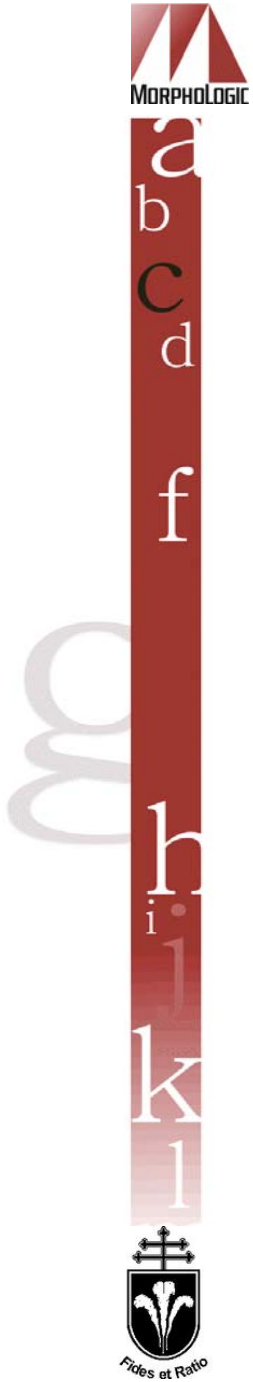
Nyelvtechnológia
4

BME, 2007. november 27.

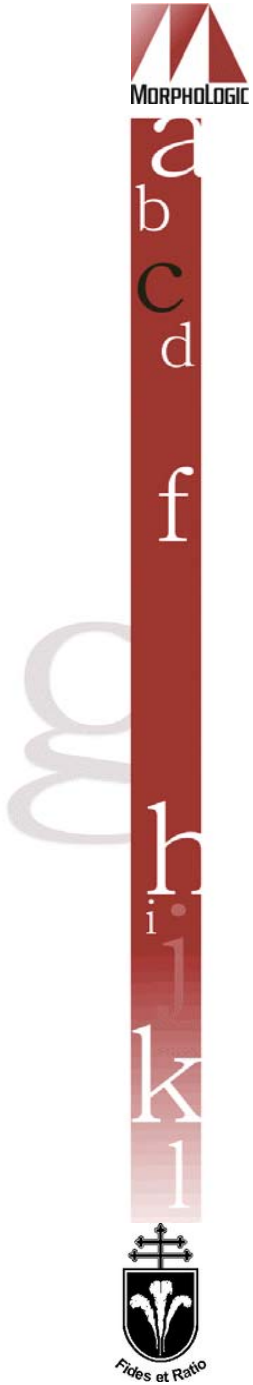


Mire elég a szintaxis?

INPUT: Jane sold a book to Bill

MEANING:

```
(S (NP  (NOUN Jane))
   (VP  (VERB  sold)
        (NP  (DET    a)
             (NOUN book))
        (PP  (PREP  to)
             (NP
                (NOUN Bill)))))
```



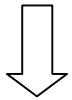
Lehet, hogy többet érne a „jelentés”?

INPUT: Jane sold a book to Bill
MEANING:

```
(sell  (agent          Jane)
       (object         a book)
       (counter-agent  Bill)
       (tense          past) )
```

Sok mondat - egy jelentés

Jane sold a book to Bill.
and
A book was sold by Jane to Bill.
and
Bill was sold a book by Jane.
and
A book was sold to Bill by Jane.



```
(sell  (agent          Jane)
       (object         a book)
       (counter-agent  Bill)
       (tense          past) )
```

Hasonló mondat - különböző jelentés

Mom baked for 3 hours

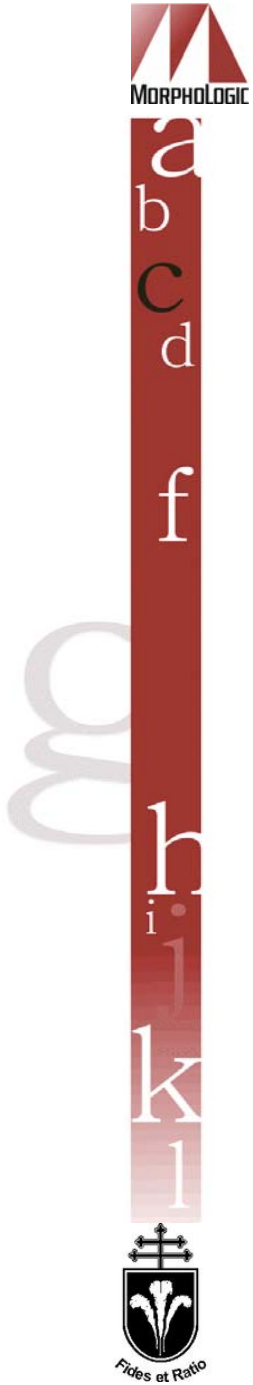
The pie baked for 3 hours

```
(baked  (agent      Mom)
        (duration   3 hours))

(baked  (object     pie)
        (duration   3 hours))
```

„Mély” esetek

agent
beneficiary
instrument
source
goal
time
duration



Híres fogalmi hálók

AZ MI kezdetén:

Quillian, Minsky, Charniak, ...

Fogalmi függőség:

Schank

Logikák:

Hendrix, Sowa (fogalmi gráfok), ...

Ontológiák:

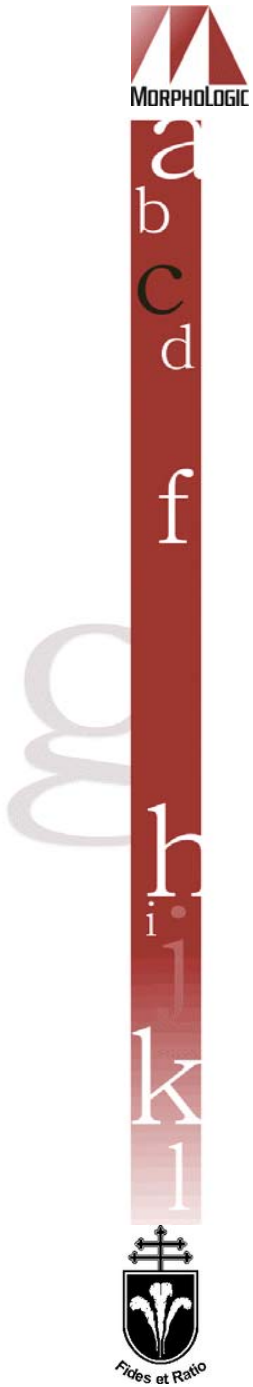
CyC, MindNet, FrameNet, ...

WordNet (pszichológusok indították):

WordNet, EuroWordNet,
eXtendedWordNet, ...

Szemantikus web:

(?)



