



Prószéky Gábor

A nyelvtechnológia alapjai

2019/2020. tanév, 1. félév

A tantárgy órái 2019-ben

- ☐ 1. óra: szeptember 11.
- ☐ 2. óra: szeptember 18.
- ☐ 3. óra: szeptember 25.
- ☐ 4. óra: október 2.
- ☐ 5. óra: október 9.
- ☐ (Pázmány-nap: október 16.)
- ☐ (A köztársaság napja: október 23.)
- ☐ (Őszi szünet: október 30.)
- ☐ 6. óra: november 6.
- ☐ 7. óra: november 13.
- ☐ 8. óra: november 20.
- ☐ 9. óra: november 27.
- ☐ 10. óra: december 4.
- ☐ 11. óra: december 11.

A tantárgy felépítése

- ❑ Előadás:
szerdánként heti 3 óra, azaz max. 135 perc
menete: két részben, egy (kis) szünettel
kezdet: 13.15, vége: 15.45 (témától függően)
- ❑ Gyakorlatok: heti 2 óra
gyakorlatvezető: Yang Zijian Győző
- ❑ Követelmény:
jelenlét előadáson (80%) és gyakorlaton is
+ aktivitás az órán
+ házi feladatok megoldása
+ 2 zárthelyi dolgozat (pótlása kritikus!)
+ kötelező vizsga az idei (!) előadások anyagából
(a diák az előadások után elérhetek lesznek)

A nyelvtechnológia

az **informatikának** az az ága,
amelynek **nyelvészeti kutatásokon alapuló** eredményei
úgy épülnek be – akár ipari méretekben is –
a különféle számítógépes rendszerekbe,
hogy ezek segítségével a gépek a felhasználók számára
a számítógépes kommunikáció folyamán
a nyelvet jól használó emberéhez hasonló
támogatást tudnak adni.



A nyelvtechnológia „elnevezései”

MT

machine translation
gépi fordítás

CL

computational linguistics
számítógépes nyelvészet

NLP

natural language processing
természetesnyelv-feldolgozás

LE

language engineering
nyelvmérnökség

HLT

human language technologies
humán nyelvtechnológiák

Miről lesz szó ebben a tárgyban?

(A tantárgy vázlatos tematikája)

- ☐ A szövegek kódolása
- ☐ Véges állapotú módszerek a nyelvtechnológiában
- ☐ Szóelemzés és -generálás
- ☐ A szóelemzés szerepe a gyakorlatban
- ☐ A mondatok szerkezete és elemzésük
- ☐ Szemantika, világábrázolás, ontológiák
- ☐ Intelligens szótárak
- ☐ Fordítástámogatás, fordítómemóriák
- ☐ Gépi fordítás
- ☐ Statisztikai módszerek
- ☐ Szóbeágyazás, neurális hálók
- ☐ ...



Források a tantárgy irodalmához

- ☐ **Computational Linguistics**
(<http://www.mitpressjournals.org/loi/coli>)
- ☐ **Natural Language Engineering**
(<https://www.cambridge.org/core/journals/natural-language-engineering#>)
- ☐ **Language and Computers**
(<http://www.brill.com/products/series/language-and-computers>)
- ☐ **Computer Speech and Language**
(<https://www.journals.elsevier.com/computer-speech-and-language>)
- ☐ **Corpus Linguistics and Linguistic Theory**
(<https://www.degruyter.com/view/j/cllt>)
- ☐ **Language Resources and Evaluation**
(<http://www.springer.com/education+%26+language/linguistics/journal/10579>)
- ☐ **Machine Translation**
(<http://www.springer.com/computer/ai/journal/10590>)
- ☐ **Journal for Language Technology and Computational Linguistics**
(<http://www.jlcl.org/?language=en>)



További irodalom

Alapkönyv:

- ❑ Daniel Jurafsky & James Martin. *Speech and Language Processing*.
Prentice-Hall, 2000/2008/2019

Rengeteg minden:

- ❑ ACL Anthology
(<http://aclweb.org/anthology-new/>)

Magyarul:

- ❑ Prószéky Gábor: *Számítógépes nyelvészet*.
Számalk, 1989
- ❑ Prószéky Gábor & Kis Balázs: *Számítógéppel emberi nyelven*.
SZAK, 1999
- ❑ Prószéky Gábor: *A nyelvtechnológia (és) alkalmazásai*.
Aranykönyv, 2005



1.

A nyelvtechnológia története

A nyelvtechnológia „evolúciója”

1950-60: ötletek

(vannak már gépek)

1960-70: kísérletek

(kialakulnak az igények)

1970-80: programok

(megjelennek a „használható” gépek)

1980-90: termékek

(a gépek kapacitása megnő)

1990-től: technológia

(a kommunikációs helyzet megváltozik)

2000-től: ipar

(egyre több a feldolgozandó szöveg)

2010-től: internet

(mindenhol)

A gépi nyelvészet „történelmének” kezdetei

- ☐ Általános tapasztalat: a nyelv változik
- ☐ Ezért: a nyelvészet a 20. századig = történeti nyelvészet
- ☐ A deskriptív nyelvészet „mechanikus segédért kiált”
(ami a „preskriptív” nyelvészethoz is jól jön majd!)
- ☐ A számítógép és a gépi fordítás gondolata egyaránt a világháború hozadéka
- ☐ Booth és Weaver: sifrírozás/desifrírozás (1947-49)
- ☐ Bar-Hillel (1951): „a teljesen automatikus gépi fordítás megvalósítható”
- ☐ A gépi fordítás bemutatkozik (némi kormánytámogatással): Georgetown/IBM
- ☐ Szovjetunió és Kelet-Európa: ott inkább matematikai nyelvészet
- ☐ Magyarországon is megindul a gépi fordítás kutatása

Modern nyelvelméletek és elméletmentes gépi kísérletek

- ☐ Generatív grammatika: Chomsky: *Syntactic Structures* (1957)
- ☐ Probléma: a transzformációk nem invertálhatók
- ☐ A számítógépes elemezni akar elsősorban, és nem generálni
- ☐ Amit inkább használnak: kategoriális (Bar-Hillel 1953),
füzér- (Harris 1955) és függősegi (Tesnière 1957) leírások
- ☐ Bar-Hillel (1959): „a teljesen automatikus gépi fordítás
(FAMT) nem lehetséges”
- ☐ Kis (tudománypolitikai) kitérő: hidegháború és holdraszállási
program (1961)
- ☐ Katz & Fodor (1963): megjelenik a szemantika!
- ☐ Chomsky *Aspects of the Theory of Syntax* (1965): szintaktikai
jegyek (következmény: szabály -> szabályosztály)
- ☐ A gépi fordítási korszak vége (1966): az ALPAC Report (ahol -
igen korán - megjelenik a fordítómemóriák alapgondolata!)

Önálló „gépi nyelvelméletek”?

- ☐ ALPAC-következmény: a számítógépes nyelvészet (computational linguistics: CL) megszületése
- ☐ Woods (1969): Lunar (holdprogram!)
- ☐ Egy korszak-meghatározó melléktermék: Augmented Transition Network (ATN)
- ☐ Winograd (1972): a nyelv procedurális közelítése (SHRDLU)
- ☐ Mesterséges intelligencia → számítógépes pszicholingvisztika
- ☐ A gépi fordítás nagy túlélői:
 - Systran (Toma, 1968 → EC, 1976 & Gachot, 1986)
 - Logos (vietnami háború → Wang/IBM/Sun, 1970)
 - Metal (Texas → Siemens, 1978)
- ☐ Gépi fordítás az USA-n kívül:
 - METEO, Eurotra, DLT és az „5. generációs japán álmom”
- ☐ Új fogalom: természetesnyelv-feldolgozás (natural language processing: NLP)

Új generatív elméletek és új gépi megoldások

- ❑ Chomsky „elejtett fonalának” felszedése (1978): a Bay Area-nyelvtanok - GPSG, LFG, HPSG
- ❑ „Frege számítógépesítése”: logikai szemantika és a „rule-to-rule” hipotézisre épülő gépi fordítás (Rosetta)
- ❑ Winograd nyelvi proceduralitása egy „kvázi-elmélet” formájában: *Language as a Cognitive Process* (1983)
- ❑ Elméleti áttörés: a reguláris nyelvtanok és a véges állapotú átmenethálók „újjászületése”: a kétszintes morfológia (Koskenniemi 1983)
- ❑ Megjelennek az első piaci alkalmazások: helyesírás-ellenőrzés, elválasztás - Macintosh, majd IBM PC (1985)
- ❑ A gépi fordítás „leszáll” a PC-re: PC Logos; Siemens Metal > Langenscheidt T1; Systran + Globalink, Kielikone, ProMT
- ❑ Függőségi és tudás-alapú paradigmák a nyelvtechnológiában: kognitív gépi nyelvelméletek, lexikális szemantika (→ WordNet 1985)

A statisztika „mindenhatósága” felé (és tovább)

- ☐ Chomsky újabb elméleteiből a transzformáció nem tűnik el
- ☐ A statisztika „beszáll” a nyelvi modellezésbe (v.ö. Chomsky)
- ☐ Szövegfeldolgozás a beszédtechnológia alapszereivel(1992)
- ☐ Van elég géppel feldolgozható szöveg: megszületik a korpusznyelvész (1995)
- ☐ Nyelv- és beszédfeldolgozás: „ebből igazi üzletet lehet csinálni!” (2000: L&H, SAIL, majd ScanSoft és Nuance)
- ☐ Kialakul a nyelvtechnológia fogalma és megjelenik a gépi beszédfordítás ígérete (2002)
- ☐ Az IBM mesterségesintelligencia- és nyelvtechnológiai „erő-demonstrációkat” tart: Deep Blue (1997) és Watson (2011)
- ☐ Egy ideig gyakorlati, majd már elvi probléma: még sincs elég adat bizonyos témákhoz és nyelvekhez (sparse data problem)
- ☐ Hibrid megoldások: lehet, hogy az ember is így csinálja?
- ☐ Vagy a szóbeágyazás és a neurális hálók oldják meg a problémát?

... de a korpusz-alapú közelítés sem annyira új, mint gondolnánk

Egy új grammatikai módszer van tehát megjelenőben, mely induktív módon halad, azaz a példákból kiindulva ismeri fel a szabályt. A grammatikát tehát az elolvasott, feldolgozott szövegek alapján építjük, úgy hogy a szabályokat a gép a példák segítségével állítja össze statisztikai következtetések útján. Ezáltal ez a módszer véget vet az előre megadott szabályok mechanikus alkalmazásának, és azt indukcióval pótolja. A szabályok így tárolódnak el a gép memóriájában, mert amit magunk találunk, azt jobban tudjuk, mint amit más mond vagy más tanultat velünk.



... de a korpusz-alapú közelítés sem annyira új, mint gondolnánk

„Simonyi új grammatikai módszert akar behozni, könyve inductive halad, azaz a példákból kiindulva tanítja a szabályt, nem pedig dogmatica. A grammaticát tehát valami olvasmány alapján akarja előadni, úgy hogy a szabályokat a tanár tanítványai közreműködésével vonhatja le ésszerű következtetések útján. Ilyenképp tehát ezen módszer véget vet a lelketlen magolásnak, és azt észfejlesztő inductióval pótolja. Eszerint a szabályok is mélyebben vésődnek be a gyermek emlékezetébe, mert amit magunk találunk, azt jobban tudjuk, mint amit más mond vagy más tanultat velünk.”

Riedl Frigyes: Simonyi kis nyelvtana (1882)



A nyelvtechnológia alapjai

2.

A szövegek kódolása és rendezése (meg egy kis nyelvstatisztika)

2019/2020. tanév, I. félév



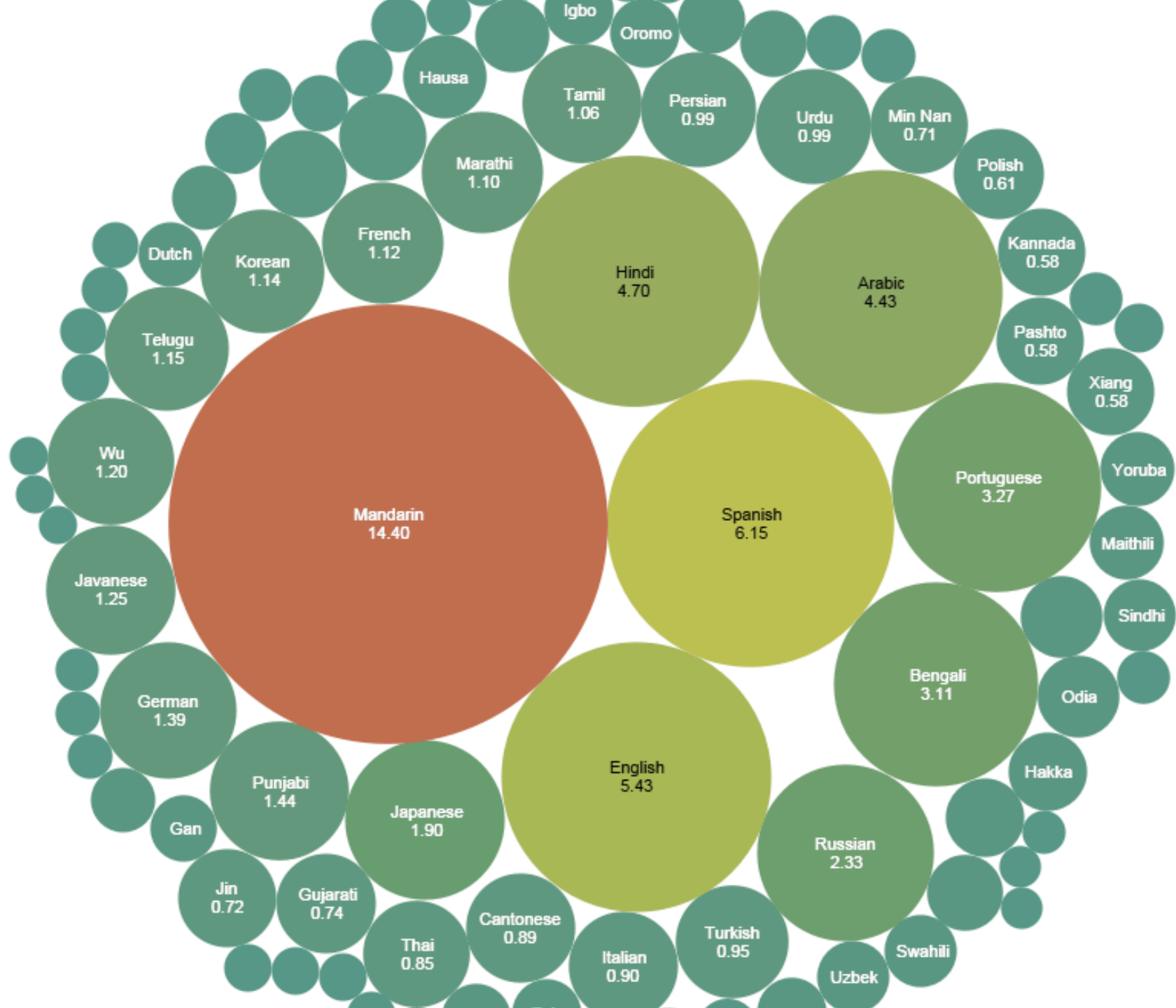
A nyelvekről



Nyelvek százmillió anyanyelvi beszélővel

(forrás: Ethnologue: Languages of the World, 2019)

	Nyelv	L1 beszélő	L1	L2 beszélő	L2	Összesen
1	Angol	379 millió	3	753 millió	1	1132 millió
2	Mandarin	918 millió	1	199 millió	5	1117 millió
3	Hindi	341 millió	4	274 millió	2	615 millió
4	Spanyol	460 millió	2	74 millió	10	534 millió
5	Francia	77 millió	14	203 millió	4	280 millió
6	Arab	-	-	274 millió	3	274 millió
7	Bengáli	228 millió	5	37 millió	14	265 millió
8	Orosz	154 millió	7	104 millió	7	258 millió
9	Portugál	221 millió	6	13 millió	16	234 millió
10	Indonéz	43 millió	29	155 millió	6	198 millió





A magyar nyelv „helye”

(forrás: Ethnologue: Languages of the World)

...

