PÁRHUZAMOS IGEI SZERKEZETEK KÖZVETLEN KINYERÉSE PÁRHUZAMOS KORPUSZBÓL

Sass Bálint
sass.balint@nytud.hu

MTA Nyelvtudományi Intézet, Budapest

MSZNY2010
Szeged, 2010. december 2-3.

MEGLÉVŐ MÓDSZER EGY NYELVRE

tevékenység	anyag	példa
	korpusz	*Vess egy pillantást a térképre.*
tagmondatra bontás NP-chunking	↓	
	reprezentáció	`ige`:vet `-t`:pillantás `-rA`:térkép
jellegzetes igei szerkezetek kinyerése	↓	
	igei szerkezetek	*pillantást vet vmire*

Előnyök

- felismeri, hogy adott bővítményi elem lexikálisan kötött, vagy kitöltetlen, vonzatszerű
  (pl.: *pillantást vet vmire* ↔ *szemére vet vmit*)
- egyszerre állapítja meg a kollokátumokat és a vonzatokat, így *teljes* szerkezeteket eredményez

MEGLÉVŐ MÓDSZER EGY NYELVRE

tevékenység	anyag	példa
	korpusz	*Vess egy pillantást a térképre.*
tagmondatra bontás NP-chunking	↓	
	reprezentáció	`ige`:vet `-t`:pillantás `-rA`:térkép
jellegzetes igei szerkezetek kinyerése	↓	
	igei szerkezetek	*pillantást vet vmire*

JELLEGZETES IGEI SZERKEZETEK KINYERÉSE

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, *vmi jön létre*, *társasház jön vmire*, *vmi jön vmire*.
2. Hossz szerint csökkenő sorba rendezés.
   Hossz (h) = esetek száma + kötött szavak száma.
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. A megmaradó szerkezetek gyakorisági érték szerint rendezett listája adja az összegyűjtött igei szerkezeteket.

JELLEGZETES IGEI SZERKEZETEK KINYERÉSE

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, *vmi jön létre*, *társasház jön vmire*, *vmi jön vmire*.
2. Hossz szerint csökkenő sorba rendezés.
   Hossz (h) = esetek száma + kötött szavak száma.
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. A megmaradó szerkezetek gyakorisági érték szerint rendezett listája adja az összegyűjtött igei szerkezeteket.

MEGLÉVŐ MÓDSZER EGY NYELVRE

tevékenység	anyag	példa
	korpusz	*Vess egy pillantást a térképre.*
tagmondatra bontás NP-chunking	↓	
	reprezentáció	`ige`:vet `-t`:pillantás `-rA`:térkép
jellegzetes igei szerkezetek kinyerése	↓	
	igei szerkezetek	*pillantást vet vmire*

→ Igei szerkezetek szótára

Ötlet

Hogyan lehetne ezt párhuzamos korpuszra alkalmazni?

. . . és így „párhuzamos szerkezeteket”
(szerkezeteket és fordításaikat) kinyerni?

Ötlet

Hogyan lehetne ezt párhuzamos korpuszra alkalmazni?

. . . és így „párhuzamos szerkezeteket”
(szerkezeteket és fordításaikat) kinyerni?

Trükk: metakorpusz.

. . . a kétnyelvű korpuszt egynyelvűnek „álcázzuk”,
és *közvetlenül* futtatjuk az eredeti eljárást.

