Nyelvtechnológia
4

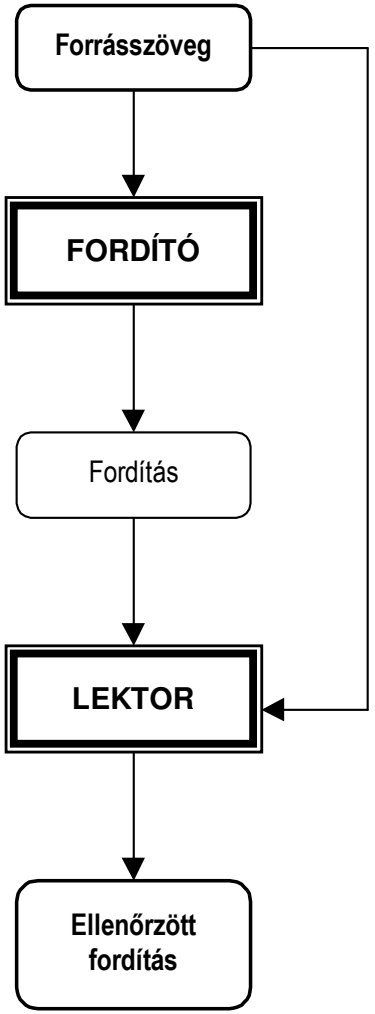
BME, 2005. október 27.

