# A duplakocka modell és az igei szerkezeteket kinyerő „ugrik és marad” módszer nyelvfüggetlensége, valamint néhány megjegyzés az UD annotáció univerzalitásáról

Sass Bálint
MTA Nyelvtudományi Intézet

2020. január 24.
MSZNY 2020, Szeged

# motiváció

A nyelv alapegységei nem a szavak, hanem a szerkezetek.
A szó csak szélső eset: olyan szerkezet, ami egy elemből áll.

:)

# *Stanisław Lem:* Hogyan maradt meg a világ?

stannum = ón

# *Stanisław Lem:* Hogyan maradt meg a világ?

stannum = ón

„– Drága öregem – jelentette ki a gép –, ha én mindent meg tudnék csinálni, ami valamilyen nyelven *s* betűvel kezdődik, akkor én lennék A Gép, Amelyik a Világon Mindent Meg Tud Csinálni, mert annyi nyelv van, hogy akármire gondolsz, valamelyik nyelven biztosan *s*-sel kezdődik a neve.”

# *Stanisław Lem:* Hogyan maradt meg a világ?

stannum = ón

„– Drága öregem – jelentette ki a gép –, ha én mindent meg tudnék csinálni, ami valamilyen nyelven *s* betűvel kezdődik, akkor én lennék A Gép, Amelyik a Világon Mindent Meg Tud Csinálni, mert annyi nyelv van, hogy akármire gondolsz, valamelyik nyelven biztosan *s*-sel kezdődik a neve.”

. . . valamelyik nyelven simán lehet, hogy **csak szerkezet van rá!**

magyar *‘egymás’* ↔ angol *‘each other’*
wolof *‘gëmm’* ↔ magyar *‘behunyja a szemét’*
magyar *‘krumpli’* ↔ francia *‘pomme de terre’*

szótár → szerkezettár

# alapok = modell & módszer

példa:

*‘Részt vesz az akcióban.’*

`vesz ACC:rész INE:akció`

# alapok = modell & módszer

példa:

*‘Részt vesz az akcióban.’*

**vesz** `ACC:rész INE:akció`

→ ige...

# alapok = modell & módszer

példa:

*'Részt vesz az akcióban.'*

`vesz` **`ACC`**`:rész` **`INE`**`:akció`

→ ige + helyek...

# alapok = modell & módszer

példa:

*'Részt vesz az akcióban.'*

`vesz ACC:rész INE:akció`

→ ige + helyek + kitöltők

# alapok = modell & módszer

példa:

*'Részt vesz az akcióban.'*

`vesz ACC:rész INE:akció`

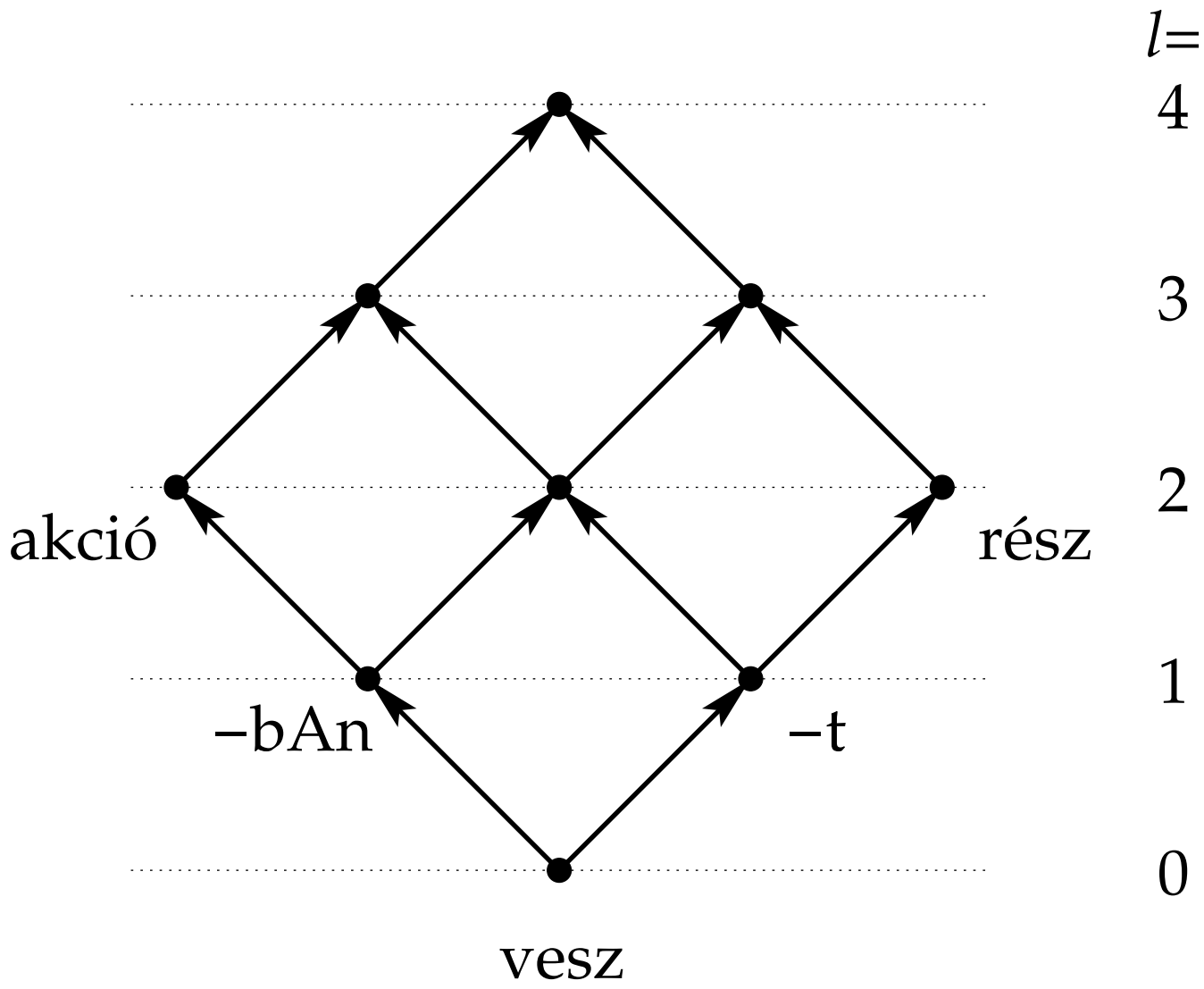
→ ige + helyek + kitöltők

cél:

