Magyar szerkezettár

bemutató
2024. november 21.

Sass Bálint

Magyar szerkezettár projekt – OTKA K 147452

HUN-REN Nyelvtudományi Kutatóközpont
Lexikológiai Intézet
sass.balint@nytud.hu

bevezető

- koncepció, lényeg, hol tartunk – mit tud, hogyan használható
- *kiindulópont:* szavak – szerkezetek, szótár – szerkezettár
- az előadás felépítése: *„fordítva”*
  (1) bevezető
  (2) demó sok példával
  (3) magyarázat és háttér
- a **Magyar szerkezettár** egy szótárszerű online eszköz
  szövegdoboz ← magyar szó vagy rövid szövegrészlet
  működése: a rendszer „kiszedi” belőle
  a szerkezeteket, konstrukciókat és bemutatja a felhasználónak
- első verzió (work in progress) – próbálják ki!
- *kísérlet:* egy szótár anyagából kiindulva – minden inputra válasz

demó

https://ccn-dev.nytud.hu

(szerkezettár = constructicon, ccn)

magyarázat és háttér

- *asztal* – eredmény: a szótári szócikk

- *asztalos* – szintén, külön címszóként, önálló jelentéssel,
  ← bár két elem: *asztal* + *-s*

- *asztalokra* – nem szerepel a szerkezettárban
  → a rendszer megállapítja, milyen **konstrukciókból**
  (önálló jelentéssel bíró egységekből) áll

  háttér: automatikus **morfológiai elemzés** bontja részekre a szót

magyarázat és háttér

konstrukció $\stackrel{\text{def}}{=}$ **forma–jelentés (vagy forma–funkció) pár, ami nem kikövetkeztethető, hanem megtanulandó**

– szavak, toldalékok

– fix + szabad elemekkel bíró egységek (*'részt vesz valamiben'*)

– absztrakt nyelvtani szerkezetek (ige + tárgy)

asztalos = 1 konstrukció, mivel önálló jelentése van

asztalokra = 3 konstrukció: *asztal* + *-k* + *-ra* kombinációja

magyarázat és háttér

- *faasztal* – nem szerepel, összetett szó, 2 konstrukció (vö. *asztalokra*)

- *asztalfiókba* – nem szerepel

  háttér: morfológiai elemzés + **összevonás**
  asztalfiókba → *asztal* + *fiók* + *-ba* → *asztalfiók* + *-ba*
  ← mert az *asztalfiók* szerepel önállóan a szerkezettárban

  3 elem, 2 konstrukció

magyarázat és háttér

- *fehér asztal* – többszavas

  teljes jogú, önálló egységként szerepel a szerkezettárban
  ← mivel önálló jelentése van