Prószéky Gábor

MorphoLogic

Fides et Ratio

Az emberi fordítás gépi támogatása

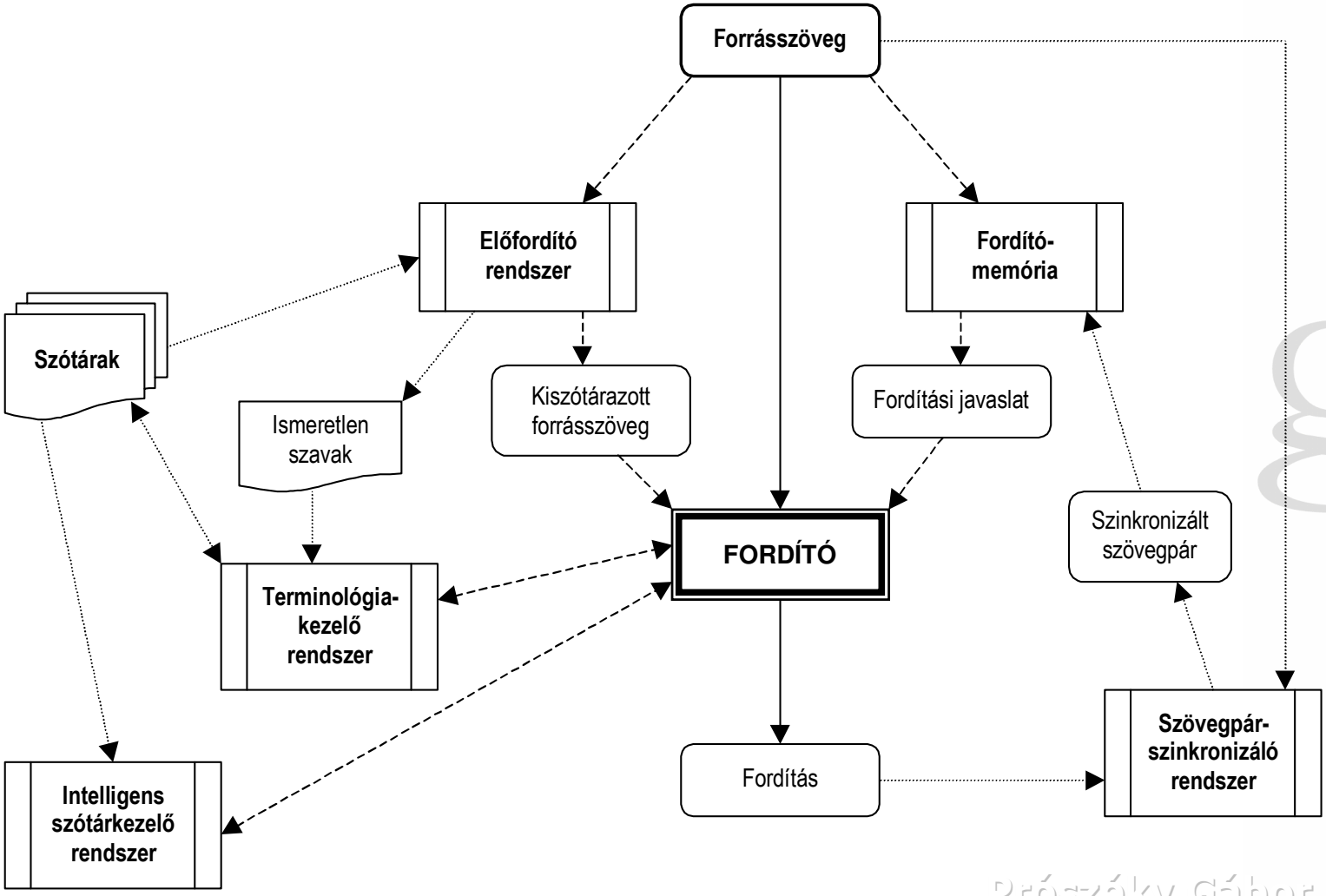


Prószéky Gábor

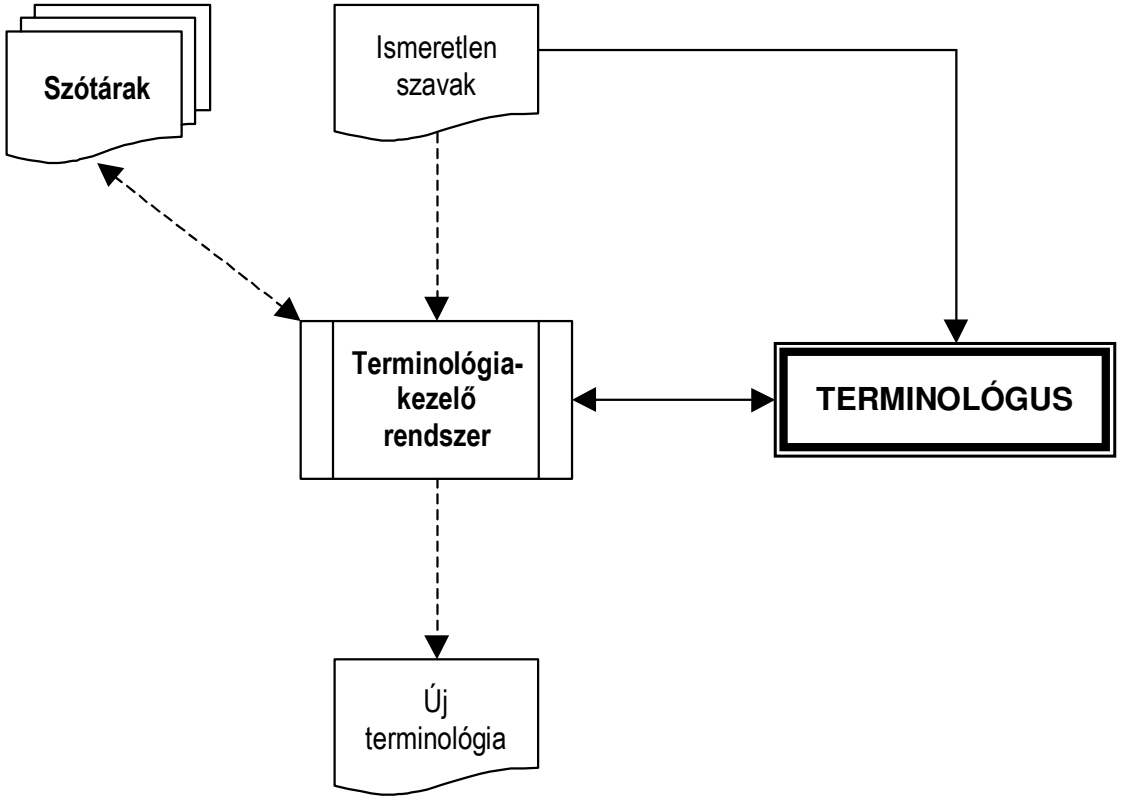
MorphoLogic

Fides et Ratio

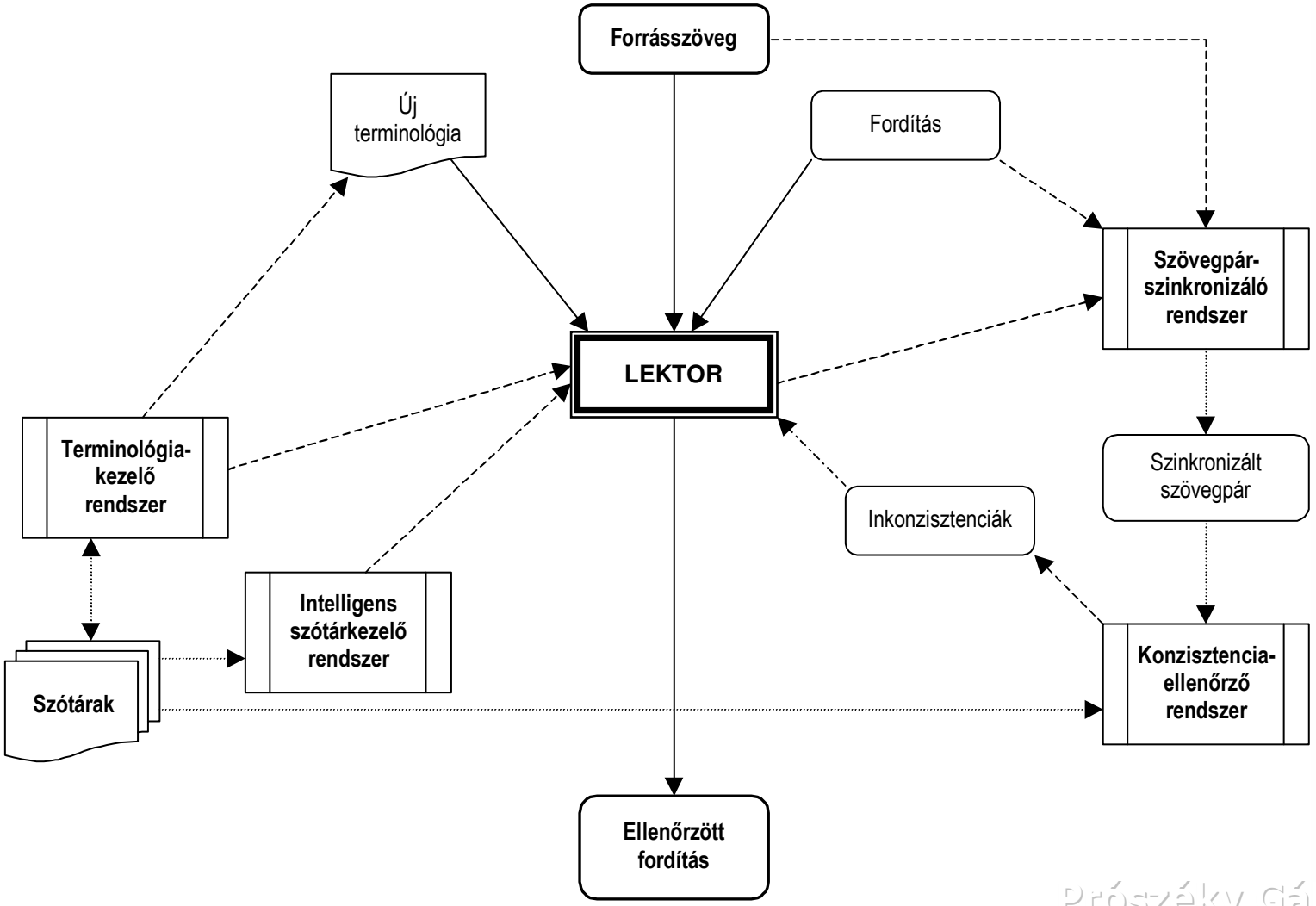
Az emberi fordítás gépi segédeszközei



A terminológus helye a folyamatban



A terminológussal kiegészített munkafolyamat



Szótárak és terminológiakezelés

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

Prószéky Gábor

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**

Prószéky Gábor

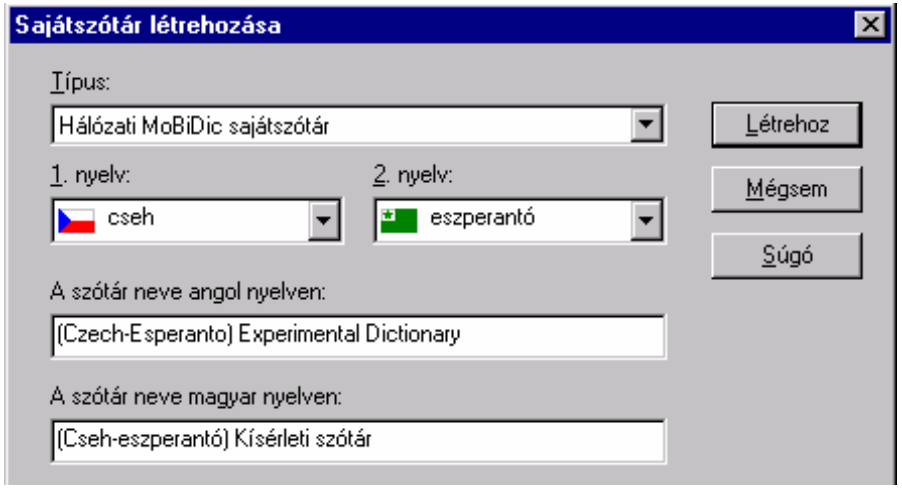
Címszavak

- élő köznyelvi szavak
- idegen szavak
- kifogásolható szavak: valaha helytelenek, durvák, illetlenek, nyelvhelyességileg nem elfogadottak
- peremszókincs: régiek, argó, nagyon újak

