Egy általános modellnek megfelelő szerkezetek kinyerése korpuszból

Sass Bálint
sass.balint@nytud.hu

MTA Nyelvtudományi Intézet
Budapest, 2011. június 16.

Bevezető – jelentőség

Példák

‘*részt vesz vmiben*’, ‘*beleüti az orrát vmibe*’, ‘*szó van vmiről*’, ‘*hasznot húz vmiből*’, ‘*kétségbe von vmit*’, ‘*kockán forog vmi*’, ‘*górcső alá vesz vmit*’ . . .

- Mik ezek?
  Vonzatos komplex igék:
  kollokációk *és* vonzatkeretek egyszerre
- gyakoriak, sokszor idiomatikus jelentéssel
- lexikai adatbázisba kellenek:
  tanulói szótárba *és* nyelvtechnológiai alkalmazások
  (pl.: gépi fordító) háttér-adatbázisába egyaránt

A VONZATOS KOMPLEX IGÉK FELÉPÍTÉSE

szerkezet	LKB	LSzB
‘*részt vesz vmiben*’,	-t	-bAn
‘*beleüti az orrát vmibe*’,	-t	-bA
‘*szó van vmiről*’,	∅	-rÓl
‘*hasznot húz vmiből*’,	-t	-bÓl
‘*kétségbe von vmit*’,	-bA	-t
‘*kockán forog vmi*’,	-n	∅
‘*górcső alá vesz vmit*	alá	-t

- *lexikálisan kötött bővítmény (LKB)*
- *lexikálisan szabad bővítmény (LSzB)*

A kollokátumot és a vonzatot
ugyanazok a nyelvi eszközök – esetrag, névutó – jelölik.

A VONZATOS KOMPLEX IGÉK FELÉPÍTÉSE

A kollokátumot és a vonzatot
ugyanazok a nyelvi eszközök – esetrag, névutó – jelölik.

Egy igén „belül” is előfordul a szerepek felcserélődése:

szerkezet	LKB	LSzB
‘*pillantást vet vmire*’	-t	-rA
‘*szemére vet vmit*’	-rA	-t

Feladat: ezek szétválasztása, az LKB-k, LSzB-k azonosítása.

Követelmények

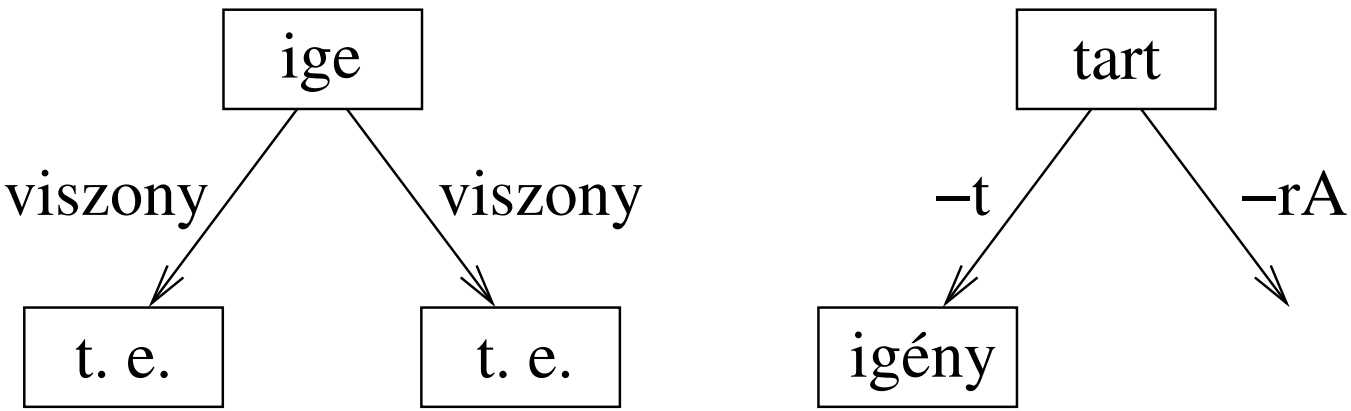
→ Egy olyan lexikai kinyerő eljárásra van szükség, mely…

- képes korpusz alapján az igei szerkezeteket azonosítani;
- 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.

MODELL

- *szerkezetmodell:*
  szerkezet = ige + bővítmények
  bővítmény = viszonyjelölő + tartalmi elem



- egyúttal *tagmondatmodell* is:
  1 tagmondat $\sim$ 1 igei szerkezet

  tagmondat-reprezentáció: csak LKB!

  pl.: '*India egész Kasmírra igényt tart.*'