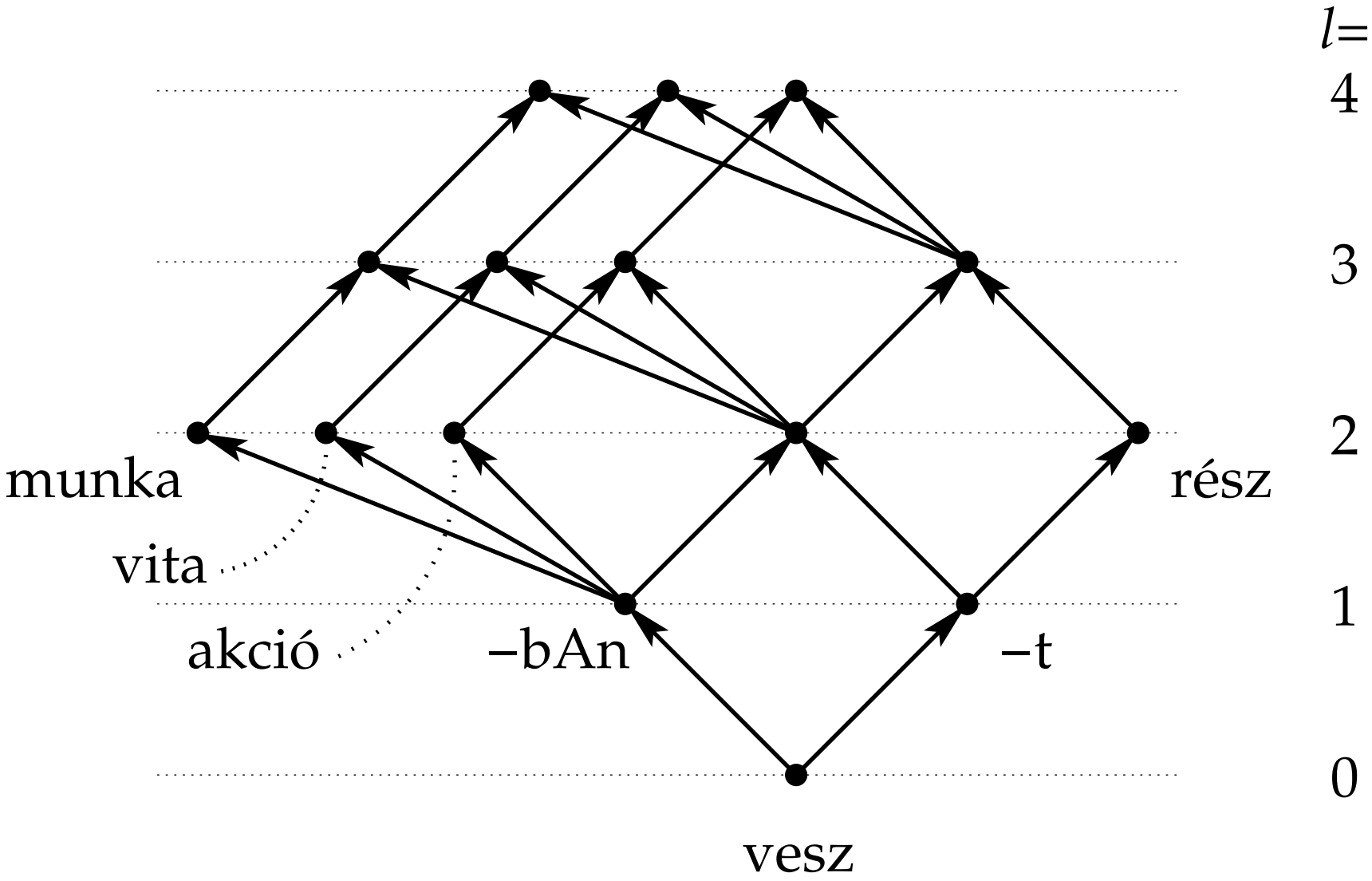
*'vesz részt vmiben'*

= teljes és tiszta **valódi igei szerkezet**ek
= lexikográfiailag hasznos igei szerkezetek