  a kiinduló szótárban a *fehér* alatt szerepelt, innen emeltük ki

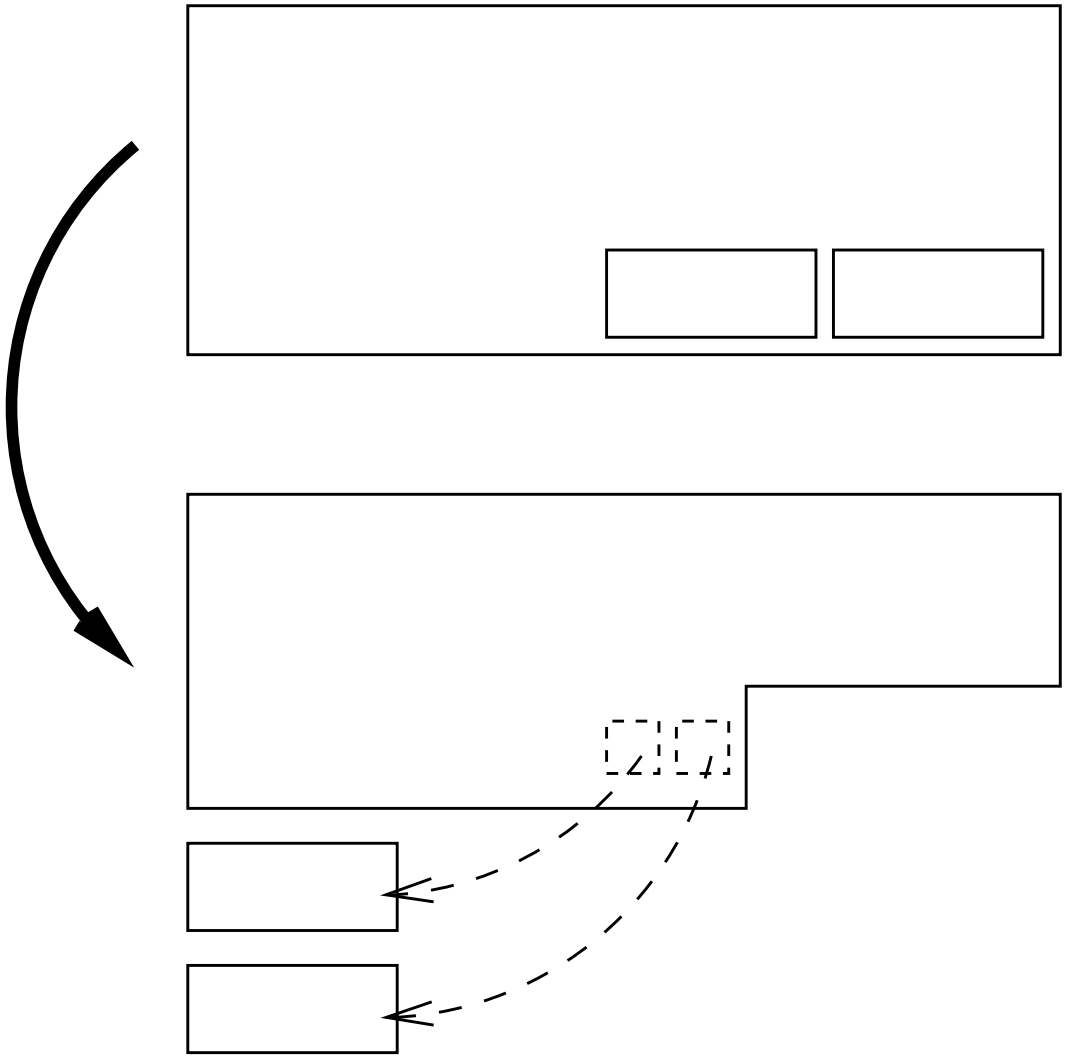
  megközelítésünk: **kiinduló szótár** ↷ szerkezettár
  Értelmező kéziszótár ↷ Magyar szerkezettár

  – **kiemeljük** a konstrukciókat (XML-részfákat) a szócikkekből
  – önálló szócikket készítünk a konstrukciónak (definícióval)
  – visszalinkeljük az eredeti szócikkhez

  Ami a szótárban kifejezésként, szókapcsolatként szerepel,
  azt elfogadjuk konstrukciónak.

  Fókusz a működésen. Jelenleg csak ÉKSz., további szótárak integrálhatók.

szerkezetek kiemelése avagy „lifting”



magyarázat és háttér

- *fehér asztal* – van, 1 konstrukció
- *sárga asztal* – nem szerepel, nincs önálló jelentése, 2 konstrukció
  két, formailag azonos, de konstrukciók szempontjából más példa

- *fehér asztal mellett* – nem szerepel, összevonás (vö. *asztalfiókba*)

  háttér: **tokenizálás** + morfológiai elemzés + ez is összevonás
  – csak most nem morfémák, hanem **szavak szintjén!**
  fehér asztal mellett → *fehér* + *asztal* + *mellett*
  → *'fehér asztal'* + mellett