A fogalmi függőség igeosztályai

ATRANS	abstract transfer
PTRANS	physical transfer
MTRANS	mental transfer
MBUILD	build new information
MOVE	move a body part
PROPEL	apply force to an object
GRASP	grasp an object
INGEST	take something into body
EXPEL	expel something from body
SPEAK	make a voiced sound
ATTEND	focus a sense organ
DO	unspecified act

Eseményábrázolás a FF elméletében

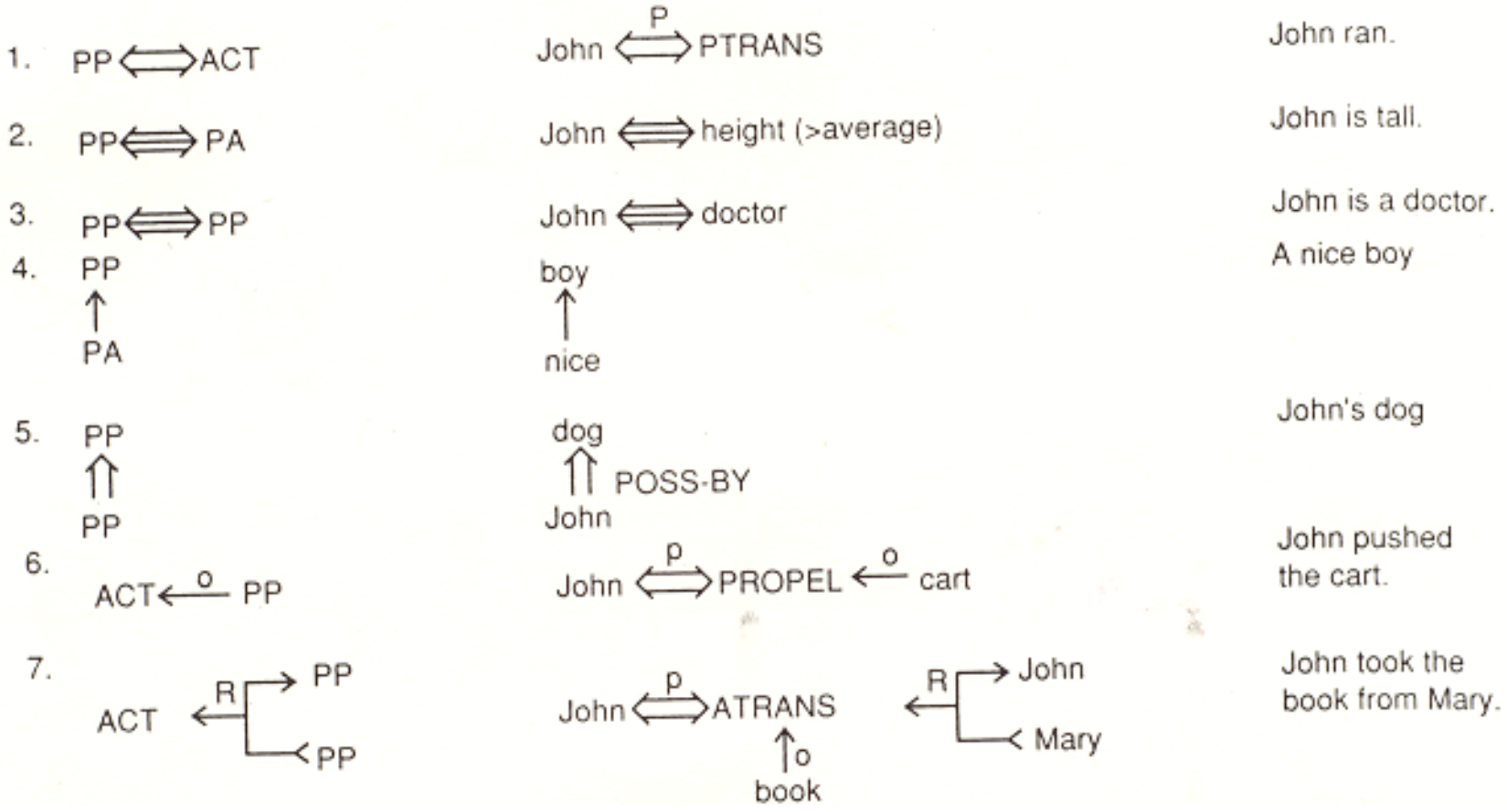
Mary took a book from John.

```
((actor      Mary)
 (action     MTRANS)
 (object     book)
 (direction  (to   Mary)
             (fromJohn)))
```

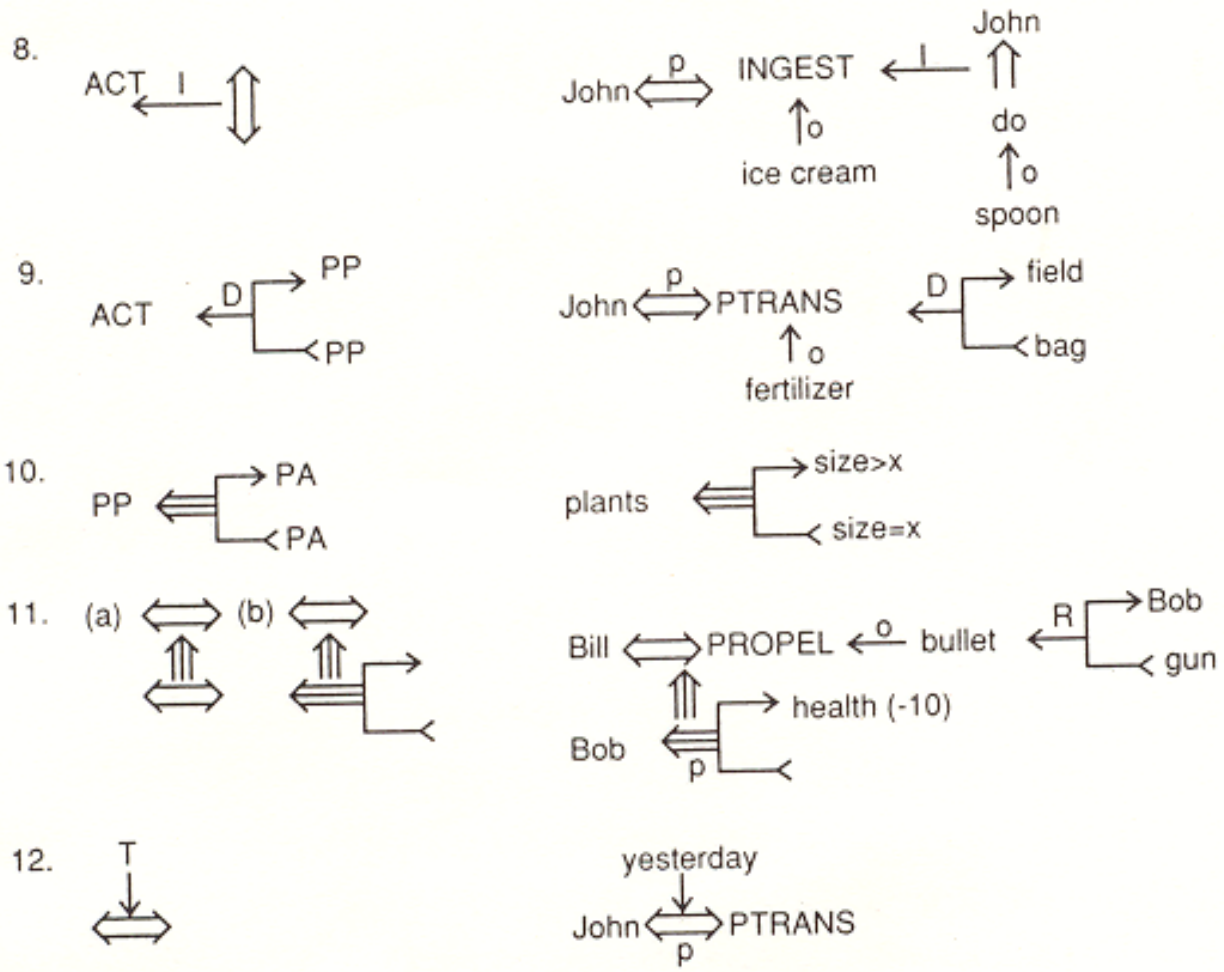
A fogalmi függőség állapotosztályai

state	*scale*	*examples*
Health	-10 to 10	-10 dead -3 sick
Fear	-10 to 0	-9 terrified -2 anxious
Mental state	-10 to 10	-7 depressed -2 sad 3 happy 7 euphoric 9 ecstatic
Hunger	-10 to 10	-8 starving -6 ravenous 5 full 8 stuffed

Schank (1)


1. PP ⟺ ACT
John —P⟺ PTRANS
John ran.
2. PP ⟺ PA
John ⟺ height (>average)
John is tall.
3. PP ⟺ PP
John ⟺ doctor
John is a doctor.
4. PP
↑
PA
boy
↑
nice
A nice boy
5. PP
⇑
PP
dog
⇑ POSS-BY
John
John's dog
6. ACT ←o— PP
John —p⟺ PROPEL ←o— cart
John pushed the cart.
7. ACT ←R— PP, PP
John —p⟺ ATRANS
↑o
book
R → John
R → Mary
John took the book from Mary.