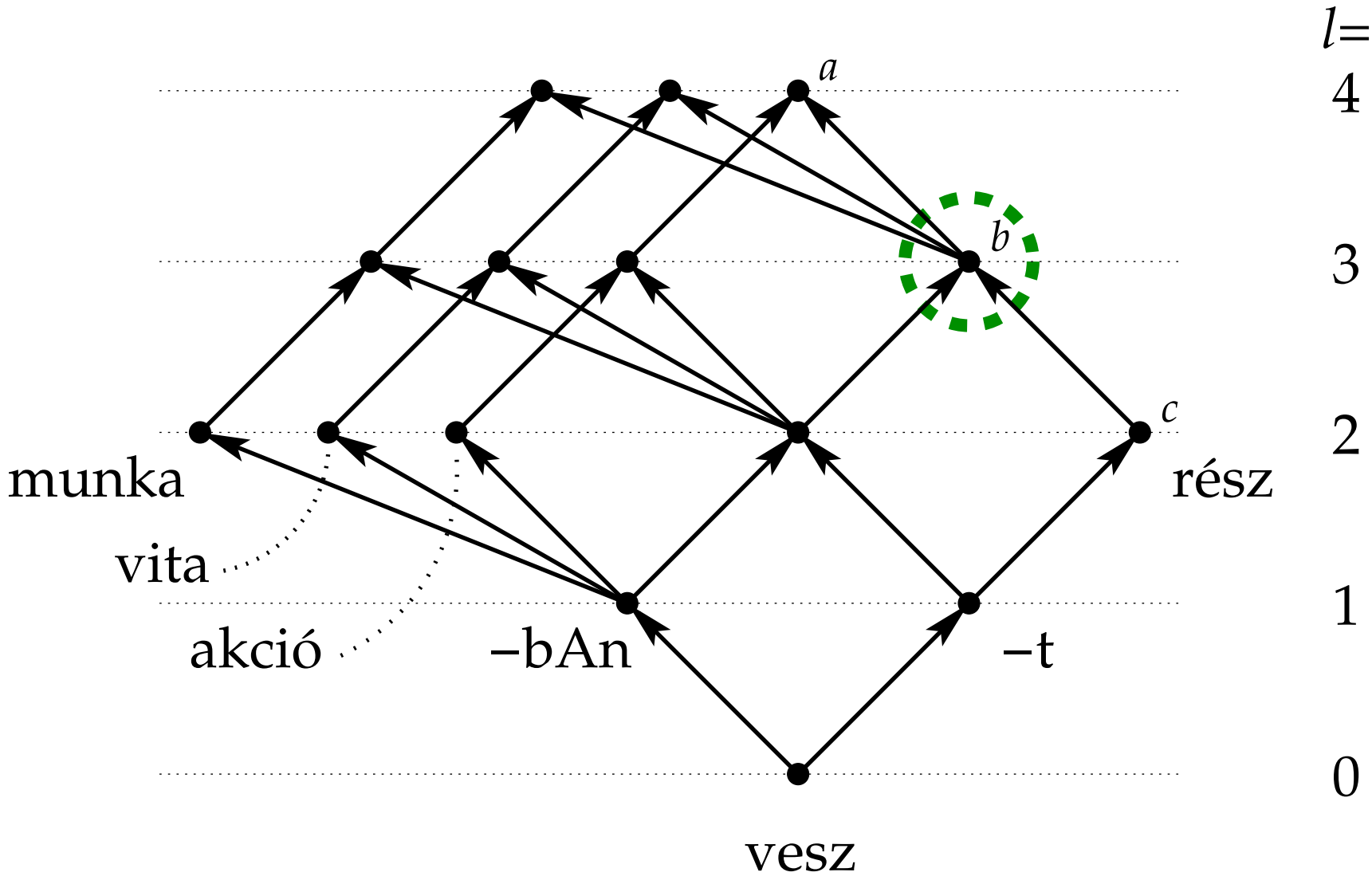
# duplakocka



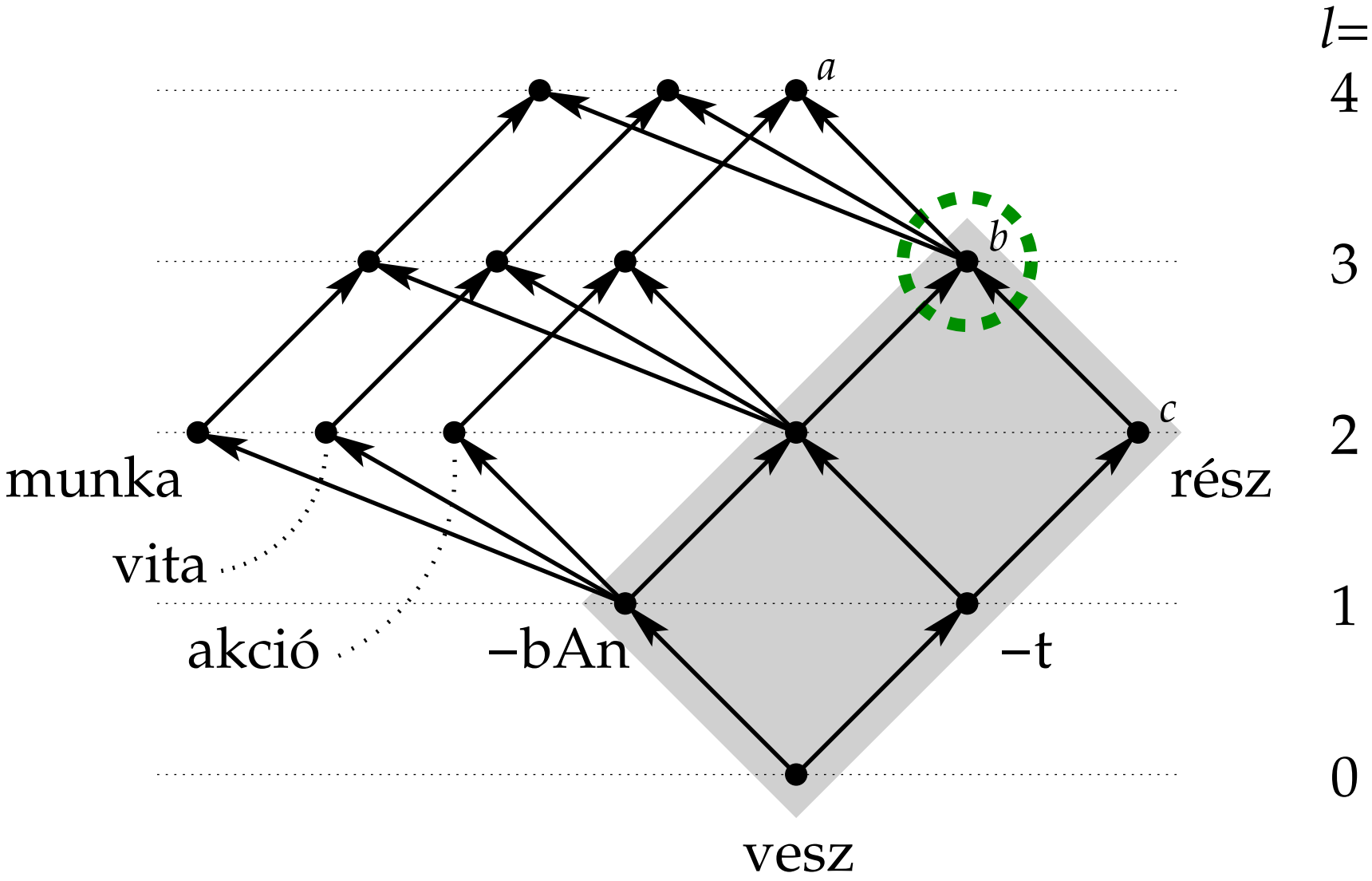
# korpuszháló



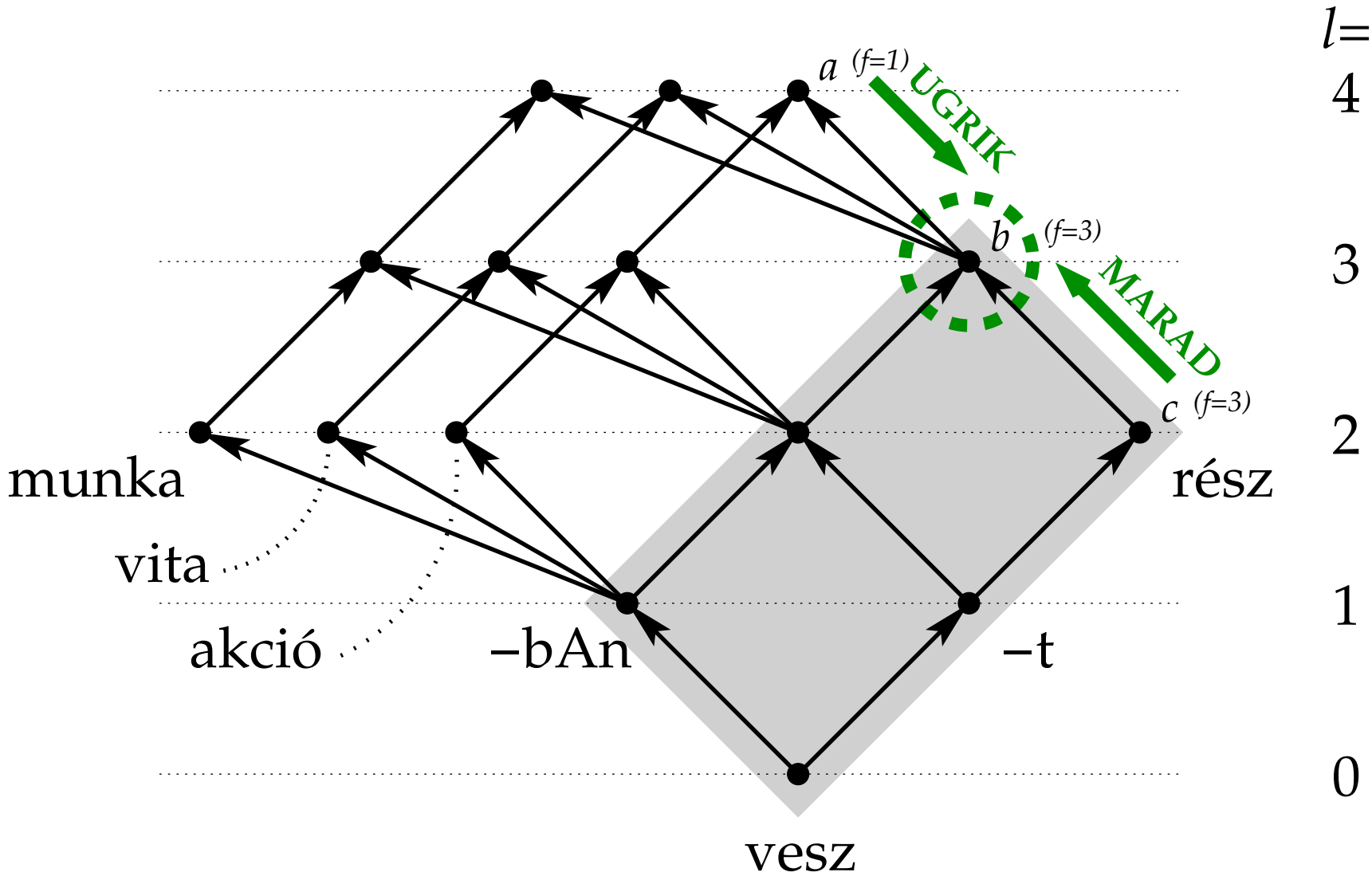
# valódi igei szerkezet

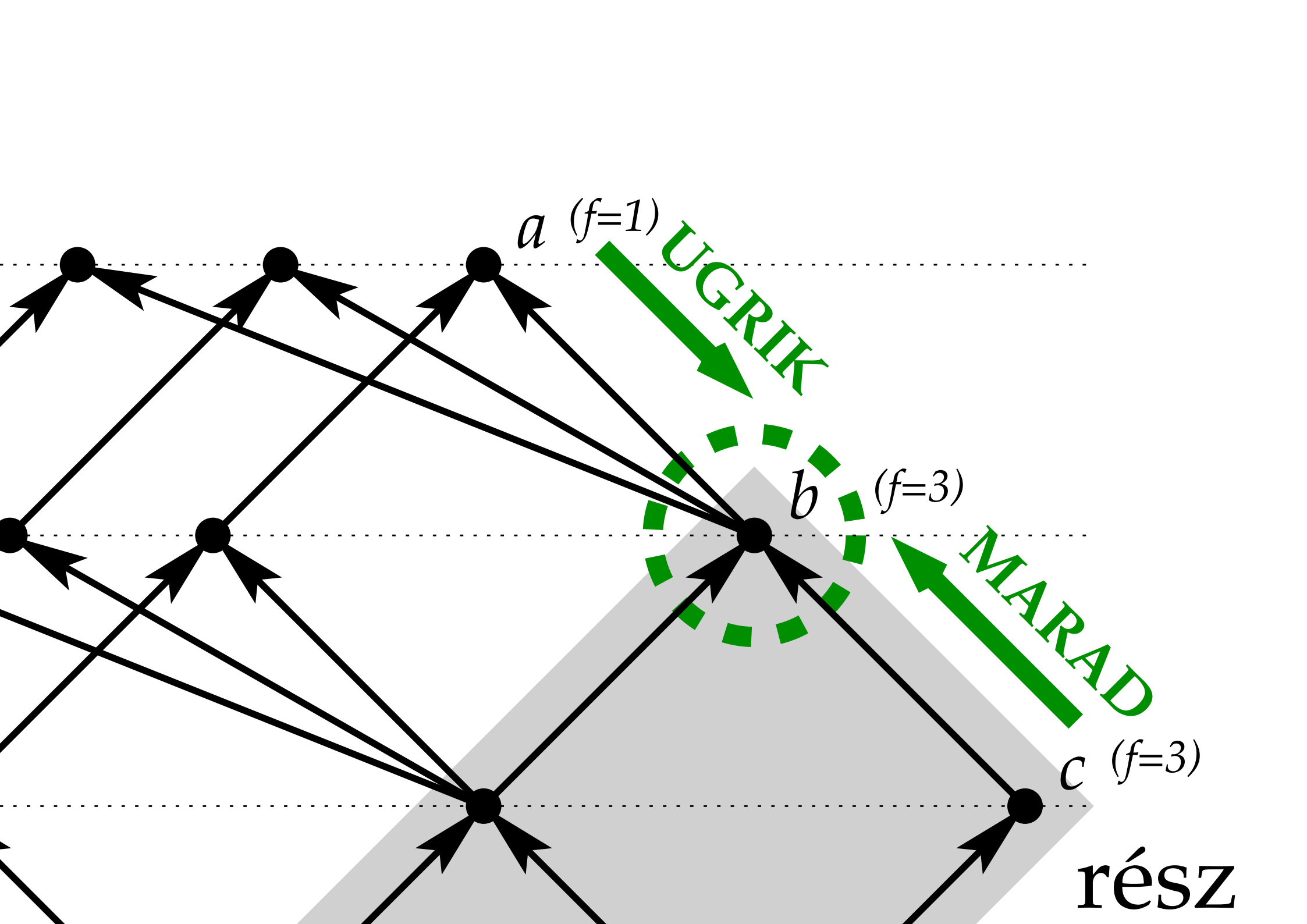


# *f* függvény



# „ugrik és marad” módszer




a
(f=1)
UGRIK
b
(f=3)
MARAD
c
(f=3)
rész

# cél

(1) a módszer tetszőleges nyelvre alkalmazható?

(2) a módszer alkalmazásához szükséges adatok
függőségileg elemzett korpuszból könnyen származtathatók?

# nyersanyag: UD korpuszok

cseh, német, angol, finn, magyar, holland, norvég, török, wolof

