IGEI SZERKEZETEK GYAKORISÁGI SZÓTÁRA

EGY AUTOMATIKUS LEXIKAI KINYERŐ ELJÁRÁS ÉS ALKALMAZÁSA

című doktori (Ph.D.) disszertáció nyilvános védése

Sass Bálint
sass.balint@itk.ppke.hu

PPKE ITK
Budapest, 2011. október 14.

IGEI SZERKEZETEK

‘*részt vesz vmiben*’, ‘*górcső alá vesz vmit*’

- ‘*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).

egyszerre igei vonzatkeretek és kollokációk
$\leftrightarrow$ két elkülönült terület

IGEI SZERKEZETEK

‘*részt vesz vmiben*’, ‘*górcső alá vesz vmit*’

igei szerkezet = ige + névszói csoport bővítmények

A példákban *két* bővítmény szerepelt:

- konkrét kötött szó
  – lexikálisan kötött bővítmény (LKB)
- bővítményi hely (esetrag)
  – lexikálisan szabad bővítmény (LSzB)

Ezt a két típusú bővítményt ugyanazokkal a nyelvi eszközökkel fejezzük ki: esetragokkal, elöljárókkal, szórendi megkötéssel.

‘*pillantást vet vkire*’ $\leftrightarrow$ ‘*szemére vet vmit*’

IGEI SZERKEZETEK

‘*részt vesz vmiben*’, ‘*górcső alá vesz vmit*’

igei szerkezet = ige + névszói csoport bővítmények

A példákban *két* bővítmény szerepelt:

- konkrét kötött szó
  – lexikálisan kötött bővítmény (LKB)
- bővítményi hely (esetrag)
  – lexikálisan szabad bővítmény (LSzB)

Ezt a két típusú bővítményt ugyanazokkal a nyelvi eszközökkel fejezzük ki: esetragokkal, elöljárókkal, szórendi megkötéssel.

‘*pillantást vet vkire*’ ↔ ‘*szemére vet vmit*’

IGEI SZERKEZETEK

- az ilyen összetett igei szerkezetek gyakoriak, sokszor idiomatikus jelentéssel
- lexikai adatbázisokban szerepelniük kell
- szükség van egy olyan automatikus eljárásra, mely megállapítja, hogy mikor melyik bővítmény LKB/LSzB, ezáltal képes arra, hogy korpuszból kinyerje a jellegzetes igei szerkezeteket

A dolgozat fő eredménye ez az *algoritmus* illetve az ennek közvetlen felhasználásával készülő magyar, egynyelvű igeiszerkezet-*szótár*.

Áttekintés

IGEI SZERKEZETEK MODELLJE

‘*hisz vmiben*’ 1 LSzB
‘*igényt tart vmire*’ LSzB + LKB
‘*pontot tesz a végére*’ 2 LKB
. . .

Cél: modell (magyar nyelvre),
mely az igei szerkezetek összes típusát ábrázolni képes.

Megoldás: függőségi elemzés alapú gráf

IGEI SZERKEZETEK MODELLJE

1. TÉZIS

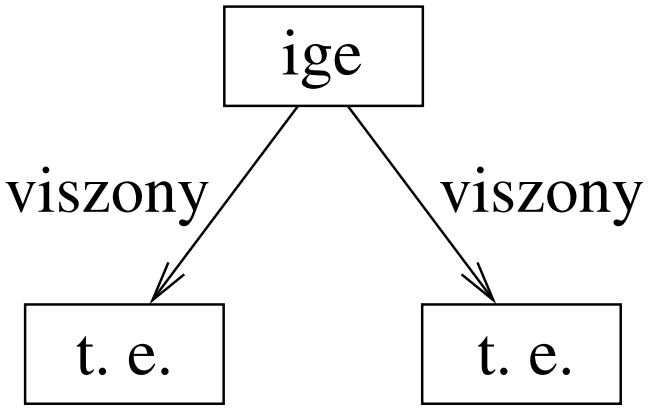
Kidolgoztam magyar nyelvre egy olyan modellt, mely képes a tagmondatok, illetve a bennük rejlő formailag nagy mértékben különböző igei szerkezetek egységes reprezentálására. Alapegység a tagmondat, mely egy központi ige és a hozzá tartozó bővítmények összessége. A bővítményeket legfontosabb tartalmi elemükkel (névszói csoport bővítmény esetén a bővítményt 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**

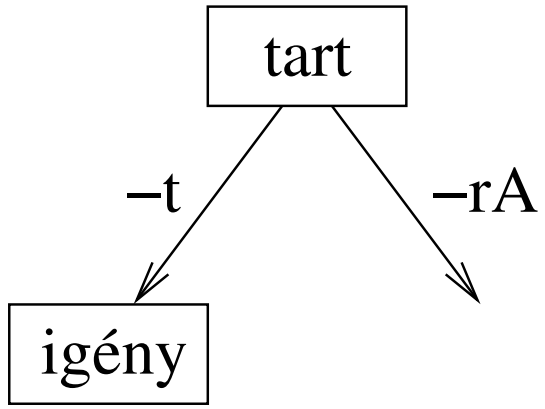
(Sass, 2009c), (Sass, 2009a), (Sass, 2008), (Sass, 2005)

A MODELL MEGJELENÍTÉSE

A modell megjelenítése függőségi fával.



a modellnek megfelelő általános függőségi fa



az '*igényt tart vmire*' reprezentációja

Alternatív forma:

`ige=`tart `-t=`igény `-rA`

A REPREZENTÁCIÓ MEGVALÓSÍTÁSA

A további kutatáshoz egy nagy méretű korpusz modell szerinti reprezentációjára volt szükségem.

Lehetőségek:

- függőségileg elemzett korpuszból levezetni
- függőségi elvű szintaktikai elemző felhasználásával

Korpusz: Magyar Nemzeti Szövegtár (187 millió szó)

Módszer: szabályalapú megközelítés egyszerű szabályokkal

Eredmény: a tagmondatra bontás és a szükséges részleges szintaktikai elemzés (igeazonosítás és névszói csoport bővítmények azonosítása) is megfelelő minőségben megoldható így.

PÉLDÁK A SZABÁLYOKRA

- tagmondatra bontás
  módszer: szabályok (reguláris kifejezések)
  pl.:

```
[,|-] @ [kötőszó|határozószó]? [vonatkozó névmás]
```

- részleges szintaktikai elemzés =
  igeazonosítás + névszói csoportok azonosítása
  módszer: többszintű reguláris nyelvtan
  pl.:

```
NP <- msd.postag='Det'
   [msd.postag='A' msd.postag='Num']*
   msd.postag='N'
```

A főnévi csoport legegyszerűbb típusát felismerő szabály:
névelő (`Det`) + tetszőleges számú melléknév (`A`) vagy számnév (`Num`) + egy főnév (`N`)

A REPREZENTÁCIÓ MEGVALÓSÍTÁSA

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.

(Sass, 2006b), (Sass, 2005)

A *Mazsola* KORPUSZLEKÉRDEZŐ

A létrejött speciális korpusz olyan lekérdezésekre ad lehetőséget, melyek egy korpuszlekérdezőnél nem megszokottak:

az igei szerkezeteket
szórendjüktől függetlenül
egységesen vizsgálhatjuk.

→ „Mazsola” korpuszlekérdező:
igék, illetve igei keretek mellett megjelenő jellegzetes bővítmények vizsgálata

Megjeleníti a lekérdezésben megjelölt bővítményi helyen megjelenő tipikus szavakat, és a hozzájuk tartozó korpuszpéldákat.