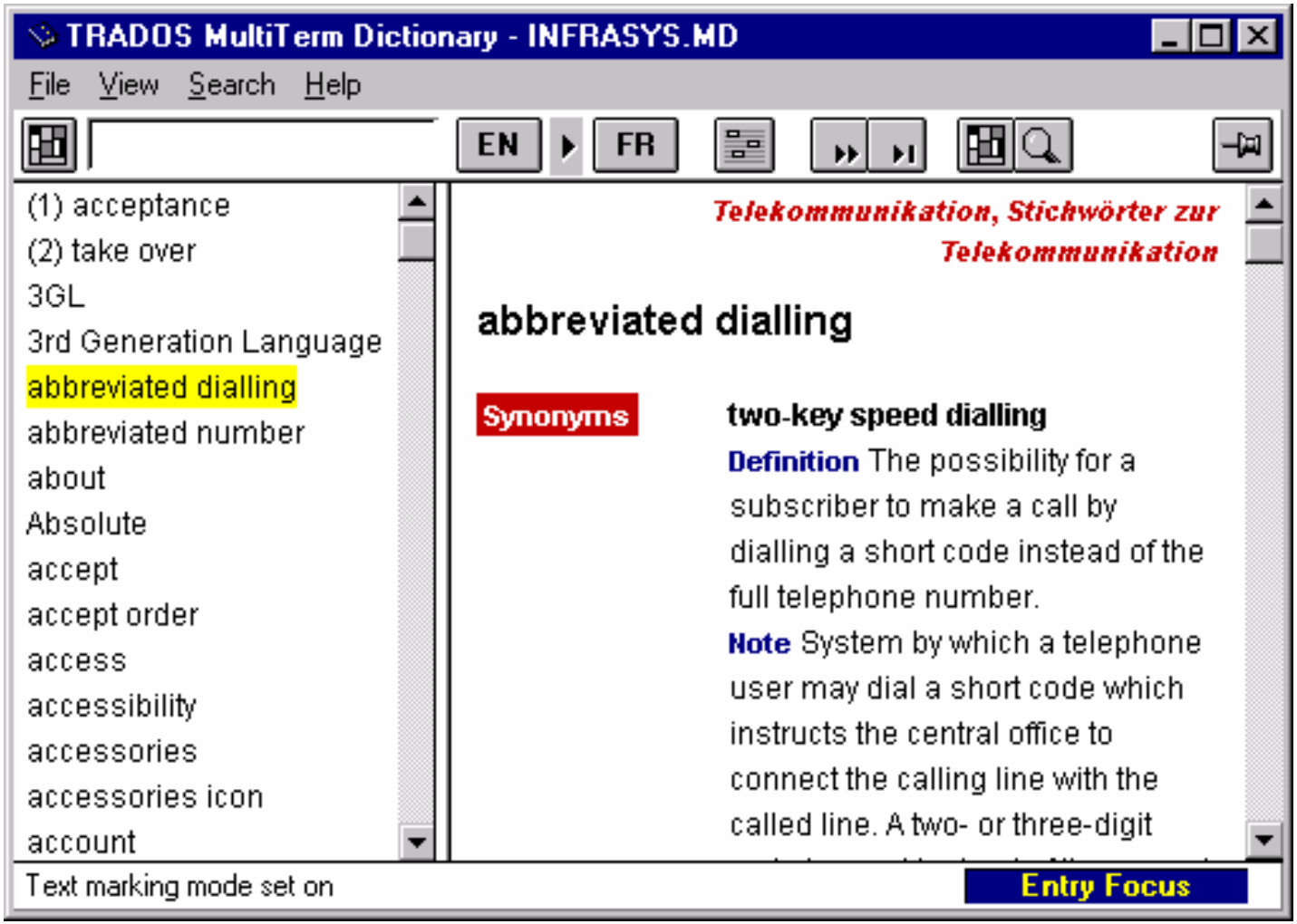
Prószéky Gábor

Segéd- és szakszótárak

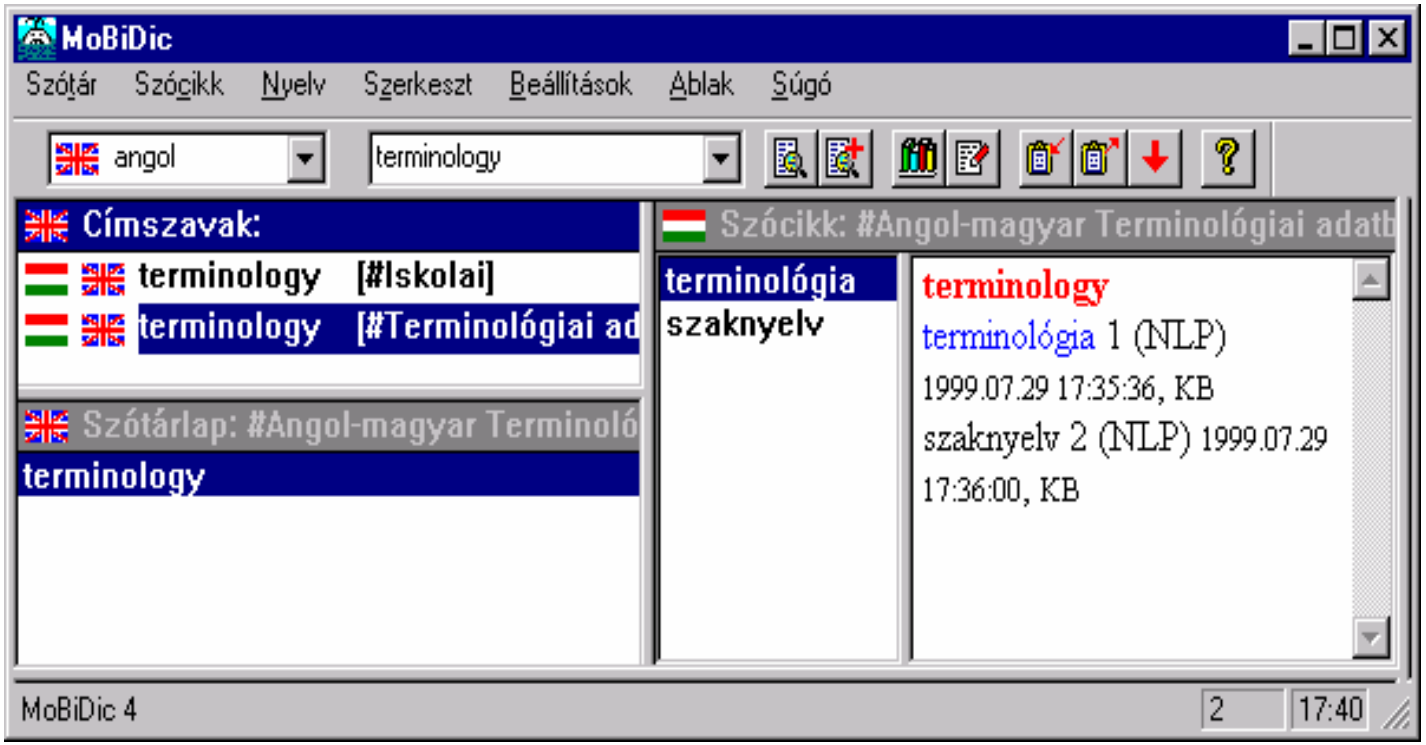
- **megváltozott szerepük az elektronikus világban**
- **értelmező és tájnyelvi szótárak**
- **szakmai szótárak (enciklopédiák?)**
- **terminológia**