  3 elem, 2 konstrukció

a „szavankénti” keresés algoritmusa

szétbont + összevon

1. a lekérdezés egy konstrukció? → ✓	*‘asztal’, ‘asztalos’, ‘fehér asztal’*
2. tokenizálás	*‘sárga asztal’*
3. szavak összevonása → ✓	*‘fehér asztal mellett’*
4. a megmaradó szavakra …	
5. morfológiai elemzés	*‘asztalokra’, ‘faasztal’*
6. morfémák összevonása → ✓	*‘asztalfiókba’*
7. megmaradó morfémák → ✓	

- az input **lebontása** nyelvi elemekre: 2. & 5.
- a releváns konstrukciók **összeépítése** az elemekből: 3. & 6.

magyarázat és háttér

szótár és szerkezettár működésének összevetése

- *avokádó*: **van**/van – és azonos!
- *fűbe harap*: nincs/van – 1 konstrukció (elrejtve ott van)
- *fűbe*: nincs/van – 2 konstrukció kombinációja: *fű* + *-be*
- *fű*: **van**/van – konstrukciók kiemelve önálló szócikké
- *aki fűbe harap*: nincs/van – szöveg részeként is felismeri
- *nem harapott fűbe*: nincs/van – **ragozva, más szórenddel** is felismeri
- *füvet harap*: nincs/van – 3 konstrukció

az online szótári felület általánosítása

1. **online szótár felülete:**
   (1) egyszavas input (kanonikus alak),
   (2) 1 db címszónak tekinti (a címszó „azonosítása” triviális),
   (3) kikeresi.
2. **szerkezettár felülete:**
   (1) tetszőleges szöveges input (~~kanonikus alak~~),
   (2) az összes konstrukció azonosítása (elemzési lépések),
   (3) az összes konstrukció kikeresése.

általánosítás:
(1) több címszó (vö. *sárga asztal*);
(2) konstrukciók: többszavas input → összefonódott konstrukciók

cél: **mindenki számára hozzáférhető legyen:**
nem várjuk el a konstrukciók valamiféle kanonikus alakjának ismeretét.

szótár helyett szerkezettár

Minden szó egyben szerkezet is
$\rightarrow$ a szótár a szerkezettár részhalmaza.

A szerkezettár mindent tud, amit a szótár – sőt többet.

A szerkezettár egy továbbfejlesztett szótár.

„Mindent” magában foglal, mert „minden” szerkezet.

Szótár $\rightarrow$ helyett $\rightarrow$ szerkezettár!

szótár

szerkezettár

magyarázat és háttér

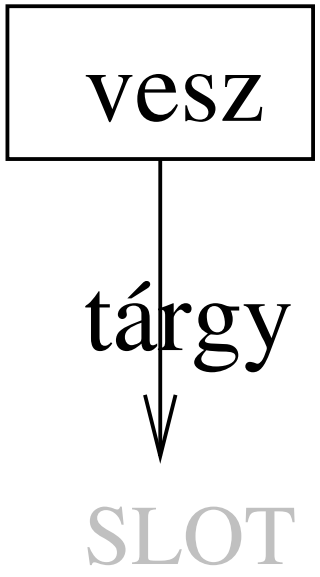
- *nem harapott fűbe*
  – felismeri a ragozott, más szórendű formát
- *a tanár részt vesz az akcióban*
  – a *'részt vesz valamiben'* igés konstrukciót ismeri fel a rendszer

  Ez az elemek egymásutániságára épülő eljárással nem megy!
  Mert a konstrukciók sokszor
  alakilag nem fixek, nem rögzített szórendűek és nem folytonosak.

  – *fűbe harap* vs *harapott fűbe*
  – *a tanár részt vett az akcióban*
  – *a tanár vesz részt az akcióban*
  – *a tanár vesz az akcióban részt*

  → ennek kezeléséhez **más megközelítés** szükséges ...