A *Mazsola* VÁLASZKÉPERNYŐJE

Korpusz: MNSZ: Magyar Nemzet

Igető: kér

Nem: Eset/névutó: -t Nem: Vonzattő:

Nem: Eset/névutó: -tól Nem: Vonzattő:

Nem: Eset/névutó: Nem: Vonzattő:

Nem: Szó:

Teljes mondatlefedés:

Mehet

Eloszlás:

1010 találat. bocsánat [51] segítség [53] elnézés [32] az [136] engedély [32] tájékoztatás [21] támogatás [25] pénz [20] felmentés [12] válasz [16] tanács [13] forint [16] magyarázat [9] igazolás [8] állásfoglalás [8] kiadatás [7] normakontroll [6] információ [9] felfüggesztés [7] kihallgatás [6] megállapítás [7] türelem [6] garancia [6] felülvizsgálat [6] ez [12] vizsgálat [7] védelem [6] ami [7] ki [6]

adat
  Az információk ellenőrzésére persze a kórháztól kér adatokat,
  Feladatai teljesítéséhez adatokat kérhet a bíróságtól, az ügyészségtől, a nemzetbiztonsági szolgálatoktól, a társadalombiztositási igazgatási szervektől.
  ha maximum öt főről kér adatot a hivataltól.
  hogy a rendőrség az adóhatóságtól, telefonszolgáltatóktól, bankoktól ügyészi jóváhagyás nélkül kérjen adatokat,
  hogy kérje tőlük a művelődési tárca a bérfejlesztéshez szükséges adatokat.
adatgyűjtés
  hanem Orbán Viktor és Deutsch Tamás ellen is adatgyűjtést kértek az ügyben érintett magánnyomozótól.
adóigazolás
  amely szerint a gépkocsi átíratásakor adóigazolást kértek a polgároktól.
adókedvezmény
  Korábban Budapest és a vidéki nagyvárosok különféle adókedvezményeket is kértek az előző kabinettől,
adószám
  Ha a bérbeadó magánszemély, akkor adószámot kell kérnie az APEH-tól,
  ha a magánszemély adószámot kér az APEH-tól
aggregátor
  A hadseregtől kértek aggregátort,
ajánlat
  A helyreállitással megbizott Szabolcs-Szatmár-Bereg Megyei Közútkezelő Kht. tiz vállalkozótól másfél milliárd forintos felújitási munkára kért a közelmúltban kivite
  Az eredménytelen pályázatok után a bizottság új ajánlatot kért a pályázóktól.
  és cserébe a pályázóktól - meghatározott müszaki paraméterekkel rendelkező - új kocsikra kért ajánlatot.

A *Mazsola* MŰKÖDÉSE

Kétféle tipikus bővítményt szolgáltat:

1. „szó szerinti” értelmű szavak, melyek sok esetben szemantikailag egységes csoportot alkotnak

   pl.: ‘*eszik vmit*’ tárgyi bővítményeként megjelenő különféle ételek (‘*kenyér*’, ‘*hús*’, ‘*hal*’, ‘*leves*’ stb.)

2. idiomatikus, komplex igék, szólások elemét alkotó szavak

   pl.: ‘*kása*’ $\rightarrow$ ‘*nem eszik olyan forrón a kását*’

A *Mazsola* KORPUSZLEKÉRDEZŐ

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.

(Sass és Pajzs, 2010b) (Sass, 2009b) (Sass, 2008) (Sass, 2006b)

A *Mazsola* KORPUSZLEKÉRDEZŐ

- A rendszer tetszőleges modell szerinti reprezentációjú korpuszra alkalmazható.
- A Magyar Nemzeti Szövegtár anyagát tartalmazó magyar változat keresőfelülete szabadon elérhető a `http://corpus.nytud.hu/mazsola` internetes címen. Kipróbálható. Felhasználói név: `vendeg`, jelszó: `mazsola`.
- Százmillió szavas korpuszméret mellett a lekérdezések feldolgozási ideje mindössze néhány másodperc.

JELLEGZETES IGEI SZERKEZETEK KINYERÉSE

A mai korpuszméreteknél szükség van olyan eszközökre, melyek automatikusan összegzik a korpuszból kinyerhető információt.

Dolgozatom legfontosabb eredménye az az automatikus módszer, mely képes korpuszból kinyerni a jellegzetes igei szerkezeteket.

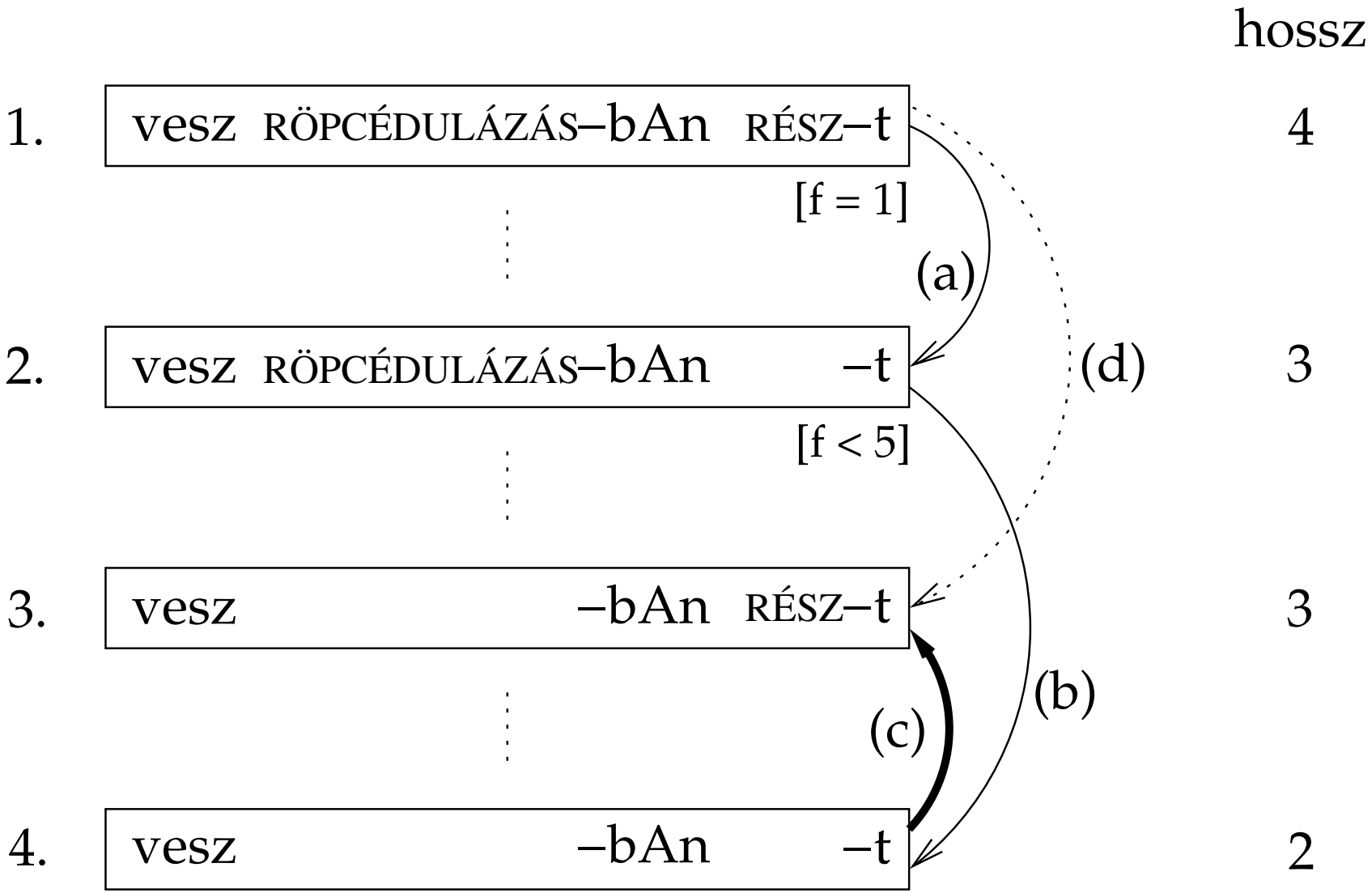
A kézi lekérdezőhöz képest egy nagyon fontos lépéssel tovább megy: meghatározza, hogy *egyáltalán mik* egy ige jellegzetes bővítménykeretei, és ezeket veszi számba.