Terminológiakezelők (1)



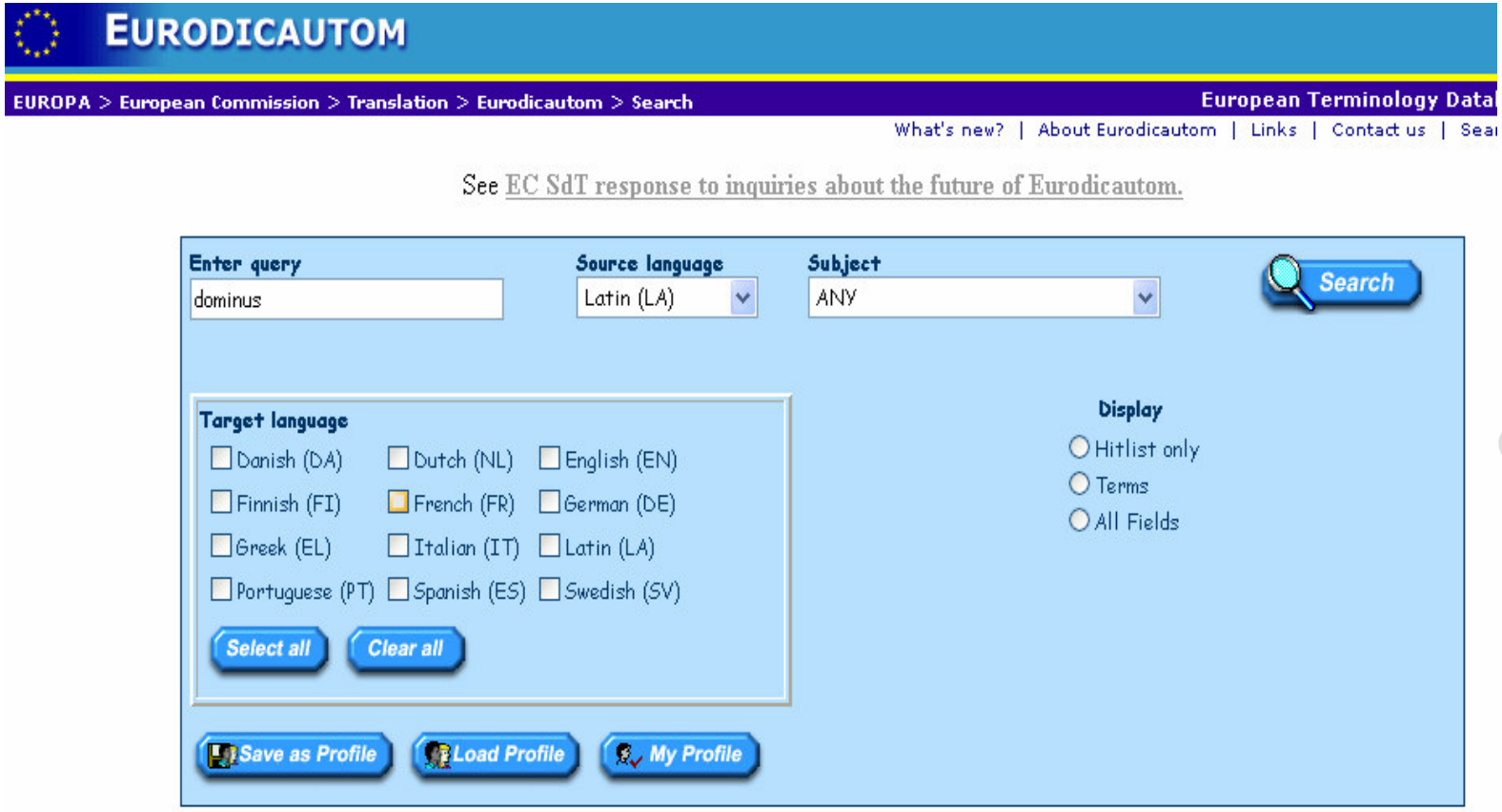
Terminológiakezelők (2)



Prószéky Gábor

Terminológiakezelők (2)