- 57. Cebuano (Fülöp-szk)
- 58. Sinhala (Sri Lanka)
- 59. Rangpuri (Bangladesh)
- 60. Északkelet-thai (Thaiföld)
- 61. Zhuang (Kína)
- 62. Malgas (Madagaszkár)
- 63. Nepáli (Nepál)
- 64. Szomáli (Szomália)
- 65. Khmer (Kambodzsa)
- 66. Madura (Indonézia)
- 67. Görög (Görögország)
- 68. Csittagóniai (Bangladesh)
- 69. Haryanvi (India)
- 70. Magahi (India)
- 71. Dekkan (India)
- 72. Magyar

...



A nyelvek területi megoszlása

(forrás: Ethnologue: Languages of the World)

Földrész	Élő nyelvek		Beszélők		
	száma	%	száma	%	ny -kénti átlag
Afrika	2 146	30,2	789 138 977	12,7	367 726
Amerika	1 060	14,9	51 109 910	0,8	48 217
Ázsia	2 304	32,4	3 742 996 641	60,0	1 624 565
Európa	284	4,0	1 646 624 761	26,4	5 797 975
Óceánia	1 311	18,5	6 551 278	0,1	4 997
Összesen	7 105	100.0	6 236 421 567	100.0	877 751

A nyelvek anyanyelvi beszélők szerinti megoszlása

(forrás: Ethnologue: Languages of the World)

Populáció	Élő nyelvek			Beszélők		
	száma	%	Σ %	száma	%	Σ %
100 000 000 to 999 999 999	8	0,1	0 ,%	2 308 548 848	40,20	40,20%
10 000 000 to 99 999 999	82	1,2	1,3%	2 480 078 977	39 42	79,62%
1 000 000 to 9 999 999	304	4,3	5,5%	951 916 458	14,55	94,18%
100 000 to 999 999	895	13,0	18,6%	283 116 716	4 75	98,84%
10 000 to 99 999	1 824	26,4	45,0%	60 780 797	1 01	99,86%
1 000 to 9 999	2 014	29,2	74,1%	7 773 810	0 13	99,99%
100 to 999	1 038	15,0	89,2%	461 250	0 00	99,99%
10 to 99	339	4,9	94,1%	12 560	0 00	99,99%
1 to 9	133	1,9	96,0%	521	0 00	100,00%
Ismeretlen	277	4,0	100,0%			
Összesen	6 909	100,0		5 959 511 717	100,0	



A nyelvi gépi feldolgozásának szintjei

A szó

- ❑ Definíciója nincs
- ❑ **Fonológiai szó**
(pl. *ház+am* fonológiai szó,
de az összetett szavak nem ilyenek)
- ❑ **Lexikai szó (lexéma)**
(pl. a *számítógépezés* szó nem lexéma,
mert levezethető alkotóelemei jelentéséből)
- ❑ **Morfológiai szó**
(pl. *láttam volna*)
- ❑ **Szintaktikai szó**
(pl. *szép+ség+e+im*,
de az igekötős ige nem ilyen,
mert a szintaxis megváltoztatja az alakját)

Még nehezebben azonosítható nyelvi egységek

- ❑ **Morféma:** a nyelv legkisebb jelentéssel bíró alkotórésze; nincs természetes határa
- ❑ **Szószerkezet, frázis:**
nincs természetes határuk; egyrészt lexémák, másrészt mondat alatti egységek
- ❑ **Megnyilvánulás, mondat:**
írásban egyszerűbb (nem nehéz a formális definíció)
- ❑ **Szöveg:**
mondatok sora, melyben az összefüggések túlmutatnak a mondatnyelvtanokon
- ❑ Megkülönböztetendő: **type** és **token**



Hang, fonéma, betű, karakter

Fonéma

❑ Szembenállás

Ha két beszédhang egyikének a helyére behelyettesítjük a másikat, és más szótagi egységet kapunk eredményül, a két beszédhangot két különböző fonémához soroljuk:

géz, kéz, méz, néz, réz -> /g/, /k/, /m/, /n/, /r/

❑ Kiegészítő eloszlás

Ha két beszédhang mindig csak egymást kölcsönösen kizáró környezetekben fordul elő, feltételesen ugyanazon fonémához sorolhatjuk őket:

pl. *hó : doh* -> ún. allofónok

Fonéma (2)

❑ Fonetikai hasonlóság

Az előző két ismerv önmagában nem zárja ki a téves fonémaazonosítás lehetőségét,
pl. a veláris nazális nemcsak a dentális nazálissal, hanem más hangokkal is a kiegészítő eloszlás viszonyában van, mégsem soroljuk sem pl. a /h/-hoz vagy a /j/-hez, hanem csakis az /n/-hez

❑ Szabad váltakozás

Ha két beszédhang ugyanazon környezetben egymással helyettesíthető anélkül, hogy a szóban forgó szótári elem azonossága veszélyben forogna, ugyanazon fonéma szabad változataival állunk szemben

Fonetikus ábécé

- ☐ Az IPA a latin (angol) ábécén alapszik
- ☐ A legtöbb latin betűs nyelv ugyanúgy ejt többet is: [b], [d], [f], [h], [k], [l], [m], [n], [p], [t], [v] stb.
- ☐ Vannak közöttük olyanok, melyek kiejtése a legtöbb latin betűvel író nyelvben eltér: [j], [r], [y] stb.
- ☐ Vannak olyan jelek, melyeket a latin betűk átalakításával hoztak létre, vagy írott formájukat használják: [ɛ], [ɜ] stb.
- ☐ Bizonyos elemeket a görög ábécéből vettek: [β], [ɣ], [ɛ], [θ], [χ], [v] stb.
- ☐ A magyar nyelv 14 magánhangzó- és 24 mássalhangzó-fonémája az IPA rendszerben is megjeleníthető

IPA-karakterkészletek

- ❑ SIL: Summer Institute of Linguistics
- ❑ Charis SIL: teljes IPA, tonális mellékjelek és sok más kevésbé általános fonetikai jel is
- ❑ Doulos SIL (a Charis SIL, csak bold és italics nélkül)
- ❑ SIL93: a korábbi SIL IPA93 betűtípusok
(Manuscript és Sophia) Unicode-ban újrakódolva

Az „extra kódolás” igénye

- ❑ Nem elég a latin ábécé
- ❑ Diakritikus jelek: „kidíszítjük” a meglevőket
- ❑ Multigráfok: kombináljuk a meglevőket
- ❑ Pl. amit a magyar „s” betűvel jelöl:
 - s magyar
 - sh angol, albán, cigány
 - ch francia, portugál
 - sch német
 - si gael, indonéz
 - sci olasz
 - skj svéd, norvég
 - š cseh, szlovák, horvát, szlovén, lett, litván, észt
 - sz/ś lengyel
 - ş román, török
 - Ŝ eszperantó

Európai ékezetes betűk (1)

Nyugat

à á â ã ä å æ ã è é ê ë ì í î ï ò ó ô

õ ö ø œ ù ú û ü û w ý ÿ ŷ

ç ħ ŋ ŕ ſ ſ ð þ ß

Európai ékezetes betűk (2)

Kelet

áâäāąéëēėęĕĭîĩıóôõö
ōóúûüūŭůųý
çćĉčċďđġĝġġġĥĥĵķĺłł
ńņñŕŕŕŕŝŝŝŝt'žžžž

Diakritikumok multigráfos kódolásai

❑ e-mail:

arvizturo tukorfurogep
(fokabel???)

❑ távirati:

aarviiztuerooe tuekoerfuuroogeep

❑ repülő ékezetes:

a'rví'ztu"ro" tu:ko:rfu'ro'ge'p

❑ számkódos:

a1rví1ztu3ro3 tu2ko2rfu1ro1ge1p

❑ TeX:

\'arv\'ízt\Hur\Ho t\'uk\'orf\'ur\'og\'ep

❑ HTML:

árvíztűrő

Karakter, karakterkód, karakterkészlet

- ☐ **Karakter:** absztrakt objektum, a számítógépen tárolt szövegek legkisebb, tovább nem bontható egysége, Egy karakter egy betű, szám, írásjel, speciális jel, illetve egyetlen egyszerűbb szövegformázó parancs ábrázolására alkalmas
- ☐ **Karakterkód:** a karakternek mint absztrakt objektumnak megfeleltetett számkód, amely számítógépen közvetlenül ábrázolható
- ☐ **Karakterkészlet:** karaktereket és számkódjaikat összerendelő táblázat, Meghatározza, hogy milyen karakterek használhatók, és azt is, hogy az egyes karakterkódokat miképpen kell értelmezni. A karakterkészlet megjeleníthető karaktereket és speciális műveleteket előíró vezérlőkaraktereket tartalmaz
- ☐ **Kódlap:** több karakterkészletet is kezelő számítógéprendszerben egy saját számkódjával azonosított karakterkészlet
- ☐ **Betűkészlet:** olyan táblázat, amely adott karakterkészlet(ek) megjeleníthető karaktereihez karakterképeket rendel, amelyek a számítógép képernyőjén vagy nyomtatásban jelennek meg. Az alkalmazott betűkészlet a szöveg számítógépbeli értelmezésére – például a szöveget nyelvi szempontból feldolgozó programokra – nincs befolyással, csak annak megjelenítésekor használatos

Kódlapok

ASCII kódtábla:

ASCII-Zeichensatz										
+	0	1	2	3	4	5	6	7	8	9
30				!	"	#	\$	%	&	'
40	(	)	*	+	,	-	.	/	0	1
50	2	3	4	5	6	7	8	9	:	;
60	<	=	>	?	@	A	B	C	D	E
70	F	G	H	I	J	K	L	M	N	O
80	P	Q	R	S	T	U	V	W	X	Y
90	Z	[	\	]	^	_	`	a	b	c
100	d	e	f	g	h	i	j	k	l	m
110	n	o	p	q	r	s	t	u	v	w
120	x	y	z	{		}	~			

Kódlapok (2)

Az ISO 8859-2 (Latin-2) kódtábla:

- ☐ magyar
- ☐ német
- ☐ horvát
- ☐ lengyel
- ☐ román
- ☐ szlovák
- ☐ szlovén
- ☐ szerb
- ☐ cseh
- ☐ albán

iso-8859-2										
+	0	1	2	3	4	5	6	7	8	9
160		À	Á	Â	Ã	Ä	Å	Æ	Ç	Š
170	Ð	Ė	Ž	-	Ž	Ž	°	ª	¸	ł
180	ˆ	İ	Š	ˆ	ˆ	š	š	ť	ž	˘
190	ž	ž	Ř	Á	Â	Ĥ	Ä	Ĺ	Ć	Ç
200	Č	É	Ê	Ë	Ě	Í	Î	Ď	Đ	Ñ
210	Ň	Ó	Ô	Õ	Ö	×	Ř	Ů	Ú	Ů
220	Ü	Ý	Ť	ß	ř	ă	â	ă	ä	î
230	ć	ç	č	é	ę	ë	ě	í	î	ď
240	đ	ñ	ň	ó	ô	õ	ö	÷	ř	ů
250	ú	ü	ü	ý	ı	'				

A Unicode és a UCS

- ❑ UCS: Universal Character Set
- ❑ Unicode szabvány: 16 biten tárolt síkokra osztja a – szabvány legutóbbi változatában rögzített – kb. 100 ezer karaktert
- ❑ Minden beszélt, holt és egyéb nyelv (pl. Braille) karakterei (betűk, számok, írásjelek stb.)
- ❑ Minden egyes karakternek egyedi száma és neve van:
 - U+0041 Latin capital letter A
 - U+0151 Latin small letter o with double acute
- ❑ Eredetileg 2 bájton (16 biten) tárolta az adatokat (Basic Multilingual Plane) , de ez nem volt elég, ezért 4 bájtusra (32 bitesre) bővítették

A Unicode és a UCS

- ❑ UTF-8 (Unicode Transformation Format): változó hosszúságú kódolással (1–6 bájt) képezi le a Unicode karaktertáblát, pl. 1 bájton tárolt kódjai az ASCII-nak felelnek meg, így a latin betűs UTF-8 kódolású szövegek a régi ASCII környezetben is olvashatóak maradnak
- ❑ UTF-16: 1 112 064 karaktert kódol
- ❑ UCS-2 (elavult), 2 bájton (UTF-16 részhalmaza)
- ❑ UCS-4: 4 bájton (31-bit) = UTF-32, ISO 10646

A Unicode és a UCS

- ❑ 1980-as évek: igény egy egységes, minden karaktert leíró kódtáblára
 - **ISO 10646** – az ISO szervezet szabványa
 - **Unicode** – amerikai cégek konzorciuma
- ❑ 1991: egyesítik a két projekt eredményeit, de megmarad a két szervezet
- ❑ Teljesen kompatibilisek azóta is!
- ❑ Különbségek is vannak: az ISO 10646 csak karakterkészlet, a Unicode-ban viszont léteznek további kiegészítések, pl. egyeztetési szabályok

Egészen pontosan...