Az algoritmus vázlata

1. Vesszük a korpusz *tagmondatait* a reprezentáció szerint.
   Maximum *két* bővítmény esetén: *váltakozó törlés*
   '*Társasház jön létre.*' (`ige`=jön $-\emptyset$=társasház $-$`rA`=lét) $\rightarrow$
   '*társasház jön létre*', '$\emptyset$ *jön létre*', '*társasház jön -rA*', '$\emptyset$ *jön -rA*'.
2. Hossz szerint csökkenő sorba rendezés.
   Hossz (h) = |LSzB| + |LKB|·2
3. A leghosszabbtól kezdve sorra elhagyjuk
   a ritka ($f < 5$) szerkezeteket.
   Az elhagyott szerkezetek gyakoriságát az *első* olyan rövidebb keret gyakoriságához adjuk hozzá, mely illeszkedik az eredeti keretre.
   pl.: '*társasház jön létre*' ($h = 4$) $\rightarrow$ '*vmi jön létre*' ($h = 3$)
4. *Visszaellenőrzés* (köv. dia)
5. A megmaradó szerkezetek gyakorisági érték szerint rendezett listája adja az összegyűjtött igei szerkezeteket.

Az algoritmus vázlata

1. Vesszük a korpusz *tagmondatait* a reprezentáció szerint.
   Maximum *két* bővítmény esetén: *váltakozó törlés*
   '*Társasház jön létre.*' (`ige`=jön −∅=társasház −`rA`=lét) →
   '*társasház jön létre*', '∅ *jön létre*', '*társasház jön -rA*', '∅ *jön -rA*'.
2. Hossz szerint csökkenő sorba rendezés.
   Hossz (h) = |LSzB| + |LKB|·2
3. A leghosszabbtól kezdve sorra elhagyjuk
   a ritka ($f < 5$) szerkezeteket.
   Az elhagyott szerkezetek gyakoriságát az *első* olyan rövidebb keret gyakoriságához adjuk hozzá, mely illeszkedik az eredeti keretre.
   pl.: '*társasház jön létre*' ($h = 4$) → '*vmi jön létre*' ($h = 3$)
4. *Visszaellenőrzés* (köv. dia)
5. A megmaradó szerkezetek gyakorisági érték szerint rendezett listája adja az összegyűjtött igei szerkezeteket.


hossz
1.
vesz RÖPCÉDULÁZÁS–bAn RÉSZ–t
4
[f = 1]
(a)
2.
vesz RÖPCÉDULÁZÁS–bAn –t
3
(d)
[f < 5]
3.
vesz –bAn RÉSZ–t
3
(b)
(c)
4.
vesz –bAn –t
2

JELLEGZETES IGEI SZERKEZETEK KINYERÉSE

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.

(Sass, 2010d), (Sass és Pajzs, 2010b), (Sass, 2009c)

Az algoritmus jelentősége

A módszer…

- képes korpusz alapján az igei szerkezeteket azonosítani (alkalmazkodik az igei szerkezet elemszámához);
- képes felismerni, elkülöníteni, hogy mikor melyik esetrag melyik szerepnek felel meg: azaz melyik bővítmény LKB és melyik LSzB;
- egyszerre állapítja meg a kollokátumokat és a vonzatokat, így *teljes* szerkezeteket eredményez.

Következmény: Az algoritmus LKB-ket és LSzB-ket tetszőleges kombinációban tartalmazó szerkezeteket szolgáltat:
így kollokációkat (csak LKB) és vonzatkereteket (csak LSzB) is.

IGEI SZERKEZETEK SZÓTÁRA

- Jellegzetes igei szerkezetek listája
  → igék köré rendezve → nyers szócikkek.
- Manuális lexikográfiai munka szükséges.

  Alacsony munkaigény : ellenőrzés és példaválasztás

  A szótár gyorsan és kis költségvetéssel előállítható.

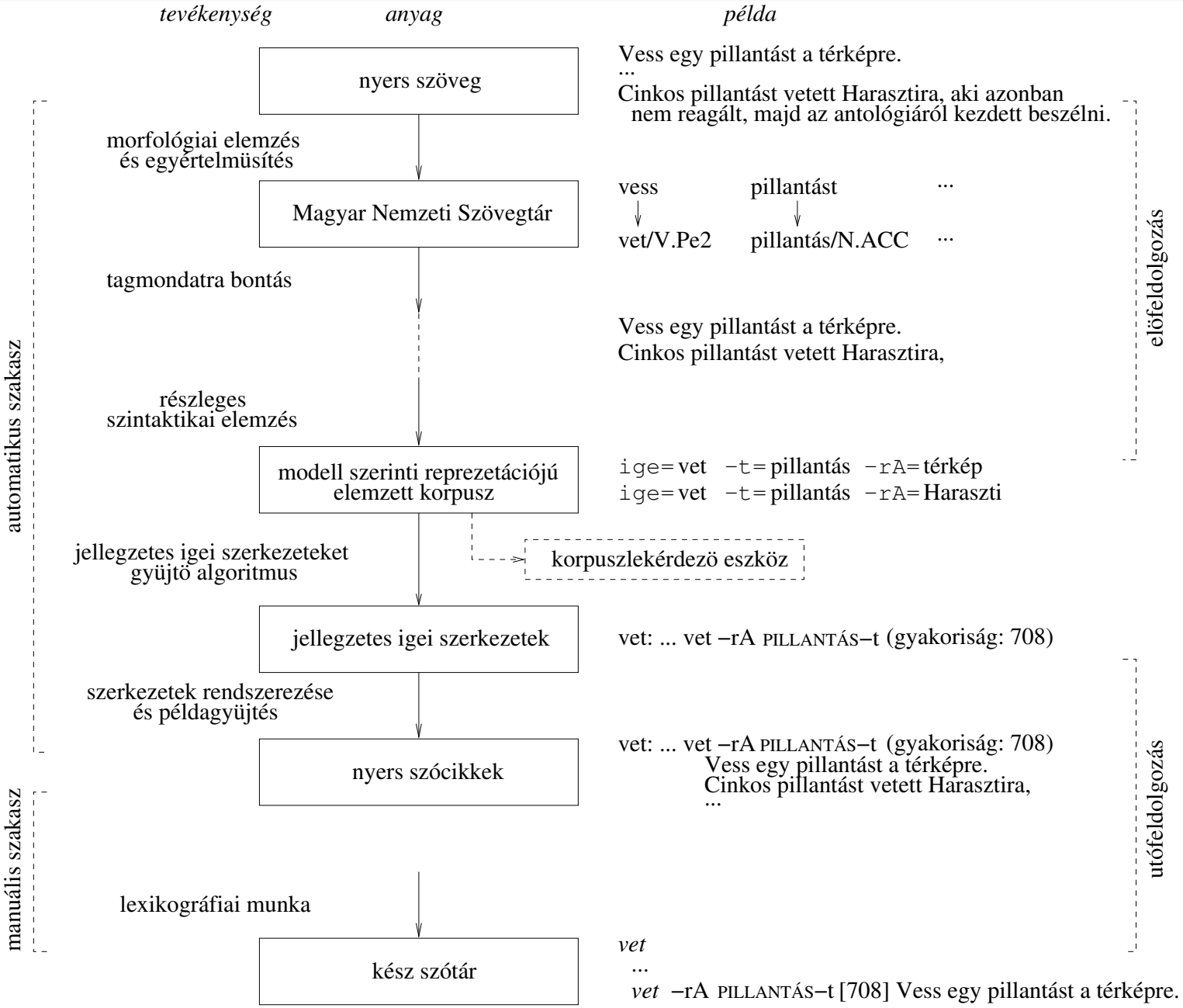
- *Egyszerre :*
  vonzatkeretszótár + kollokációs szótár + gyakoriségi szótár
- Lehetővé teszi az igei szerkezetek
  összevetését a különféle mutatók révén.

