelhetők marginális esetként. Sokszor bírnak a részeikből nem levezethető, azaz nem kompozicionális, idiomatikus jelentéssel, kiemelt fontosságú részét kell hogy képezzék az emberi felhasználásra szánt szótáraknak és az automatikus nyelvfeldolgozó eszközök nyelvi erőforrásainak egyaránt. Fordításaikat a legtöbb esetben érdemes külön egységként tárolni, mert gyakran nem megjósolható elemeket tartalmaznak.

Szükség van tehát egy olyan adatközpontú számítógépes eljárásra, mely rendet vág a bővítménykijelölő nyelvi eszközök egymást átfedő erdejében, szétválasztja a konkrét, kötött szót tartalmazó és a szabadon kitölthető bővítményeket. Megállapítja, „felfedezi", hogy az egyes igei szerkezeteknek melyik bővítményi szó szorosan vett része kollokációként, és mely szükséges vonzati helyek kapcsolódnak még hozzá, azaz képes arra, hogy korpuszból kinyerje a jellegzetes igei szerkezeteket. A dolgozat fő eredménye ez az algoritmus (3.3. rész a dolgozatban), illetve az ennek közvetlen felhasználásával készülő magyar, egynyelvű igeiszerkezet-szótár (4.2. rész a dolgozatban).

A szótár – mely az igei szerkezetek legegyszerűbb modelljére építve készült – kézzelfoghatóvá teszi az igei szerkezeteket kinyerő algoritmus hasznosságát. A módszer igazi jelentőségét azonban az adja, hogy több irányban is kiterjeszthető. Egyrészt a modell nyelvfüggetlensége miatt megfelelő nyelvspecifikus előfeldolgozást követően számos nyelvre módosítás nélkül alkalmazható a kinyerő eljárás, így különféle nyelvű hasonló szótárak állíthatók elő. Másrészt nagyobb bonyolultságú szerkezetek – ld. például a fent említettekhez képest még egy jelzői kollokátumot is tartalmazó '*gyenge lábakon áll*' vagy '*száraz lábbal kel át vmin*' szerkezeteket –, valamint főnévi, melléknévi stb. központú szerkezetek feltérképezésére is alkalmas. Harmadrészt pedig –

a modell speciális alkalmazása révén – ugyanez az említett kinyerő algoritmus alkalmassá tehető párhuzamos igei szerkezetek, azaz igei szerkezetek és fordításaik azonosítására is. Ezen a módon az eljárás képes felfedni az egymásnak megfelelő, de formailag egymásra nem hasonlító aszimmetrikus szerkezetpárokat is, valamint a szerkezetek különféle (szinonim) idegen nyelvű megfelelőit és azok gyakorisági viszonyait is.

Módszertan

A lexikográfia egyik aktuális kérdése az, hogy a számítógép segítségével mennyire tudjuk *automatizálni* a szótárírás egyes lépéseit. A szótár anyagát automatikusan *korpuszból* kiindulva állíthatjuk elő, gyűjthetjük össze. Kutatásomban a szigorúan *korpuszvezérelt* megközelítés szerint járok el. A korpuszt nem csupán segédeszközként, vagy előzetes hipotézisek alátámasztására/cáfolatára használom, hanem korpuszt hitelesnek és reprezentatívnak elfogadva az igei szerkezetekről szóló nyelvi tudást kizárólag korpuszmegfigyelések alapján állítom össze. A korpuszvezérelt anyaggyűjtés során automatikusan dől el, hogy mik a jellegzetes igei szerkezetek, és közülük – korpuszgyakoriság alapján – melyik kerül be a szótárba és melyik nem. A mai nagyméretű korpuszok már biztos alapot nyújtanak a ritkább jelenségek karakterizálásához is.