- ☐ ISO/IEC 10646-1:1993 $\approx$ Unicode 1.1
- ☐ ISO/IEC 10646-1:2000 $\approx$ Unicode 3.0
- ☐ ISO/IEC 10646-2:2001 $\approx$ Unicode 3.2
- ☐ ISO/IEC 10646:2003 $\approx$ Unicode 4.0
- ☐ ISO/IEC 10646:2003 és az 1. kiegészítés $\approx$ Unicode 4.1
- ☐ ISO/IEC 10646:2003 és
az 1. kiegészítés,
a 2. kiegészítés, és részben
a 3. kiegészítés $\approx$ Unicode 5.0
- ☐ ...
- ☐ Unicode 12.0 (2019. június)
- ☐ Unicode Character Database:
<http://www.unicode.org/Public/UNIDATA/>
- ☐ Továbbá: <http://www.decodeunicode.org/> (137 374 karakter)



Ábécék és rendezési elvek

A nyelvtechnológia alapjai - 2019. szeptember 18.

Ábécék és rendezés

- A kiejtés szerinti írás (?)
- Az ábécébe rendezés elvei
- Átjárás ábécék között: transzliterálási problémák
- A könyvtári rendezés

Az olyan sajátos célú munkák (lexikonok, enciklopédiák, atlaszok és térképek névmutatói stb.), amelyekben magyar és idegen nyelvű szóanyag erősen keveredik egymással, rendszerint az úgynevezett általános latin betűs ábécét követik: *a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z*. Ebben a rendszerben mind a magyar, mind az idegen többjegyű betűknek minden egyes eleme külön, önálló egységnek számít, és a besorolás nincs tekintettel sem a magyar ékezetekre, sem az idegen betűk mellékjeleire. Az ilyenfajta munkák szavainak betűrendbe sorolása bonyolult feladat, itt például a több szóból álló egységeket vagy az egész kapcsolatot, vagy csak az első elem alapján szokás besorolni, ahogyan a könnyebben kezelhetőség érdeke vagy egyéb szempont kívánja. A követendő eljárást a könyvtárügy és a szakirodalmi tájékoztatás (dokumentáció) területén szabvány (MSZ 3401) írja elő, illetőleg belső szakmai útmutatók szabályozzák.



A magyar akadémiai rendezés (1)

A különböző betűvel kezdődő szavakat az első betűk ábécébeli helye szerint állítjuk rendbe, illetőleg keressük meg

Példa:

*acél, cukor, csók, gép, hideg, kettes, olasz,
öröm, Nagy, nyúl, remény, sokáig, szabad,
Tamás, vásárol*

A magyar akadémiai rendezés (2)

Az egyjegyű betűt teljesen elkülönítjük az azonos írásjeggyel kezdődő, de külön mássalhangzót jelölő kétjegyű (ill, háromjegyű) betűtől: mindig az egyjegyű betű van előbb

Példa:

*cudar, cukor, cuppant, csalit, csata, Csepel
dac, domb, duzzog, dzsem, dzsungel, Zoltán, zongora,
zúdul, zsalu, zseni*

A magyar akadémiai rendezés (3)

A többjegyű betűk kettőzött változatait sohasem az egyszerűsített alakok szerint soroljuk be a betűrendbe, hanem a megkettőzött betűt mindig két külön betűre bontjuk, s így soroljuk a szót a megfelelő helyre ($ccs = cs + cs$; $ggy = gy + gy$; $ddzs = dzs + dzs$)

Példa:

*kas, Kassák, kastély, kaszinó, kassza, kaszt
mennek, menü, menza, Menyhért, mennyi,
nagy, naggyá, nagygyakorlat, naggyal, nagyít*

A magyar akadémiai rendezés (4)

A magánhangzók rövid és hosszú változatát jelölő betűk (*a* – *á*, *e* – *é*, *i* – *í*, *o* – *ó*, *ö* – *ő*, *u* – *ú*, *ü* – *ű*)

a kialakult szokás szerint mind a szavak elején, mind pedig a szavak belsejében azonos értékűnek számítanak a betűrendbe sorolás szempontjából.

A magánhangzó hosszú változatát tartalmazó szó tehát meg is előzheti a rövid változatút

Példa:

ír, Irak, Irán, írandó, íránt, író, iroda, irónia

A magyar akadémiai rendezés (5)

A rövid magánhangzós szó kerül viszont előbbre olyankor, ha két szó betűsora csak az azonos magánhangzók hosszúsága tekintetében különbözik

Példa:

*égbolt, Eger, egér, éger, égés, egész,
keres, kérés, koros, kóros, kórós,
szel, szél, szelel, szeles, szelés, széles, szelet*

A magyar akadémiai rendezés (6)

A különírt elemekből álló szókapcsolatokat és az egybeírt vagy kötőjellel kapcsolt összetételeket minden tekintetben olyan szabályok szerint soroljuk betűrendbe, mint az egyszerű szavakat, A szóhatárokat tehát nem vesszük figyelembe. Ugyanez a szabály érvényes a közzavak közé besorolt tulajdonnevekre is

Példa:

*Kis részben, kissé, Kiss Ernő, kis sorozat,
kissorozat-gyártás, kis számban, kistányér, kis virág,
tiszafa, Tiszahát, Tisza Kálmán, Tisza menti, Tiszántúl, Tisza-
part, tiszavirág, tiszt*

A magyar akadémiai rendezés (7)

A betűrendbe soroláskor a szokásostól némiképp eltérően kezeljük a régies írású magyar családnevekben, valamint az idegen szavakban és tulajdonnevekben előforduló régi magyar, illetőleg idegen betűket

Példa:

Czibere, Czuczor, Cházár, Császár, Cselényi

Régies és idegen hangjelölés

- ☐ *aa* = á: *Gaal* [e : gál]
ch = cs: *Madách* [e : madács]
ee = é: *Veér* [e : vér]
cz = c: *Czuczor* [e : cucor]
eő = ö: *Eötvös* [e : ötvös]
s = zs: *Jósika* [e : józsika]
ew = ö: *Thewrewk* [e : török]
th = t: *Csáth* [e : csát]
oo = ó: *Soós* [e : sós]
ts = cs: *Takáts* [e : takács]
y = i: *Kölcsey* [e : kölcsei] stb
w = v: *Wesselényi* [e : veselényi] stb
- ☐ angol: *Greenwich* [e : grínics], *joule* [e : dzsúl]
- ☐ cseh: *Dvořák* [e : dvorzsák], *Škoda* [e : skoda]
- ☐ francia: *Eugène* [e : özsen], *Nîmes* [e : nim]
- ☐ német: *Schäfer* [e : séfer], *Werther* [e : verter]
- ☐ olasz: *Bologna* [e : bolonya], *quattrocento* [e : kvatrocsentó]
- ☐ portugál: *Guimarães* [e : gimarajs], *piranha* [e : piránya]
- ☐ román: *Tîrgoviște* [e : tirgoviste], *piața* [e : piaca]
- ☐ szerbhorvát: *Đurić* [e : gyúrity], *Živogošće* [e : zsvogostye]

Továbbá...

1-szer 1 az 1

II. János Pál

10 kérdés a személyi jövedelemadóról

11. Távhő-konferencia

XIX. század

88 kérdés az AIDS-ről

8000 germanizmus

ABALKIN, L. I.

Aczél György

ACÉLOK



Nyelvstatisztikai alapok

Betűstatisztikák

- ❑ Hangstatisztikák, betűstatisztikák
- ❑ Titkosírás-megfejtések
- ❑ Morse-ábécé, írógép-billentyűzet
- ❑ A nyelvek tipizálhatók n-gráfjaik alapján (pl. mássalhangzó-torlódással kezdődő szavaink régen egyáltalán nem voltak, és ma is csak bizonyosak vannak)
- ❑ Információ (Shannon): kisebb valószínűségű üzenetnek az információtartalma nagyobb, és független események együttes információtartalma az események információtartalmának összege

Egy kis információelméleti háttér

- ❑ Shannon (1949): tfh egy kommunikációs folyamatban veszünk részt, melynek csatornáján az X halmaz jeleiből összetevődő véges sorozatok, üzenetek áramlanak
- ❑ Ha elég sok üzenet áll rendelkezésre, akkor mérni tudjuk azt a $p(x)$ valószínűséget, hogy adott $x \in X$ elem milyen gyakran fordul elő várhatóan egy üzenetben
- ❑ Az x jel információtartalma: $I(x) := -\log_2[p(x)]$
- ❑ Az $\underline{X}$ üzenet információtartalma is:
ha $\underline{X} = (x_1, x_2, \dots, x_j) \in X^j$ jelek sorozata, akkor
$$H(\underline{X}) = H(x_1, x_2, \dots, x_j) = I(x_1) + I(x_2) + \dots + I(x_j) =$$
$$= -\log_2 p(x_1) - \log_2 p(x_2) - \dots - \log_2 p(x_j) =$$
$$= -\log_2 [p(x_1)p(x_2)\dots p(x_j)]$$
- ❑ Shannon (1951): Prediction and Entropy of Printed English

n-gramok

- ❑ n-gram: egymás után következő n db karakter egy szövegben
- ❑ Az n-gramok lehetnek szavak is, de nem feltétlenül azok
- ❑ Az n-gramok használatának előnye, hogy nyelvfüggetlen és hibatűrő megoldások megvalósítására igen alkalmas
- ❑ Stochastic Language Models (N-Gram) Specification:
<http://www.w3.org/TR/ngram-spec/>
- ❑ Microsoft Web N-gram Services:
<http://research.microsoft.com/en-us/collaboration/focus/cs/web-ngram.aspx>
- ❑ Google N-gram Viewer: <http://books.google.com/ngrams>

Magyar betűstatisztika (1-gramok)

	Gyakoriság	Információ bitben
A	9,35	3,43
Á	3,72	4,77
B	1,72	5,87
C	0,60	7,40
D	1,71	5,90
E	9,71	3,37
É	3,87	4,71
F	0,88	6,87
G	3,55	4,83
H	1,23	6,37
I	4,39	4,53
J	1,21	6,39
K	5,35	4,24
L	6,30	4,00
M	3,92	4,69

	Gyakoriság	Információ bitben
N	5,47	4,21
O	4,47	4,50
Ö	2,14	5,57
P	1,04	6,61
Q	0,00	?
R	4,22	4,58
S	6,57	3,94
T	7,87	3,68
U	1,29	6,30
Ü	0,93	6,77
V	1,81	5,81
W	0,00	?
X	0,01	13,33
Y	2,21	5,52
Z	4,46	4,50

n-gramok a gyakorlatban

Küpfmüller (1954) szöveggenerálása

1-, 2-, 3- és 4-gramokkal:

1. EME GKNEET ERS TITBL VTZEN
2. AUSZ KEINU WONDINGLIN DUFRN
3. PLANZEUNDGES PHIN INE UNDEN
4. ICH FOLGEMASZIG BIS STEHEN DISPONIN

n-gramok a magyarban

1. nórmígenytnaeőettesy sdrúk üze neée issgnis k, őentetnsmrddtca.
2. Bégiz érij, ercszépót, ekalfönembenyett kmáb os alőrre.
3. Ezenlapoltike lehátszökmég hog, aztradon két. Korcsant ván.
4. Hárone – Az a léhez, meg, akarózsák hogy öregész egény látán, járnagyon kérdeket egy emberet. Tordárd sem tan.
5. Senki adsaadott, hogy egy ki tornak, a kapta a mielőtte szor ismerni most vele aótsap a eipővel. - Ó, biztosítás úr csak társallatság vele. Hallj ott.
6. Történik. Telt év óta meglátta bevallott kissé lehajtott és mindig, légez erligyúneáenl eltűnt.

Megfigyelhető: a 4. után itt nem javul igazán a minőség, Miért?

Mert kicsi tanítókorpusz esetén (esetünkben 8 Mbájt) a négyesnél nagyobb csoportok eloszlásának megbízhatósága már nem elegendő

Zipf-törvény

- ❑ G. K. Zipf az 1930-as években vette észre: ha egy szövegben összeszámoljuk az egyes szavak előfordulási gyakoriságát, majd sorba rendezzük őket gyakoriságuk alapján, és ábrázoljuk a gyakoriságot rang szerint (azaz annak a függvényében, hogy az illető szó hányadik a sorban), akkor (közelítéssel) egy hatványfüggvény szerint lecsengő görbét kapunk -1-hez közeli kitevővel
- ❑ Tehát a lista i -edik elemének gyakorisága: $R_i \sim \frac{1}{i^\alpha}$
(α a korpuszra jellemző, 1 körüli konstans)
- ❑ Következmény: ha a leggyakoribb szavakat tartjuk csak meg, a korpusz nagy része megmarad
- ❑ A jelenség az emberi nyelvek univerzális tulajdonságának bizonyult: pl. hasonló viselkedést kapunk, ha nem a szavakat, hanem az összefüggő szöveg n -gramjait vizsgáljuk

Heaps-törvény

- ❑ Heaps (1978): egy N szóból álló korpusz esetén a különböző szavak száma:

$$V \approx K \cdot N^{\beta}$$

K : a szövegtől függő konstans ($10 \leq K \leq 100$)

β : a szövegtől függő konstans

(angolnál: $0,4 \leq \beta \leq 0,6$ - magyarnál: $0,6 \leq \beta \leq 0,7$)

- ❑ Következmény: egy korpuszban a különböző szavak száma új dokumentumok hozzáadásával folyamatosan növekszik (pl. az elírásoknak, a tulajdonneveknek és az új szavaknak köszönhetően), de ez a növekedés szublineáris, azaz az új szavak nem lesznek gyakori szavak



További statisztikai nyelvtörvények

- ☐ Mandelbrot: összefüggés van a szavak hosszúsága és rangja között is, ui. a rövidebb szavak minden nyelvben sokkal nagyobb gyakorisággal fordulnak elő, mint a hosszabbak, méghozzá a rangjuk a hosszal fordítottan arányos
- ☐ Észrevétel: ha a történelmi fejlődés során egy szó gyakorisága megnő, rendszerint meg is rövidül
- ☐ Következmény: nyelveink legősibb rétegét (pl. a testrészek neveit) szinte kivétel nélkül egytagú, két-három fonémából álló szavak alkotják
- ☐ „*Matematikai igazság, hogy kinek két szótagú neve van, fél annyi kortessel választhatik meg, mint az, kinél négy szótagot kell kiáltani*”

(Eötvös: A falu jegyzője)