IGEI SZERKEZETEK SZÓTÁRA

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. A jellegzetes igei szerkezeteket kinyerő algoritmus (4. tézis) 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 kifejezéseket tartalmazza. Olyan tanulói szótár jött így létre, mely a legfontosabb igei jelentéseket megvilágítja, elősegíti az „idiomatikus”, a nemcsak nyelvtanilag helyes, hanem magyarul megszokott kifejezésmódot.

(Sass et al., 2010a) (Sass és Pajzs, 2010b) (Pajzs és Sass, 2010) (Sass és Pajzs, 2010c)


tevékenység
anyag
példa
nyers szöveg
Vess egy pillantást a térképre.
...
Cinkos pillantást vetett Harasztira, aki azonban nem reagált, majd az antológiáról kezdett beszélni.
morfológiai elemzés és egyértelmüsítés
Magyar Nemzeti Szövegtár
vess pillantást ...
vet/V.Pe2 pillantás/N.ACC ...
tagmondatra bontás
Vess egy pillantást a térképre.
Cinkos pillantást vetett Harasztira,
részleges szintaktikai elemzés
modell szerinti reprezetációjú elemzett korpusz
ige=vet −t=pillantás −rA=térkép
ige=vet −t=pillantás −rA=Haraszti
korpuszlekérdezö eszköz
jellegzetes igei szerkezeteket gyüjtö algoritmus
jellegzetes igei szerkezetek
vet: ... vet −rA PILLANTÁS−t (gyakoriság: 708)
szerkezetek rendszerezése és példagyüjtés
nyers szócikkek
vet: ... vet −rA PILLANTÁS−t (gyakoriság: 708)
Vess egy pillantást a térképre.
Cinkos pillantást vetett Harasztira,
...
lexikográfiai munka
kész szótár
vet
...
vet −rA PILLANTÁS−t [708] Vess egy pillantást a térképre.
automatikus szakasz
manuális szakasz
elöfeldolgozás
utófeldolgozás

A SZÓTÁR FELHASZNÁLÁSA

Nyelvtanulás támogatása:
jellegzetes, gyakori ige–névszó + névszó–ige kollokációk

Mi a '*meet the requirements*' magyar megfelelője?

Ismert: '*követelmény*'

Mi a hozzá társítandó ige?

Kötött szavak szerinti mutató $\rightarrow$ '*megfelel követelménynek*'


SASS BÁLINT – VÁRADI TAMÁS
PAJZS JÚLIA – KISS MARGIT
Magyar
igei szerkezetek
A leggyakoribb vonzatok
és szókapcsolatok szótára
TINTA KÖNYVKIADÓ

NYELVFÜGGETLENSÉG

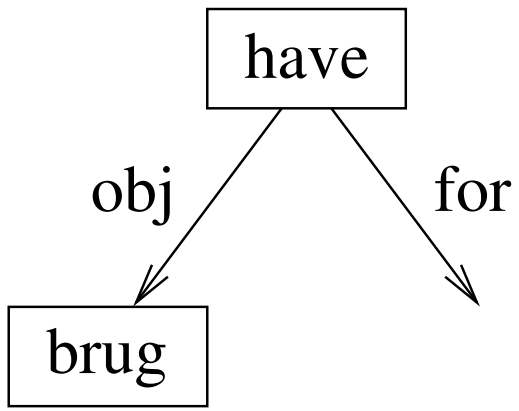
Állítás: a modell nyelvfüggetlen.
A magyaron kívül számos nyelvre előállítható a modell szerinti reprezentáció, és kinyerhetők a fenti típusú igei szerkezetek.

Reprezentáció előállítása: viszonyjelölők meghatározása

→ dán és szerb: *elöljárók*, alany és tárgy esetén *sorrendiség*

dán példa: '*have brug for*'
= szüksége van vmire

szerb példa: '*ići u prilog*'
= támogat („haszonba megy")



NYELVFÜGGETLENSÉG

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.

Előállítható a korpuszlekérdező (3. tézis), a 4. tézisben leírt algoritmus futtatható, segítségével kinyerhetők a jellegzetes igei szerkezetek. A szükséges manuális munka befektetésével az 5. tézisben bemutatott szótár is elkészíthető.

(Sass, 2009d)

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

Ötlet

Párhuzamos korpusz és párhuzamos igei szerkezetek
(igei szerkezetek és fordításaik)
reprezentálhatók a modell szerint?

Speciális reprezentáció: metakorpusz.

. . . a kétnyelvű korpuszt egynyelvűnek „álcázzuk”

Ebből a reprezentációból a *változatlan* kinyerő eljárás
közvetlenül párhuzamos szerkezeteket gyűjt.