http://europa.eu.int/eurodicautom/controller



Prószéky Gábor

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

Prószéky Gábor

Többszavas kifejezések keresése

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

Prószéky Gábor

Nyelvfüggő szótárproblémák

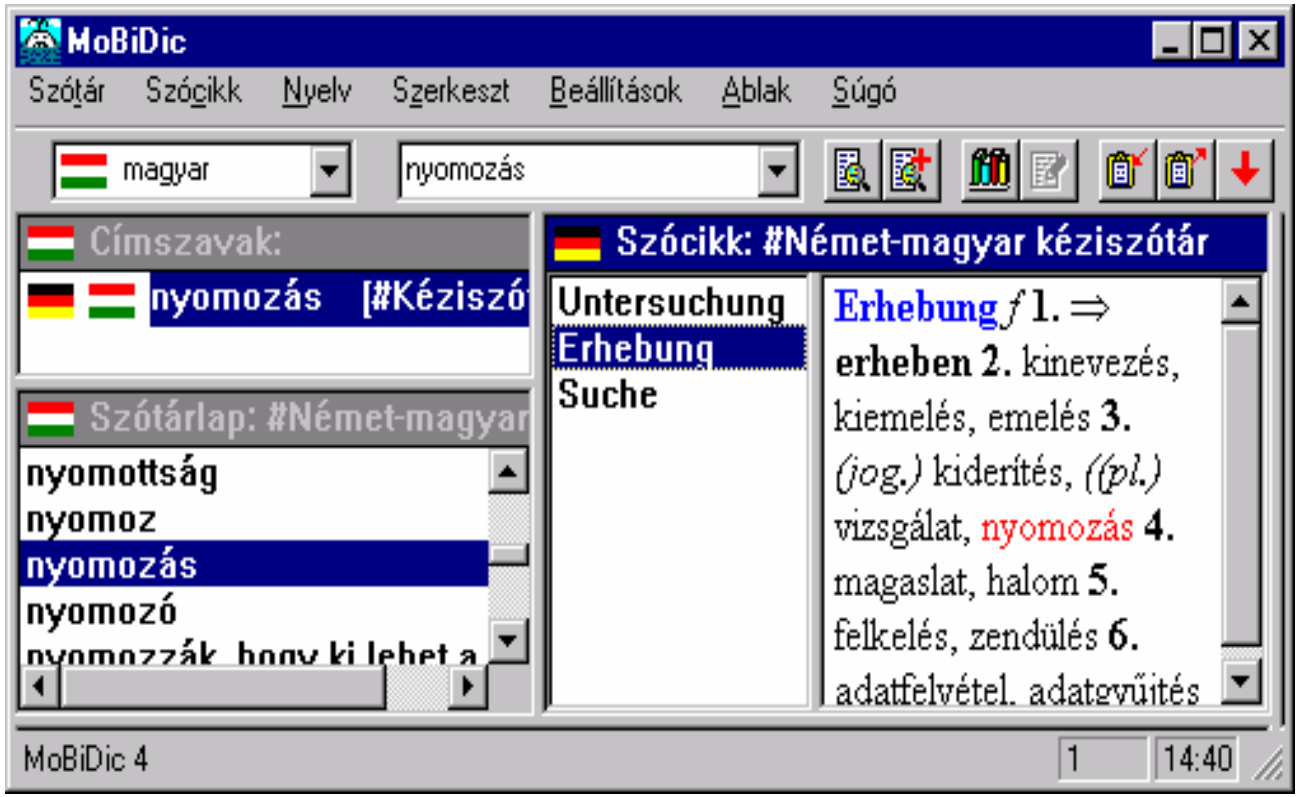
- A forrás- és célnyelv ábécéjének ismerete
- A forrás- és a célnyelv ábécérendjének ismerete
- A fonetikai információ kezelése
- Nyelvi keresénél a szótár grammatikai információival való kompatibilitás

Prószéky Gábor

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**

A szótárak megfordíthatósága



Prószéky Gábor

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**

Az „ablakos” kommunikáció nehézségei

- **kilépni az adott alkalmazásból**
- **elindítani**
- **kinyitni vagy felnagyítani**
- **beírni**
- **klikkelni**
- **átmozgatni**
- **lekicsinyíteni vagy bezárni**
- **visszalépni az eredeti alkalmazásba**

Prószéky Gábor

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

Prószéky Gábor

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.

A fordítómemória a következő, nagyon hasonló mondatot tartalmazza (a különbségeket a fenti és az alábbi mondatban is megjelöltük):

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

Ennek az adatbázisbeli fordítása (amelyet a fordítómemória felajánl):

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

A fordító ebből a következő - pontos - fordítást állítja elő:

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

Prószéky Gábor

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)

Prószéky Gábor

Szövegszinkronizálás

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

Prószéky Gábor

Nem feltétlenül 1-1 értelmű

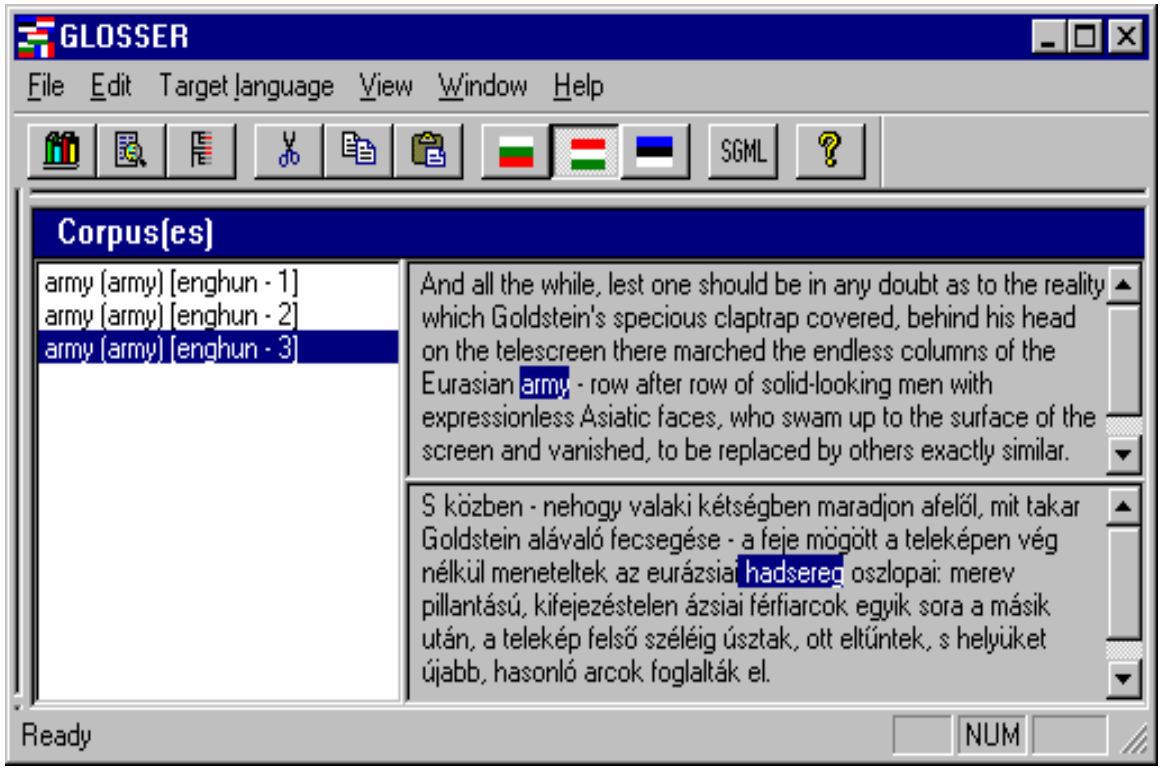
(1 = 1,2) O stylographe à la plume de platine, que ta course rapide et sans heurt trace sur le papier au dos satiné les glyphes alphabétiques qui trans-mettront aux hommes aux lunettes étin-ce-lantes le récit narcissique d'une double ren-contre à la cause autobusilistique.