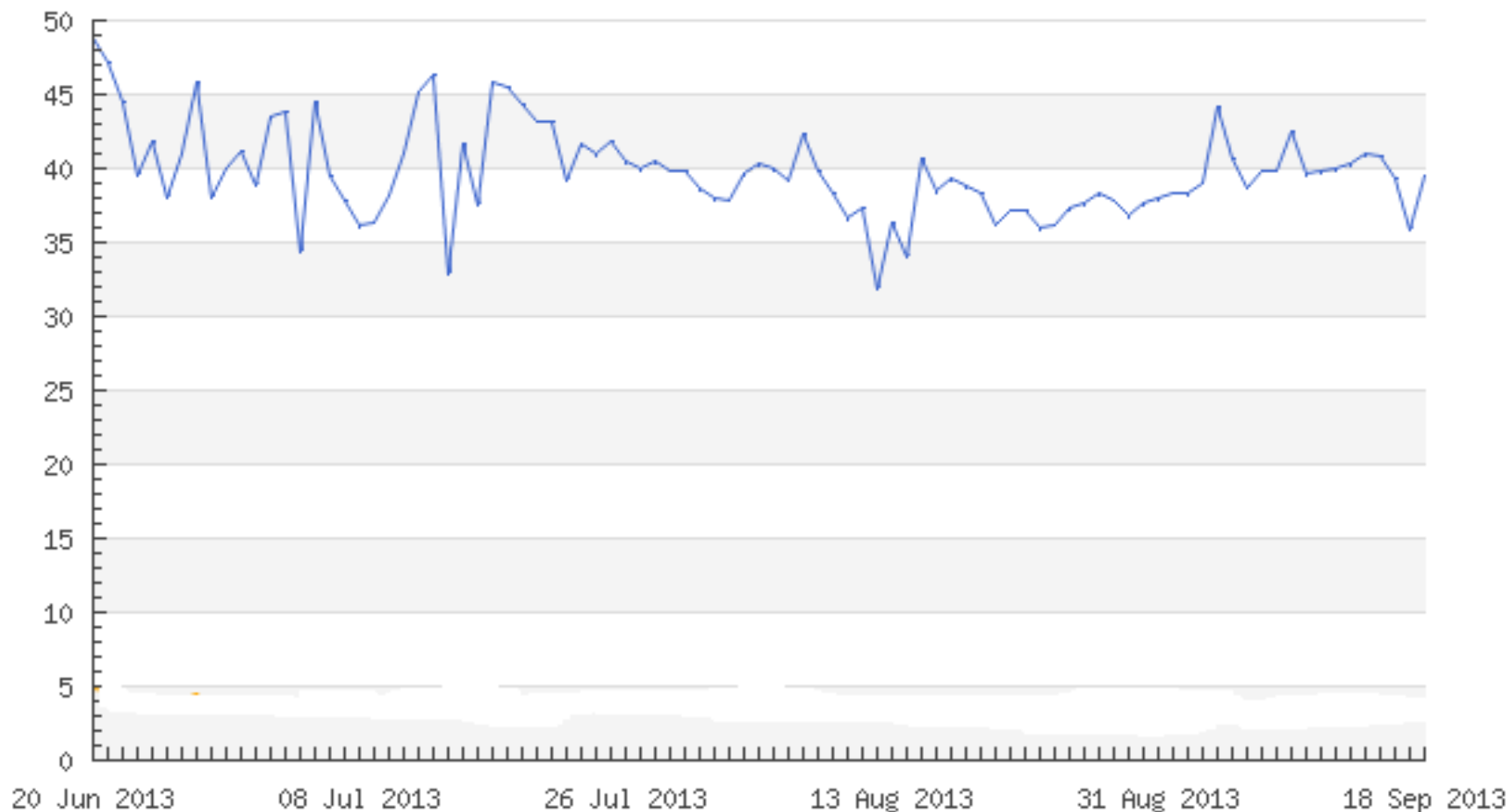
Egy érdekességek

Egy angol nyelvemléktudásai szerint nem számít
melyik szerzőnek vannak a betűk egy szóban,
az nyelvben fontos dolog, hogy az első és az
utolsó betűk a helyükön legyenek. A többi betű
lehet teljesen összevisszaság, mégis probléma
nélkül olvasható a szöveg. Ennek oka, hogy nem
olvassuk el minden betűt egyenként, hanem a
szót olvasunk a maga egészében.



Szövegek és nyelvek a weben

Az indexelt web mérete (milliárd weboldalba)

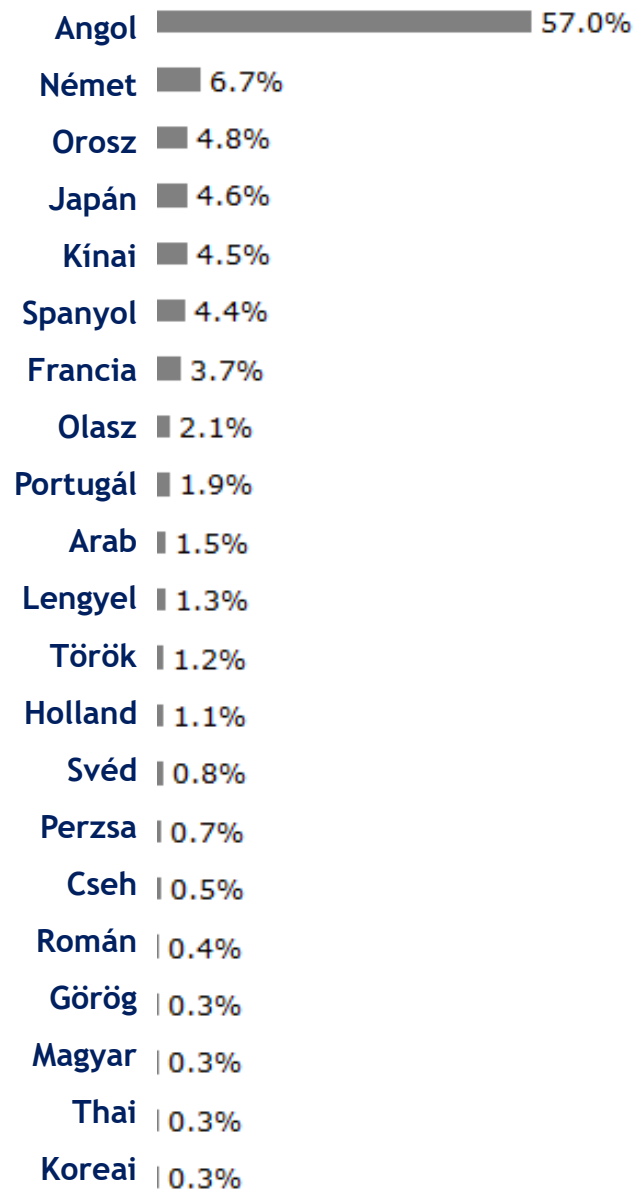


A web és a nyelvek

Nyelv	Internet- használók	Internet- penetráció	Növekedés (2000-)	Internet- használók a világ %-ában	Népesség
Angol	565 004 126	43,4 %	301,4 %	26,8 %	1 302 275 670
Kínai	509 965 013	37,2 %	1478,7 %	24,2 %	1 372 226 042
Spanyol	164 968 742	39,0 %	807,4 %	7,8 %	423 085 806
Japán	99 182 000	78,4 %	110,7 %	4,7 %	126 475 664
Portugál	82 586 600	32,5 %	990,1 %	3,9 %	253 947 594
Német	75 422 674	79,5 %	174,1 %	3,6 %	94 842 656
Arab	65 365 400	18,8 %	2501,2 %	3,3 %	347 002 991
Francia	59 779 525	17,2 %	398,2 %	3,0 %	347 932 305
Orosz	59 700 000	42,8 %	1825,8 %	3,0 %	139 390 205
Koreai	39 440 000	55,2 %	107,1 %	2,0 %	71 393 343
Σ	1 615 957 333	36,4 %	421,2 %	82,2 %	4 442 056 069
A többi	350 557 483	14,6 %	588,5 %	17,8 %	2 403 553 891
Összesen	2 099 926 965	30,3 %	481,7 %	100,0 %	6 930 055 154



A weblapok nyelvei



Következmények

1. Az internetes tartalom több mint a fele angol
2. Az internethasználók háromnegyedének az anyanyelve nem angol
3. $1+2 \rightarrow$ A nem-angol anyanyelvűek világában a web angol tartalmának egyre nagyobb része jelenik meg
4. $3 \rightarrow$ Aki a webet olvassa, az növekvő számban nem-angol anyanyelvű; aki pedig a webes tartalmat közzéteszi, az csökkenő részben angol anyanyelvű
5. $4 \rightarrow$ A nem-anyanyelvűek nagy része nem ismeri fel a nyelvi hibákat és pontatlanságokat

Hosszú távú következmény:

mi lesz a statisztikai alapú nyelvi programokkal?



3.

Számítógépes morfológia

2019/2020. tanév, I. félév



Mini-nyelvészet

Morfológiatipológia és számítógépes közelítése

- ☐ A szóalakok a jelentésviszonyokat a világ nyelveiben többféle módon fejezik ki
- ☐ Izoláló nyelv
(pl. kínai, de már lassan az angol is!)
- ☐ Flektáló nyelv
(pl. német, szláv nyelvek – „túlterhelt” szóalakok)
- ☐ Agglutináló nyelv
(toldaléksorok)
- ☐ Inkorporáló (poliszintetikus) nyelv
pl. csukcs, aleut, inuktikut:
Parismunngaujumaniralauqsimanngittunga
Paris+mut+nngau+juma+niraq+lauq+si+ma+nngit+jun
(=Sose mondtam, hogy Párizsba akartam volna menni.)
- ☐ Konkatenatív és nem-konkatenatív morfológiák
(pl. arab, héber, máltai)

Morfológiai alapfogalmak

- ☐ morf
bagoly, ló, megy, ...
-hoz, -nak, -ért, ...
- ☐ allomorf
bagoly-/bagly-, ló-/lov-, megy-/men-/me-/mé-, ...
-hoz/-hez/-höz, -nak/-nek, (-ért), ...
- ☐ morféma
bagOly, lŌV, mEGY, ...
-hŌz, -nAk, -ért, ...
- ☐ szuppletív alakok
go – went, gut – besser – best, volt – van – lesz
- ☐ részlegesen szuppletív alakok
France – French – Franco-

Morfológiai alapfogalmak (2)

- ☐ inflexió = ragozás
house+s, ház+ak
- ☐ deriváció = képzés
dis+establish+ment, pázmány+os+od+ó
- ☐ igeragozás
(konjugáció)
- ☐ névszóragozás
(deklináció)
- ☐ paradigma
(maga a toldalékok egy adott tőtípushoz kapcsolódó rendszere)
- ☐ morfológiai osztályok
(a tövek viselkedési osztályai inflexióik szerint)
- ☐ lexikális osztályok
(nem feltétlenül formai kritériumokra épülő osztályok,
pl. a nemek a franciában)

Mit jelölünk morfológiai úton?

- ☐ szám
- ☐ személy
- ☐ eset
- ☐ nem
- ☐ hasonlítás: alapfok, közép fok, felső fok, túlzó fok
- ☐ idő (rég- és közel)múlt (sőt: elbeszélő, összetett, *-nd*), jelen, jövő (?)
- ☐ mód: kijelentő, felszólító, feltételes, kötő, ...
- ☐ cselekvő, műveltető, szenvedő, visszaható, ...
- ☐ igenevesítés (nem-finit alakok létrehozása): gerundium, participium, nomen actionis, infinitivus (főnévi, melléknévi, határozói, sőt igei igenevek)
- ☐ de pl. az aspektus a magyarban ritkán jelölt morfológiailag
 - nem jelöltek: a folyamatos cselekvő (ami megszakítható), pl. *keres, csinál*), a folyamatos nem-cselekvő (ami nem megszakítható), pl. *virágzik, tud*
 - lehetnek jelöltek: a befejezett ige (az igekötő „perfektál”), pl. *megindul, kitalál*, de: *villant*
- ☐ ...

Hogy jelöl a morfológia?

- ☐ affixum (=toldalék)
- ☐ szuffixum (=végződés)
- ☐ infixum
(valami hasonló: *ház+a+m*, *ház+a+i+m*)
- ☐ cirkumfixum
(valami hasonló: *ge+wander+t*, *leg+nagy+obb*)
- ☐ reduplikáció
tagalog: *sulat (ír)*, *susulat (írni fog)*
- ☐ klitikumok:
proklitikum (*a*, *dr.*)
enklitikum (*-e*)
- ☐ összetételek
Leben+s+versicherung+s+gesellschaft+s+angestellter

Atipikus/ritkább/bonyolultabb esetek a lexiko-morfológiában

- ☐ Inflexió: *oxen*, *teeth*, *formulae*, *cherubim*, *criteria*, *indices*, *mafiosi*
- ☐ Deriváció: *jarda*, *lövölde*, *óvoda*; *bölcsőde*
- ☐ Összetétel: *blueberry*, *strawberry*, *raspberry*, *cranberry*; *esernyő*
- ☐ Furcsaságok (morfológiai idiómák?): *man-of-war*, *ládafia*
- ☐ Zárójelezési paradoxon a morfológia és a szintaxis határán:
((*electrical engineer*))ing)
((un(*grammatical*)))ity)
((*magyar nyelv*))ű)
- ☐ ((*barokk fuvol*))ista) vs. (*első* ((*fuvol*))ista))
((*haza(ad)*))ás) vs. (*szabad*((*rúg*))ás))
- ☐ A szemantikai viszonyok nehezen felismerhetők az összetételekben:
mosó|nő, *mosó|ruha*, *mosó|szappan*, *mosó|teknő*, *mosó|konyha*, ...



A morfológiai jelenségek formális leírása felé

A számítógépes morfológia célja

- ❑ A számítógépes morfológia olyan számítógépes technológiák és algoritmusok kialakításával foglalkozik, melyek segítségével különféle nyelvű toldalékolt szóalakok elemzése és generálása megoldható
- ❑ A számítógépes morfológia az írott alakokkal foglalkozik
- ❑ 1983: Koskenniemi vs. Winograd
- ❑ Szóalak-felismerés: a program visszaadja a toldalékolt szóalakból a szótári tövet (lemmatizálás), annak szófaját és a megjelenített nyelvtani információt
- ❑ Szóalak-generálás : a helyes szóalak előállítása a szótő és a morfológiailag releváns nyelvtani információ segítségével

Számítógépes morfológiai elemzés és generálás

❑ Példa morfológiai elemzésre:

1. Tövesítés (lemmatizálás): **dogs > dog**
2. Szófaj: **Haus > Haus/Noun**
3. Morfoszintaktikai jegyek: **went > go/Verb + Past**
4. Szóképzés (pl. török):

Finlandiyalılaştırıramadıklarımızdanmışsınızcasına >

Finlandiya/Noun + Prop + A3sg + Pnon + Nom

['(behaving) as if you have been one of those whom we could not convert into a Finn(ish citizen)/someone from Finland']

5. Összetettség-elemzés:

számítógépesmorfológia-oktatás >

számítógép/Noun + N2Adj | morfológia/Noun | oktat/Verb + V2N

❑ Példa szóalak-generálásra:

kesztyű/Noun + PS-sg2-pl + Ade > kesztyűidhez

Miért kell morfológiai elemzés?