A KORPUSZTÓL A SZERKEZETEKIG

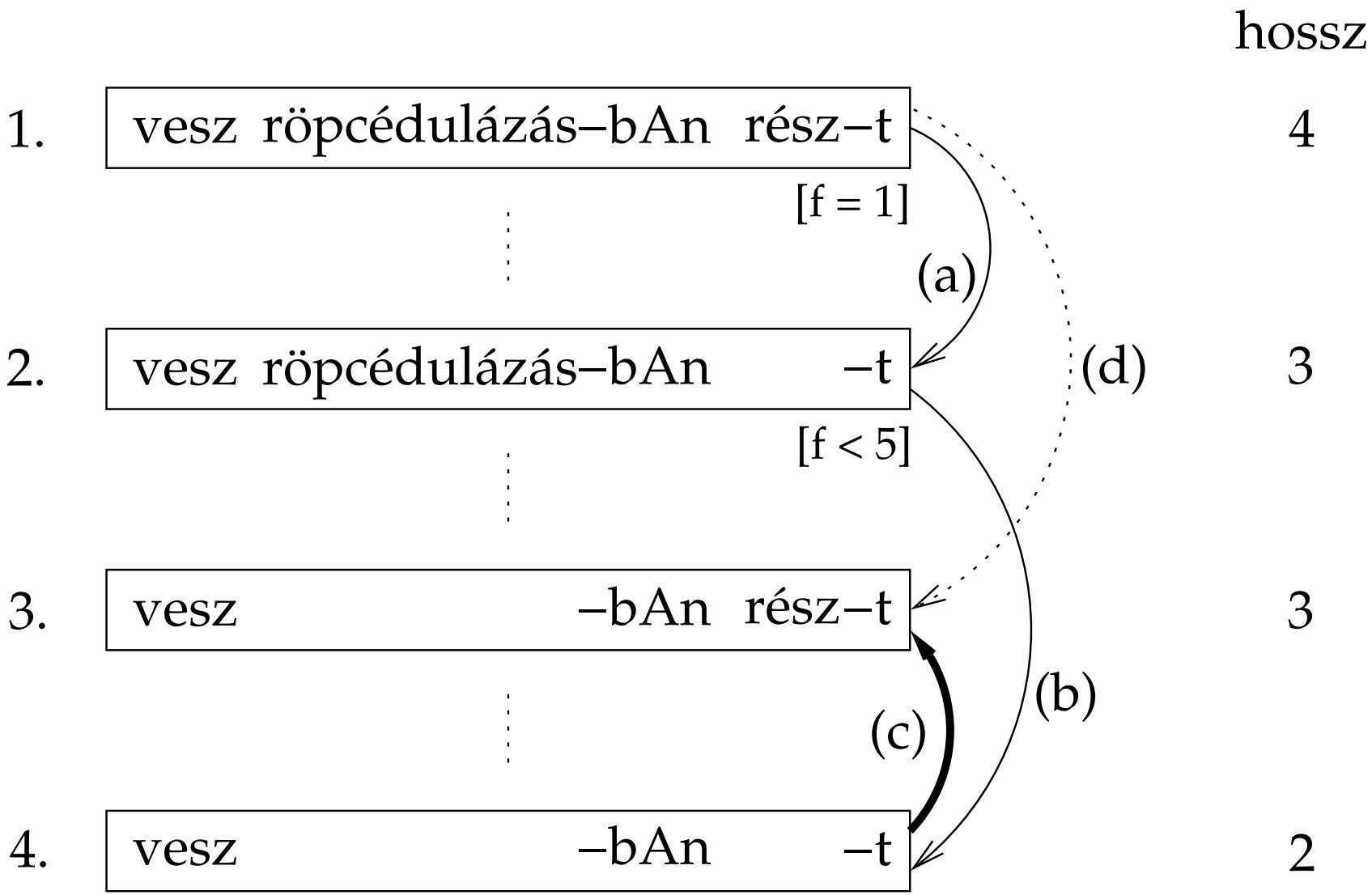
tevékenység	anyag	példa
	korpusz	‘*Vess egy pillantást a térképre.*’
tagmondatra bontás szintaktikai elemzés	↓	
	reprezentáció	`ige`=vet `-t`=pillantás `-rA`=térkép
jellegzetes igei szerkezetek kinyerése	↓	
	igei szerkezetek	‘*pillantást vet vmire*’

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.

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
(d)
3
[f < 5]
3.
vesz –bAn rész–t
3
(b)
(c)
4.
vesz –bAn –t
2

A KORPUSZTÓL A SZERKEZETEKIG

tevékenység	anyag	példa
	korpusz	‘*Vess egy pillantást a térképre.*’
tagmondatra bontás szintaktikai elemzés	↓	
	reprezentáció	`ige`=vet `-t`=pillantás `-rA`=térkép
jellegzetes igei szerkezetek kinyerése	↓	
	igei szerkezetek	‘*pillantást vet vmire*’

→ Igei szerkezetek szótára

NYELVFÜGGETLENSÉG

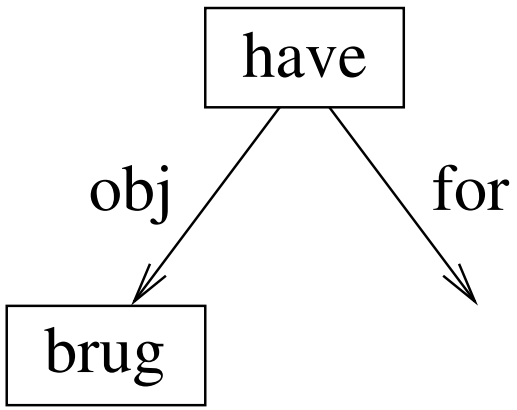
Állítás: a modell nyelvfüggetlen.
A magyaron kívül számos nyelvre előállítható a modell szerinti reprezentáció, és kinyerhetők a fenti típusú igei szerkezetek.