Az utóbbi évtizedekben a korpuszvezérelt lexikográfia eredményei sok tekintetben forradalmasították a szótárkészítést. Az egyik fontos eredmény a *több szóból álló lexikai egységek* – kollokációk, frazémák, idiomatikus kifejezések, állandósult szókapcsolatok – jelentőségének felismerése és a korábbinál sokkal hangsúlyozottabb megjelenítése az új szótárakban. Sinclair szerint „a legtöbb jelentés realizációjához szükséges, hogy egynél több szó jelenjen meg a szövegben.” Kutatásomban a formailag különböző szerkezeteket, az egyszavas és többszavas nyelvi elemeket – az igéket és az igei szerkezeteket – egységes keretben kezelem. A szótár készítése során a többszavas igei kifejezéseket, szerkezeteket teljes jogú lexémákként a szótárkészítési folyamat kö-

zéppontjába állítom, amint ezt a bevezetőben említett példák is mutatják. *Típusfüggetlen* megközelítésem lehetővé teszi, hogy minden esetben a teljes szerkezetet reprezentálhassam, azaz ne maradhasson el a szerkezet egésze szempontjából lényeges elem. A szerkezetek teljessége a kiértékelés során is hangsúlyos követelményként szerepel.

A magyar nyelv szórendje szabad, legalábbis abban az értelemben, hogy a mondatban az ige és bővítményei szinte tetszőleges sorrendben elhelyezkedhetnek, közéjük egyéb szereplők ékelődhetnek. Más szóval: az igei szerkezetek lehetnek folytonosak és megszakítottak, bármilyen sorrendi variánsban előfordulhatnak. A szórendi variabilitás kezelése úgy oldható meg hatékonyan, ha a magyar nyelv leírására a nyelv természetéhez jól illeszkedő *függőségi nyelvtan* nyelvelméleti keretet választjuk. A függőségi leírásban általában szavak szoktak lenni az alapelemek. Kutatásomban ezzel szemben a *morfémát* választottam alapelemnek, hogy a szavakon kívül az ige és a bővítmény közötti viszonyt kifejező elemeket (az esetragokat) önálló elemként értelmezhessem. Az igei szerkezetek gyűjtésekor nem a szokásos megközelítést követem, mely csak a szavak egymás-mellettiségét tekinti, hanem jelen esetben egy szerkezet elemei mindig konkrét *függőségi viszonyban* vannak egymással. Ezek a függőségi viszonyok maguk is teljes jogú elemei lesznek az igei szerkezeteknek, ezáltal az említett egységes keret magában foglalja a kollokátumot nem tartalmazó igei szerkezeteket – köztük az igei vonzatkereteket – is.

Új tudományos eredmények

A dolgozat jellegzetes igei szerkezetek korpuszból való kinyerésével foglalkozik. Elsősorban azokra az igei szerkezetekre koncentrál, melyek egyszerre többszavas kifejezések és vonzatkeretek, azaz a vonzattal rendelkező komplex igékre. Ilyen például a '*hasznot húz vmiből*', az '*igényt tart vmire*' vagy az '*lehetővé tesz vmit*'. Ezek a szerkezetek lexikálisan szabad bővítményt, LSzB-t ('*vmiből*', '*vmire*', '*vmit*'), és lexikálisan között bővítményt, LKB-t ('*hasznot*', '*igényt*', '*lehetővé*') is tartalmaznak.

Az első feladat az volt, hogy kidolgozzak egy olyan modellt magyar nyelvre, mely az igei szerkezetek összes típusát – különös tekintettel a fent említett típusra – ábrázolni képes. Erre egy speciális függőségi elemzés alapú gráf volt a legalkalmasabb.

A modell kialakításával a dolgozat 2.1. részében foglalkozom, az új eredményeket a következőképpen foglalhatjuk össze:

1. tézis.
Kidolgoztam magyar nyelvre egy olyan modellt, mely képes a tagmondatok, illetve a bennük rejlő formailag nagy mértékben különböző igei szerkezetek egységes reprezentálására. A reprezentáció alapegysége a tagmondat, mely egy központi ige és a hozzá tartozó bővítmények összességét jelenti. A bővítményeket legfontosabb tartalmi elemükkel (névszói csoport bővítmény esetén a bő-

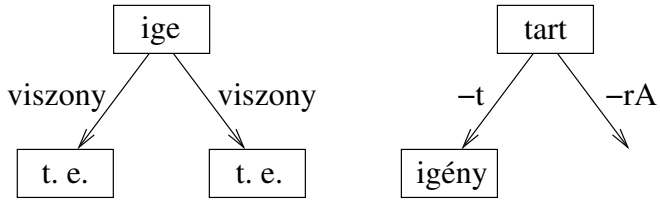
vítményt képviselő csoport feje) és a bővítményt az igéhez kapcsoló függőségi viszonnyal (névszói csoport bővítmény esetén az esetrag vagy névutó) jellemzem. Összefoglalva:

tagmondat = ige + bővítmények halmaza

bővítmény = viszonyjelölő + tartalmi elem

A tézishez kapcsolódó publikáció:
(Sass, 2009c), (Sass, 2009a), (Sass, 2008), (Sass, 2005)