(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 mondatnál alacsonyabb szintekről

- **a szófordítások gépi megtalálásának problémája**

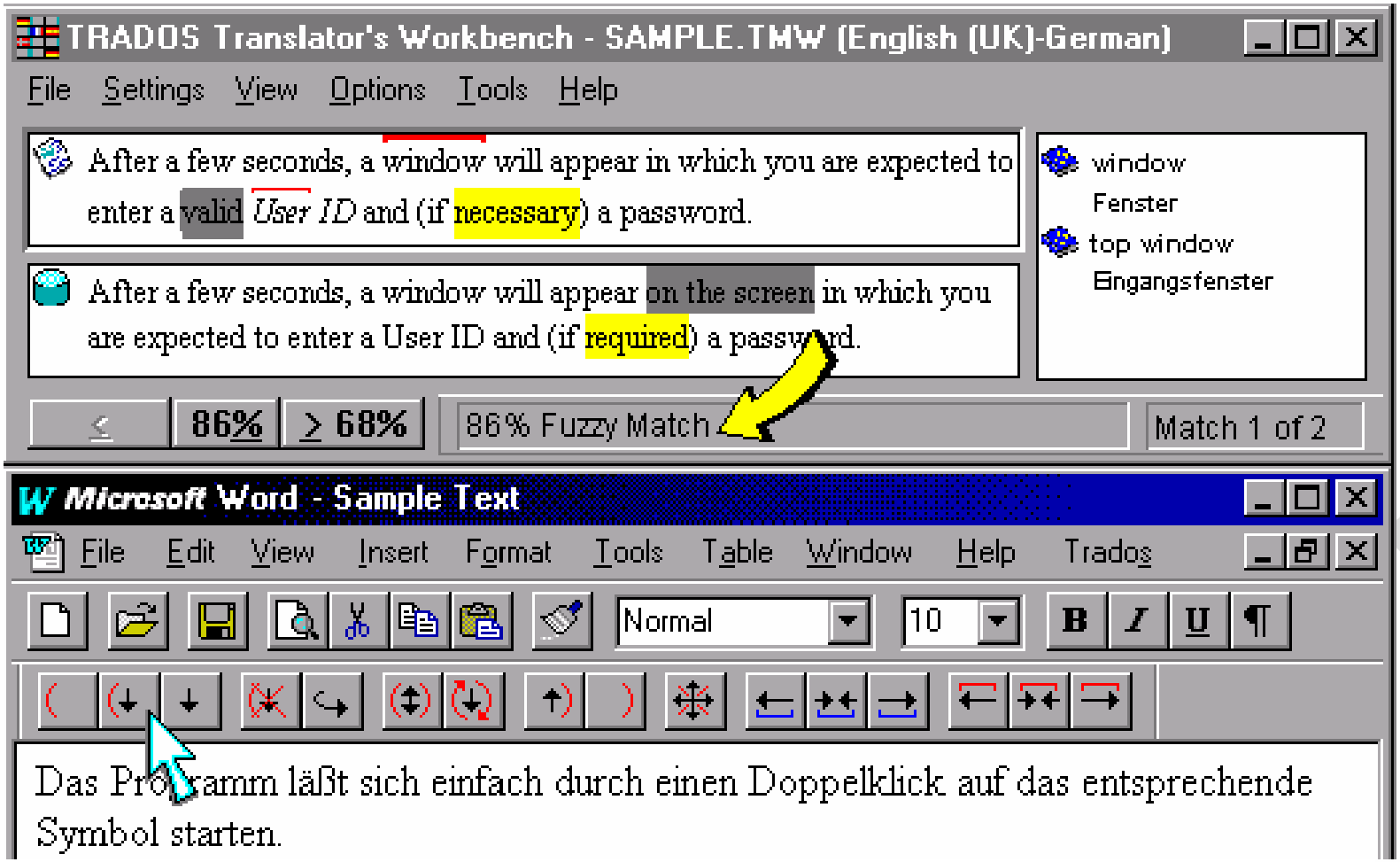


A nyelvi szerkezetek hasonlóságáról

- 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

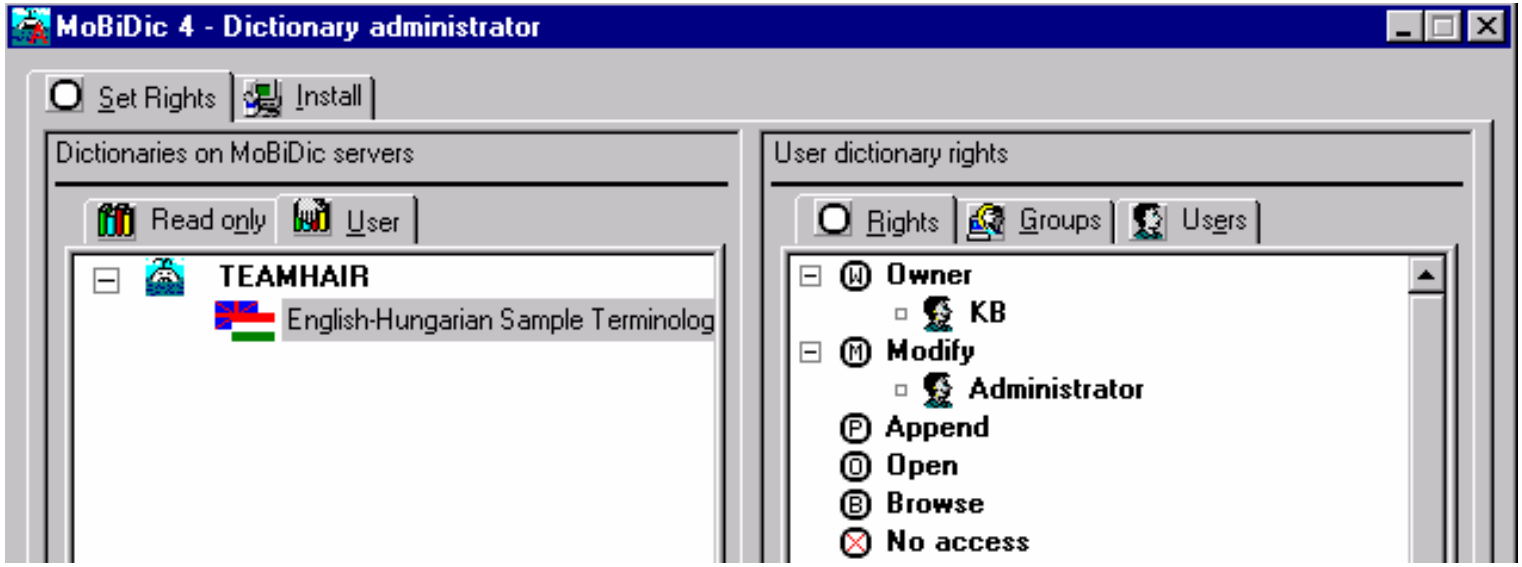
Prószéky Gábor

A fordítómemória mint eszköz



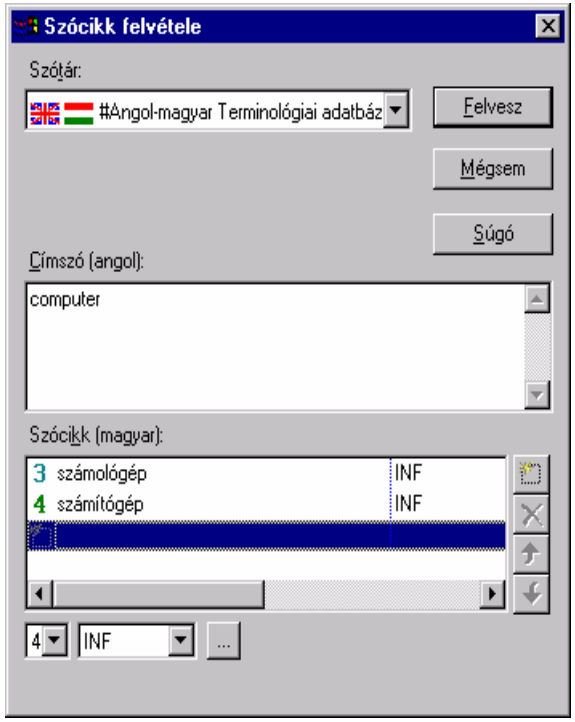
Fordítócsoportok támogatása

- osztott terminológiakezelő rendszer
- írás, bővítés: jogosultságok kezelése



Fordítócsoportok támogatása (2)

- elektronikus korrektúrakezelés

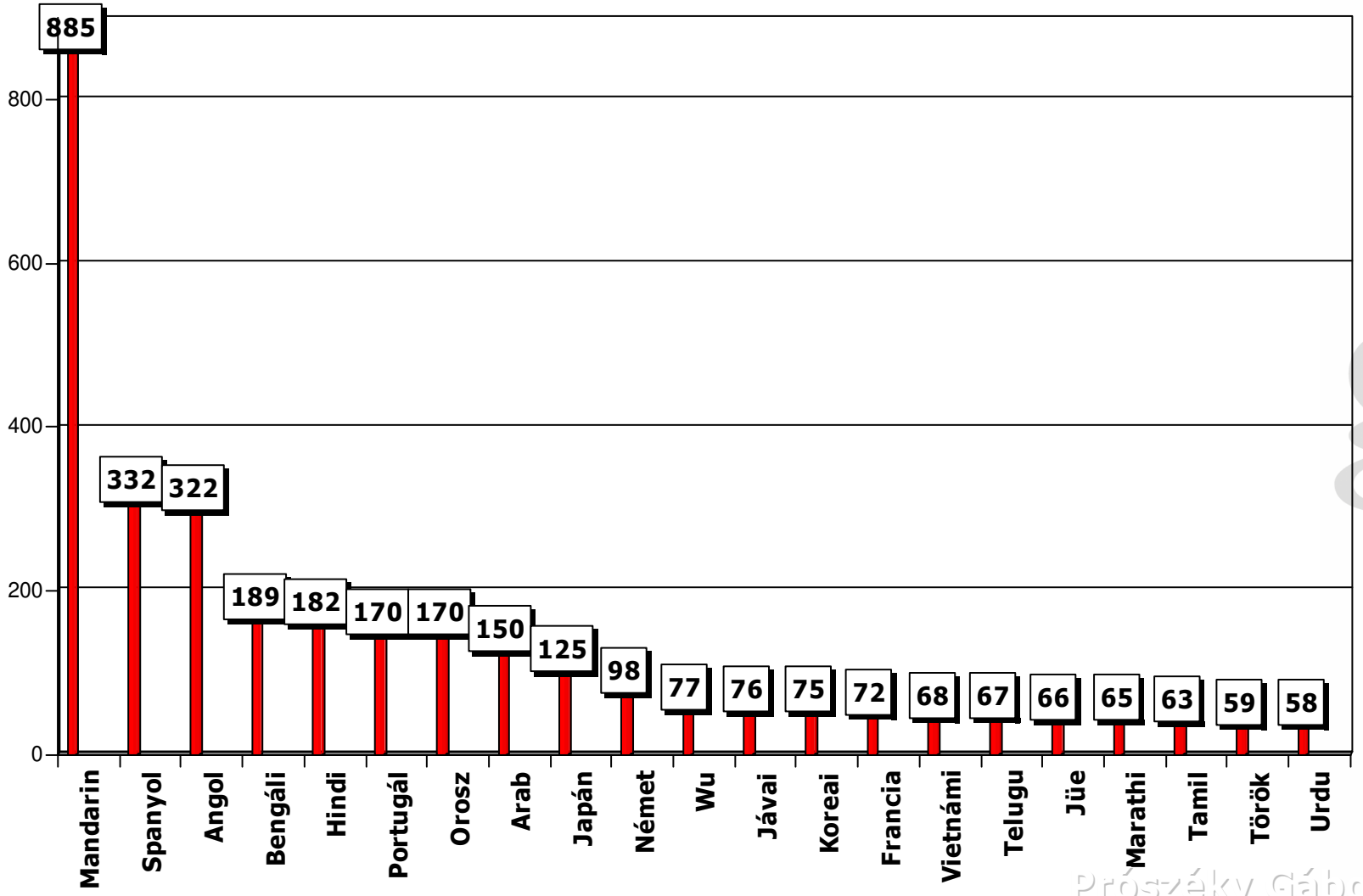


- számítógépes konzisztencia-ellenőrzés

A gépi fordítás (MT) alapmódszerei

- közvetlen fordítás
- közvetítőnyelves fordítás
- transzfer rendszerek

A világ nyelvei



Az internethasználók nyelvei

(millió főben)	1999	2002	2005	Beszélők	internetezők
angol	85	165	231	500	46%
kínai	10	75	220	885	25%
japán	20	64	105	125	84%
