28 millió szintaktikailag elemzett mondat és 500000 igei szerkezet

Sass Bálint
sass.balint@nytud.mta.hu

MTA Nyelvtudományi Intézet

NLP Meetup
Budapest, 2015. április 29.

„Magyar igei szerkezetekről fogok beszélni.”

„Magyar igei szerkezetekről fogok beszélni.”

„Nagyon örülök, hogy részt vehetek ezen a meetup-on.”

„Magyar igei szerkezetekről fogok beszélni.”

„Nagyon örülök, hogy részt vehetek ezen a meetup-on.”

„Köszönöm a figyelmet!”

„Magyar igei szerkezetekről fogok beszélni.”
→ **beszél -rÓl**

„Nagyon örülök, hogy részt vehetek ezen a meetup-on.”

„Köszönöm a figyelmet!”

„Magyar igei szerkezetekről fogok beszélni.”
→ **beszél -rÓl**

„Nagyon örülök, hogy részt vehetek ezen a meetup-on.”
→ **vesz rész-t -n**

„Köszönöm a figyelmet!”

„Magyar igei szerkezetekről fogok beszélni.”
→ **beszél -rÓl**

„Nagyon örülök, hogy részt vehetek ezen a meetup-on.”
→ **vesz rész-t -n**

„Köszönöm a figyelmet!”
→ **köszön figyelem-t**

„Magyar igei szerkezetekről fogok beszélni.”
→ **beszél -rÓl**

„Nagyon örülök, hogy részt vehetek ezen a meetup-on.”
→ **vesz rész-t -n**

„Köszönöm a figyelmet!”
→ **köszön figyelem-t**

„Magyar igei szerkezetekről fogok beszélni.”

→ **beszél -rÓl**

lehetőség, terv, jövő, téma, eredmény, kultúra, kudarc, lényeg, konkrétum

„Nagyon örülök, hogy részt vehetek ezen a meetup-on.”

→ **vesz rész-t -n**

„Köszönöm a figyelmet!”

→ **köszön figyelem-t**

„Magyar igei szerkezetekről fogok beszélni.”

→ **beszél -rÓl**

lehetőség, terv, jövő, téma, eredmény, kultúra, kudarc, lényeg, konkrétum

„Nagyon örülök, hogy részt vehetek ezen a meetup-on.”

→ **vesz rész-t -n**

„Köszönöm a figyelmet!”

→ **köszön figyelem-t**

„Magyar igei szerkezetekről fogok beszélni.”

→ **beszél -rÓl**

lehetőség, terv, jövő, téma, eredmény, kultúra, kudarc, lényeg, konkrétum

„Nagyon örülök, hogy részt vehetek ezen a meetup-on.”

→ **vesz rész-t -n**

konferencia, ünnepség, megbeszélés, tüntetés, olimpia, szertartás, fesztivál

„Köszönöm a figyelmet!”

→ **köszön figyelem-t**

„Magyar igei szerkezetekről fogok beszélni.”

→ **beszél -rÓl** szabad hely (vonzat)

lehetőség, terv, jövő, téma, eredmény, kultúra, kudarc, lényeg, konkrétum

„Nagyon örülök, hogy részt vehetek ezen a meetup-on.”

→ **vesz rész-t -n** kombináció

konferencia, ünnepség, megbeszélés, tüntetés, olimpia, szertartás, fesztivál

„Köszönöm a figyelmet!”

→ **köszön figyelem-t** kötött elem (szókapcsolat)

„Magyar igei szerkezetekről fogok beszélni.”

→ **beszél -rÓl** szabad hely (vonzat)

lehetőség, terv, jövő, téma, eredmény, kultúra, kudarc, lényeg, konkrétum

„Nagyon örülök, hogy részt vehetek ezen a meetup-on.”

→ **vesz rész-t -n** kombináció

konferencia, ünnepség, megbeszélés, tüntetés, olimpia, szertartás, fesztivál

„Köszönöm a figyelmet!”