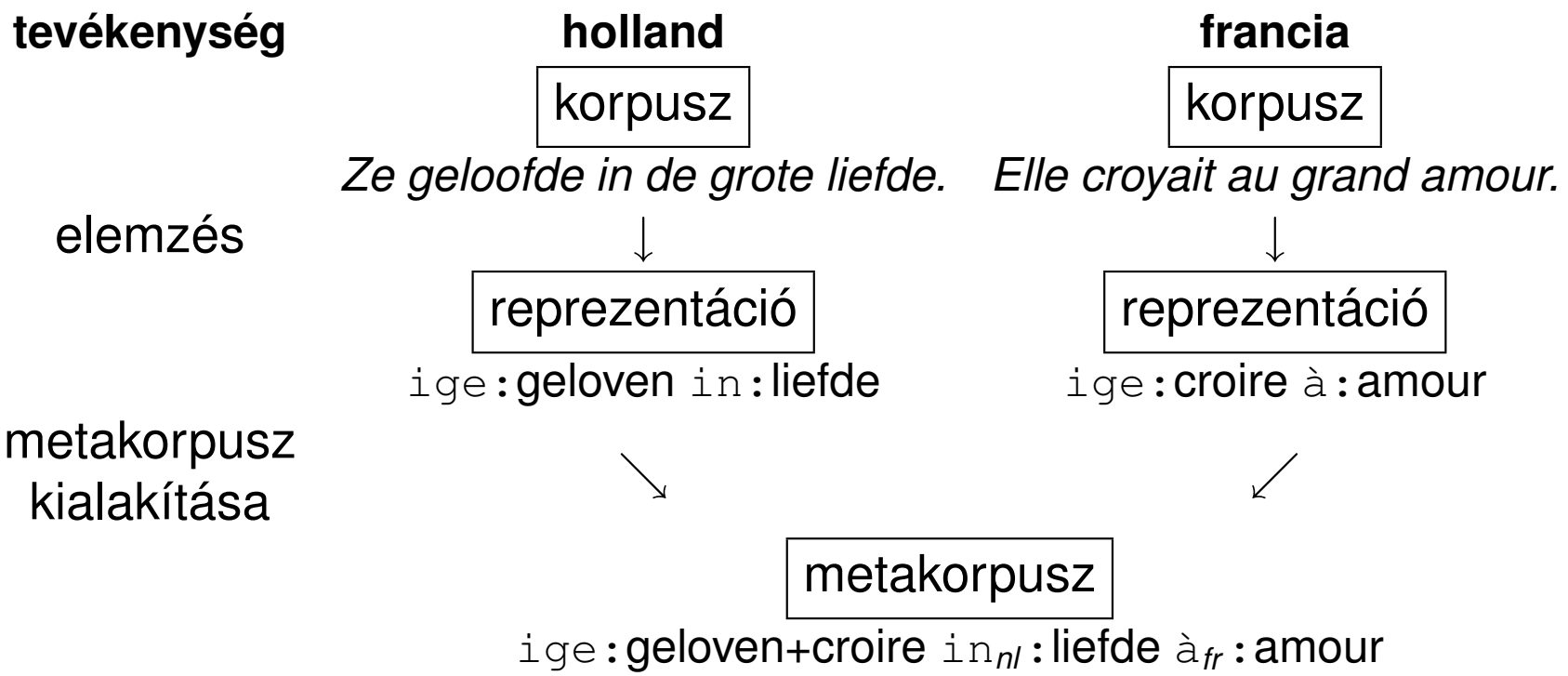
A METAKORPUSZ KIALAKÍTÁSA

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

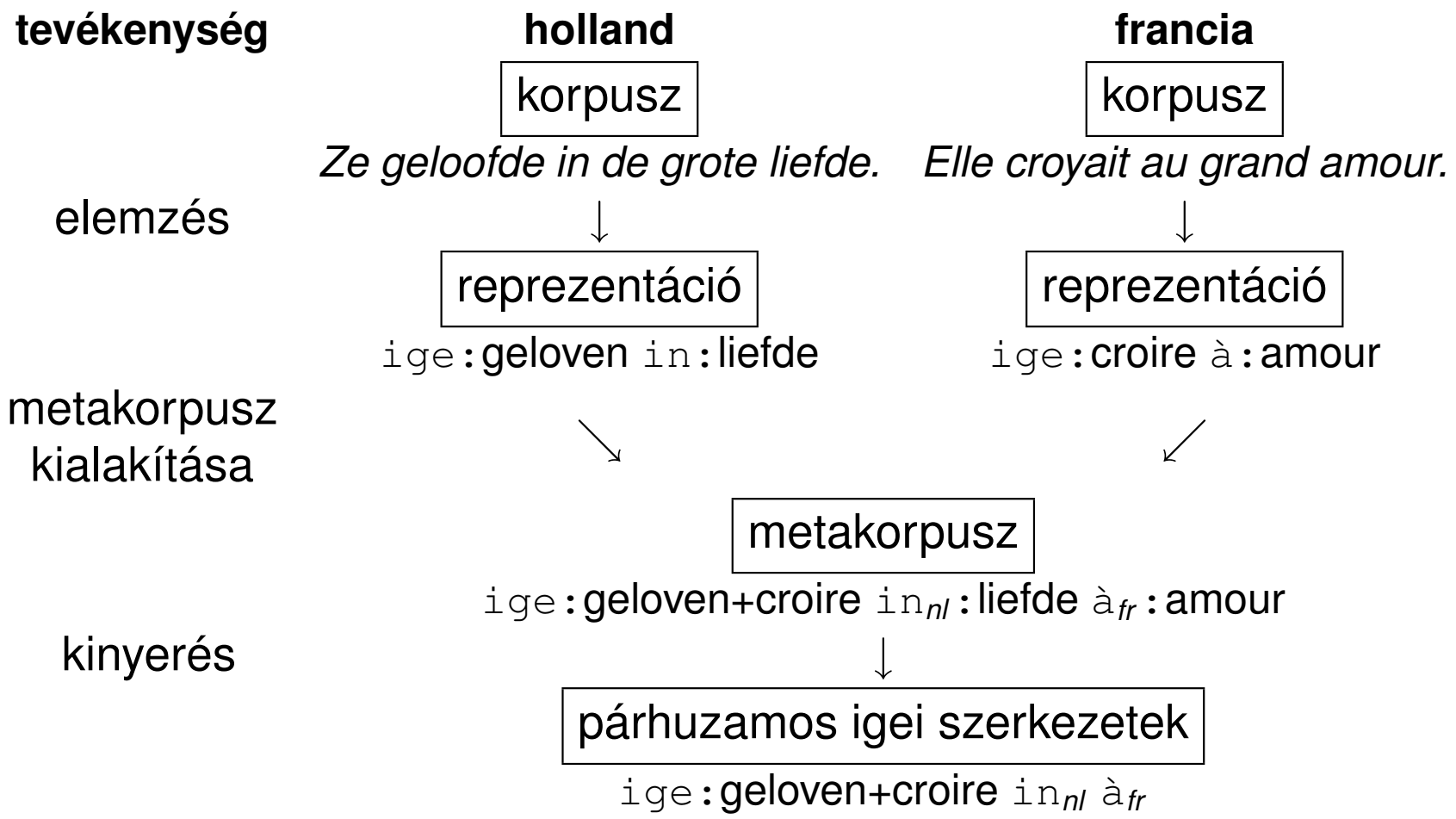
1. Tagmondat-szintű illesztés: a tagmondatokat fordítási egységenként sorra egymáshoz rendeltük.
2. Az egymáshoz rendelt tagmondatok holland ill. francia igéjéből: igepár. (pl.: *gaan+aller* 'megy')
3. A tagmondatpárban található bővítményeket (mindkét nyelvűeket) egy halmazként soroltuk fel az igepár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