A METAKORPUSZ KIALAKÍTÁSA

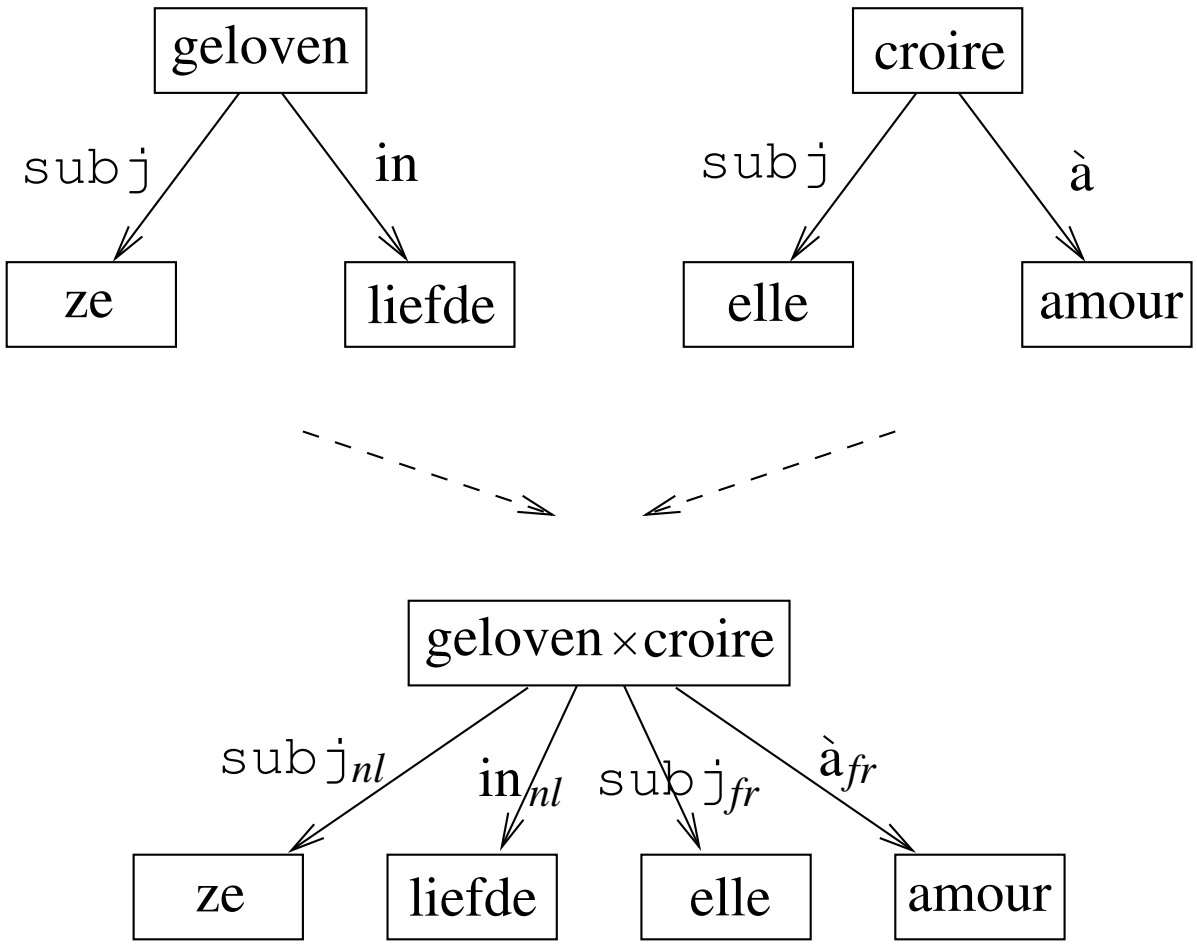
korpusz: Dutch Parallel Corpus, holland–francia (3,5 mió token)
elemzés: nyelvenként külön, tagmondatra bontás és részleges szintaktikai elemzés egyszerű szabályokkal

1. Tagmondat-szintű illesztés: a tagmondatokat fordítási egységenként sorra egymáshoz rendeltük.
2. Az egymáshoz rendelt tagmondatok holland ill. francia igéjéből: igepár. (pl.: '*gaan*×*aller*' 'megy')
3. A tagmondatpárban található bővítményeket (mindkét nyelvűeket) egy halmazként soroltuk fel az igepár mellett.

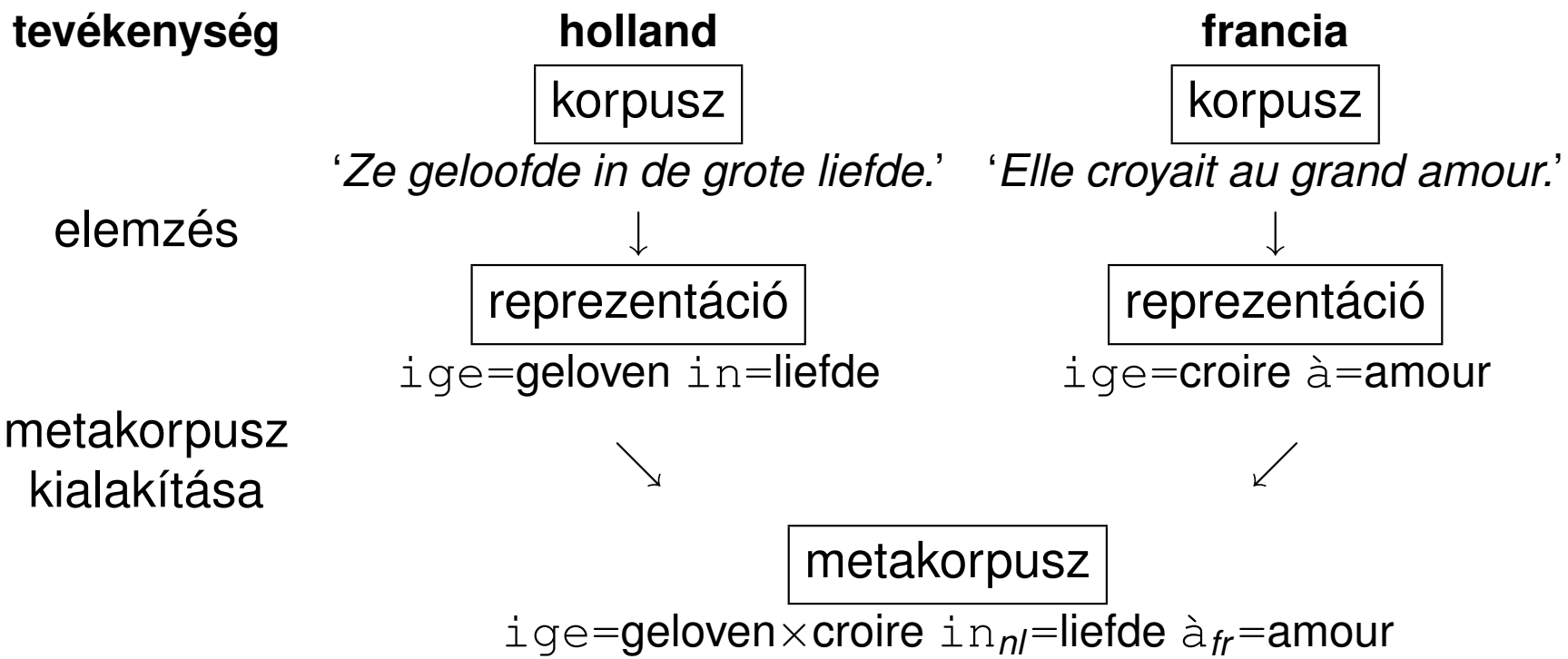
holland tagmondat:	'*Ze geloofde in de grote liefde.*'
francia tagmondat:	'*Elle croyait au grand amour.*'
magyar fordítás:	'Hitt a nagy szerelemben.'
reprezentáció:	`ige`=geloven×croire `in`$_{nl}$=liefde `à`$_{fr}$=amour