spanyol	13	49	80	332	24%
német	14	44	71	98	72%
koreai	5	30	50	75	67%
francia	10	24	49	72	68%
olasz	10	25	42	57	74%
portugál	4	20	38	170	22%
skandináv*	8	14	15	19	75%
holland	6	13	15	20	73%
egyéb	6	63	140	3500	4%

Angol és magyar szövegek a weben

Language	Web Size	Language	Web Size
Albanian	10,332,000	Catalan	203,592,000
Breton	12,705,000	Slovakian	216,595,000
Welsh	14,993,000	Polish	322,283,000
Lithuanian	35,426,000	Finnish	326,379,000
Latvian	39,679,000	Danish	346,945,000
Icelandic	53,941,000	Hungarian	457,522,000
Basque	55,340,000	Czech	520,181,000
Latin	55,943,000	Norwegian	609,934,000
Esperanto	57,154,000	Swedish	1,003,075,000
Roumanian	86,392,000	Dutch	1,063,012,000
Irish	88,283,000	Portuguese	1,333,664,000
Estonian	98,066,000	Italian	1,845,026,000
Slovenian	119,153,000	Spanish	2,658,631,000
Croatian	136,073,000	French	3,836,874,000
Malay	157,241,000	German	7,035,850,000
Turkish	187,356,000	English	76,598,718,000