a munka legnagyobbrészt: a korpuszok előfeldolgozása volt

*feladat:*
az igék, valamint az ige közvetlen bővítményeit képviselő helyek és kitöltők meghatározása

tagmondat? → minden egyes igei alak gyökér
*'He didn't think he **needed** to **know** anything about South Asia.'*

# UD korpuszok: helyek kinyerése

megfelelő UD relációkkal:
`nsubj`, `obj`, `iobj`, `obl`, `case` `xcomp`;
valamint: `Case` feature

plusz: `Acc=in` ← „*in* + tárgyeset”

*gond:*

német ‘*am*’ → egyedi szabályok

angol ‘*to*’ / holland ‘*te*’ / wolof ‘*ci*’: `PART` + `mark` → önálló szófaj?

# UD korpuszok: kitöltők kinyerése

közvetlen dependens kisbetűsített lemmája

névmástörlés, kivéve:
‘*maga*’ (`Reflex=Yes`), ‘*egymás*’ (`PronType=Rcp`)

*gond:*

német ‘*sich*’ – másképp jelölve → egyedi módon

cseh ‘*navzájem*’, német ‘*einander*’, török ‘*birbiri*’
– más más kódolás → egyedi módon

‘*aufeinander*’ (!) ↔ ‘*each other*’ (!)

# UD korpuszok: igék kinyerése

igekötő hozzákapcsolása: igekötő + ige → *'upbreak'*

*gond:*

`compound:prt` ↔ **magyar** `compound:preverb`

*'stir up'* `compound:prt` ↔ *'go away'* `advmod`

**magyar** *'fel+használ'* ↔ **német** *'aufklären'*

# UD korpuszok: tanulságok

nagyon hasznos!

nem felelnek meg maradéktalanul annak, hogy
- ugyanazon jelenséget mindig ugyanúgy jelöljünk,
- eltérő jelenséget pedig mindig eltérően jelöljünk

= azaz mindig minden egységesen, ugyanúgy működjön,
amennyire csak lehet, ne kelljen nyelvfüggő lépéseket végezni

*tipp*: '*to*', '*egymás*' → legyen önálló szófaj!

*valójában*: formai függőségek → **„funkcionális"** függőségek
= az azonos *funkciójú* elemek és relációk kapnának azonos jelölést
az elvégzett átalakító lépések mind ebbe az irányba hatnak

→ ezek után: futtattuk az eredeti „ugrik és marad" módszert ...