- A különböző szóalakok nagy száma miatt (típusok)
10 millió szavas angol szövegtörzs esetén $< 100\,000$,
10 millió szavas finn szövegtörzs esetében $> 800\,000$



- Toldalékoló nyelvek esetében tetszőleges szövegtörzshöz igaz az alábbi állítás: a szövegtörzsben aktuálisan előforduló szövegszó-típusok száma kisebb, mint azoknak a lehetséges szótípusoknak a száma, melyek nincsenek benn az aktuális törzsben

A számítógépes morfológia fontosabb technikai ismérvei

- ☐ A (minél nagyobb) szókészlet
- ☐ A (lehetőleg teljes) toldalékkészlet
- ☐ Az ismeretlen alakok kezelése
- ☐ Az elemzéshez választott módszer
- ☐ A lexikonok ábrázolásának módja

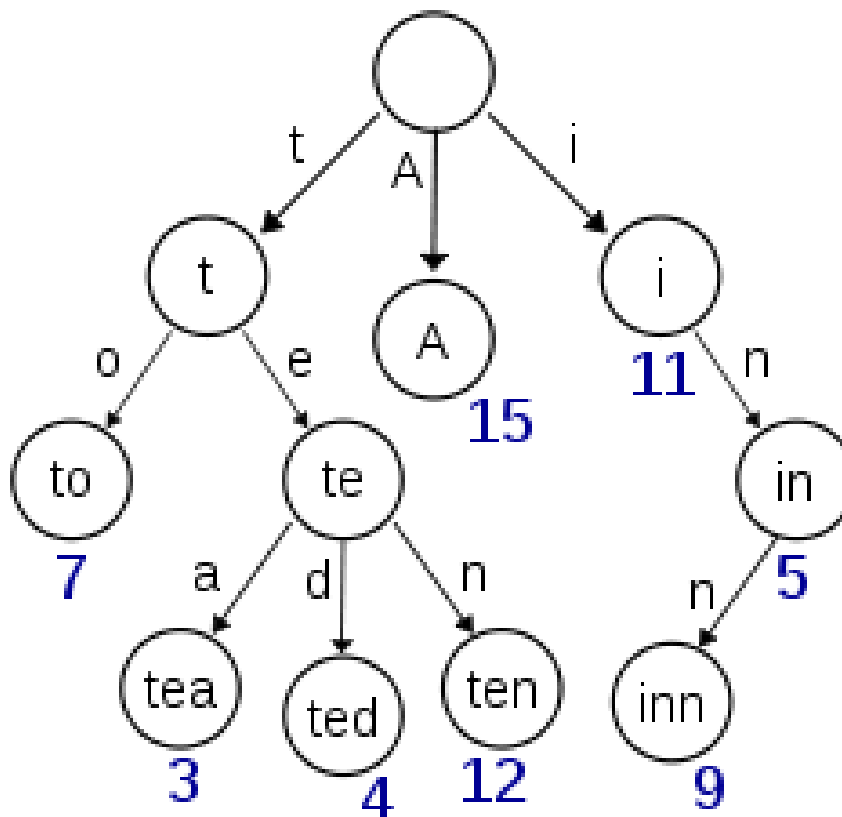


Néhány szó a szótárak tárolásáról

A szófa (=trie)

(A, i, in, inn, to, tea, ted, ten)

A szófa (Fredkin 1960) egy olyan, a szavak rákövetkező karaktereivel címkézett élsorozatokot tartalmazó fa, amelyben egy szót úgy találunk meg, hogy végigjárjuk karakterenként.

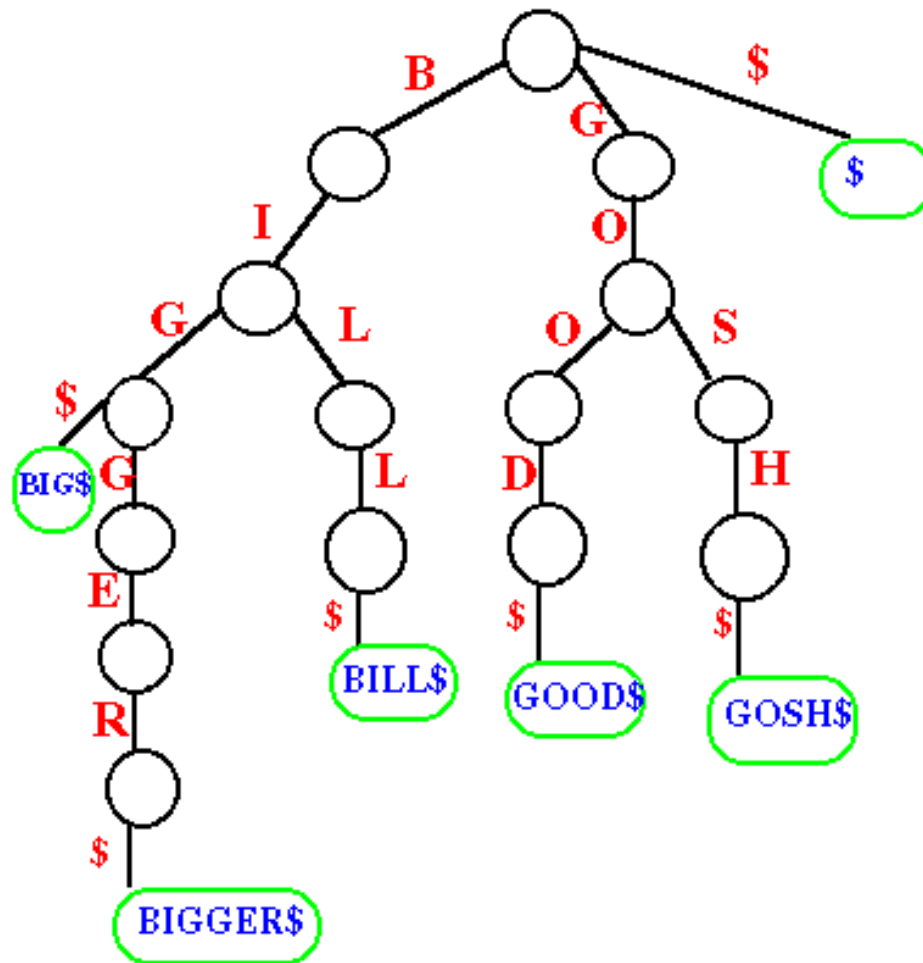


A szófák általános tulajdonságai

- ☐ Bináris keresőfák: $O(\log(n))$ (n a fában tárolt elemek száma)
- ☐ Szófa: m hosszúságú kulcs megtalálása max. $O(m)$
- ☐ Nagy számú rövid füzér tárolása esetén a szófa kevesebb helyet igényel, mint a bináris keresőfa (ui. a kulcsokat nem tároljuk, a csomópontokat meg közösen használják az egyforma kezdőszeletű füzérek kulcsai)
- ☐ Hasítótáblák helyett is használható (bár néha a szófák lassabbak)
- ☐ Nem mindenre jó: vannak füzérként nehezen ábrázolható kulcsok (pl. a lebegőpontos számok)
- ☐ De: szótárak ábrázolására alkalmas
- ☐ ADFSA (körmentes determinisztikus véges automata) a szófánál is jobb, de csak ha nincs kiegészítő információ (csak puszta szólista)

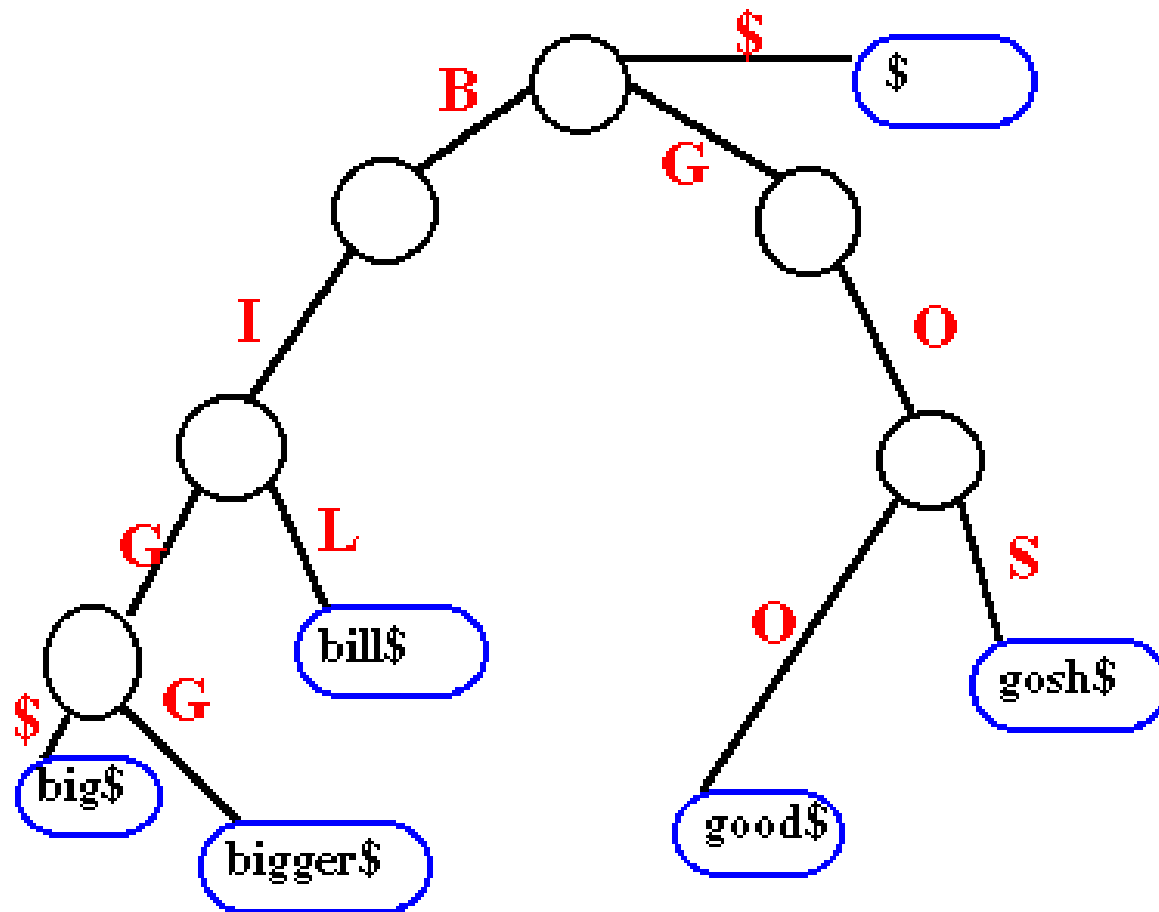
Általános szófa

(big, bigger, bill, good, gosh)



Módosított (kompakt) szófa

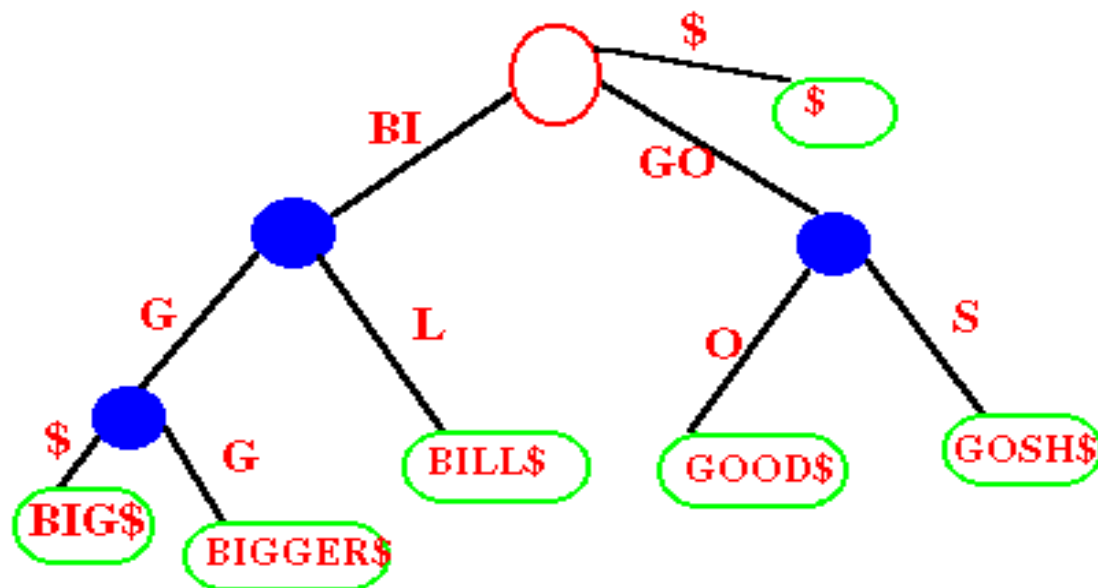
(big, bigg/er, bil/l, goo/d, gos/h)



Erősen módosított (PATRICIA) szófa

(big, bigg/er, bil/l, goo/d, gos/h)

- ❑ PATRICIA = Practical Algorithm to Retrieve Information Coded in Alphanumeric (Donald R. Morrison, 1968)



- ❑ Bármely élen több karakter is lehet, pl. az előtagok (igekötők, *re-*, *pre-*, *anti-* stb.) vagy a tipikus és ritka kezdő betűpárok
- ❑ Angol: a $26^2=676$ indító betűpárból csak 309 létezik (amiből 88 csak 15-nél kevesebb szó elején)

A Kay-féle szóábrázolás

(*alma, alom, anya, anyag, apa, apad, aránytalanság*)