Schank (2)


8.
ACT
I
John
p
INGEST
I
John
o
do
ice cream
o
spoon
John ate ice cream.
9.
ACT
D
PP
PP
John
p
PTRANS
D
field
o
bag
fertilizer
John fertilized the field.
10.
PP
PA
PA
plants
size>x
size=x
The plants grew.
11.
(a)
(b)
Bill
PROPEL
o
bullet
R
Bob
gun
health (-10)
Bob
p
Bill shot Bob.
12.
T
yesterday
John
PTRANS
p
John ran yesterday.

Forgatókönyvek

Jane was hungry. She decided to go to a restaurant. She ordered spaghetti and a Pepsi. The waitress brought it quickly so when she left she left her a large tip.

QUESTION:
Did Jane eat anything?

Az „étterem” forgatókönyve (a tipikus eseménysor)

entering

seating

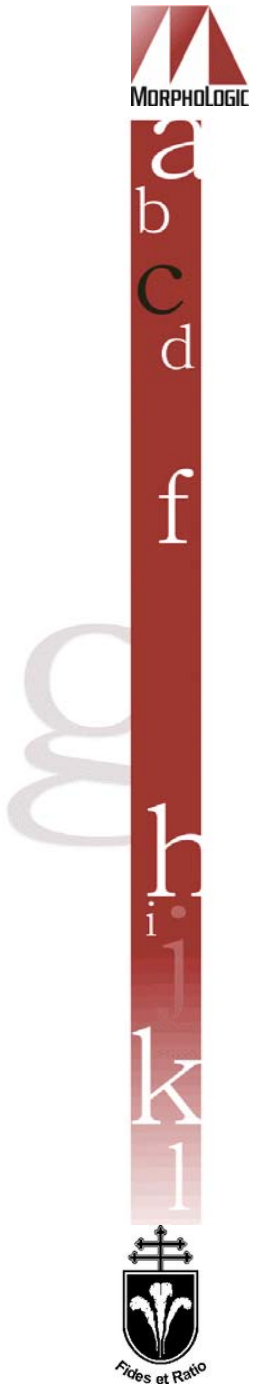
ordering

serving

eating

paying

leaving



Az „étterem” forgatókönyve (alapismeretek)

ROLES	***constraints***	***defaults***
PATRON	*HUMAN*	*ADULT*
WAITER	*HUMAN*	*FEMALE*
		ADULT
COOK	. . .	
MANAGER	. . .	

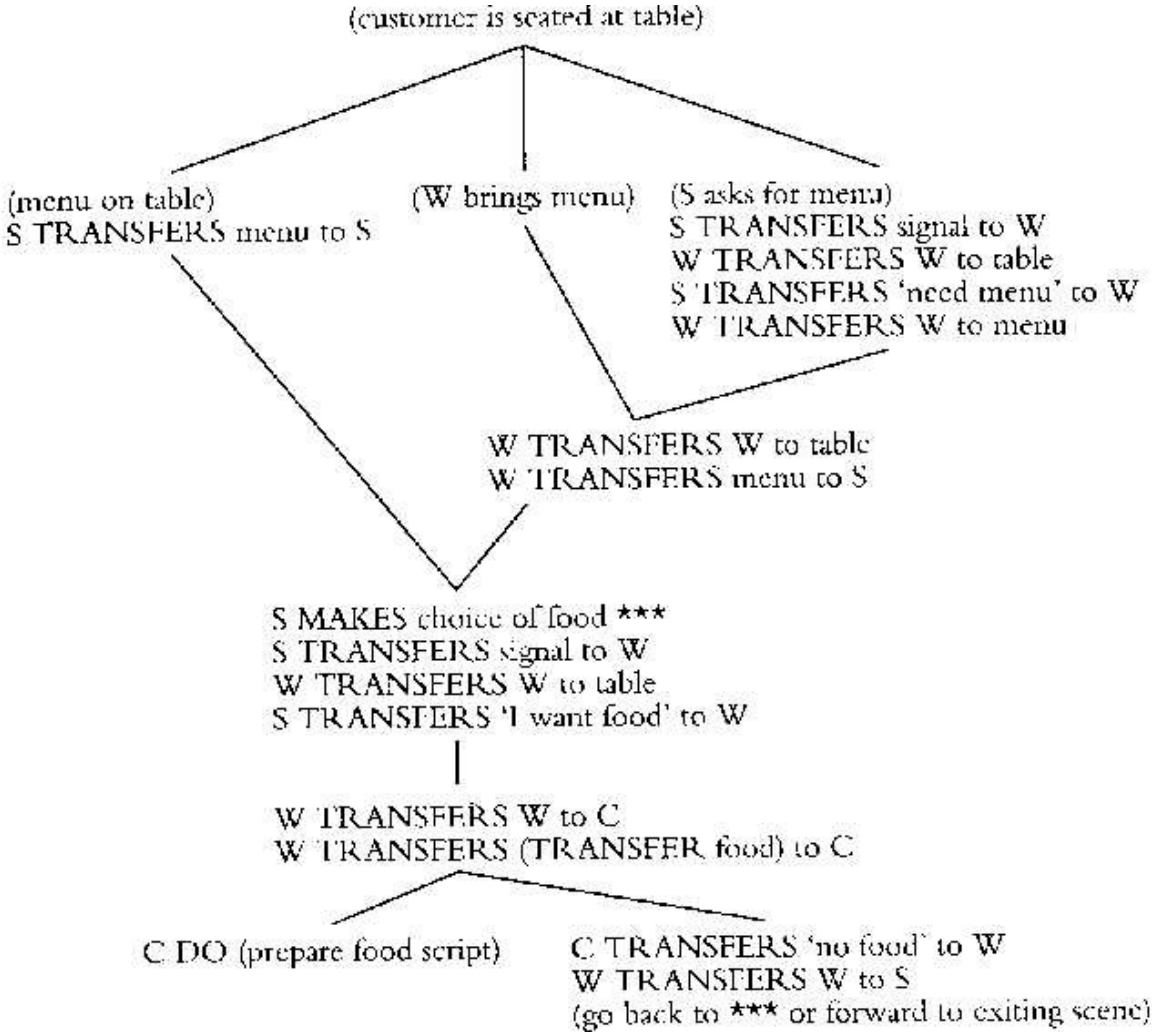
HEADER

PLANNER	PATRON
GOAL	SATISFY(HUNGER)
	SATISFY(SOC-INTERACTION)