A modell legszemléletesebben 1-mélységű függőségi fával ábrázolható, melynek az ige a gyökere, az élek a viszonyjelölők, a csomópontok pedig a tartalmi elemek. Az 1. ábrán látható a modellnek megfelelő általános függőségi fa, és az egyik fenti szerkezet konkrét reprezentációja.



1. ábra. A modell megjelenítése függőségi fával. Bal oldalon a modellnek megfelelő általános függőségi fa látható viszonyjelölőkkel és tartalmi elemekkel (t. e.), jobb oldalon pedig egy konkrét szerkezet, az '*igényt tart vmire*' reprezentációja. Az LSzB-hez (esetünkben ez a '*vmire*') tartozó tartalmi elem nem része a szerkezetnek.

—•—

A következő kérdés nyilván az, hogy hogyan alakítható ki egy korpusznak a fenti modell szerinti reprezentációja. Természetesen előállítható ez a forma egy függőségileg elemzett korpuszból (treebank-

ból), vagy függőségi elvű szintaktikai elemző felhasználásával. Megfelelő méretű függőségileg elemzett korpusz, illetve függőségi elemző magyar nyelvre nem állt rendelkezésre. Dolgozatomnak nem célja egy magyar függőségi elemző kialakítása (ez egy önálló dolgozat tárgya lehetne), a további kutatáshoz egy nagy méretű korpusz megfelelő minőségű reprezentációjára volt szükségem.

Reprezentatív magyar nyelvű korpuszként a 187 millió szavas Magyar Nemzeti Szövegtárat választottam, és azt vizsgáltam meg, hogy közelítő módszerrel, szabályalapú megközelítéssel, egyszerű szabályokkal elő lehet-e állítani a szükséges reprezentációt. Kiderült, hogy a tagmondatra bontás és a szükséges részleges szintaktikai elemzés (lényegében igeazonosítás és névszói csoport bővítmények azonosítása) is megfelelő minőségben megoldható így.

A korpusz feldolgozását a dolgozat 2.2. részében tárgyalom, a fejezet tanulságát a következő tézis mondja ki:

> **2. tézis.**
> **Megmutattam, hogy morfoszintaktikailag annotált korpuszból szabályalapú tagmondatra bontással és szabályalapú részleges szintaktikai elemzéssel, viszonylag egyszerű szabályrendszerrel megbízható modell szerinti reprezentációjú korpusz állítható elő.**
>
> A tézishez kapcsolódó publikáció:
> (Sass, 2006b), (Sass, 2005)

Természetesen a jövőben egy valódi függőségi elemző felhasználásával a reprezentáció minősége javítható, de mostani állapotában is elegendően jó ahhoz, hogy a további kutatásnak alapanyaga lehessen.

—•—

Az így létrehozott reprezentáció önmagában értékes erőforrás. Mint speciális korpusz különböző olyan lekérdezésekre ad lehetőséget, me-

lyek egy korpuszlekérdezőnél nem megszokottak: elvonatkoztathatunk a szórendtől, az igei szerkezeteket az adott korpuszmondatban épp megjelenő szórendjüktől függetlenül egységesen vizsgálhatjuk. Ezért készítettem el a Mazsola elnevezésű korpuszlekérdező rendszert, melynek segítségével az igék, illetve igei keretek mellett megjelenő jellegzetes bővítményeket vizsgálhatjuk. Megjeleníti a lekérdezésben megjelölt bővítményi helyen megjelenő tipikus szavakat, és a hozzájuk tartozó megfelelő korpuszpéldákat is.