- ☐ Kay (1977): tömörítés numerikus prefixekkel

alma – 0

alom – 2

anya – 1

anyag – 4

apa – 1

apad – 3

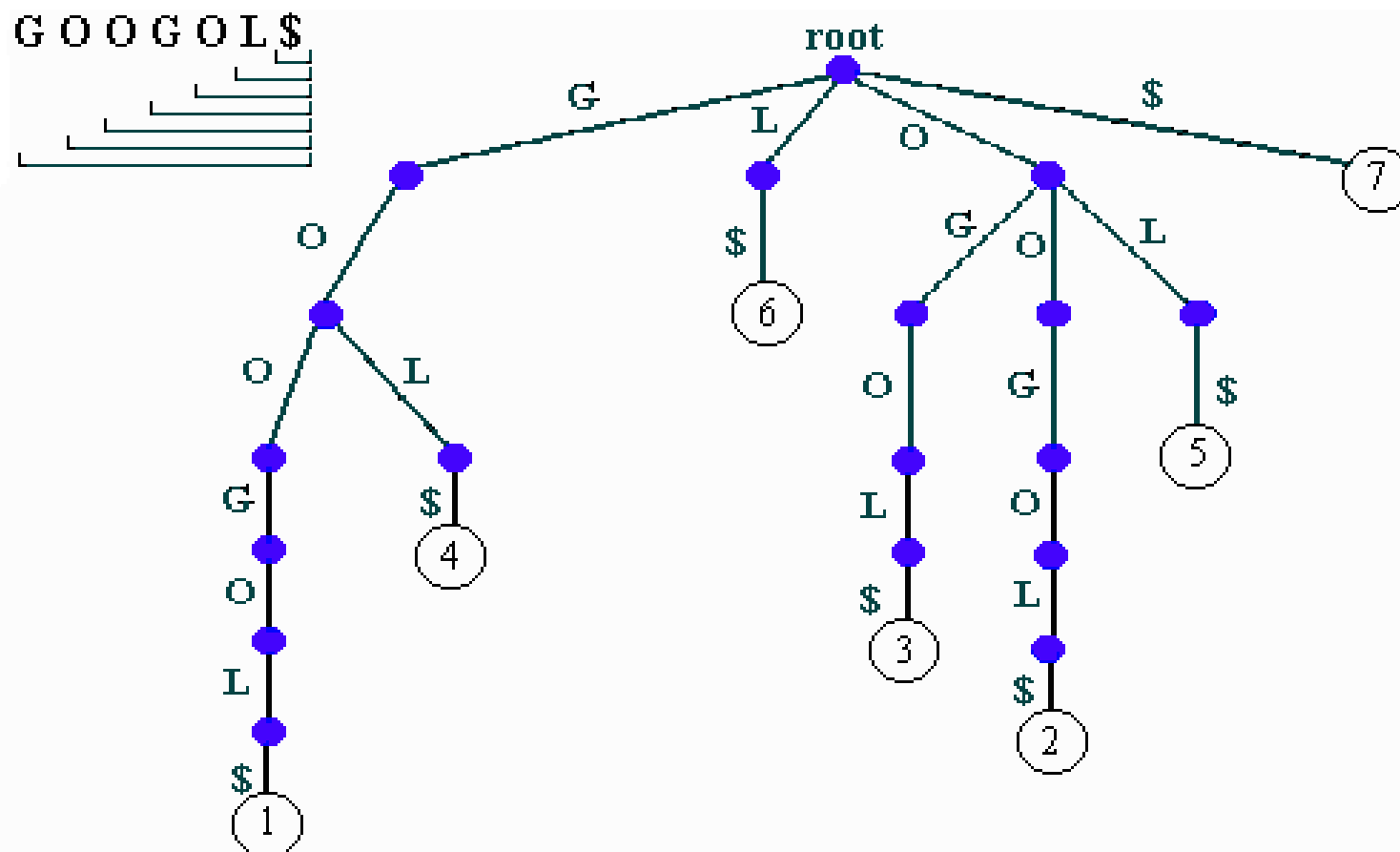
aránytalanság – 1

- ☐ Tehát a szótár:

alma, 2om, 1nya, 4g, 1pa, 3d, 1ránytalanság

- ☐ Akkor éri meg, ha hasonlítanak a szókezdetek
(nagy szótár esetén mindig!)

Szuffixum-szófa



- ☐ Egy füzér minden lehetséges végszelete tárolva
- ☐ A szuffixum-szófának n levele van és a magassága is n

A szuffixum-szófa néhány tulajdonsága

Egy n hosszú S füzérre épített általánosított szófára (többek közt) az alábbiak állnak:

- ☐ $O(m)$ időben eldönthető, hogy egy m hosszú P füzér a részfüzére-e
- ☐ $O(m)$ időben megtalálható egy tetszőleges m hosszú P részfüzérének első előfordulása
- ☐ $O(m+z)$ időben megtalálható mind a z darab előfordulása egy m hosszú részfüzérének
- ☐ Az S_i és az S_j füzérek leghosszabb közös részfüzére megtalálható $O(n_i + n_j)$ idő alatt



Automaták és véges állapotú morfológiák

Nyelvek, nyelvtanok, automaták

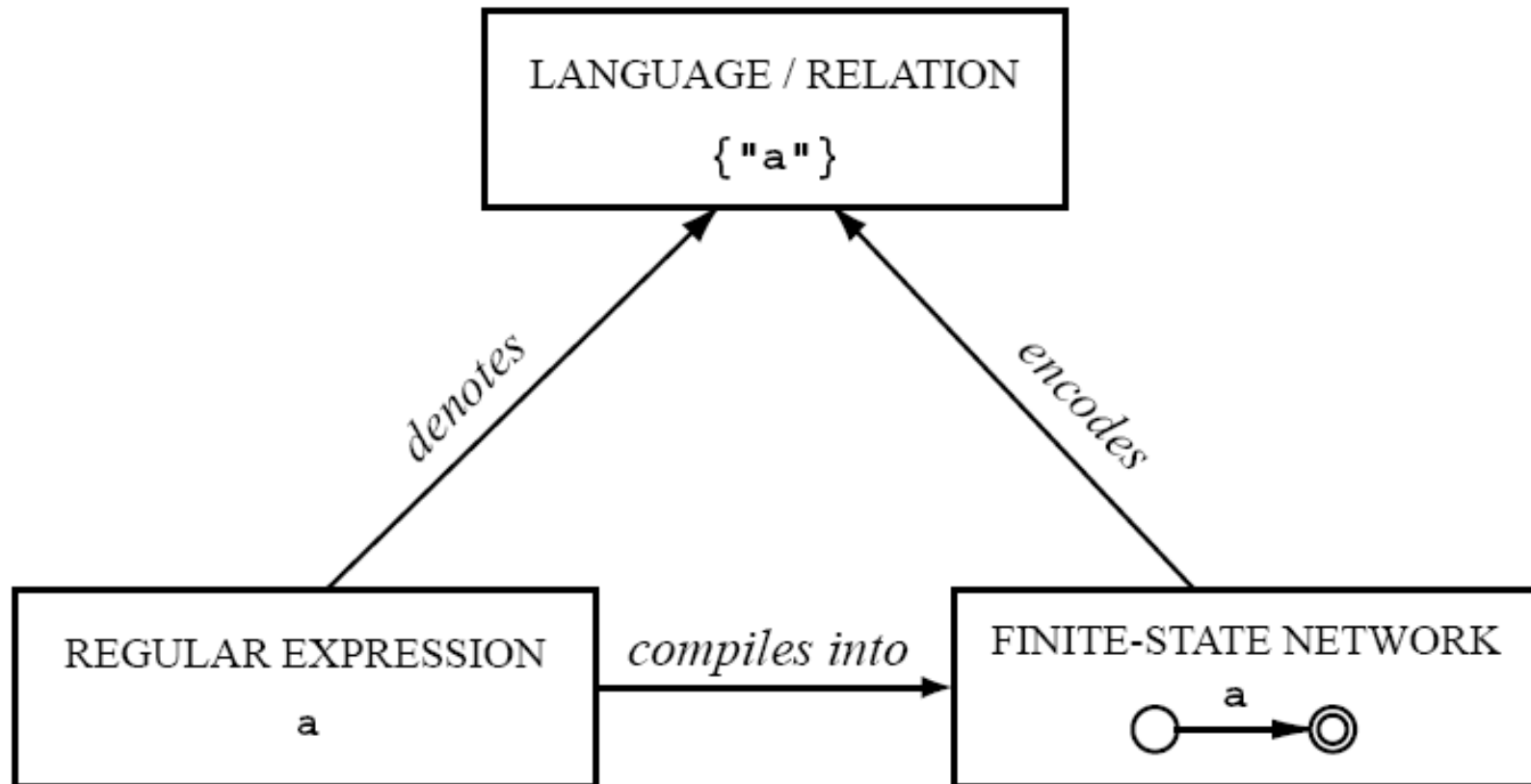
- ❑ **A formális nyelv** egy formális nyelvtan által leírt füzérek halmaza
- ❑ $G=(N, T, R, S)$ egy **formális nyelvtan**, ahol
 - N nemterminális szimbólumok véges halmaza
 - T terminális szimbólumok véges halmaza
 - R $(T \cup N)^* N (T \cup N)^* \rightarrow (T \cup N)^*$ alakú szabályok véges halmaza
 - $S \in N$ mondatszimbólum
- ❑ Formális nyelvek és formális nyelvtanok **Chomsky-hierarchiája**
ahol $A, B \in N$; $a \in T$; $\alpha, \beta, \gamma \in (T \cup N)^*$:
 - 0-típusú nyelvtan $(\alpha \rightarrow \beta)$: rekurzív megszámlálható nyelvek
 - 1-típusú nyelvtan $(\alpha A \beta \rightarrow \alpha \gamma \beta)$: környezetfüggő nyelvek
 - 2 -típusú nyelvtan $(A \rightarrow \gamma)$: környezetfüggetlen nyelvek
 - 3 -típusú nyelvtan $(A \rightarrow a \text{ and } A \rightarrow aB)$: reguláris nyelvek
- ❑ Az **elfogadó automaták hierarchiája**:
 - 0: Turing-gép
 - 1: Lineárisan kötött automata
 - 2: Veremautomata – $O(n^3)$
 - 3: Véges állapotú automata – $O(n)$



Véges állapotú automaták

- ❑ **Véges állapotú automata:** $A = (S, \Sigma, s, F, T)$ ötös, ahol
 - S állapotok véges halmaza
 - Σ egy ábécének nevezett véges halmaz
 - $s \in S$ kiinduló állapot
 - $F \subseteq S$ a végállapotok halmaza
 - $T: S \times (\Sigma \cup \{\varepsilon\}) \rightarrow S$ átmenetfüggvény
- ❑ Az $X = x_0 x_1 \dots x_n$ Σ ábécéből alkotott füzért **A elfogadja**, ha létezik S -ben az átmenetek $r_0, r_1, \dots, r_n$ sorrendje a következő feltételekkel:
 - i. $r_0 = s$
 - ii. $r_{i+1} = T(r_i, x_i) \quad (i = 0, \dots, n-1)$
 - iii. $r_n \in F$
- ❑ **Reguláris nyelv:** a véges állapotú automata által elfogadott füzérek halmaza
- ❑ **Reguláris kifejezés:** olyan formula, mely konkatenáció, unió és iteráció használatával meghatároz egy reguláris nyelvet

Reguláris kifejezés – reguláris nyelv – véges gép



lytechnológia alapjai - 2019. szeptember 25.

Véges átalakítók (FST)

- ❑ **Véges átalakító:** $T = (S, \Sigma, \Gamma, s, F, \delta)$ hatos, ahol
 - S az állapotok véges halmaza
 - Σ a bemenő ábécének nevezett véges halmaz
 - Γ a kimenő ábécének nevezett véges halmaz
 - $s \in S$ a kezdő állapot
 - $F \subseteq S$ az elfogadó állapotok halmaza
 - $T: S \times (\Sigma \cup \{\varepsilon\}) \times (\Gamma \cup \{\varepsilon\}) \rightarrow S$ átmenetfüggvény
- ❑ A T **átalakítja** az $\alpha \in \Sigma^*$ füzért $\beta \in \Gamma^*$ füzérbe (röviden: $\alpha[T]\beta$) ha létezik út a kezdőállapotból egy végállapotba α bemenősorozat és β kimenősorozat mellett
- ❑ Az **FSA** és a **FST** különbsége: az FSA kimenetén egy Boole-válasz jön létre, míg az FST egy füzért ad eredményül (a másik szalag tartalmát)

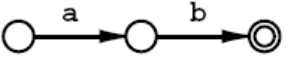
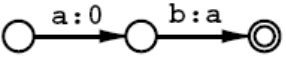
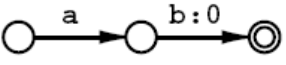
A reguláris nyelvek és a véges automaták helye

- ❑ $G=(N, T, R, S)$ egy **formális nyelvtan**, ahol
 - N nemterminális szimbólumok véges halmaza
 - T terminális szimbólumok véges halmaza
 - R $(T \cup N)^* N (T \cup N)^* \rightarrow (T \cup N)^*$ alakú szabályok véges halmaza
 - $S \in N$ mondat-szimbólum
- ❑ Formális nyelvek és formális nyelvtanok **Chomsky-hierarchiája**
ahol $A, B \in N$; $a \in T$; $\alpha, \beta, \gamma \in (T \cup N)^*$:
 - 0-típusú nyelvtan $(\alpha \rightarrow \beta)$: rekurzív megszámlálható nyelvek
 - 1-típusú nyelvtan $(\alpha A \beta \rightarrow \alpha \gamma \beta)$: környezetfüggő nyelvek
 - 2 -típusú nyelvtan $(A \rightarrow \gamma)$: környezetfüggetlen nyelvek
 - 3 -típusú nyelvtan $(A \rightarrow a \text{ and } A \rightarrow aB)$: reguláris nyelvek
- ❑ Az **elfogadó automaták hierarchiája**:
 - 0: Turing-gép
 - 1: Lineárisan kötött automata
 - 2: Veremautomata – $O(n^3)$
 - 3: Véges állapotú automata – $O(n)$

Műveletek véges átalakítókkal

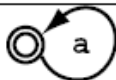
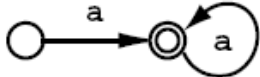
Konkatenáció:

 $w[T]y$ és $x[S]z$
 $acsa$
 $wx[TS]yz$

Expression	Language / Relation	Network
$a \circ b$	$\{“ab”\}$	
$a : 0 \quad b : a$	$\{<“ab”, “a”>\}$	
$a \quad b : 0$	$\{<“ab”, “a”>\}$	

Unió:

 $x[T]y$ vagy $x[S]y$
 $acsa$
 $x[T \cup S]y$

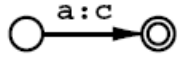
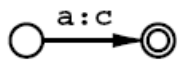
Expression	Language / Relation	Network
a^*	$\{“”, “a”, “aa”, \dots\}$	
a^+	$\{“a”, “aa”, \dots\}$	

Iteráció:

 $w[T^*]y$ és $x[T]z$
 $acsa$
 $wx[T^*]yz$ és $\varepsilon[T^*]\varepsilon$

Metszet:

 $x[T]y$ és $x[S]y$
 $acsa \quad x[T \times S]y$

Expression	Language / Relation	Network
$a : b \quad .o. \quad b : c$	$\{<“a”, “c”>\}$	
$a : b \quad .o. \quad b \quad .o. \quad b : c$	$\{<“a”, “c”>\}$	

Kompozíció:

 $x[T]y$ és $y[S]z$
 $acsa$
 $x[T \circ S]z$



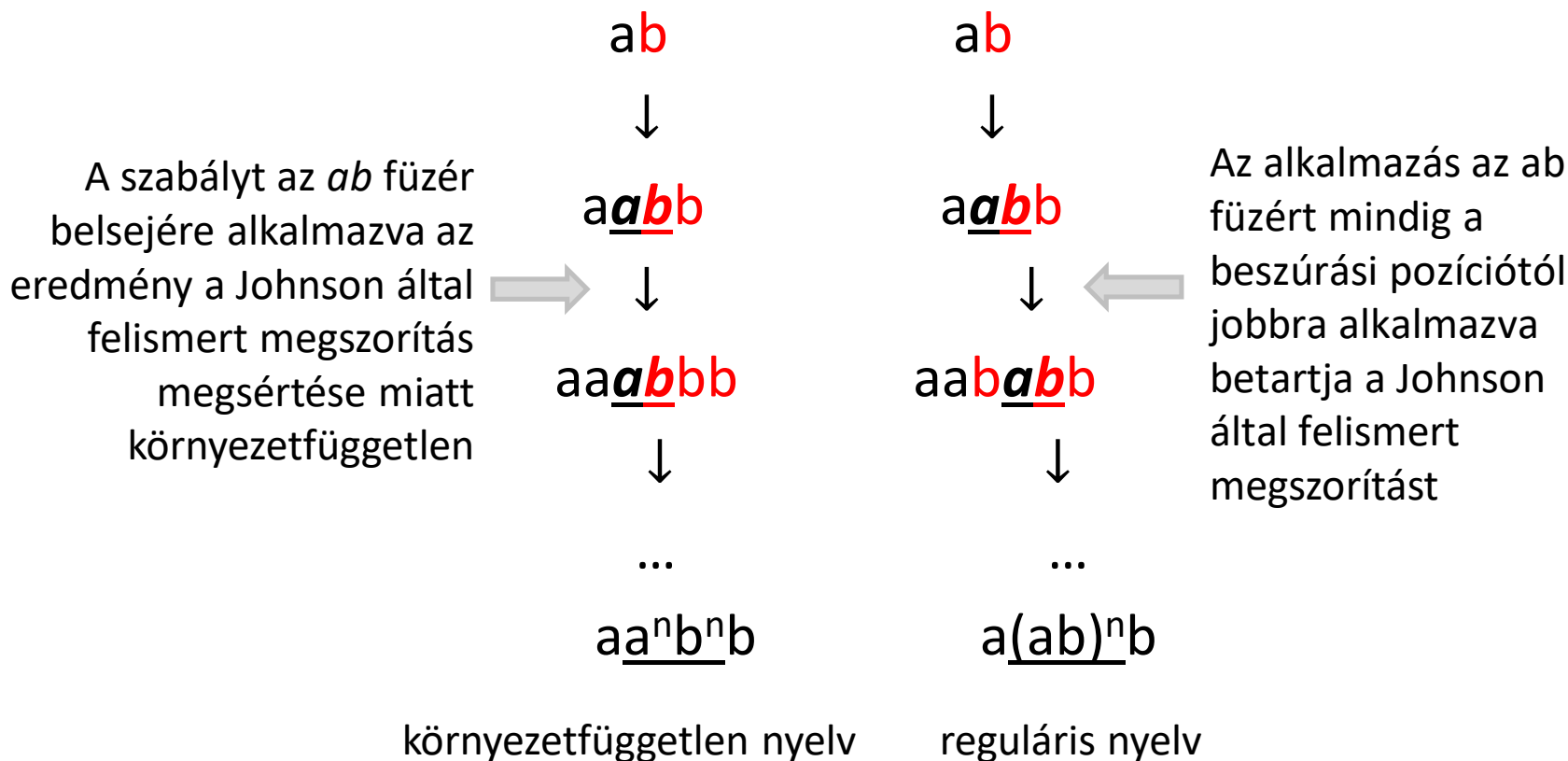
A kétszintes leírás felé...

Johnson felfedezése (1972)

- ❑ A generatív fonológiában, ha az $\alpha \rightarrow \beta / \gamma \_ \delta$ szabállyal állítjuk elő a $\gamma\beta\delta$ füzért a $\gamma\alpha\delta$ füzérből, akkor a szabály bármely rákövetkező alkalmazása a β füzért érintetlenül hagyja, azaz csak γ és δ füzéreket érinti
- ❑ C. Douglas Johnson a *Formal Aspects of Phonological Description* (1972) című művében észrevette, hogy egyazon generatív szabály ismételt alkalmazása esetén nem szokás érinteni az előző alkalmazás kimenetét, hanem az újabb alkalmazás attól vagy balra vagy jobbra helyezkedik el
- ❑ Következmény: ha egyazon ciklusban nem alkalmazzuk az újraíró szabályt saját kimenetére, akkor a bemenet-kimenet párok leírhatók egy **reguláris relációval** (példa a következő dián!)
- ❑ Amint tudjuk korábbról, a reguláris reláció megfeleltethető egy reguláris nyelvnek, ami ekvivalens egy véges állapotú átalakítóval
- ❑ Johnson megmutatta, hogy a generatív fonológia szabályai sokkal kevésbé „erősek”, mint ahogy ezt a leírásukhoz használt (környezetfüggő) formalizmus sugallja: generatív képességük a fenti megszorítás általános elfogadottsága miatt az 1-es Chomsky-nyelvosztály helyett a 3-as nyelvosztályba sorolja őket

Johnson felismerése egy példán

$$\varepsilon \rightarrow ab / \_ b$$



Kaplan & Kay (1981/1994)

- ☐ Kaplan és Kay (1981): egy hatékony elemző irányában kutatva leírták az újraíró szabályok átalakítókba való fordításának algoritmusát
- ☐ Azt vették észre, hogy a reguláris relációk zártak a soros kompozícióra nézve: ha veszünk két olyan szabályt, amit átalakítóval modellálunk, akkor ha az első átalakító kimenete a másik bemenete, a kompozíció művelete segítségével helyettük **egyetlen ekvivalens átalakítót** kaphatunk
- ☐ A kompozíció eredményeként kapott gép az első átalakító bemenetét a második kimenetére úgy képezi le, hogy **nem generál semmiféle köztes eredményt**
- ☐ Tetszőleges számú fonológiai újraíró szabályt sorban alkalmazva reguláris relációt kapunk, az alkalmazott **szabályok számától függetlenül**
- ☐ Megjegyzés: nincs olyan művelet az újraíró szabályok eredeti világában, ami ugyanerre volna képes

A szabálymegfordítás problémája

Egy (nem létező nyelvből származó) példa:

labiális realizáció p előtt: $N \rightarrow m$,

dentális realizáció egyébként: $N \rightarrow n$;

m után: $p \rightarrow m$

Az újraíró szabályok tehát:

$N \rightarrow m / \_\_ p$; elsewhere, n .

$p \rightarrow m / m \_\_$

Lexikálisból felszíni alak (generálás):

$kaNpat \Rightarrow kammatt$

Felszíniből lexikális alak (elemzés):

$kammatt \Rightarrow \{kaNpat, kampat, kammatt\}$

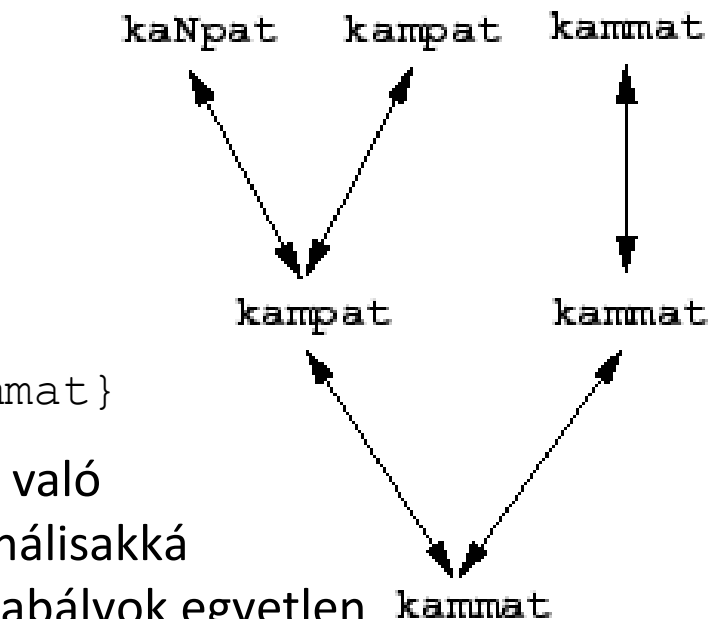
Ugyanazon generáló szabályok elemzésre való

használatakor a kötelező szabályok opcionálisakká

válhatnak: a $kaNpat$ lexikális füzérből a szabályok egyetlen $kammatt$

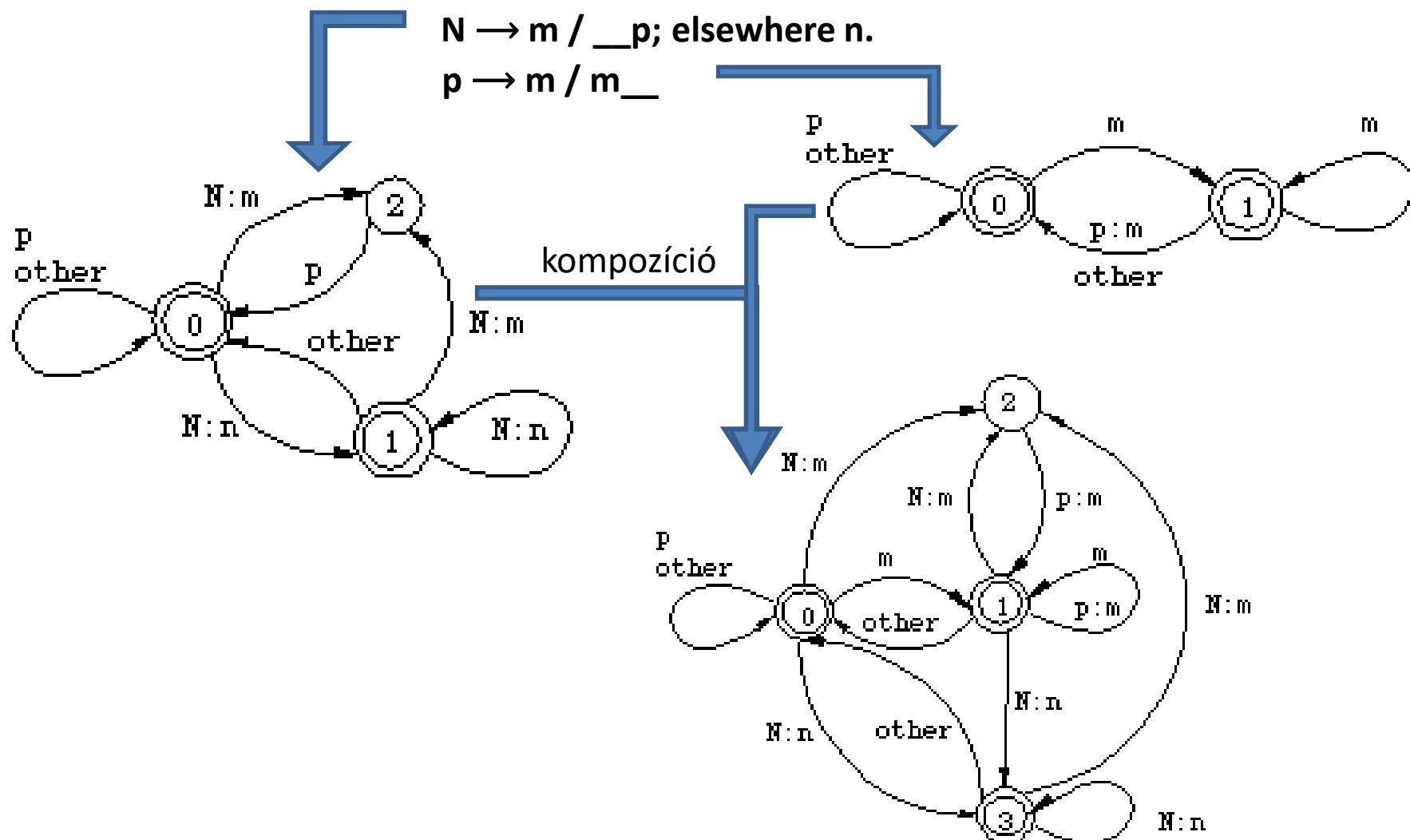
felszíni alakot generálnak, de a felszíni alakból az inverz

leképezés „egy a sokhoz” típusú is lehet



Figures from <http://www.ling.helsinki.fi/~koskenni/esslii-2001-karttunen/>

A 'kaNpat' automata



A korábbi beszűrőszabály véges automatája

$\varepsilon \rightarrow ab / \_ b$

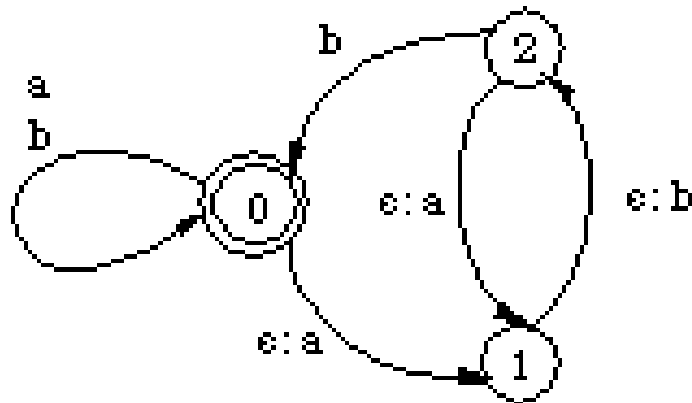
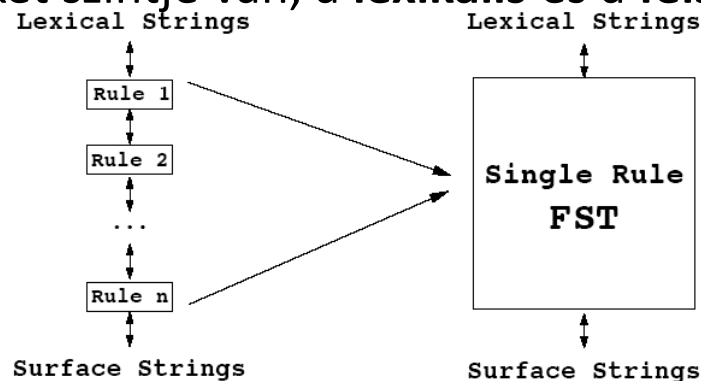


Figure from <http://web.stanford.edu/~laurik/publications/fsc-91/fsc91.html>

Összegzés: véges átalakítók soros kapcsolása

- ❑ A fonológiai levezetések köztes állapotai mindig kiküszöbölhetők az egyes szabályokból kapott átalakítók kompozíciójával: az eredményül kapott átalakítónak csak két szintje van, a **lexikális** és a **felszíni**