reprezentáció: filler-slot fák



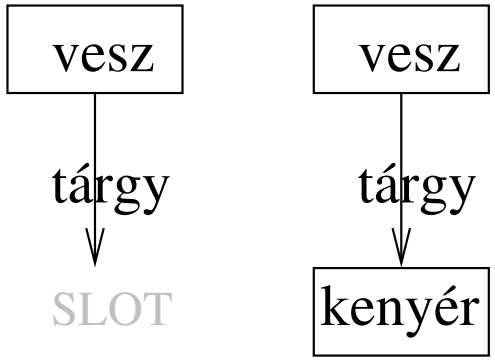
a konstrukciók felfoghatók így

gráf-reprezentáció

a függőségi fák nagyjából ilyenek ← ezeket fogjuk használni

szükséges lépés: **függőségi elemzés**

a reprezentáció alkalmazása



1. konstrukciók reprezentálására – *szabad helyekkel*
   → így tároljuk őket a szerkezettárban
   valami kezelése
2. szöveg reprezentálására – *minden hely kitöltött*
   → az inputot menet közben elemezzük

→ konstrukciók azonosítása az input szövegben:
a szerkezettárban lévő konstrukciók illesztése az input reprezentációjára

a TF (tree fragments) algoritmus működése

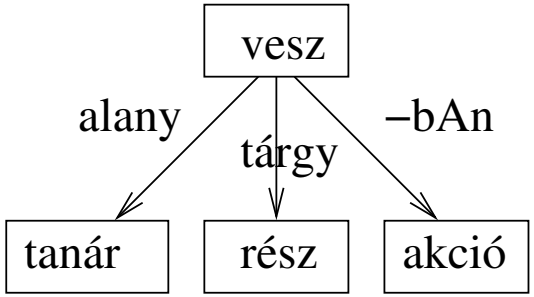
1. példa

bemenet: *a tanár részt vesz az akcióban*

a TF algoritmus működése

1. példa

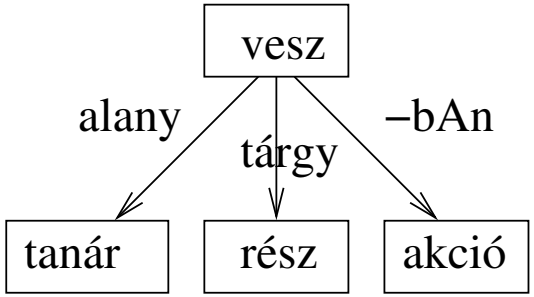
bemenet: *a tanár részt vesz az akcióban*