Reprezentáció előállítása: viszonyjelölők meghatározása

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

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

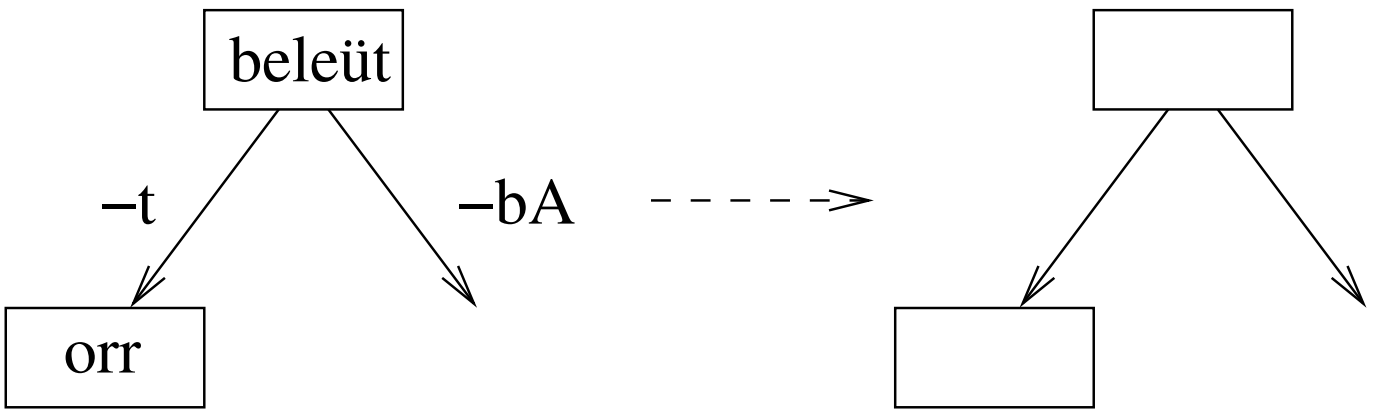
szerb példa: '*ići u prilog*'
= támogat („haszonba megy")



AZ ABSZTRAKT MODELL

A kinyerő algoritmus működésének feltétele: *pusztán* a fenti gráfstuktúrának megfelelő formájú reprezentáció.

Megtehetjük tehát, hogy elvonatkoztatunk az eddigi tartalomtól, és egyéb szerkezeteket próbálunk meg ebben a struktúrában reprezentálni.

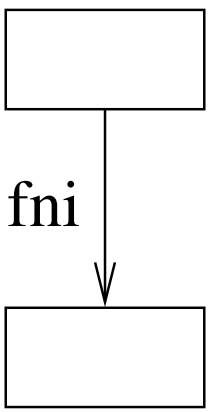


bővítmény → *jegy*
viszonyjelölő → *él* vagy *címke*
tartalmi elem → *csomópont*

AZ ABSZTRAKT MODELL ALKALMAZÁSA 1/3

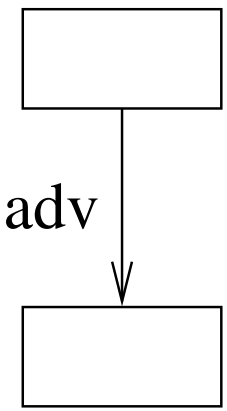
Névszói csoporttól különböző bővítmények kezelése.

Főnévi igenév bővítmény.



→ ‘*kell csinálni*’, ‘*kell leperkálni*’ . . .

Határozószó bővítmény.



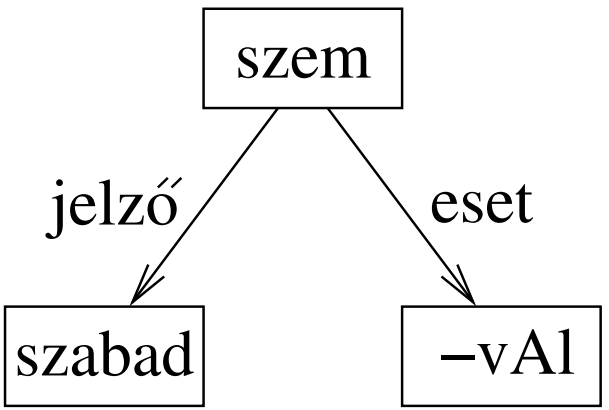
→ ‘*mindig akad*’, ‘*együtt él*’ . . .

Helyhatározó, időhatározó hasonlóan kezelhető,
ha van ilyen annotáció a korpuszban.

AZ ABSZTRAKT MODELL ALKALMAZÁSA 2/3

Nem ige-központú szerkezetek kezelése:
jellegzetes *főnévi* csoportok kinyerése.