# eredmények

| # | nyelv | igei szerkezet | magyar megfelelő |
|---|---|---|---|
| 1. | cs | **být SUBJ:rozdíl mezi** | (van különbség vmi között) |
| 2. | cs | **investovat do** | (befektet vmibe) |
| 3. | cs | čekat **se** | (várandós) |
| 4. | de | fallen SUBJ:aktie auf | (esik részvény vmire) |
| 5. | de | finden sich SUBJ:info auf | (megtalálható információ vhol) |
| 6. | de | handeln sich um | (arról van szó) |
| 7. | en | **do IOBJ OBJ:favor** | (szívességet tesz vkinek) |
| 8. | en | get in:touch with | (kapcsolatba lép vkivel) |
| 9. | en | make sure | (meggyőződik) |
| 10. | en | take OBJ:care of | (vigyáz vmkire) |
| 11. | fi | ottaa **Ill**:huomio OBJ | (figyelembe vesz vmit) |
| 12. | fi | ottaa **Ill**:käyttö OBJ | (használatba vesz vmit) |

# eredmények

| # | nyelv | igei szerkezet | magyar megfelelő |
|---|---|---|---|
| 13. | fi | ottaa Ill:käsi OBJ | (kézbe vesz vmit) |
| 14. | hu | lesz SUBJ:szükség -rA | |
| 15. | hu | tesz lehetővé -t | |
| 16. | nl | zien OBJ:kans **te** | (lát lehetőséget vmit csinálni) |
| 17. | no | få OBJ med:seg | (magával visz vmit) |
| 18. | no | få OBJ:gjennomslag i | (áttörést ér el vmiben) |
| 19. | no | få OBJ **på:seg** | (felvesz vmit (ruhafélét) magára) |
| 20. | no | få **OBJ:tillit** | (visszanyeri az önbizalmát) |
| 21. | no | få OBJ i:løp | (futtat vmit (szoftvert)) |
| 22. | no | ha OBJ **på:seg** | (vmi (ruhaféle) van rajta) |
| 23. | wo | am OBJ:kàttan **ci** | (van energiája vmit csinálni) |
| 24. | wo | wax IOBJ OBJ | (mond vkinek vmit) |

# eredmények

*mind jó:*

look → look, look good/great, look like
deal → deal with
go → go to (fn), going to (ige), go crazy

*szerkezetek kontrasztív megfeleltetése:*

| | | |
|---|---|---|
| cs | mluvit/hovořit | o |
| de | sprechen | von |
| en | talk | about |
| fi | puhua | Ela (-stA) |
| hu | beszél | Del (-rÓl) |
| nl | praten | over |
| no | snakke | om |

# eredmények

*helyek (slot-ok) kontrasztív megfeleltetése:*

| cs | de | magyar megfelelő |
|---|---|---|
| čekat Acc=na | warten Acc=auf | (vár -rA) |
| **Acc=na** | **Acc=auf** | |
| mít Dat=k:dispozice | stehen Dat=zu:Verfügung | (rendelkezésre áll) |
| patřit Dat=k | gehören Dat=zu | (tartozik vmihez) |
| **Dat=k** | **Dat=zu** | |

# irányok, ötletek, tervek

1. jobb módszer, hibák kiküszöbölése: *'få øye på løve'* $\leftrightarrow$ *'húz ACC közé'*
2. nagy (nem magyar) dep elemzett korpuszra futtatni: DEPCC
3. plusz egy szint: *'fontos szerepet játszik vmi vmiben'*
4. minden root kezelése $\rightarrow$ névszói állítmányok
5. kiértékelés … de hogyan?
6. triplakocka – pl. élő/élettelen
7. klaszterezés – *'eszik ACC:<ÉTEL>'*
8. igeiszerkezet-embedding – *'participate in'* $\sim$ *'take part in'*
9. hely-embedding (slot-embedding):
   *'få **OBJ** på:seg'* $\sim$ *'ha **OBJ** på:seg'*, *'fő **NOM**'* $\sim$ *'főz NOM **ACC**'*

…

# összefoglalás

predikátum-argumentum struktúra kell csak
$\rightarrow$ működik a modell is és az algoritmus is

*valóban:* előállítható elemzett korpuszból a szükséges bemenet

az UD korpuszok feldogozása nem volt trivi
jó lenne: „ugyanazt ugyanúgy, mást máshogy jelölni” :)

*valóban:* nyelvfüggetlen a modell és a módszer
lexikográfiailag is hasznos valódi igei szerkezeteket kapunk

*elérhető:*

`github.com/sassbalint/double-cube-jump-and-stay-multilingual`

Köszönöm a figyelmet!

`sass.balint@nytud.hu`