A rendszer alapvetően kétféle tipikus bővítményt szolgáltat. Egyrészt a „szó szerinti" értelmű szavakat, melyek sok esetben szemantikailag egységes csoportot alkotnak; ilyenek például az '*eszik vmit*' tárgyi bővítményeként megjelenő különféle ételek ('*kenyér*', '*hús*', '*hal*', '*leves*' stb.). Másrészt viszont az idiomatikus, komplex igék, vagy szólások elemét alkotó szavakat; ilyen a szintén az '*eszik vmit*' lekérdezés eredményében szereplő '*kása*', mely nem azért kerül a jellegzetes szavak közé, mert manapság olyan tipikus étel lenne, hanem pontosan a '*nem eszik olyan forrón a kását*' szólás miatt.

A Mazsola korpuszlekérdezőt a dolgozat 3.2. részében ismertetem, jellemzőit az alábbi tézisben fogalmazom meg:

3. tézis.
Létrehoztam a Mazsola elnevezésű speciális korpuszlekérdező eszközt. Segítségével feltérképezhetjük az igék bővítményszerkezetét, megállapíthatjuk igék, illetve igei keretek lényeges bővítményeit, beleértve a komplex igéket is. Hasznos segédeszköz a korpuszalapú nyelvészeti kutatásban, lexikai adatbázisok kézi építésekor, és igei szerkezetekre való példák keresésekor.

A tézishez kapcsolódó publikáció:
(Sass és Pajzs, 2010b) (Sass, 2009b) (Sass, 2008) (Sass, 2006b)

A rendszer tetszőleges modell szerinti reprezentációjú korpuszra alkalmazható. A Magyar Nemzeti Szövegtár anyagát tartalmazó eredeti

magyar változat keresőfelülete szabadon elérhető a `http://corpus.nytud.hu/mazsola` internetes címen, ki is próbálható a `vendeg` ideiglenes felhasználói névvel és a hozzá tartozó `mazsola` ideiglenes jelszóval. Százmillió szavas korpuszméret mellett a lekérdezések feldolgozási ideje mindössze néhány másodperc.

—•—

A mai korpuszok elérték azt a méretet, mikor a kézi lekérdezők mellett szükség van olyan eszközökre is, melyek automatikusan összegzik a korpuszból kinyerhető információt. A Mazsola ebből a szempontból a kézi lekérdezőnek felel meg, képes konkrét igei keret konkrét bővítményi helyén megjelenő tipikus szavakat bemutatni.

Dolgozatom legfontosabb eredménye az az automatikus módszer, mely ennél egy nagyon fontos lépéssel tovább megy: képes arra, hogy korpusz alapján meghatározza, hogy *egyáltalán* mik egy ige jellegzetes bővítménykeretei, azaz automatikusan megállapítani, hogy „mi mindent érdemes" a Mazsolától kérdezni, és mintegy ezeket a lekérdezéseket „le is futtatja". Ezáltal az egyes igékhez tartozó jellegzetes igei szerkezeteket tudjuk számba venni.

Az algoritmus részletes bemutatása és kiértékelése a dolgozat 3.3. részében található, lényegét a következő tézis foglalja össze:

> **4. tézis.**
> **Kidolgoztam egy lexikai kinyerő eljárást, mely a mondatvázak gyakoriságainak speciális összegzésére épül. Ez az eljárás alkalmas arra, hogy a modell (1. tézis) szerinti reprezentációval bíró korpuszból a különféle bonyolultságú, jellegzetes igei szerkezeteket kinyerje.**
>
> A tézishez kapcsolódó publikáció:
> (Sass, 2010d), (Sass és Pajzs, 2010b), (Sass, 2009c)

A módszer újdonsága, hogy egyrészt alkalmazkodik az igei szerkezet elemszámához, azaz kettő illetve több elemű kifejezéseket egyaránt

eredményez; másrészt képes felfedezni, hogy az ige mellett egy adott fontos bővítmény esetén csak a viszony (LSzB) vagy a konkrét tartalmi elem is (LKB) lényeges: LSzB-ket és LKB-kat – akár vegyesen – tartalmazó igei szerkezeteket egyaránt szolgáltat. Az utóbbi csoportba tartoznak az 1. tézisnél említett '*hasznot húz vmiből*', '*igényt tart vmire*' és '*lehetővé tesz vmit*' vonzatos komplex igék.