Jegyek:
jelző, főnév esete, főnév száma, főnév birtokos személyragja
→ ‘*saját lábán*’, ‘*száraz lábbal*’, ‘*belső fül*’, ‘*szabad szemmel*’

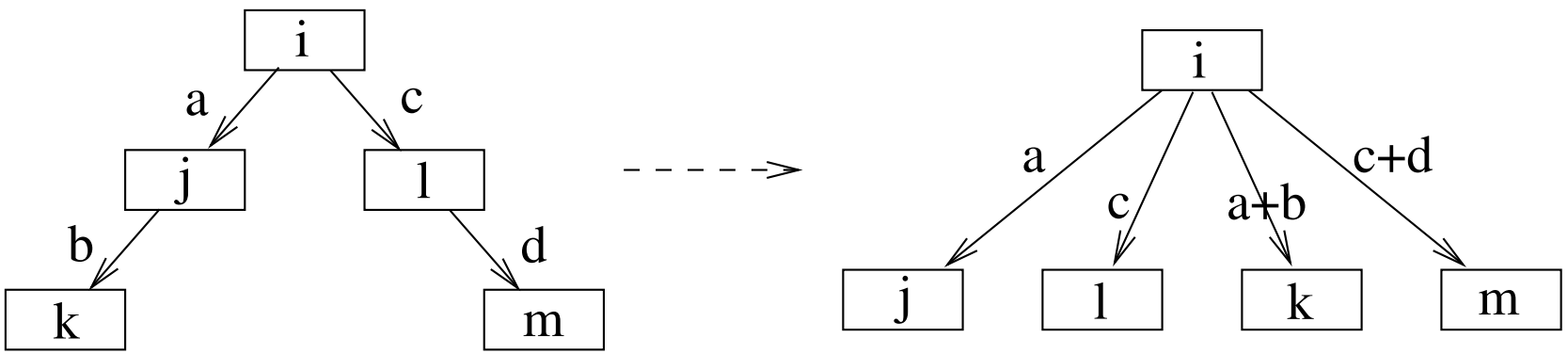


Eset megjelenése: él helyett csomópont!

AZ ABSZTRAKT MODELL ALKALMAZÁSA 3/3

Többszintű függőségi struktúra kezelése.

Tipikus eset: ige + főnévi bővítmény + főnév jelzője



Visszavezetés az egyszintű modellre: „kisimítás”

Az absztrakt modell alkalmazása 3/3

Többszintű függőségi struktúra kezelése.

Holland példák – részletesen elemzett korpuszból.

holland:	‘ige=speel obj=rol obj+ADJ=belangrijk’
magyar:	‘*jelentős szerepet játszik*’
holland:	‘ige=bewaar in=verpakking in+ADJ=oorspronkelijk’
magyar:	‘*eredeti csomagolásban tárol*’
holland:	‘ige=breng tot=einde tot+ADJ=goed obj’
magyar:	‘*sikerre („jó befejezésig”) visz vmit*’

Az absztrakt modell alkalmazása 3/3

Többszintű függőségi struktúra kezelése.

Holland példák – részletesen elemzett korpuszból.

holland:	‘ige=speel obj=rol obj+ADJ=belangrijk’
magyar:	‘*jelentős szerepet játszik*’
holland:	‘ige=bewaar in=verpakking in+ADJ=oorspronkelijk’
magyar:	‘*eredeti csomagolásban tárol*’
holland:	‘ige=breng tot=einde tot+ADJ=goed obj’
magyar:	‘*sikerre („jó befejezésig”) visz vmit*’

Ötlet

Párhuzamos korpusz és párhuzamos igei szerkezetek
(igei szerkezetek és fordításaik)
szintén reprezentálhatók a fenti módon.

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

Hogyan lehet a módszert párhuzamos korpuszra alkalmazni?

Ötlet

Párhuzamos korpusz és párhuzamos igei szerkezetek
(igei szerkezetek és fordításaik)
szintén reprezentálhatók a fenti módon.

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

Hogyan lehet a módszert párhuzamos korpuszra alkalmazni?

Speciális reprezentáció: metakorpusz.

. . . a kétnyelvű korpuszt egynyelvűnek „álcázzuk”