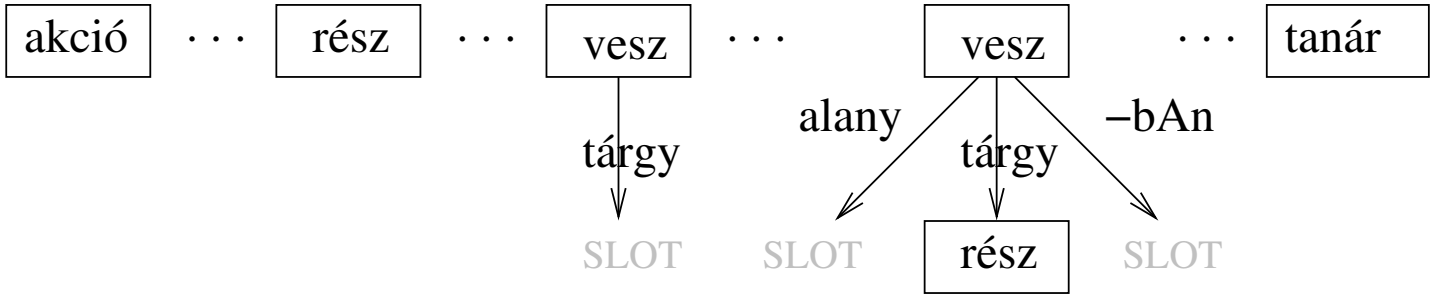
a TF algoritmus működése

1. példa

bemenet



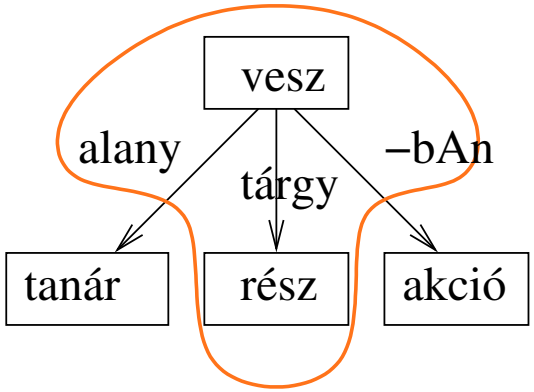
konstrukciók a szerkezettárban



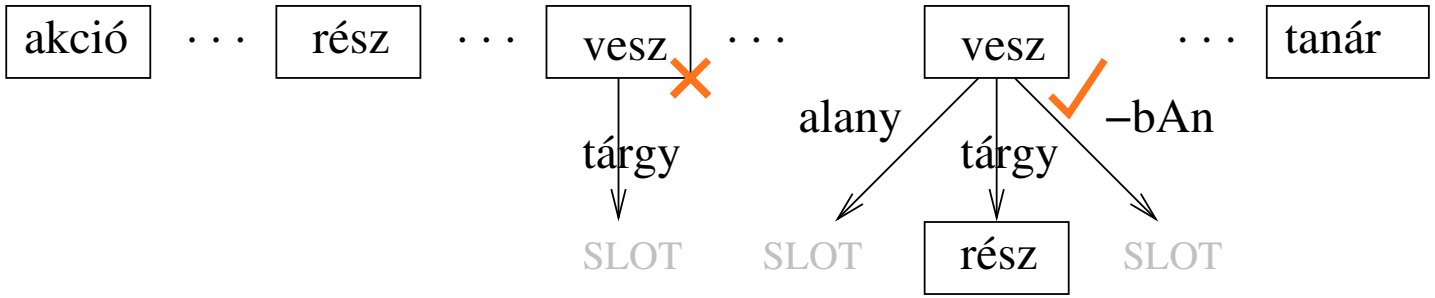
a TF algoritmus működése

1. példa

bemenet



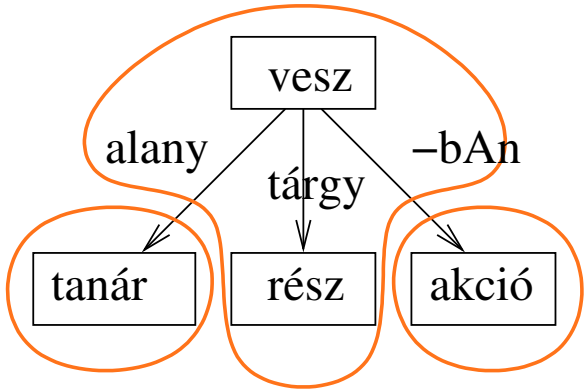
konstrukciók a szerkezettárban



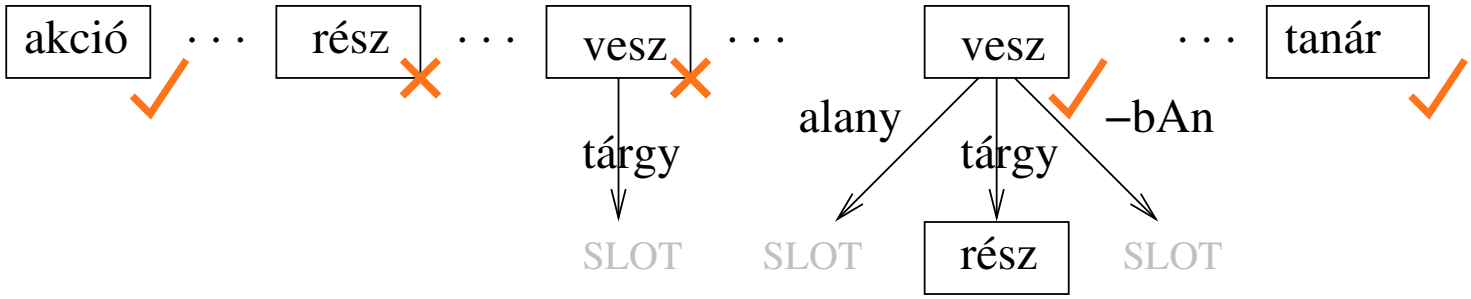
a TF algoritmus működése

1. példa

bemenet



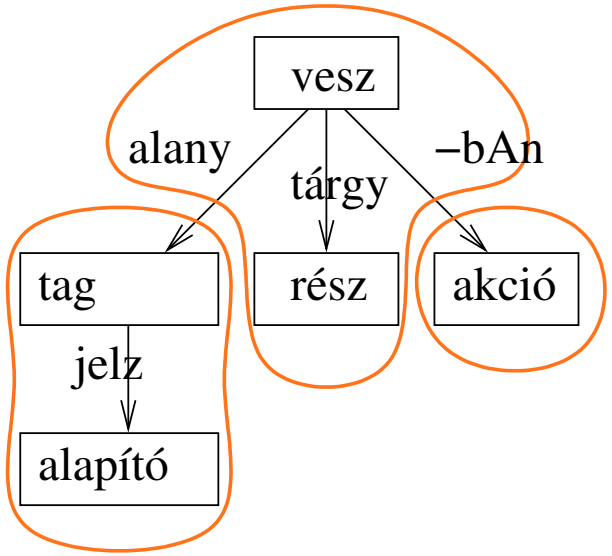
konstrukciók a szerkezettárban



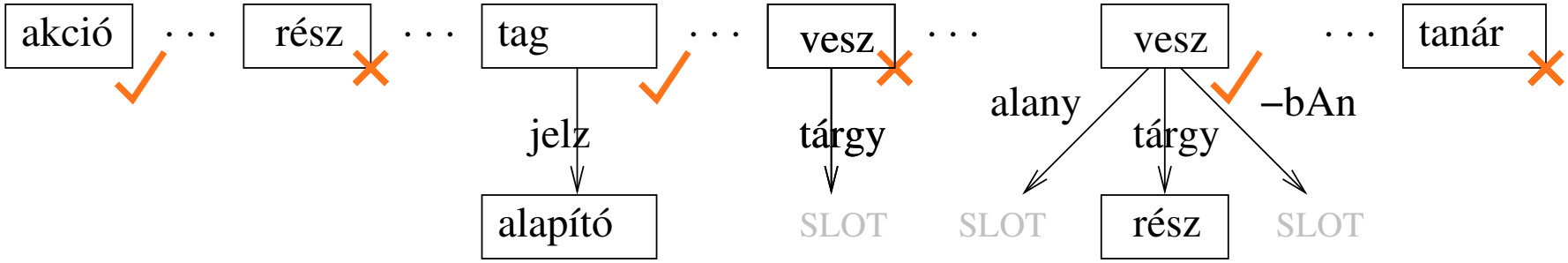
a TF algoritmus működése

4. példa

bemenet: *az alapító tag részt vesz az akcióban*



konstrukciók a szerkezettárban



a TF algoritmus

1. induljunk el a gyökér-csomópontnál;
2. keressük ki szerkezettárból az itt **illeszkedő** konstrukciókat;
3. ha több van, akkor vegyük közülük a leghosszabbat;
4. **távolítsuk el** a megtalált konstrukciót a fából; (egy csomópont kezelése)
5. így a fa kisebb fa-töredékekre esik szét;
6. vegyük ezeket, és futtassuk le az algoritmust mindegyiken.

$\rightarrow$ a beazonosított konstrukciók
az input fa-reprezentáció különböző **töredékeinek** fognak megfelelni

más megközelítés – mégis hasonló felépítés: *asztalfiókba* vs *a tanár* ...