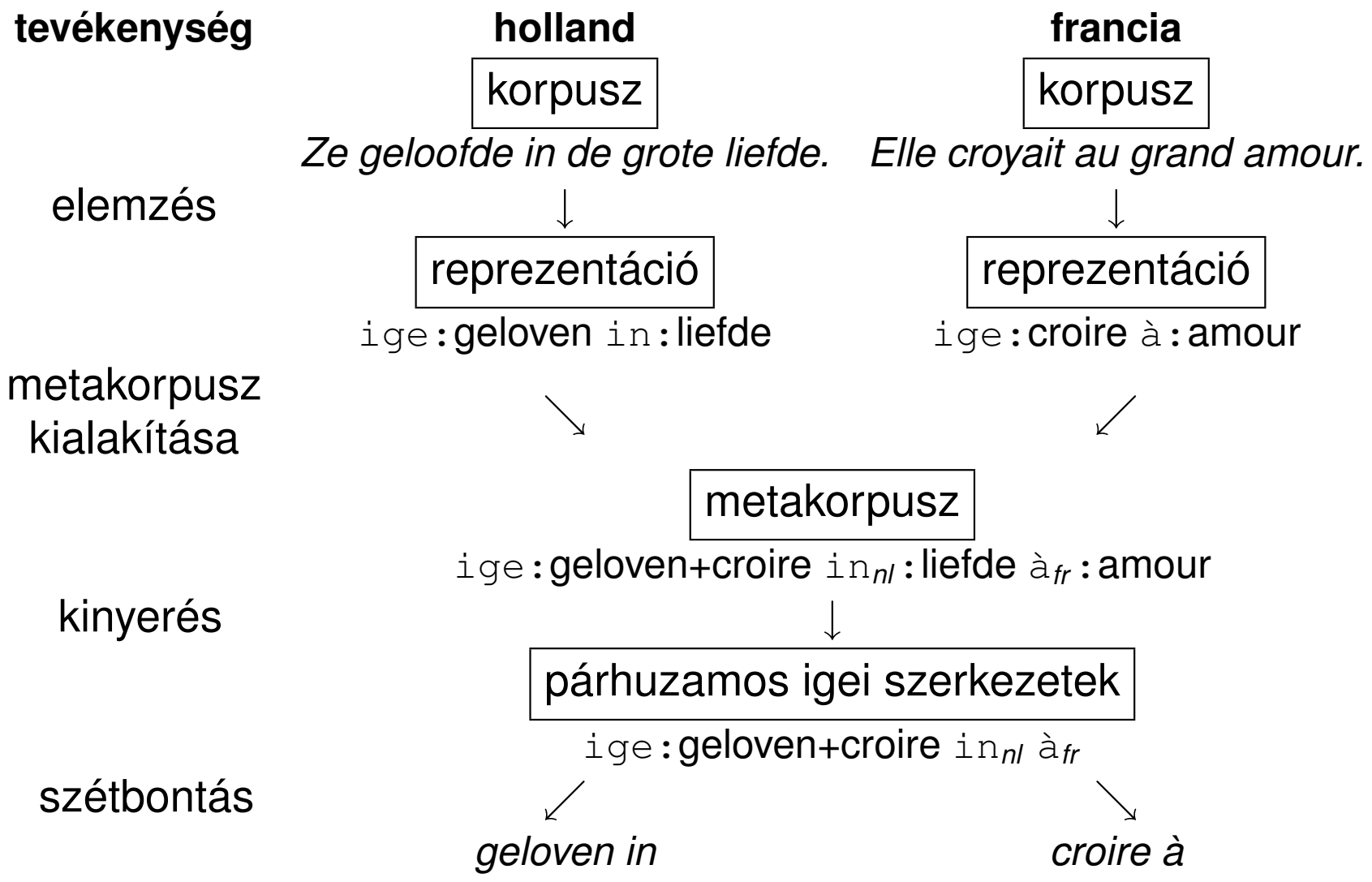
A MÓDSZER KÉT NYELVRE



A MÓDSZER KÉT NYELVRE



A MÓDSZER KÉT NYELVRE



KIÉRTÉKELÉS

futtatás: $f \geq 20$ előforduló 1356 db igepárra
kiértékelés tárgya: vonzatos komplex igék (pl.: *részt vesz vmiben*)

jó = értelmes, *teljes* szerkezet, megfelelő fordítás

Engedmények:

- birtokos szerkezet: holland *van* ill. francia *de*
- alany és tárgy
- határozószó hiánya

eredmény: 58 db legalább 15-ös gyakorisági értékű szerkezet, melyben vonzat és lexikálisan kötött bővítmény is volt.

Ebből 34 bizonyult helyesnek: **pontosság = 58,6%**

(Korábbi egynyelvű pontosság magyarra, 50 db-ra: 94%)

PÉLDÁK – ASZIMMETRIA

eredmény: aszimmetrikus szerkezetek

Def: más felépítésű (pl.: *krumpli* = *pommes de terre*)

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

Def: más felépítésű (pl.: *krumpli* = *pommes de terre*)

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

Def: más felépítésű (pl.: *krumpli* = *pommes de terre*)

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* 'fellebbez vkihez'

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* 'fellebbez vkihez'

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* 'fellebbez vkihez'

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* 'fellebbez vkihez'

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'

Felhasználás

Egy gépi fordítónak az ilyen fajta szerkezeteket ismernie kell: legalább a leggyakoribbakat.

Google fordító

Het gaat om een andere kwestie.
→ It is a different issue. (!)

Il s' agit d' une autre question.
→ It is a question of another question. :)

Érdemes a leggyakoribbakat külön (kézi?) szabállyal kezelni?

Összefoglalás

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

Összefoglalás

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

Köszönöm a figyelmet!