A METAKORPUSZ KIALAKÍTÁSA

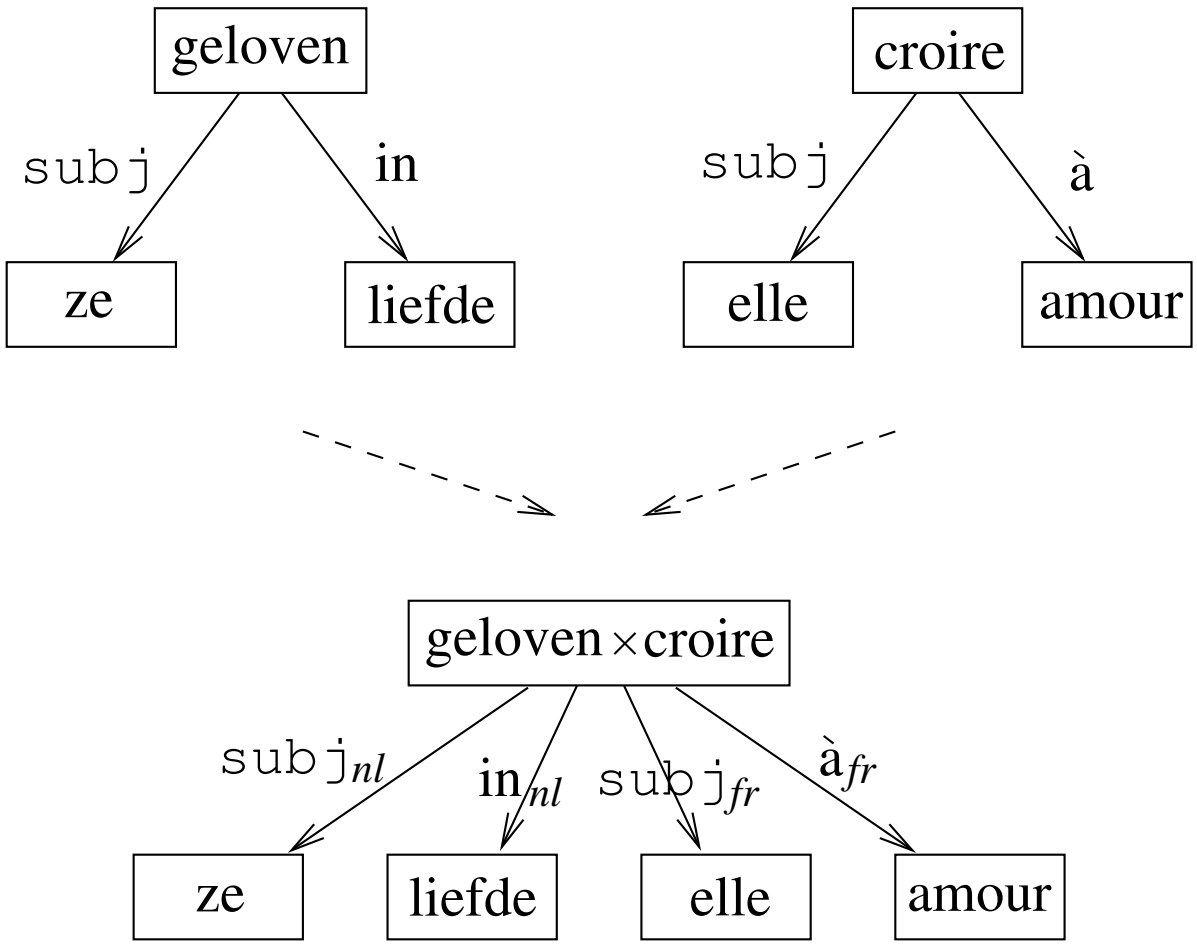
korpusz: Dutch Parallel Corpus, holland–francia (3,5 mió token)
elemzés: nyelvenként külön, tagmondatra bontás és részleges szintaktikai elemzés egyszerű szabályokkal

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.

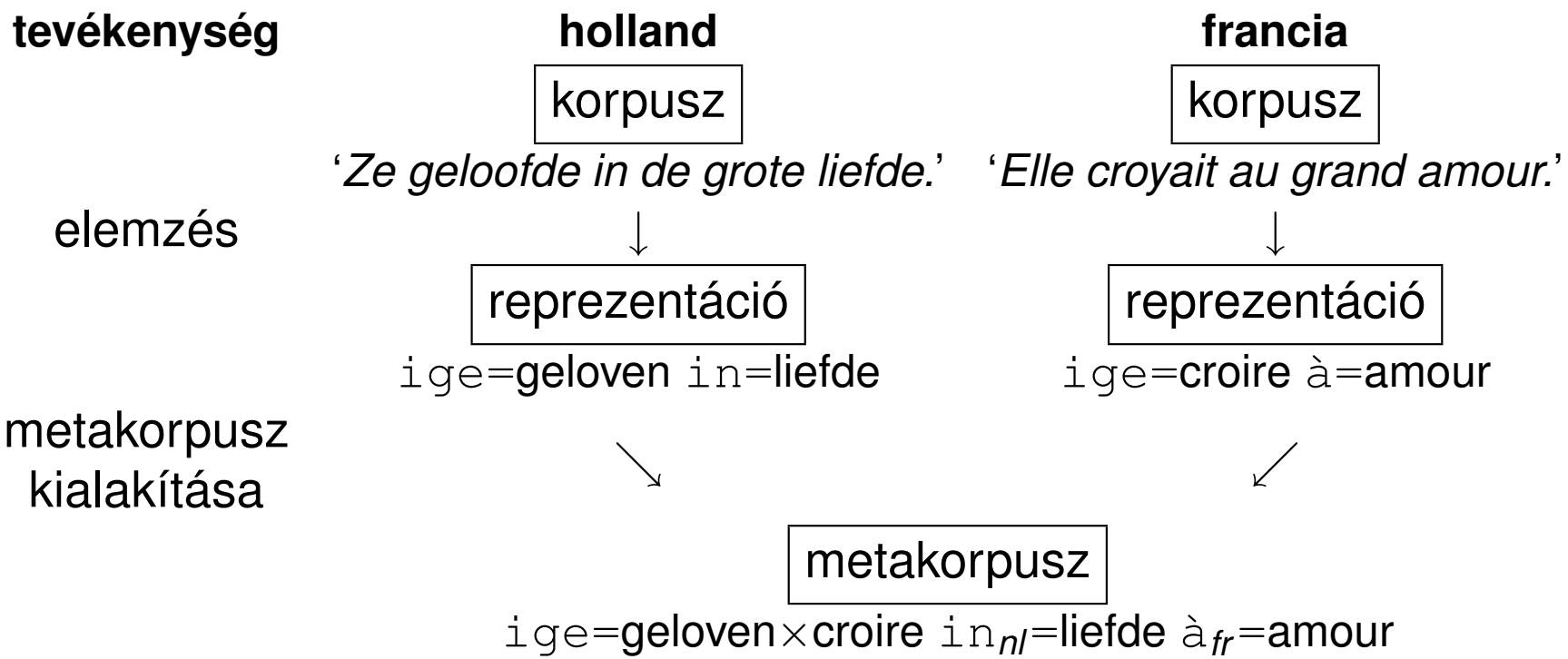
holland tagmondat:	'*Ze geloofde in de grote liefde.*'
francia tagmondat:	'*Elle croyait au grand amour.*'
magyar fordítás:	'Hitt a nagy szerelemben.'
reprezentáció:	`ige`=geloven×croire $\texttt{in}_{nl}$=liefde $\texttt{à}_{fr}$=amour