BODY

EVENT_CHAIN

Az „étterem” teljes forgatókönyve

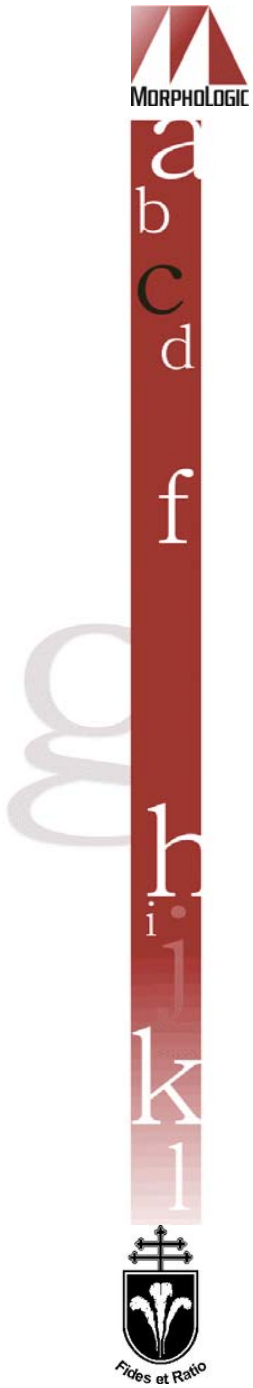


Szótárak és terminológiakezelés

- nyomtatott szótárak és elektronikus szótárak
- terminológiai adatbázisok
- közvetlen és közvetett elektronikus szótárak
- egynyelvű, kétnyelvű és többnyelvű szótárak
- a forrásnyelv és a célnyelvek aszimmetriája

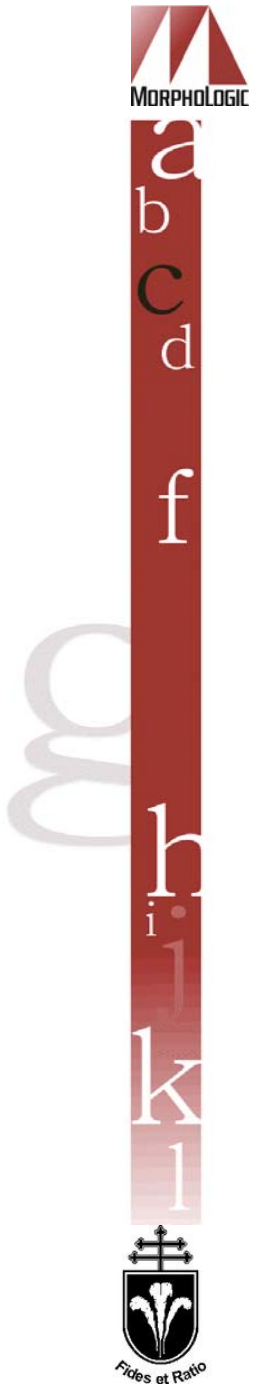
Szerkesztési elvek

- Az (önálló ill. utaló) szócikkek és felépítésük
- A szócikkfej: címszó, homonimák és álhomonimák, alak- és írásváltozatok, kiejtés, elválasztás, szófaj, főbb toldalékos alakok, nyelvtani megjegyzés, stílusminősítés
- Jelentéscsoportok (alapjelentés és jelentésárnyalatok): értelmezések (ekvivalensek) és példák
- Szóláshasonlatok, közmondások, más szavakkal alkotott összetételek, származékszók



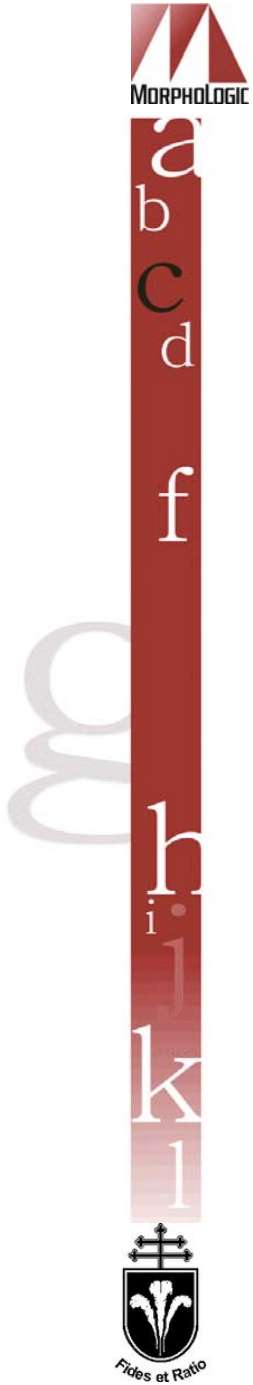
Keresés a szótár(ak)ban

- betű szerint
- csonkolt keresés
- hasonlósági keresés (fuzzy, spell)
- nyelvi alapú keresés a bemeneti oldalon
- nyelvi alapú keresés a találati oldalon
- a kifejezések kezelésének problémái: alcímszók, kulcsszó-választás, indexek, egyazon kifejezés több címszó alatt
- „könyvespolc”: egységes felület
- egyidejű használat: párhuzamos(nak tűnő) keresés



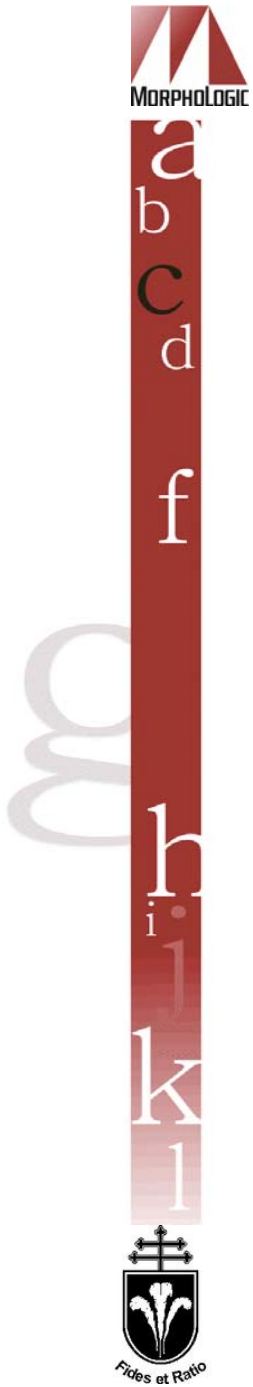
Többszavas kifejezések keresése

- csak címszóként
- betű szerint
- teljes szövegű kereséssel
- reguláris kifejezésként
- tőindexekkel: készítéskor vagy elemzési időben (is)