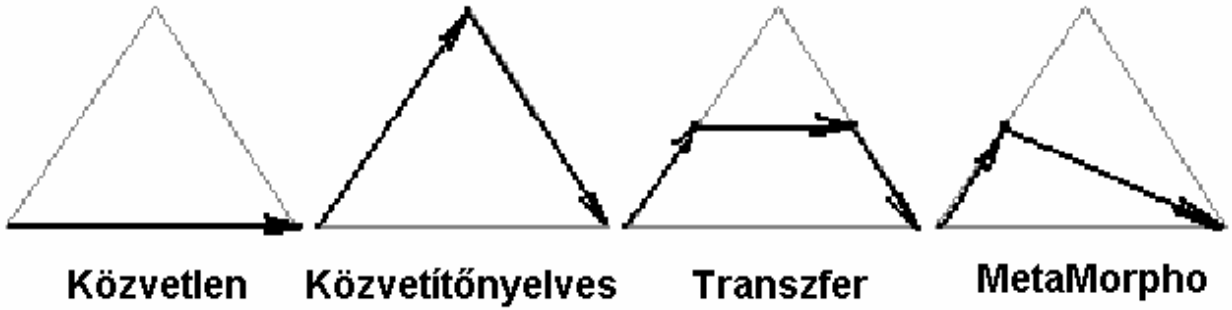
… milyen „minőségű” szövegek vannak a weben?

	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

Prószéky Gábor

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

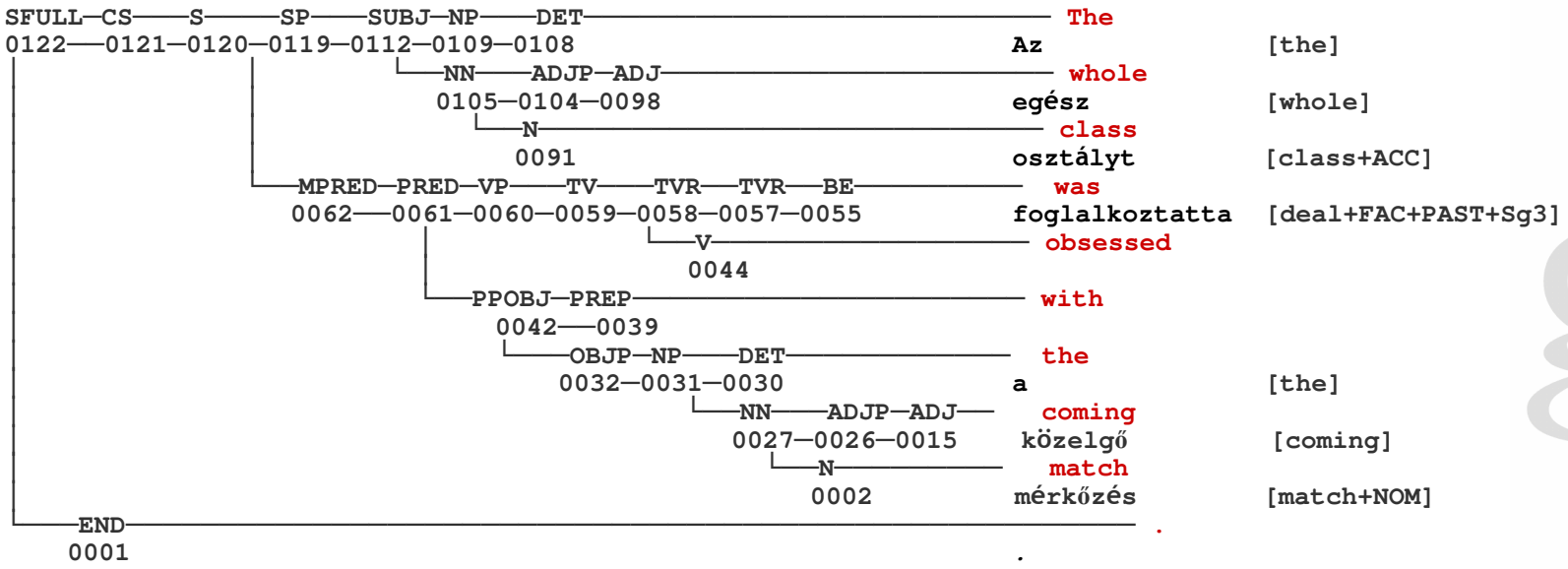
Prószéky Gábor

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

A MetaMorpho „belülről”

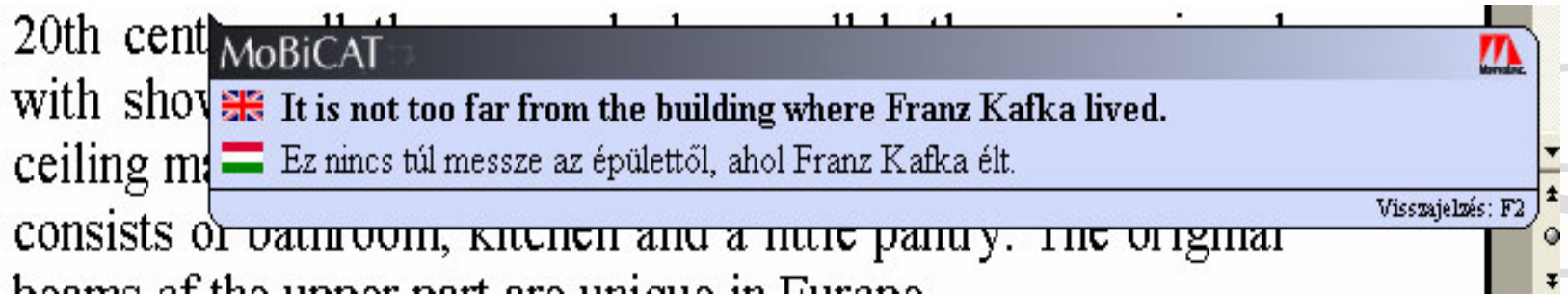
EN: The whole class was obsessed with the coming match.



HU: Az egész osztályt foglalkoztatta a közelgő mérkőzés.

Angol-magyar gyorsfordító szolgáltatás

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



Tetszőleges szöveg fordítása: MorphoWord

How to use the program?
Hogyan használjuk a programot?

▶	Press the blue triangle in the toolbar. The program translates this text from English into Hungarian. Nyomjad a kék háromszöget az eszköztárban. A program ezt a szöveget fordítja le angolról magyarra.
▷	Click the empty triangle in the toolbar if you wish the English text to disappear. Kattints az üres háromszögre az eszköztárban, ha azt kívánod, hogy az angol szöveg tűnjön el.