A METAKORPUSZ KIALAKÍTÁSA

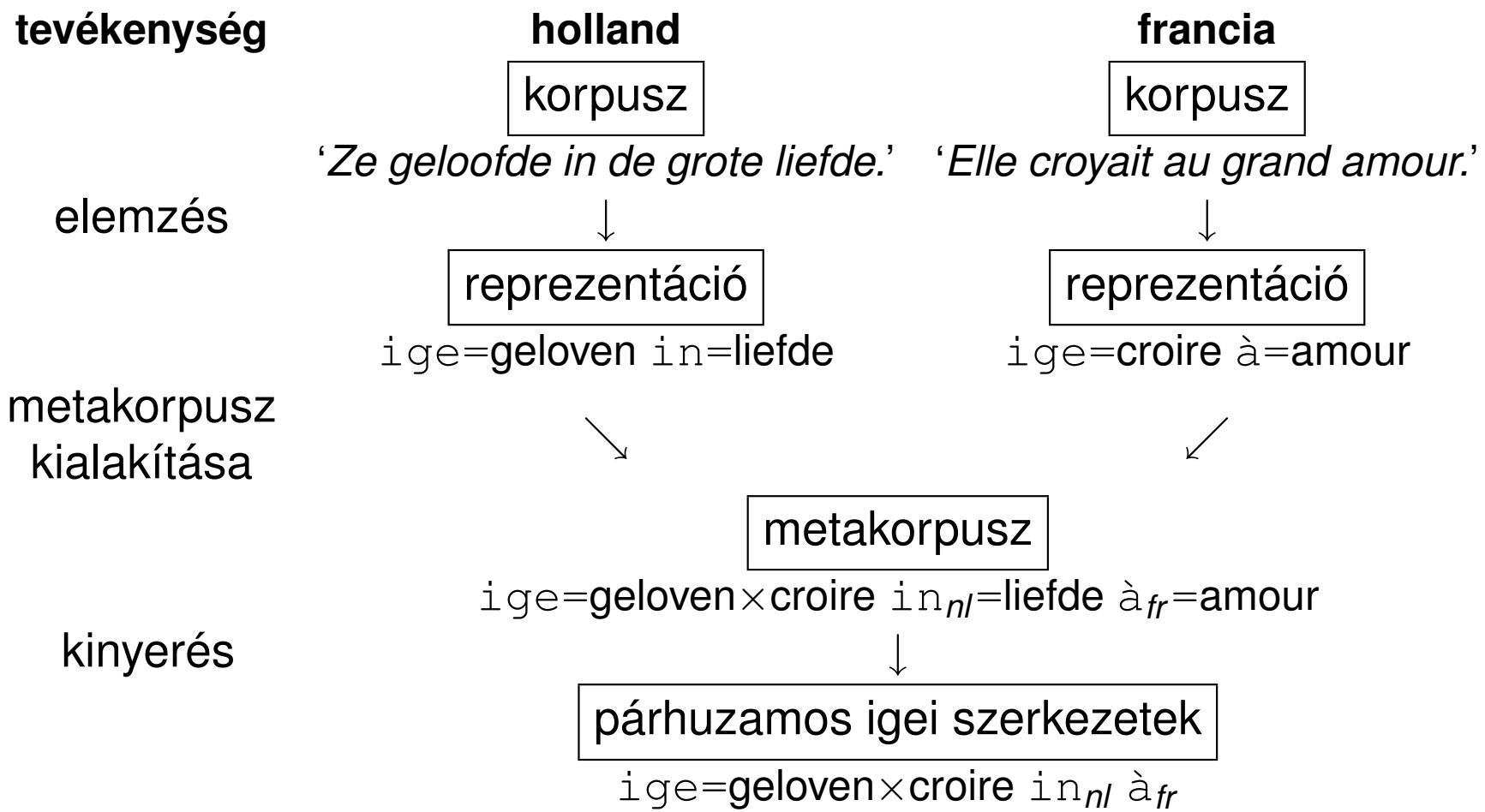
Visszavezetés az eredeti modellre: „összefésülés”