Alkalmazás

Az algoritmus által szolgáltatott, igei szerkezeteket tartalmazó lista közvetlenül alkalmazható egy igei szerkezeteket tartalmazó szótár készítése során. Az igei szerkezeteket az igék köré rendezve automatikusan előállított nyers szócikkekhez jutunk. Ahhoz, hogy ebből egy kiadható szótár álljon elő manuális lexikográfiai munkára van szükség. A lexikográfiai munkaigény alacsony, a munka az ellenőrzésre és példaválasztásra korlátozódik, a szótár gyorsan és kis költségvetéssel előállítható. A szótár vonzatkeretszótár, kollokációs szótár és gyakorisági szótár egyszerre, valamint a szofisztikált mutatók révén lehetővé teszi az igei szerkezetek összevetését számos szempont szerint.

A szótárkészítés lépéseit, magát a szótárt, és lehetséges felhasználásait a dolgozat 4.2. részében tárgyalom, jelentőségét az alábbi tézis fogalmazza meg:

> **5. tézis.**
> **Létrehoztam egy új típusú szótárt, melynek alapelemei nem szavak, hanem szószerkezetek: az igei szerkezetek. A puszta szövegtől a nyers szócikkekig tisztán automatikus nyelvfeldolgozó eszközökkel jutottam el, melyek közül kiemelendő a jellegzetes igei szerkezeteket kinyerő algoritmus (4. tézis), mely a szótári anyaggyűjtést automatizálja. Megmutattam, hogy ez a lexikai kinyerő eljárás jól alkalmazható a szótárkészítésben: az elkészült szótár valóban a nyelvre jellemző vonzatokat és igei ki-**

fejezéseket tartalmazza. Olyan tanulói szótár jött így létre, mely a legfontosabb igei jelentéseket megvilágítja, elősegíti az „idiomatikus", a nemcsak nyelvtanilag helyes, hanem magyarul megszokott kifejezésmódot.

A tézishez kapcsolódó publikáció:
(Sass et al., 2010a) (Sass és Pajzs, 2010b) (Pajzs és Sass, 2010) (Sass és Pajzs, 2010c)

Hogyan használhatjuk a szótárt a nyelvtanulás támogatására, ha külföldiként magyarul akarunk megnyilatkozni? Segítségével feltérképezhetjük az ige–névszó kollokációkat: meghatározhatjuk az igékhez társítható névszókat, és (a kötött szavak szerinti mutató segítségével) a névszókhoz társítható igéket is. Ha angolként a magyarul akarunk megszólalni, és a '*meet the requirements*' megfelelőjét keressük, akkor a '*követelmény*' szónál meg fogjuk találni, hogy az ehhez illeszkedő ige a '*megfelel*', és nem a '*találkozik*' vagy valami hasonló.

A kész szótár (Sass et al., 2010a) hozzáférhető, megjelent a Tinta Könyvkiadó gondozásában.

— • —

Külön jelentőséget ad egy automatikus nyelvfeldolgozó eljárásnak, ha *nyelvfüggetlen*. A mi megközelítésünk nyelvfüggetlensége a reprezentáció előállíthatóságának nyelvfüggetlenségén múlik. A reprezentációra épülő eszközök, eljárások (a korábbi tézisekben ismertetett korpuszlekérdező, az igei szerkezeteket kinyerő eljárás, a szótárkészítés automatikus része) a reprezentáció automatikus folyományai. Mivel a reprezentáció lényegében csak arra támaszkodik, hogy van a nyelvekben prédikátum–argumentum struktúra, az várható, hogy a reprezentáció számos nyelvre előállítható. Ezt a sejtést a magyartól különböző szerkezetű dán és szerb, nyelvvel végzett kísérletek révén támasztottam alá.

A módszer nyelvfüggetlenségét a dolgozat 5.1. részében tárgyalom, a fejezet eredményét a következő tézis tartalmazza:

6. tézis.
Megmutattam, hogy az 1. tézis szerinti egységes reprezentáció nyelvfüggetlen, számos nyelvre kialakítható. Ez lényegében azon múlik, hogy a nyelvek megnyilatkozásai felbonthatók igéből és az ige bővítményeiből álló egységekre (tagmondatokra), valamint megadható az egyes bővítmények és az ige közötti függőségi viszony. A korpuszlekérdező (3. tézis) elkészítése alig igényel plusz munkát, egyszerűen beilleszthetjük az új korpuszt az eddigiek közé. A 4. tézisben leírt algoritmus tetszőleges egységes reprezentációjú korpuszon ugyanúgy futtatható, ezáltal az igei szerkezetek gyűjtése nyelvfüggetlen módon megvalósítható. Végeredményben az erre épülő, az 5. tézisben bemutatott szótár is előállítható, korlátozott mennyiségű manuális lexikográfiai munka befektetésével.