A METAKORPUSZ KIALAKÍTÁSA

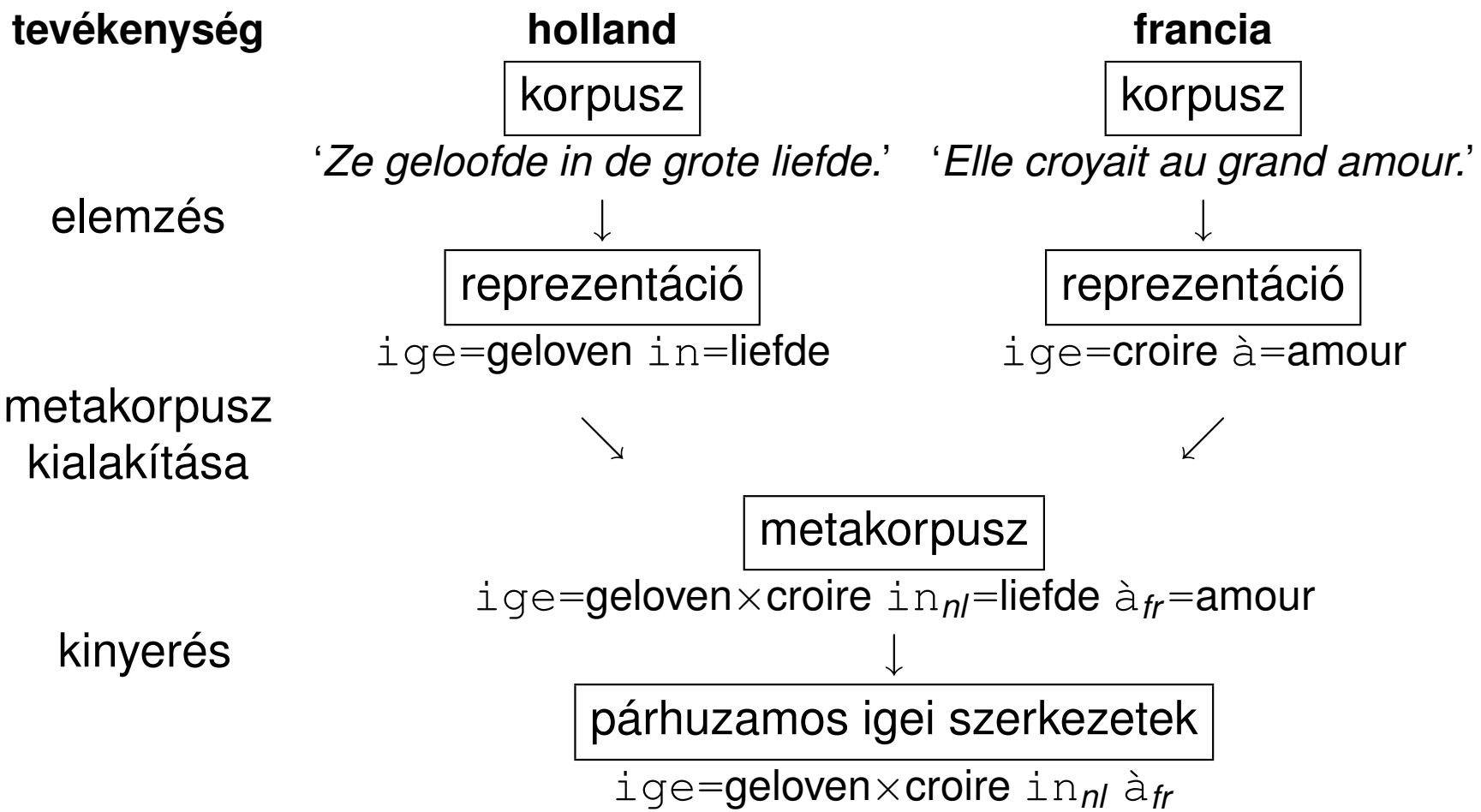
Visszavezetés az eredeti modellre: „összefésülés”



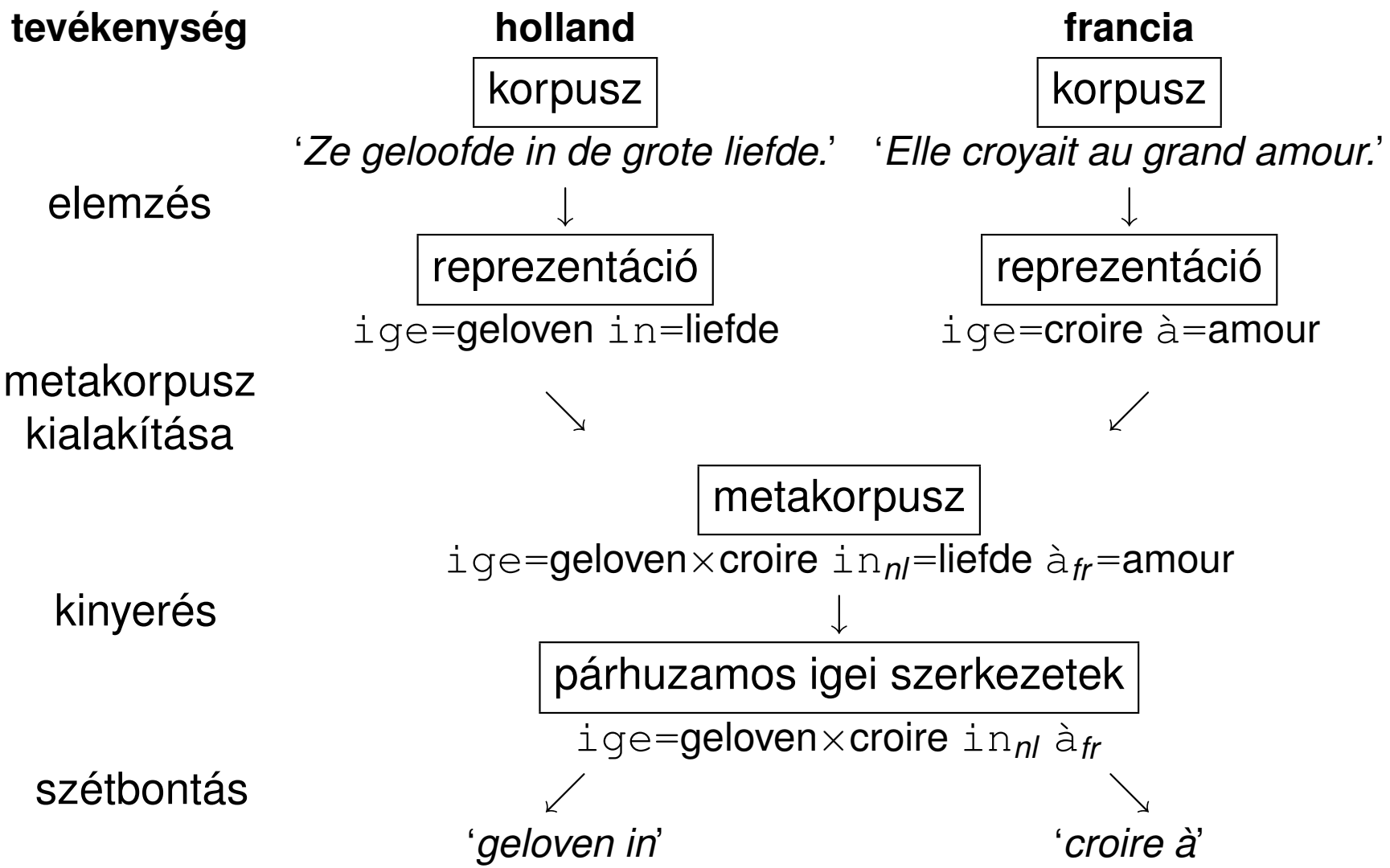
A MÓDSZER ALKALMAZÁSA KÉT NYELVRE



A MÓDSZER ALKALMAZÁSA KÉT NYELVRE



A MÓDSZER ALKALMAZÁSA KÉT NYELVRE



PÁRHUZAMOS IGEI SZERKEZETEK KINYERÉSE

7. TÉZIS

Megmutattam, hogy egy párhuzamos tagmondat közös reprezentációja kialakítható olyan módon, mely formailag megegyezik egy egynyelvű tagmondat eredeti modell szerinti reprezentációjával. Az igei szerkezeteket kinyerő eljárást az így reprezentált párhuzamos korpuszon 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.