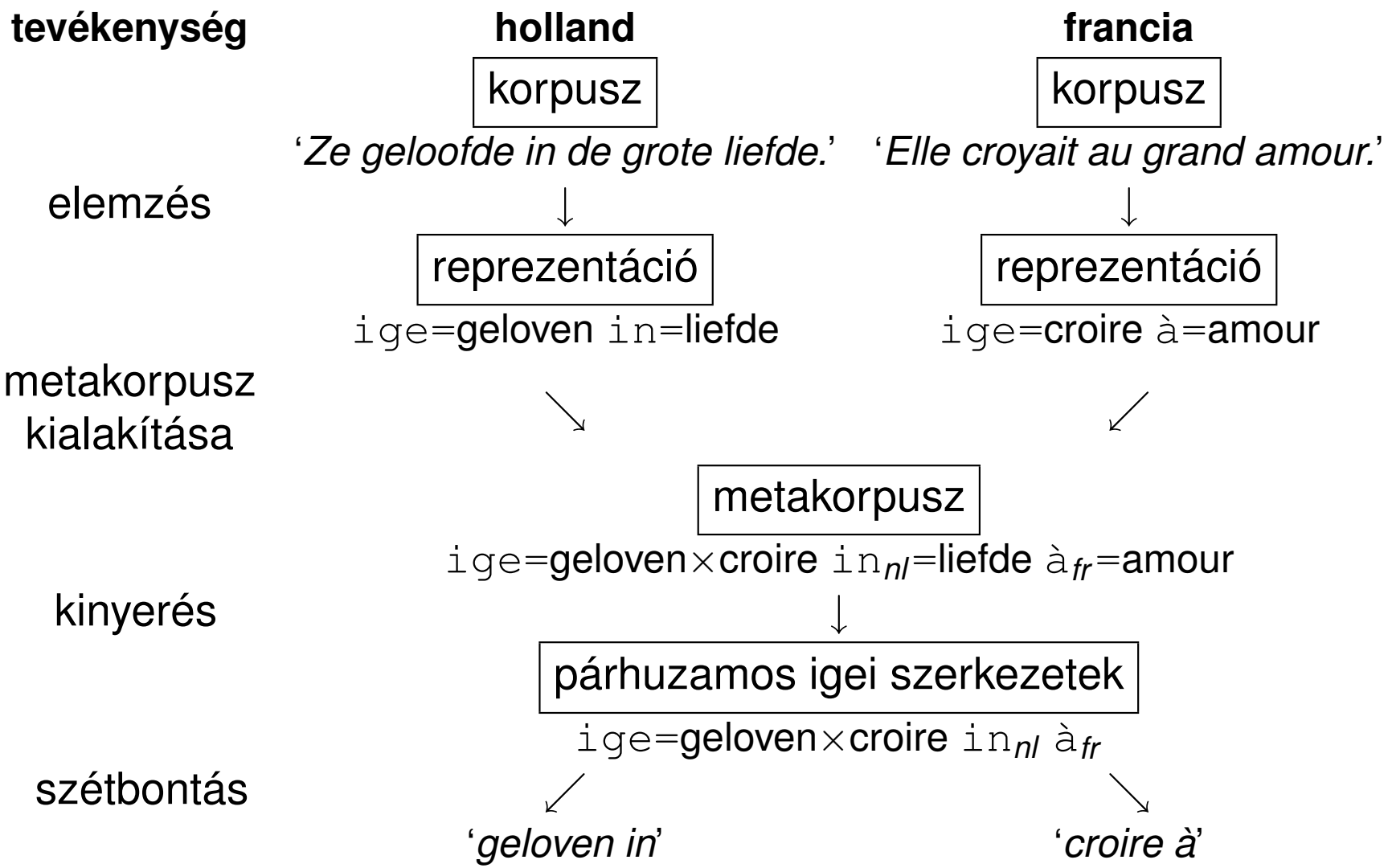
A MÓDSZER ALKALMAZÁSA KÉT NYELVRE



A MÓDSZER ALKALMAZÁSA KÉT NYELVRE



A MÓDSZER ALKALMAZÁSA KÉT NYELVRE



Példák – aszimmetria

eredmény: aszimmetrikus szerkezetek

gyenge (tartalmi) aszimmetria

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

erős (formai) aszimmetria

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

‘*zijn van toepassing op*’ = ‘*appliquer se à*’ ’vonatkozik vmire’