A tézishez kapcsolódó publikáció:
(Sass, 2009d)

A jövőben a módszerrel az előző tézisben bemutatott magyar nyelvű szótárhoz hasonló nyelvtanulást segítő szótárak készülhetnek egyéb – hazánkban keresett – idegen nyelvekre is.

— • —

A modellt (1. tézis) többféle módon is kiterjeszthetjük, pontosabban többféle bonyolultabb struktúrát visszavezethetünk az 1. ábrán (10. oldal) is látható 1-mélységű függőségi fa szerkezetre. A legizgalmasabb kérdés az, hogy elő tudunk-e állítani olyan reprezentációt, mely *párhuzamos* korpusz alapján készül, párhuzamos tagmondatokat, és ezáltal párhuzamos szerkezeteket (szerkezeteket és megfelelő fordításaikat) tartalmaz; de emellett megfelel az eredeti modellnek, következésképpen a kinyerő algoritmusunk futtatható rajta. Ezen a módon egy olyan eljárást nyernénk, mely a változatlan kinyerő eljárás alkalmazásával párhuzamos szerkezeteket eredményezne: az igei szerkezetekhez megkapnánk másik nyelvű fordításaikat is.

A modell kiterjesztéseit a dolgozatban a 5.2. és a 5.3. fejezetben tárgyalom, a módszernek a párhuzamos igei szerkezetek kinyerésére való alkalmazásáról a dolgozat 5.4. részében számolok be, az alábbi tézis összegzi ezt az ígéretes irányt:

7. tézis.
Megmutattam, hogy egy párhuzamos tagmondat (azaz két különböző nyelvű, egymásnak megfelelő tagmondat) közös reprezentációja kialakítható az eredeti modell szerinti formában: a központi elem a két (különnyelvű) igéből alkotott pár lesz, a bővítményeket pedig egy összesített halmazként rendelem e központi elem mellé. Ezzel előáll a párhuzamos korpuszok olyan reprezentációja, mely formailag megegyezik az egynyelvű korpuszok eredeti modell szerinti reprezentációjával. Az igei szerkezeteket kinyerő eljárást ezen a reprezentáción közvetlenül futtatva kétnyelvű, párhuzamos igei szerkezeteket, azaz szerkezeteket és a másik nyelvű megfelelőiket tudtam kinyerni. A módszer képes arra, hogy párba állítson olyan szerkezeteket is, melyek aszimmetrikusak, azaz a két nyelven teljesen eltérő felépítésűek.

A tézishez kapcsolódó publikáció:
(Sass, 2010d)

A párhuzamos szerkezetekre vonatkozó vizsgálatokat egy holland–francia korpuszon végeztem. Az eredményben megkaptam például a holland '*nemen deel aan*' és a francia '*participer à*' alkotta aszimmetrikus párt (jelentésük: '*részt vesz vmiben*'). Látjuk, hogy amit a holland összetett igével fejez ki, azt a francia itt egy szóval, egy egyszerű igével.

A módszer segítségével a jövőben olyan nyelvtanulást segítő kétnyelvű szótárak állíthatók elő, melyek a használatból nyert egymásnak megfeleltetett igei szerkezetek révén elősegítik a jobb nyelvhasználatot, az anyanyelvi beszélők számára is természetes nyelvi produkciót.

A kétnyelvű szótárak ilyen előállításának kidolgozása a jövő feladata, dolgozatom egy fontos lépés ebben az irányban.

Köszönetnyilvánítás

Köszönöm feleségemnek, *Dórinak*, a gyerekeknek, *Micinek*, *Csöpinek*, *Lencsinek* és *Jáninak*, és tágabb családomnak az állandó támogatást és biztatást.