(Sass, 2010d)

PÉLDÁK

- *aszimmetria:*

‘*houden van*’ = ‘*aimer OBJ*’ ’szeret vmit’

‘*nemen deel aan*’ = ‘*participer à*’ ’részt vesz vmiben’

- *idiomatikus megfelelők:*

‘*maken deel van*’ = ‘*faire partie de*’ ’részét képezi vminek’

‘*doen beroep op*’ = ‘*faire appel à*’ ’támaszkodik vmire’

Példák

- *aszimmetria:*

‘*houden van*’ = ‘*aimer OBJ*’ ’szeret vmit’

‘*nemen deel aan*’ = ‘*participer à*’ ’részt vesz vmiben’

- *idiomatikus megfelelők:*

‘*maken deel van*’ = ‘*faire partie de*’ ’részét képezi vminek’

‘*doen beroep op*’ = ‘*faire appel à*’ ’támaszkodik vmire’

TOVÁBBLÉPÉS

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 lépés ebben az irányban.

TOVÁBBLÉPÉS

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 lépés ebben az irányban.

Köszönöm a figyelmet!

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.

Könyvfejezet

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.

VÁLASZOK

ALEXIN ZOLTÁN KÉRDÉSEIRE

„A tagmondatokra bontást F-mérték segítségével mérte. A 171 mondat felhasználásával kapott F-mérték: 85% volt. Kritikaként felvethető, hogy a teszteléshez felhasznált szöveg mérete viszonylag kicsi volt, ami nem ad megbízható predikciót az algoritmus működésre vonatkozóan nagy mennyiségű szöveg esetére.”

„A főnévi szerkezetek az igék vonzatainak meghatározására egy főnévi csoport nyelvtant használt, amelyet más kutatók publikáltak. A többszintű reguláris nyelvtannal megadott definíciót a Magyar Nemzeti Szövegtár 147 millió szavas szövegállományán tesztelte. Mivel ehhez az állományhoz nem tartozik egy referencia elemzés – egy teszt korpusz – ezért a nyelvtan pontosságát, hatékonyságát nem tudta meghatározni. Pedig ez a feldolgozási lépés kulcsszerepet játszik a későbbi algoritmusokban.”

VÁLASZOK ALEXIN ZOLTÁN KÉRDÉSEIRE

- a dolgozat lényegi eredménye:
  modell (1. tézis) + *algoritmus* (4. tézis)

  Fő mondanivaló: bemutatott modell–algoritmus páros alkalmas arra, hogy segítségükkel
  korpuszból kinyerjük a jellegzetes szerkezeteket.
- Egy nagy méretű korpusz megfelelő minőségű reprezentációja *szükséges előfeltétel* volt ahhoz, hogy az algoritmust kipróbálhassam.
- A reprezentáció *előállítása* (2. tézis) tehát szempontomból másodlagosnak tekinthető. Így nem szorosan vett témája a dolgozatnak a nyelvi elemző lépések elemzése, egyenkénti kidolgozása, tökéletesítése.

VÁLASZOK ALEXIN ZOLTÁN KÉRDÉSEIRE

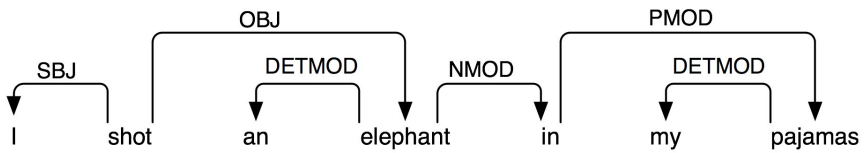
- A reprezentáció előállítása: *közelítő módszerekkel* történt.
  Nem állítom, hogy a nyelvi elemző lépések megvalósítása kiemelkedő minőségű. Azt mutatom be, hogy már az ilyen nem tökéletes bemenetből is jó eredmények születnek.
- Az egyes lépések (pl.: függőségi elemzés) kidolgozása önmagukban önálló PhD dolgozatok témáját adhatják.
  Üzenet: érdemes jobb elemzés és jobb reprezentáció előállításán fáradozni, mert látjuk, hogy az igei szerkezeteket kinyerő algoritmus működik.
- A részleges függőségi elemzés minőségét a végeredmény – azaz a szótár – minőségén mérhetjük le.
  Egy pontossági mérőszám:
  a lexikográfusok által elfogadott szerkezetek aránya **91%**.

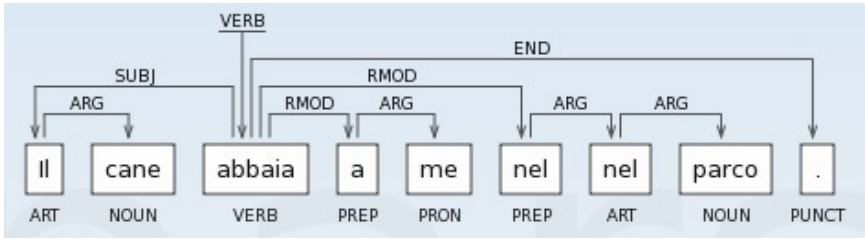
VÁLASZOK

ALEXIN ZOLTÁN KÉRDÉSEIRE

„Mi az oka annak, hogy az igei vonzatkeret modellben az ábrákon következetesen az igéktől mutat a nyíl a bővítmények felé?”

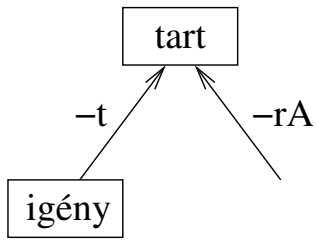
- ige–bővítmény viszony
  = fej–dependens aszimmetrikus függőségi reláció
- mindkét irány előfordul – példák:





- a fej határozza meg . . .
  - az egész szerkezet szemantikai tulajdonságait,
  - azt, hogy az adott dependensek kötelezőek-e,
  - a dependensek morfológiai alakját.

LSzB-t tartalmazó szerkezetek esetén nincsenek a semmiből mutató nyilak.



VÁLASZOK ALEXIN ZOLTÁN KÉRDÉSEIRE

„Feltételezve, hogy a program statisztikai módszerrel tud fordítani a két nyelv között igei szerkezeteket automatikusan, milyen további feladatokat kellene megoldani egy öntanuló automatikus fordítórendszer létrehozásához?”

- gépi fordító rendszer $\leftrightarrow$ lexikai adatbázis
- egy már meglévő gépi fordító rendszer lexikális erőforrását *egészítheti ki* az igei szerkezetek megfelelő fordítására vonatkozó információval – kézi vs. *automatikus* erőforrás-építés

További feladatok:

- igei szerkezetek *jelentésegyértelműsítése*
- *koordináció* kezelése $\rightarrow$ a több azonos viszonyjelölőt tartalmazó szerkezetek helyes fordítása
- több bővítmény esetén: a bővítmények egymásnak való *megfeleltetése*
- öntanulás: további korpuszok felolgozása – offline

VÁLASZOK

BÁRDOSI VILMOS KÉRDÉSEIRE

„2. A 13. oldalon szó van arról, hogy a kollokációk kezelésének igénye az Akadémiai Nagyszótár munkálatai során is felmerült korábban. [. . .] egy megjegyzés erejéig ki lehetne térni arra is, hogy azután – a nemzetközi lexikográfiai gyakorlattal sajnos ellentétes módon – e szempont háttérbe szorult, nagyban csökkentve a készülő szótár használati értékét.”