→ **köszön figyelem-t** kötött elem (szókapcsolat)

Ilyen szerkezetekből van egy nagy gyűjtemény:

Igeiszerkezet-lista: 535000 db igei szerkezet

csap

becsap -t	1248
lecsap -rA	620
mér csapás-t -rA	360
megcsappan	217
lesz csapadék	205
csap -t hón-A=alá	80
becsap ajtó-t maga=mögött	28

száll

száll -rA	610	száll =mellett sík-rA	94
száll	463	száll vonat-rA	80
száll vita-bA -vAl	359	száll maga-A-bA	72
száll -bA	292	száll -n	71
száll -ért sík-rA	150	száll sík-rA	69
száll -ért harc-bA	142	száll -bA -vAl	67
száll -bAn	141	száll -ért ring-bA	65
száll -vAl	134	száll part-rA	64
száll ring-bA	103	száll harc-bA	63
száll fej-A-bA	101	száll -rÓl -rA	61

tej

iszik tej-t	93
aprít -t tej-bA	50
van tej-A	31
felenged -t tej-vAl	21
fej tej-t	21
szív -t maga-A-bA anyatej-vAl	19
felforral tej-t	17
önt tej-t -bA	16
ad tej-t tehén	16
kifut tej	13
összekever -t tej-vAl	12
fürészt -t tej–vaj-bAn	12
ver tojás-t tej-vAl habos-rA	11

Hogy állt elő az igeiszerkezet-lista?

Automatikusan nyertük ki az igei szerkezeteket tagmondatokra bontott, elemzett korpuszból egy speciális algoritmus segítségével.

Gyakorisági alapon: a ritka („lényegtelen”) bővítményeket elhagyja.

Probléma: *szemére vet mit* $\leftrightarrow$ *pillantást vet vmire*

	-rA	-t
vet	szem	?
vet	?	pillantás

Ezt megoldja: felfedezi a lexikálisan kötött/szabad bővítményeket.

Hogy állt elő az igeiszerkezet-lista?

Automatikusan nyertük ki az igei szerkezeteket tagmondatokra bontott, elemzett korpuszból egy speciális algoritmus segítségével.

Gyakorisági alapon: a ritka („lényegtelen”) bővítményeket elhagyja.

Probléma: *szemére vet mit* $\leftrightarrow$ *pillantást vet vmire*

	-rA	-t
vet	szem	?
vet	?	pillantás

Ezt megoldja: felfedezi a lexikálisan kötött/szabad bővítményeket.

Hogy állt elő az igeiszerkezet-lista?

Automatikusan nyertük ki az igei szerkezeteket tagmondatokra bontott, elemzett korpuszból egy speciális algoritmus segítségével.

Gyakorisági alapon: a ritka („lényegtelen”) bővítményeket elhagyja.

Probléma: *szemére vet mit* $\leftrightarrow$ *pillantást vet vmire*

	-rA	-t
vet	szem	?
vet	?	pillantás

Ezt megoldja: felfedezi a lexikálisan kötött/szabad bővítményeket.

Korpusz? MNSZ1, 187 millió szövegszó.
Elemzett? Bejelölve az ige + a főnévi csoport bővítmények.

engem meg sem hallgattak .

`stem@@meghallgat ACC@@én`

A hasmenéstől szenvedő betegeknek sokat kell inniuk ,

`stem@@iszik ACC@@sok NOM@@beteg`

Az egyik támadójátékos elhúzta mellettem a labdát ,

`stem@@elhúz ACC@@labda mellett@@én NOM@@támadójátékos`

A másik erőforrás:

Mazsola adatbázis: 28 millió elemzett tagmondat

kereső hozzá: `http://corpus.nytud.hu/mazsola`

Mire lehet ez jó?

1. szabad hely – *szemantikailag koherens szóosztály!*
   eszik -t – ételek; **eltörik alany** – testrészek
   $\rightarrow$ ontológia csomópont – vonzati hely összerendelés

Mire lehet ez jó?

1. szabad hely – *szemantikailag koherens szóosztály!*
   eszik -t – ételek; **eltörik alany** – testrészek
   $\rightarrow$ ontológia csomópont – vonzati hely összerendelés
2. „kakukktojás”
   eszik -t – kása; **eltörik alany** – mécses
   $\rightarrow$ idiómák, szólások azonosítása

Mire lehet ez jó?

1. szabad hely – *szemantikailag koherens szóosztály!*
   eszik -t – ételek; **eltörik alany** – testrészek
   $\rightarrow$ ontológia csomópont – vonzati hely összerendelés
2. „kakukktojás”
   eszik -t – kása; **eltörik alany** – mécses
   $\rightarrow$ idiómák, szólások azonosítása
3. vonzat kötelezőségének vizsgálata
   -t $\gg$ **-t -rA**: *felszólít*, *tanít*
   -t $\ll$ **-t -rA**: *bíz*, *alapoz*

Mire lehet ez jó?