magyarázat és háttér

- *a tanár részt vesz az akcióban* – igés konstrukció
- *a tanár kenyeret vesz az üzletben* – azonos forma
- *az alapító tag részt vesz az akcióban* – **két többszavas konstrukció**
  rekurzivitás → a fában lejjebb lévő konstrukciókat is felismeri
  konstrukciók összefonódása, „szerkezetek szövedéke”
- *az okos gyerek részt vesz az akcióban* – azonos forma
- *a tanár részt vett az akcióban* – variáció: múlt idő
- *a tanár vesz részt az akcióban* – variáció: szórend
- *a tanár vesz az akcióban részt* – variáció: szórend
- *a tanár részt vesz a munkában* – más elem
- *a tanár kollégáival részt vesz az akcióban* – plusz bővítmény
  ez nem zavarja a felismerést, az illeszkedés ugyanúgy működik

keret és végső cél

- **konstrukciós nyelvtan**
- *konstrukció* $\stackrel{\text{def}}{=}$ forma–jelentés (vagy forma–funkció) pár,
  ami nem kikövetkeztethető, hanem megtanulandó
  – szavak, toldalékok
  – fix + szabad elemekkel bíró egységek (*'részt vesz valamiben'*)
  – absztrakt nyelvtani szerkezetek (ige + tárgy)
- *szerkezettár (konstruktikon)* $\stackrel{\text{def}}{=}$ a konstrukciók összessége,
  a végső lexikai erőforrás, ami magába foglal „mindent”:
  minden jelentéssel vagy funkcióval bíró nyelvi egységet
- **„Mindezeket magába foglalva egy szerkezettár a nyelv egészéről tud képet adni, egyesítve a szótárt és a nyelvtant.”**
  Lehetetlen? Próbáljuk meg! :)

hozzáférés

https://szerkezettar.hu

Szabadon hozzáférhető.

Első verzió, alapműködés.

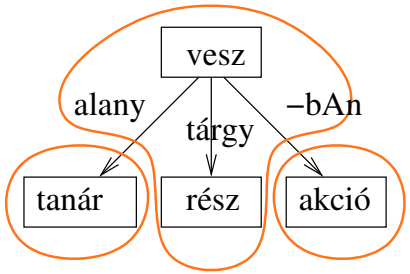
Próbálják ki, használják, jelezzenek vissza, javasoljanak!

Könnyen hivatkozható URL-ek:
https://szerkezettar.hu/fű

Kontakt:
szerkezettar@nytud.hu

összegzés

1. szerkezet (konstrukció), szerkezettár	demó, magyarázat, háttér
2. szótár $\curvearrowright$ szerkezettár	„lifting”
3. első algoritmus: „szavankénti” keresés	lebontás-összeépítés
4. általánosított online felület	szótár helyett szerkezettár
5. összetettebb szerkezetek	$\rightarrow$ gráf-reprezentáció
6. második algoritmus: TF (tree fragments)	illesztés-eltávolítás
7. keret és végső cél	
8. **https://szerkezettar.nytud.hu**	



Bálint Sass
sass.balint@nytud.hu

köszönetnyilvánítás

Az ebben az előadásban ismertetett kutatást a Nemzeti Kutatási, Fejlesztési és Innovációs Hitaval (NKFIH) OTKA (K 147452) projektje támogatta. A K 147452 számú projekt a Kulturális és Innovációs Minisztérium Nemzeti Kutatási Fejlesztési és Innovációs Alapból nyújtott támogatásával, a K_23 pályázati program finanszírozásában valósult meg.

Magyar szerkezettár

bemutató
2024. november 21.

Sass Bálint

Magyar szerkezettár projekt – OTKA K 147452

HUN-REN Nyelvtudományi Kutatóközpont
Lexikológiai Intézet
sass.balint@nytud.hu

titkos példák

Be lehet írni bármit?
(1) csak ÉKSz, (2) abból is csak a „kifejezés”-ként kategorizált dolgok

- *átverekedte magát a bokrokon*
- *amint a mellékelt ábra mutatja, jó lesz*
- *egy ember alulról szagolná az ibolyát*
- *csak falra hányt borsó*
- *keresztbe tettek a szomszédnak*
- *Jóska kezet adott a postásnak*
- *Jóska hangot adott a véleményének*
- *a kapitány jelt ad az indulásra*
- *nem csoda, hogy érezteti a hatását*