A Nagyszótárt a dolgozatban nem vizsgáltam,
ezért ettől a megjegyzéstől eltekintettem.

VÁLASZOK BÁRDOSI VILMOS KÉRDÉSEIRE

„3. A 17. oldal többszavas kifejezésekről szóló első bekezdésének végén azt olvashatjuk, hogy a TSZK-kat a "legutóbbi időkig marginális jelenségnek, kivételnek tartották". Ez a kijelentés továbbra is árnyalásra szorul (esetleg csak az angolra vonatkozik?), mivel például francia vonatkozásban a 16. századtól kezdve, különösen pedig az Enciklopédia nyelvészeti szócikkei, majd Bréal (1897) és Bally (1905) meghatározó munkái óta e nyelvi elemek kitüntetett figyelemben részesülnek.”

Igyekeztem egyértelművé tenni a kifogásolt részt: a kijelentés csakis a számítógépes nyelvészet szakirodalmára vonatkozik.

„A TSZK-k a nyelvtan és a lexikon határterületén helyezkednek el, ez lehet az oka annak, hogy a számítógépes nyelvfeldolgozásban a legutóbbi időkig marginális jelenségnek, kivételnek tartották a TSZK-kat, jelentőségüket alábecsülték (Sag et al., 2002).”

VÁLASZOK BÁRDOSI VILMOS KÉRDÉSEIRE

„4. A 17. oldal alján felsorolt [. . .] frazémaosztályok nem teljesen tükrözik a nemzetközi frazeológiai szakirodalom általánosabb és differenciáltabb kategorizálását (vö. pl. HSK 28.1. és 28.2. kötetek). Az "intézményesült kifejezések" típusnál példaként megadott "fáj a feje" szókapcsolat esetében pedig felmerülhet a kérdés, hogy valóban áll-e rá a definíció, mivel a "fáj" ige abban felcserélhető a vele rokonértelmű "hasogat, szétmegy" igékkel.”

- intézményesült kifejezések: kompozicionális szókapcsolatok, de tagjaik nem cserélhetők fel rokonértelmű szóval.
- ‘*fáj a feje*’ – elsősorban a névszói elem miatt tartom intézményesültnek: ‘*fáj a buksija*’ $\sim$ beütötte
  vö: *my hand hurts* vs. *I have a headache*
- ige: ‘*sajog a feje*’ – nem jó, ‘*szétmegy*’ rokonértelműsége.
- példa: ‘*bűncselekményt követ el*’
- példa: ‘*telefonfülke*’ = ‘*telephone booth*’, ‘*telephone box*’, ~~‘*telephone cabinet*’~~, ~~‘*telephone closet*’~~

VÁLASZOK BÁRDOSI VILMOS KÉRDÉSEIRE

„5. A 21-22. oldalakon a többmorfémás kifejezésekről írott rész (2. definíció) alkotja egy frazeológus számára a dolgozat egyetlen zavaró, vitatható részét. A frazéma [. . .] két minimális és elégséges jellemzője a polilexikalitás és a lexikalizálódás. A polilexikalitás [. . .] nem foglalja magában a klasszikus értelemben vonzatos igéknek nevezett egységek vonzatait, esetragjait (hisz vmiben, croire à/en qqch. [...]). Attól persze, hogy ez a kiterjesztett felfogás nem illik bele a frazeológia megszokott kategóriáiba, a dolgozat és főleg a végtermék [...] szempontjából a lépés érthető és logikus.”

Alapötletem: foglalkozzunk *egységes keretben*
a kollokációs és a vonzatos igékkel.

1. kollokációk kezelése – vonzatok kezelése
2. Egy nyelvtanulónak például mindegy,
   hogy egy adott nyelvi elem szó vagy frazéma.

   Az egyik nyelven szó, a másikon frazéma:
   ‘*krumpli*’ = ‘*pommes de terre*’, ‘*participer à*’ = ‘*nemen deel aan*’

VÁLASZOK

BÁRDOSI VILMOS KÉRDÉSEIRE

„6. 23. oldal alulról a 2. bekezdés: "Az ilyen típusú szerkezetek egyszerre vonzatkeretek és többszavas kifejezések: a kollokációk közül (és a kollokációs szótárakból) vonzatuk miatt, a vonzatkeretek közül (és a vonzatszótárakból) pedig a jelen lévő kollokátum miatt lógnak ki." A bekezdés utolsó mondatát esetleg lehetne úgy árnyalni, hogy e példák viszont mind benne vannak a jó frazeológiai szótárakban.”

Az idézett résznél megtartottam az eredeti szövegezést, mert itt általánosságban beszélek kollokációs szótárról és vonzatszótárról. A javasolt árnyalást a 73. oldalon tettem meg a következő mondat révén:

"Megjegyzendő, hogy a modern frazeológiai szótárak a kollokációk mellett figyelmet fordítanak a vonzatok gondos feltüntetésére is (Forgács, 2003; Bárdosi, 2009)."

VÁLASZOK BÁRDOSI VILMOS KÉRDÉSEIRE

„7. 103. oldal, a 19. táblázat 18. sorszámú szerkezetére vonatkozó megjegyzéssel kapcsolatban jelzem, hogy a francia is rendelkezik egy olyan megfelelő szerkezettel a korpusz által jelzett 'participer' mellett (= prendre part), ami biztosítja a szimmetriát."

- '*participer à*' = '*nemen deel aan*' (részt vesz -bAn)
- A példa megmutatja: az algoritmus képes nem azonos felépítésű szerkezeteket egymásnak megfeleltetni.
- A korpusz alapján az algoritmus azt az információt is szolgáltatja, hogy ahol a hollandban '*nemen deel aan*' van ott a franciában „általában" '*participer à*' „szokott" lenni.

VÁLASZOK BÁRDOSI VILMOS KÉRDÉSEIRE

„8.a. A dolgozatban kitüntetett szerepet játszó frazémákkal kapcsolatban azonban némi hiányérzete van az olvasónak, aki joggal hiányolhatja a nemzetközi frazeológiai szakirodalom néhány alapművét (pl. Harald Burger et al. által szerkesztett és a Walter de Gruyter kiadónál 2007-ben a HSK 28.1. és 28.2. köteteként Phraseologie címmel megjelent művet).”

Igyekeztem megfelelni ennek a kívánalomnak. Az említett mű helyett annak egy számomra könnyebben hozzáférhető előzménye került be az irodalomjegyzékbe:

Harald Burger: *Phraseologie. Eine Einführung am Beispiel des Deutschen.* Erich Schmidt Verlag, Berlin, 2003.

IGEI SZERKEZETEK GYAKORISÁGI SZÓTÁRA

EGY AUTOMATIKUS LEXIKAI KINYERŐ ELJÁRÁS ÉS ALKALMAZÁSA

című doktori (Ph.D.) disszertáció nyilvános védése

Sass Bálint
sass.balint@itk.ppke.hu

PPKE ITK
Budapest, 2011. október 14.