1. szabad hely – *szemantikailag koherens szóosztály!*
   eszik -t – ételek; **eltörik alany** – testrészek
   $\rightarrow$ ontológia csomópont – vonzati hely összerendelés
2. „kakukktojás”
   eszik -t – kása; **eltörik alany** – mécses
   $\rightarrow$ idiómák, szólások azonosítása
3. vonzat kötelezőségének vizsgálata
   -t $\gg$ **-t -rA**: *felszólít*, *tanít*
   -t $\ll$ **-t -rA**: *bíz*, *alapoz*
4. mondatelemző döntésének támogatása – *koordináció*

   „Ráütöttem a pecsétet és az oklevelet átadtam.”

Mire lehet ez jó?

1. szabad hely – *szemantikailag koherens szóosztály!*
   eszik -t – ételek; **eltörik alany** – testrészek
   $\rightarrow$ ontológia csomópont – vonzati hely összerendelés
2. „kakukktojás”
   eszik -t – kása; **eltörik alany** – mécses
   $\rightarrow$ idiómák, szólások azonosítása
3. vonzat kötelezőségének vizsgálata
   -t $\gg$ **-t -rA**: *felszólít*, *tanít*
   -t $\ll$ **-t -rA**: *bíz*, *alapoz*
4. mondatelemző döntésének támogatása – *koordináció*

   „Ráütöttem (a pecsétet és az oklevelet) átadtam.”

Mire lehet ez jó?

1. szabad hely – *szemantikailag koherens szóosztály!*
   eszik -t – ételek; **eltörik alany** – testrészek
   $\rightarrow$ ontológia csomópont – vonzati hely összerendelés
2. „kakukktojás”
   eszik -t – kása; **eltörik alany** – mécses
   $\rightarrow$ idiómák, szólások azonosítása
3. vonzat kötelezőségének vizsgálata
   -t $\gg$ **-t -rA**: *felszólít*, *tanít*
   -t $\ll$ **-t -rA**: *bíz*, *alapoz*
4. mondatelemző döntésének támogatása – *koordináció*

   „(Ráütöttem a pecsétet) és (az oklevelet átadtam).”

Mire lehet ez jó?

1. szabad hely – *szemantikailag koherens szóosztály!*
   eszik -t – ételek; **eltörik alany** – testrészek
   $\rightarrow$ ontológia csomópont – vonzati hely összerendelés
2. „kakukktojás”
   eszik -t – kása; **eltörik alany** – mécses
   $\rightarrow$ idiómák, szólások azonosítása
3. vonzat kötelezőségének vizsgálata
   -t $\gg$ **-t -rA**: *felszólít*, *tanít*
   -t $\ll$ **-t -rA**: *bíz*, *alapoz*
4. mondatelemző döntésének támogatása – *koordináció*

„(Ráütöttem a pecsétet) és (az oklevelet átadtam).”

	pecsét-t	oklevél-t
átad	1	66
ráüt	66	0

Összefoglalás

Két erőforrás:

1. Mazsola adatbázis = 28 millió elemzett tagmondat
   az MNSZ1 tagmondatainak sekély szintaktikai elemzéssel ellátott változata, mely a *Mazsola* lekérdező adatbázisaként szolgál
2. Igeiszerkezet-lista = 535000 igei szerkezet
   a Mazsola adatbázisból automatikusan származtatott igeiszerkezet-lista, melyből a *Magyar igei szerkezetek* szótár is született

Elérhető:

`http://corpus.nytud.hu/isz`

Kutatási célra szabadon, üzleti célra egyedi megállapodás keretében.

Összefoglalás

Két erőforrás:

1. Mazsola adatbázis = 28 millió elemzett tagmondat
   az MNSZ1 tagmondatainak sekély szintaktikai elemzéssel ellátott változata, mely a *Mazsola* lekérdező adatbázisaként szolgál
2. Igeiszerkezet-lista = 535000 igei szerkezet
   a Mazsola adatbázisból automatikusan származtatott igeiszerkezet-lista, melyből a *Magyar igei szerkezetek* szótár is született

Elérhető:

`http://corpus.nytud.hu/isz`

Kutatási célra szabadon, üzleti célra egyedi megállapodás keretében.

köszön figyelem-t