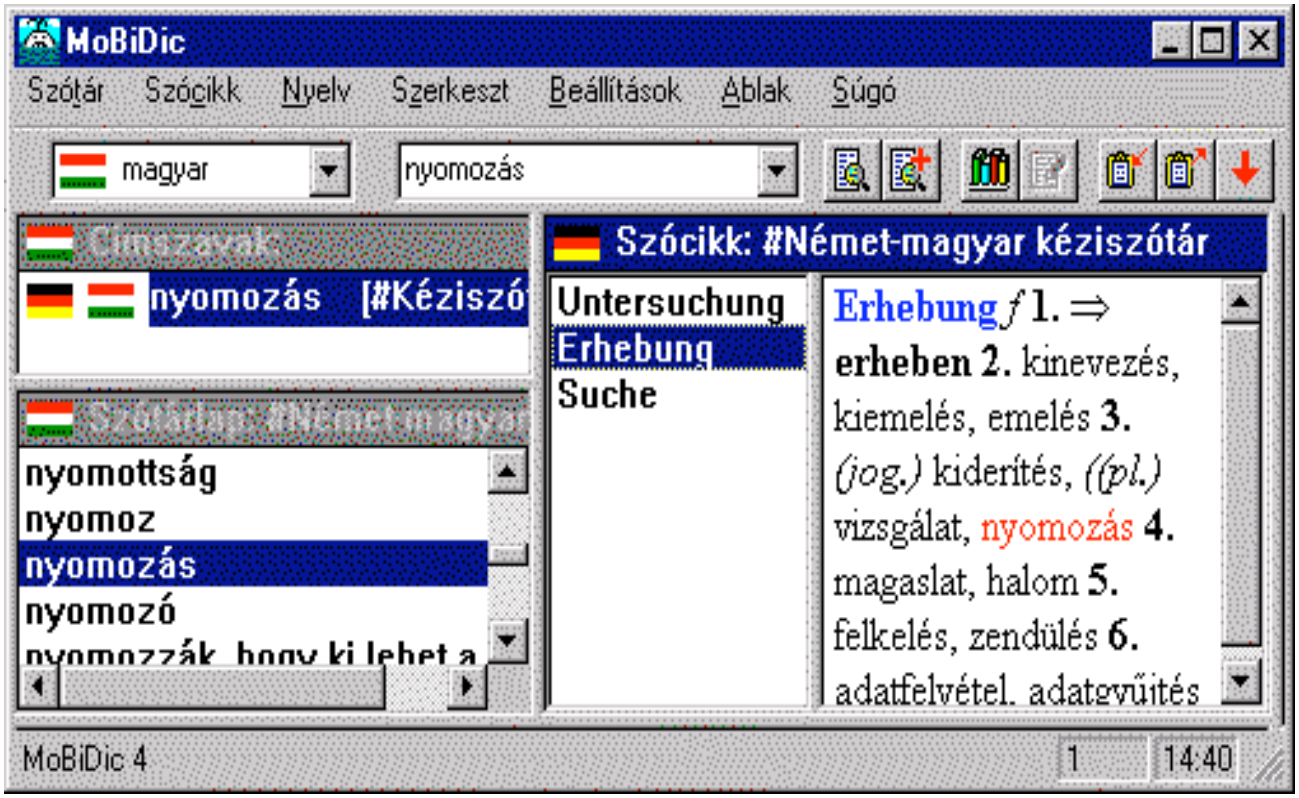


A szótári jobboldal szerepe

- papírszótárak esetében: csak tipográfiai
- elektronikusan: új lehetőség
- ábécé-környezet helyett szinonimák
- többféle jelentés kezelése a baloldali címszavak segítségével
- új találati ablak
- elektronikusan érdemes „kifordítani” a szótárakat

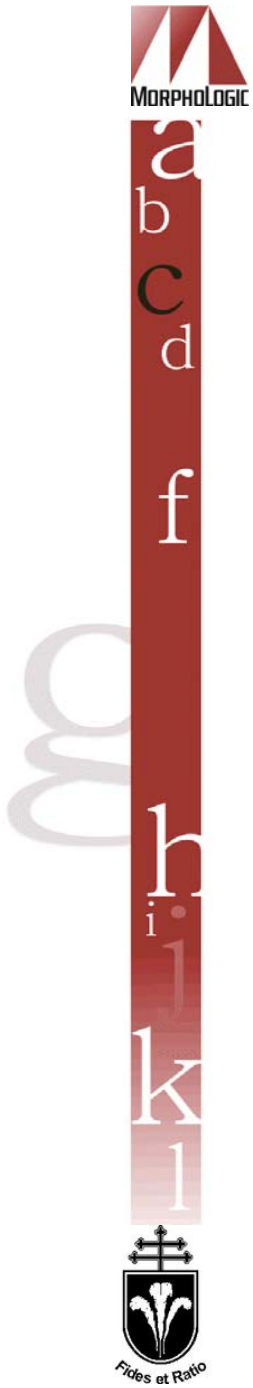


Az elektronikus szótárak megfordíthatók



Gyorsfordítók

- amikor információ kell, pl. szótári, akkor:
  csak amit kérek, nem többet,
  de azt gyorsan,
  kevés aktív művelettel
  és a lehető legautomatikusabban!
- kialakul a „pop-up” viselkedés
- a kijelölhetőség, ill. az automatikus indíthatóság szerepe



A „rávetítős” megoldás lépései

- szöveg(rész)-felismerés
- nyelvi elemzés: morfológia, lemmák, szókapcsolatok (esetleg környezetelemzés)
- szótári keresés: tövesítve vagy csak literálisan
- megjelenítés: buborékban vagy fix ablakban
- log: automatikus információgyűjtés lehetősége

A fordítómemória gondolata

A lefordítandó mondat:

After a few seconds, a window will appear in which you are expected to enter a valid User ID and (if *necessary*) a password.

Korábban már fordítottuk ezt:

After 5 seconds, a window will appear on the screen in which you are expected to enter a User ID and (if *required*) a password.

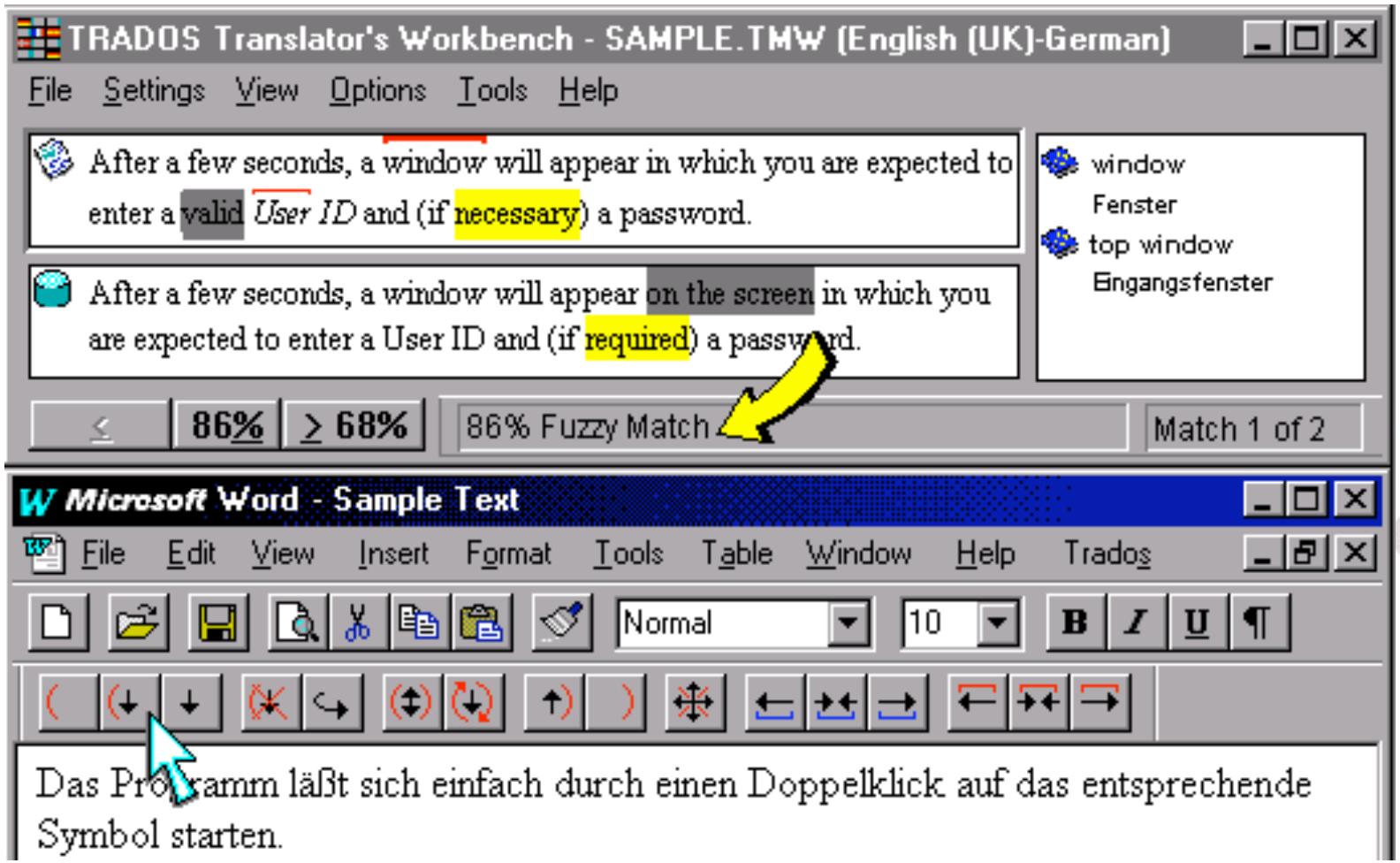
Méghozzá így:

Öt másodperc múlva egy ablak jelenik meg a képernyőn, amelybe be kell gépelni egy felhasználó-azonosítót és (ha szükséges) egy jelszót.

Ebből a következő fordítás könnyen előállhat:

Néhány másodperc múlva egy ablak jelenik meg, amelybe be kell gépelni egy érvényes felhasználó-azonosítót és (ha szükséges) egy jelszót.

A fordítómemória mint eszköz



Szövegszinkronizálás

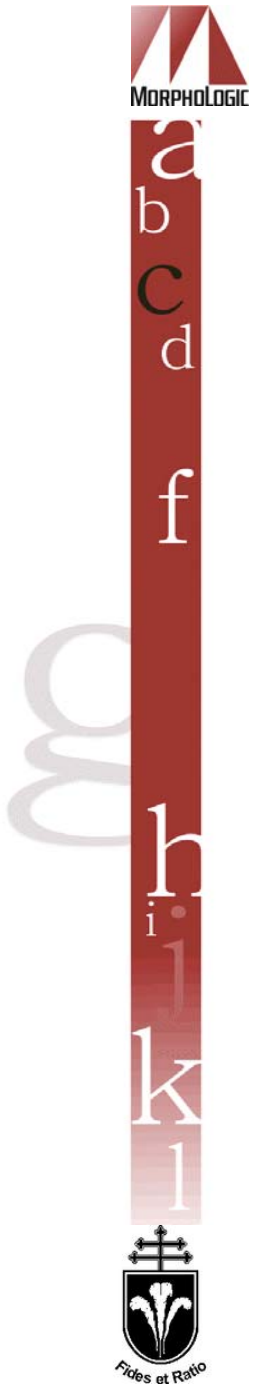
- bi-text
- párhuzamos korpuszok
- szinkronizálás: valós időben és utólag
- pl. a Biblia

„You will not surely die," the serpent said to the woman.
(Genesis 3:4)

A kígyó erre azt mondta az asszonynak: „Dehogy is haltok meg!"
(Ter 3,4)

Szövegszinkronizálási szintek

- bekezdésszint
- mondatszint
- frázis-szint (?)
- szószint (??)
- mondathatár-problémák
- horgonyok
- statisztikai módszerek

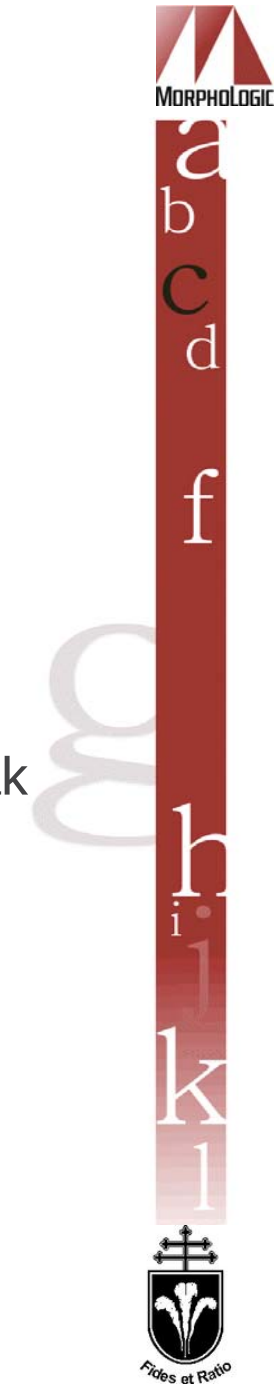


Nem feltétlenül 1-1 értelmű

(1 = 1,2) O stylographe à la plume de platine, que ta course rapide et sans heurt trace sur le papier au dos satiné les glyphes alphabétiques qui trans-mettront aux hommes aux lunettes étin-ce-lantes le récit narcissique d'une double ren-contre à la cause autobusilistique.

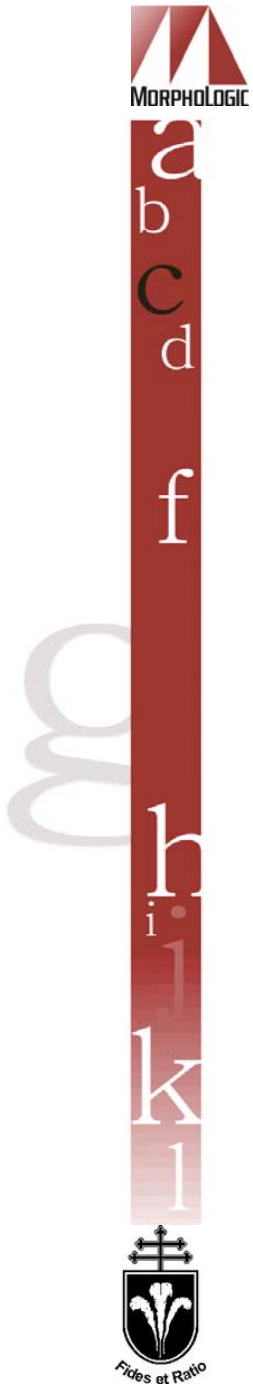
(1 = 1) Ó, platinahegyű töltőtoll!

(2 = 1) Vajha tajtékos-gyors futásod a szaténhátú papirosra róná amaz alfabéta-cikornyákat, melyek a csillogó okulárés emberek tudomására hozzák az autóbuszilisztikus-okú találkozás önbálványozó krónikáját!



A nyelvi szerkezetek hasonlóságáról

zöld kutya
zöld macska
sárga kutya
sárga macska
piros egér
kis asztal
hét kis ágy
a tegnapi buliról
elmentem a tegnapi buliról
beléptünk az EU-ba
jó napot!