Köszönöm témavezetőmnek, *Prószéky Gábornak*; főnökömnek, *Váradi Tamásnak*; legközelebbi munkatársamnak, *Oravecz Csabának*; lexikográfus kolléganőmnek, *Pajzs Júliának*; és a doktori iskola vezetőinek, *Roska Tamásnak* és *Szolgay Péternek* a szakmai támogatást és segítséget.

Köszönet barátaimnak, munkatársaimnak és mindenkinek, akik munkájukkal, ötleteikkel, tanácsaikkal, találó meglátásaikkal vagy bármilyen más módon támogattak, biztattak és segítségemre voltak a doktori évek és a dolgozatírás ideje alatt.

A szerző publikációi

Könyv

Sass Bálint – Váradi Tamás – Pajzs Júlia – Kiss Margit 2010a. *Magyar igei szerkezetek – A leggyakoribb vonzatok és szókapcsolatok szótára*. Tinta Könyvkiadó, Budapest.

Folyóiratcikk

Sass Bálint – Pajzs Júlia 2010b. Igei szerkezetek gyakorisági szótára – félautomatikus szótárkészítés nyelvtechnológiai eszközök segítségével. *Alkalmazott Nyelvtudomány*, 2010(1–2):5–32.

Könyvfejezet

Sass Bálint 2006a. Extracting idiomatic Hungarian verb frames. In Salakoski, Tapio – Ginter, Filip – Pyysalo, Sampo – Pahikkala, Tapio (eds.): *Advances in Natural Language Processing*, 303–309. Springer, Berlin Heidelberg New York. Lecture Notes in Computer Science, Vol. 4139.

Sass Bálint 2008. The Verb Argument Browser. In Sojka, Petr – Horák, Aleš – Kopecek, Ivan – Pala, Karel (eds.): *Text, Speech and Dialogue*, 187–192. Springer, Berlin Heidelberg New York. Lecture Notes in Computer Science, Vol. 5246.

Sass Bálint 2009a. Korpusznyelvészeti eszköz a magyar igék bővítményszerkezetének vizsgálatára. In Sinkovics Balázs (szerk.): *LingDok 8. – Nyelvész-doktoranduszok dolgozatai*, 143–155. JATEPress, Szeged.

Sass Bálint 2009b. „Mazsola" – eszköz a magyar igék bővítményszerkezetének vizsgálatára. In Váradi Tamás (szerk.): *Válogatás az I. Alkalmazott Nyelvészeti Doktorandusz Konferencia előadásaiból*, 117–129, MTA Nyelvtudományi Intézet, Budapest.

Sass Bálint – Pajzs Júlia 2010c. FDVC – creating a corpus-driven frequency dictionary of verb phrase constructions. In Granger, Sylviane – Paquot, Magali (eds.): *eLexicography in the 21st century: New challenges, new applications. Proceedings of eLex 2009, Cahiers du CENTAL 7. Presses universitaires de Louvain*, 263–272, Louvain-la-Neuve, Belgium.

Külföldi konferenciakötet

Pajzs Júlia – Sass Bálint 2010. Towards semi-automatic dictionary making. In *Proceedings of the XIV. EURALEX International Congress*, 453–462.

Sass Bálint 2007. First attempt to automatically generate Hungarian semantic verb classes. In *Proceedings of the 4th Corpus Linguistics conference*, Birmingham.

Sass Bálint 2009c. A unified method for extracting simple and multiword verbs with valence information and application for Hungarian. In *Proceedings of RANLP 2009*, 399–403, Borovets, Bulgária.

Sass Bálint 2009d. Verb Argument Browser for Danish. In *Proceedings of the 17th Nordic Conference of Computational Linguistics, NoDaLiDa 2009*, 263–266, Odense, Dánia.

Hazai konferenciakötet

Sass Bálint 2005. Vonzatkeretek a Magyar Nemzeti Szövegtárban. In Alexin Zoltán – Csendes Dóra (szerk.): *III. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2005)*, 257–264, Szeged.

Sass Bálint 2006b. Igei vonzatkeretek az MNSZ tagmondataiban. In Alexin Zoltán – Csendes Dóra (szerk.): *IV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2006)*, 15–21, Szeged.

Sass Bálint 2010d. Párhuzamos igei szerkezetek közvetlen kinyerése párhuzamos korpuszból. In Tanács Attila – Vincze Veronika (szerk.): *VII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2010)*, 102–110, SZTE, Szeged.