PÉLDÁK – ASZIMMETRIA

eredmény: aszimmetrikus szerkezetek

GYENGE (TARTALMI) ASZIMMETRIA

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

ERŐS (FORMAI) ASZIMMETRIA

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

‘*zijn van toepassing op*’ = ‘*appliquer se à*’ ’vonatkozik vmire’

PÉLDÁK – ASZIMMETRIA

eredmény: aszimmetrikus szerkezetek

GYENGE (TARTALMI) ASZIMMETRIA

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

ERŐS (FORMAI) ASZIMMETRIA

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

‘*zijn van toepassing op*’ = ‘*appliquer se à*’ ’vonatkozik vmire’

Példák – szinonimák

eredmény: szinonimák

adott szerkezet több megfelelője + gyakorisági viszonyokkal

‘*agir se de*’ ’szó van róla, szóban forog, illeti, vonatkozik’ szerkezet négy fordítása:

holland megfelelő	gyakorisági érték
‘*gaan om*’	114
‘*zijn OBJ*’	69
‘*betreffen OBJ*’	27
‘*gaan over*’	24

→ lexikográfiai felhasználás

→ gépi fordítás: további szabályokat lehet tanulni, hogy melyik fordítás milyen feltételek mellett alkalmazandó

PÉLDÁK – IDIOMATIKUS MEGFELELŐK

eredmény: idiomatikus megfelelők

IGÉK

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

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

ELÖLJÁRÓK

‘*nemen deel aan*’ = ‘*participer à*’ ’részt vesz vmiben’
‘*doen beroep op*’ = ‘*faire appel à*’ ’fellebbez vkihez’
‘*hebben effect op*’ = ‘*avoir effet sur*’ ’hatása van vmire’

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

PÉLDÁK – IDIOMATIKUS MEGFELELŐK

eredmény: idiomatikus megfelelők

IGÉK

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

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

ELÖLJÁRÓK

‘*nemen deel aan*’ = ‘*participer à*’ ’részt vesz vmiben’
‘*doen beroep op*’ = ‘*faire appel à*’ ’fellebbez vkihez’
‘*hebben effect op*’ = ‘*avoir effet sur*’ ’hatása van vmire’

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

PÉLDÁK – IDIOMATIKUS MEGFELELŐK

eredmény: idiomatikus megfelelők

IGÉK

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

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

ELÖLJÁRÓK

‘*nemen deel aan*’ = ‘*participer à*’ ’részt vesz vmiben’
‘*doen beroep op*’ = ‘*faire appel à*’ ’fellebbez vkihez’
‘*hebben effect op*’ = ‘*avoir effet sur*’ ’hatása van vmire’

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

PÉLDÁK – IDIOMATIKUS MEGFELELŐK

eredmény: idiomatikus megfelelők

IGÉK

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

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

ELÖLJÁRÓK

‘*nemen deel aan*’ = ‘*participer à*’ ’részt vesz vmiben’
‘*doen beroep op*’ = ‘*faire appel à*’ ’fellebbez vkihez’
‘*hebben effect op*’ = ‘*avoir effet sur*’ ’hatása van vmire’

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

Összefoglalás 1/2

Sokféle szerkezet reprezentációja megfogalmazható a bemutatott absztrakt modellben, így ez az általános keret a változatlan kinyerő eljárással együttműködve számos típusú jellegzetes szerkezet kinyerésére alkalmassá tehető.

- igei szerkezetek névszói csoport bővítményekkel
- egyéb bővítményekkel
- jellegzetes névszói csoportok
- többszintű szerkezetek
- párhuzamos szerkezetek

Összefoglalás 2/2

Párhuzamos korpuszra alkalmazva a módszer kétnyelvű igei szerkezetek hasznos gyűjteményét képes előállítani. Felfedezi a formailag egymásra nem hasonlító, de egymásnak megfelelő, egymás fordításaiként kezelendő igei szerkezeteket is.

A módszer egyszerre rendelkezik az alábbi tulajdonságokkal:

- igei kollokációkinyerés
- igei vonzatkeret-megállapítás
- megszakított és változó szórendű szerkezetek kezelése
- többnyelvű szerkezetek kinyerése párhuzamos korpuszból
- nyelvfüggetlen eljárás

A jövőben hozzájárulhat kétnyelvű szótárak korpuszvezérelt anyaggyűjtési munkálataihoz.

Összefoglalás

Párhuzamos korpuszra alkalmazva a módszer kétnyelvű igei szerkezetek hasznos gyűjteményét képes előállítani. Felfedezi a formailag egymásra nem hasonlító, de egymásnak megfelelő, egymás fordításaiként kezelendő igei szerkezeteket is.

A módszer egyszerre rendelkezik az alábbi tulajdonságokkal:

- igei kollokációkinyerés
- igei vonzatkeret-megállapítás
- megszakított és változó szórendű szerkezetek kezelése
- többnyelvű szerkezetek kinyerése párhuzamos korpuszból
- nyelvfüggetlen eljárás

A jövőben hozzájárulhat kétnyelvű szótárak korpuszvezérelt anyaggyűjtési munkálataihoz.

Köszönöm a figyelmet!