A gépi fordítás alapmódszerei

- szabály-alapú:
  közvetlen fordítás
  közvetítőnyelves fordítás
  transzfer rendszerek

- statisztikai

Becslések az európai nyelven írt internetes szövegek lehetséges méretéről

	Nyelv	Szó a weben
1	angol	76 598 718 000
2	német	7 035 850 000
3	francia	3 836 874 000
4	spanyol	2 658 631 000
5	olasz	1 845 026 000
6	portugál	1 333 664 000
7	holland	1 063 012 000
8	svéd	1 003 075 000
8	norvég	609 934 000
10	cseh	520 181 000
11	magyar	457 522 000
12	dán	346 945 000
13	finn	326 379 000
14	lengyel	322 283 000
15	szlovák	216 595 000
16	katalán	203 592 000

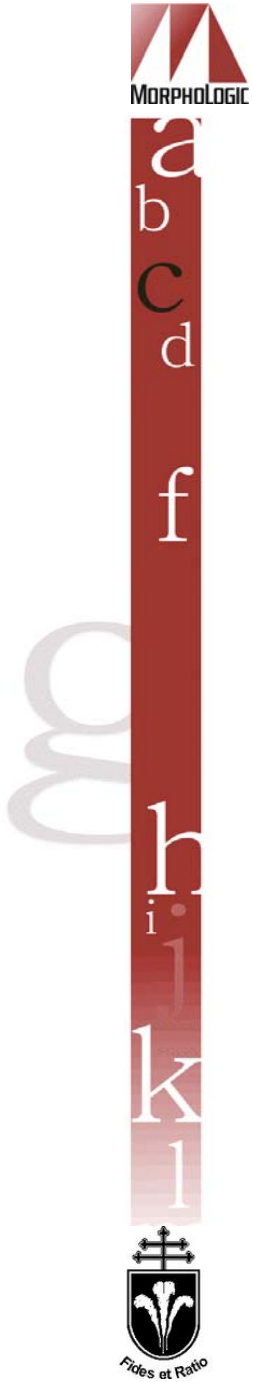
	Nyelv	Szó a weben
17	török	187 367 000
18	maláj	157 241 000
19	horvát	136 073 000
20	szlovén	119 153 000
21	észt	98 066 000
22	ír	88 283 000
23	román	86 392 000
24	eszperantó	57 154 000
25	latin	55 943 000
26	baszk	55 340 000
27	izlandi	53 941 000
28	lett	39 679 000
29	litván	35 426 000
30	velszi	14 993 000
31	breton	12 705 000
32	albán	10 332 000

MorphoLogic

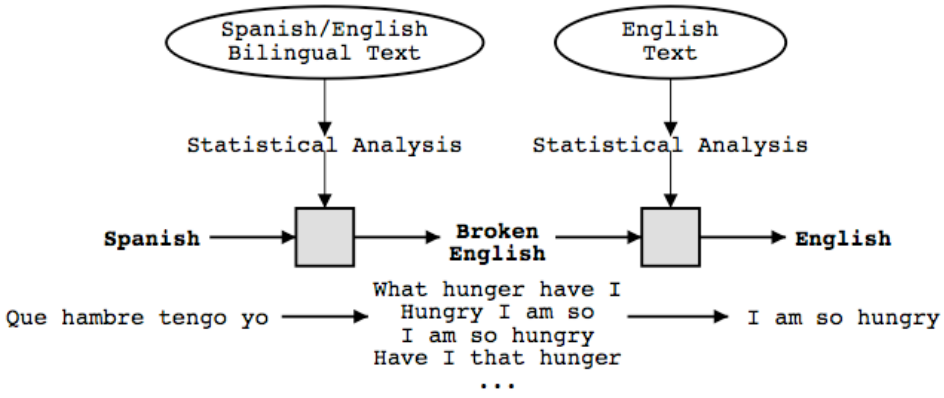
Fides et Ratio

… milyen „minőségű” szövegek vannak a weben?

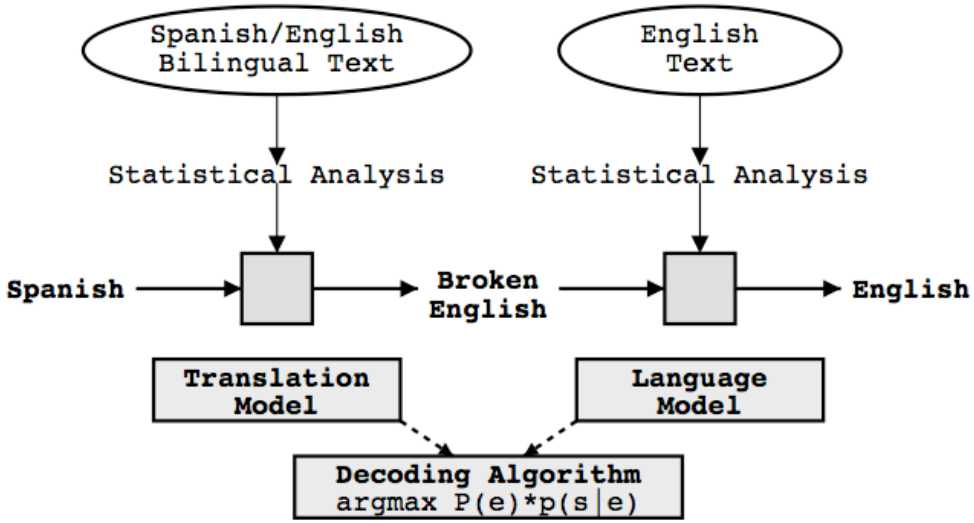
	Full
internet	2 460 000 000
internte	67 400
interent	681 000
intenret	116 000
intrenet	193 000
inetrnet	128 000
itnernet	66 400
niternet	47 700
interne.	19 200 000
intern.t	1 940 000
inter.et	19 400 000
inte.net	2 480 000
int.rnet	436 000
in.ernet	522 000
i.ternet	441 000
.nternet	1 150 000



Statisztikai gépi fordítás

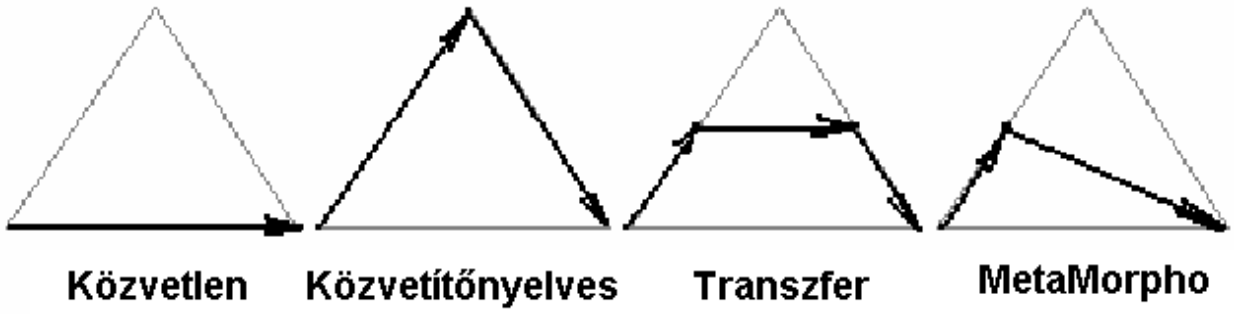


- Fordítási modell
- Nyelvmodell
- Bayes



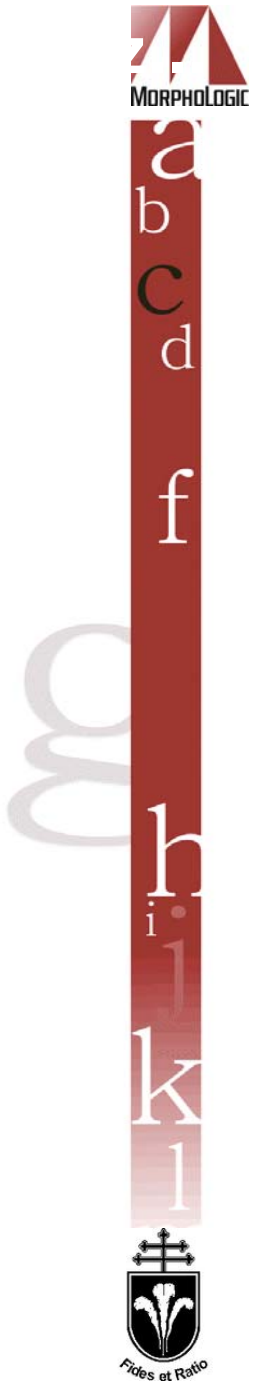
MetaMorpho-elvek

- **Nincs** külön szótár és külön nyelvtan
- **Csak minta-párok vannak**: bemenet/interpretáció szerkezet-párok
- **Egyetlen elemzési menet**: nincs rákövetkező művelet (pl. transzfer)
- Célszerkezet-generálás: **az elemzés „melléktermékeként”**
- **Új:**



Minták: általánosított nyelvészeti információk

- Rövid, specifikus minták:
  szótári címszavak
- Hosszabb, specifikus minták:
  többtagú kifejezések
- Részlegesen alulspecifikált minták:
  kollokációk, idiómák
- Teljesen alulspecifikált minták:
  nyelvészeti szabályok
- Fordítástámogató nyelv:
  minta-interpretáció párok



A MetaMorpho projekt

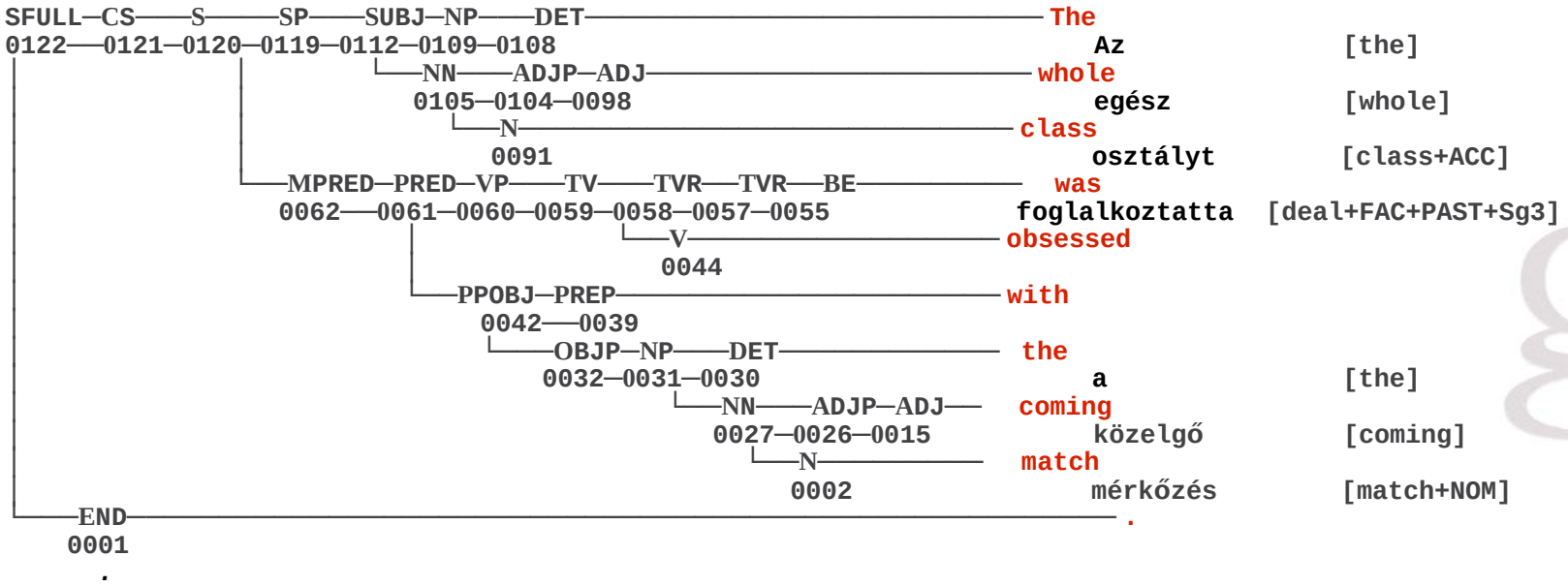
- **A projekt**: 1991-től folyamatosan készített moduljaink felhasználásával (kb. 100 emberév) 2000-ben indult, belső projektként (semmilyen külső támogatása nincs)
- **Cél**: mondatszintű fordítás – új elven: a szavak környezetének felhasználásával (egy n elemű mondatban éppen n darab (n-1) elemből álló környezet van)
- **Forrásnyelv**: angol, magyar
- **Célnyelv(ek)**: magyar, angol, ...
- **Szakterület**: nincs – de dinamikusan bővíthető
- **Minta-alapú**: példák (TM) és szabályok (MT) egységesen
- **Minták száma**: kb. 200.000
- **Lexikon**: kb. 100.000 alapszó
- **Elvárt sebesség**: 50 karakter/s
- **Felhasználói felület**: MoBiCAT, MoBiWAP, MMO-Office, MorphoWord, MoBiWeb, webforditas.hu

MorphoLogic

Fides et Ratio

A MetaMorpho „belülről”

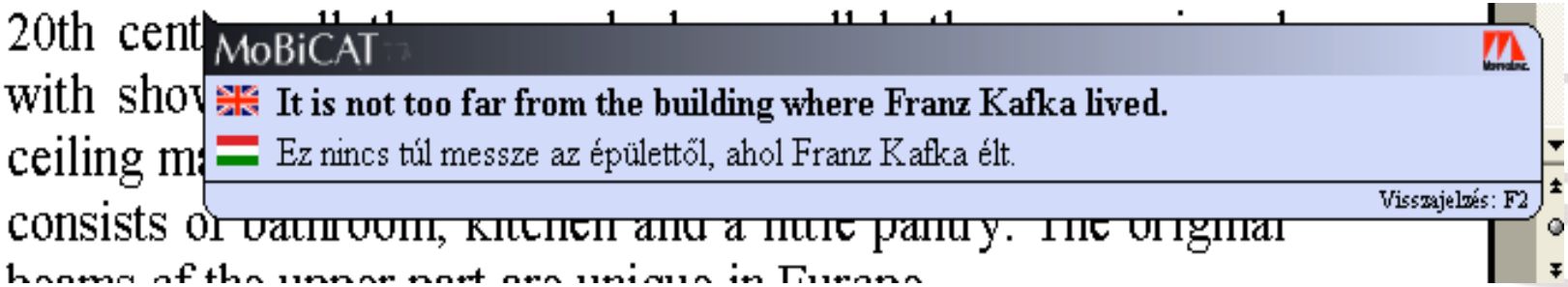
EN: The whole class was obsessed with the coming match.



HU: Az egész osztályt foglalkoztatta a közelgő mérkőzés.

Angol-magyar gyorsfordító szolgáltatás

MoBiCAT: teljes mondatok fordítása
(MoBiCAT-szerver akár intraneten vagy interneten)



Angol-magyar weblap-fordítás

(MorphoWeb, webforditas.hu)