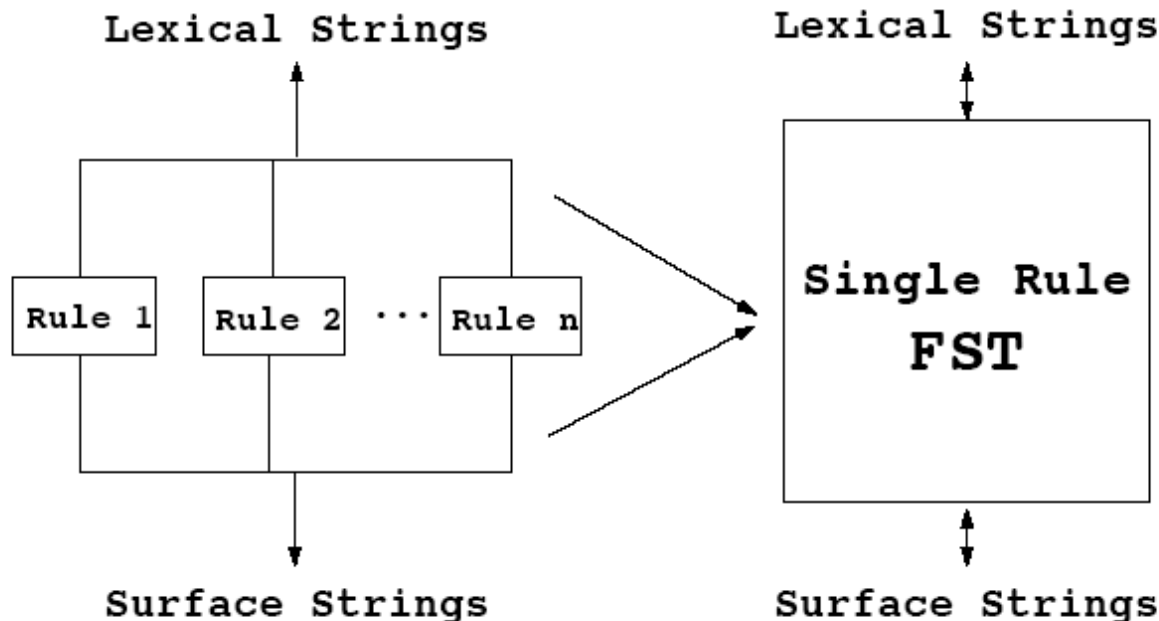
Figures from <http://www.ling.helsinki.fi/~koskenni/esslli-2001-karttunen/>

- ❑ Az egyetlen generatív átalakító használata **sokkal hatékonyabb** felismerésre is, mintha az eredeti szabályoknak megfelelő átalakítók egyenként invertált sorrendben működnének
- ❑ Kaplan és Kay megoldotta ugyan az egyes szabályok sorozata átalakítóba való fordításának problémáját, azonban a nagy szabályrendszerek egyetlen átalakítóba való kompozíciója az akkori technikai korlátok miatt a **gyakorlatban megvalósíthatatlan** volt

Véges átalakítók párhuzamos kapcsolása

Kimmo Koskenniemi PhD-értekezése: *Two-level Morphology* (1983)

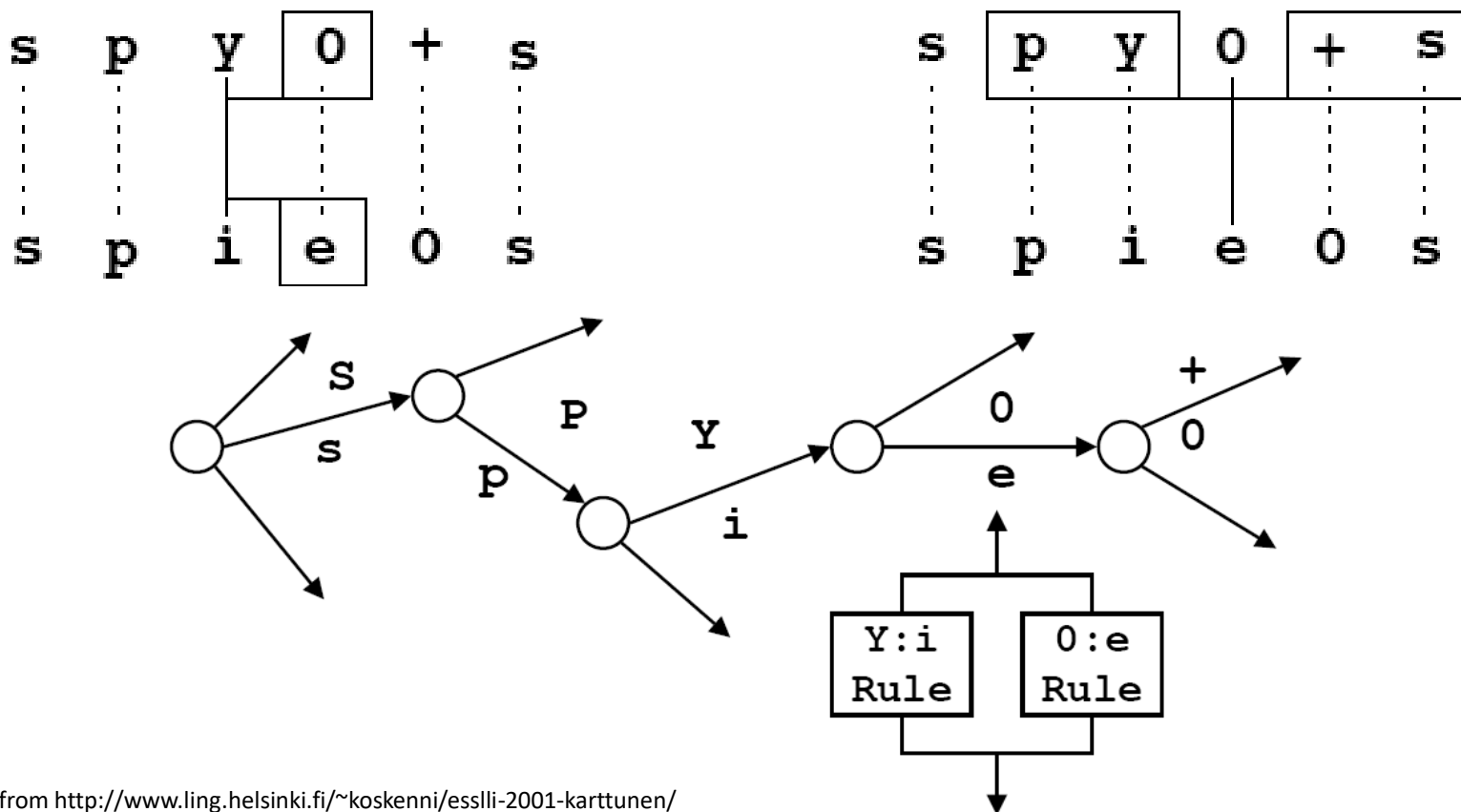
Tudva, hogy a lexikális és a felszíni alakok közötti megfeleltetés leírható reguláris relációval, Koskenniemi egy lényegi változtatást, a szabályhalmazból származó átalakítók **párhuzamos** kapcsolását javasolta



Figures from <http://www.ling.helsinki.fi/~koskenni/esslli-2001-karttunen/>

Egy „igazi” (angol nyelvi) példa

$$y:i \Leftrightarrow \_ 0:e$$



Figures from <http://www.ling.helsinki.fi/~koskenni/esslli-2001-karttunen/>

A kétszintes szabályok alakja

Egy kétszintes szabály az alábbi három részből áll:

- **megfeleltetett párok:** amiről valójában a szabály „szól”
- **környezet** (bal környezet (lc) és/vagy jobb környezet (rc))
- **szabályoperátor**

Megfeleltetett párok: egy lexikális és egy felszíni karakterből álló pár (pl. $t:c$ ahol a lexikális t megfelel a felszíni c -nek)

Környezet: az adott jelenséget körülvevő fonológiai helyzetet specifikálja (az aláhúzás jelenti a a megfeleltetett pár helyzetét az adott környezetben)

Szabályoperátor: a megfeleltetett pár és a környezete között fennálló reláció; nagyjából a formális logika feltételeket és következményeket leíró operátorainak felelnek meg:

- $\Rightarrow$ a megfeleltetés *csak akkor* áll fenn, ha ez a környezet
- $\Leftarrow$ ha ez a környezet, *akkor* a megfeleltetés fennáll
- $\Leftrightarrow$ a megfeleltetés *akkor és csak akkor* áll fenn, ha ez a környezet
- $\nLeftarrow$ a megfeleltetés *soha* nem fordul elő ebben a környezetben ($/\Leftarrow$)

Generatív vs. kétszintes szabályok

- ❑ A $t \rightarrow c / \_\_ i$ generatív szabály azt jelenti, hogy
 - a t átalakul c -be, ha i előtt áll, és miután újraírtuk t -t c -vé, t **többé nem létezik**
 - A generatív szabályok sorozata **tetszőleges számú köztes szint** segítségével konvertál beleső reprezentációkat felszíni formákra
 - A generatív szabályok **egyirányúak**: csak belső reprezentációkból tudnak felszíni alakokat létrehozni, fordított irányban nem alkalmazhatók

- ❑ Az ezzel analóg kétszintes szabály: $t : c \Rightarrow \_\_ i$
 - A lexikális t megfelel a felszíni c -nek i előtt; de nem változik át c -vé, hanem **megmarad** t -ként a szabály alkalmazása után is
 - A kétszintes szabályok egyfajta megfeleltetést fejeznek ki, nem újraírást, párhuzamosan alkalmazandók, és nem sorosan, és az újraírással szemben nem hoznak létre **semmilyen köztes reprezentációs szintet**
 - A kétszintes szabályok **kétirányúak** és deklaratívak: bizonyos megfeleltetéseket fogalmazznak meg a lexikális és a felszíni formák között

"Kizárólag, de nem mindig"

$$L:S \Rightarrow E$$

- ☐ Mit jelent?
 - L csak E-ben realizálódik S-ként
 - L S-ként való realizáció nincs megengedve $\neg$ E-ben
 - Ha L:S, akkor ennek E-ben kell lennie
 - De: L: $\neg$ S megengedett E-ben
- ☐ Logikailag: az $\Rightarrow$ operátor azt jelenti, hogy a megfeleltetésből következik a környezet, de a környezetből nem feltétlenül következik a megfeleltetés
- ☐ Példa: **t:c** $\Rightarrow$ **i** (a lexikális **t** megfelel csak **i** előtt felel meg a felszíni **c**-nek, de ebben a környezetben nem feltétlenül mindig; azaz a lexikális **t** más realizációi is előfordulhatnak ebben a környezetben, beleértve a **t:t** párt is)
- ☐ Negatív megfogalmazásban: ez a szabály letiltja a **t:c** pár minden olyan előfordulását, ami nem **i** előtt van
- ☐ A $\Rightarrow$ szabály nagyjából a generatív fonológia opcionális szabályának felel meg, és tipikusan a szabad variációk leírására használják

”Mindig, de nem kizárólag”

$$L:S \Leftarrow E$$

- ☐ Mit jelent?
 - L E-ben mindig S-ként realizálódik
 - $L \rightarrow S$ -ként való realizációja nincs megengedve E-ben
 - Ha L E-ben van, akkor L:S-nek kell lennie
 - De: L:S előfordulhat máshol
- ☐ Logikailag: az $\Leftarrow$ operátor azt jelenti, hogy a környezetből következik a megfeleltetés, de a megfeleltetésből nem feltétlenül következik a környezet
- ☐ Példa: $t:c \Leftarrow \_\_ i$ (a lexikális t i előtt mindig kötelezően megfelel a felszíni c -nek, de nem szükségszerűen csak ebben a környezetben; azaz a $t:c$ előfordulhat más környezetben is)
- ☐ Negatív megfogalmazásban: ha a $t : \neg c$ pár azt jelenti, hogy a lexikális t minden felszíni alaknak megfelelhet, ami nem c , akkor a fenti szabály megtiltja a $t : \neg c$ előfordulását az adott környezetben
- ☐ A $\Leftarrow$ szabály nagyjából a generatív fonológia kötelező szabályának felel meg, és tipikusan akkor használják, amikor a megfeleltetés kötelező egy adott környezetben, de előfordulhat más környezetben is

"Mindig és kizárólag"

$$L:S \Leftrightarrow E$$

- ☐ Az $\Leftarrow$ és az $\Rightarrow$ operátor kombinációja
- ☐ Mit jelent?
 - L S-ként akkor és csak akkor realizálható, ha E a környezet
 - Mind $L:S \Rightarrow E$, mind $L:S \Leftarrow E$ fennáll
 - L:S kötelező E-ben, de máshol sehol
- ☐ Példa: **t:c** $\Leftrightarrow$ **i** (a lexikális **t** akkor és csak akkor felel meg a lexikális **c**-nek, ha **i** előtt áll)
- ☐ A $\Leftrightarrow$ szabályt akkor használják, ha egy megfeleltetés **kötelező** egy adott környezetben (v.ö. $\Leftarrow$ operátor) és **semmilyen más környezetben nem** fordul elő (v.ö. $\Rightarrow$ operátor)
- ☐ Ekvivalens a bikondicionális logikai operátorral és azt jelenti, hogy egy megfeleltetés **akkor és csak akkor** megengedett, ha az az adott környezetben van

"Soha"

$L:S \nLeftarrow E$

- ☐ Az $\nLeftarrow$ operátort tipikusan egy általános szabály alóli kivételek kezelésére használják"
- ☐ Az $\nLeftarrow$ operátort gyakran így írják: $/\Leftarrow$
- ☐ Mit jelent?
 - L soha nem realizálódik S-ként E-ben
 - L S-ként való realizációja nem megengedett E-ben
 - Ha L E-ben van, akkor L:-S-nek kell fennállnia
- ☐ Példa: $t:c \nLeftarrow \_\_ i:\hat{e}$ (a lexikális **t** nem felelhet meg a felszíni **c**-nek **i:** $\hat{e}$ előtt)
- ☐ A szabály által megfogalmazott megfeleltetés az adott környezetben le van tiltva
- ☐ Megengedi tehát a **tatê** és **catê** felszíni alakokat, de tiltja a **tacê** és **cacê** alakot
- ☐ A $\nLeftarrow$ operátor hasonló az $\Leftarrow$ operátorhoz abban, hogy nem tiltja meg az adott megfeleltetést más környezetekben

Konvenciók, speciális szimbólumok

- ❑ **Alapértelmezett** (pl.. **t:t**, **i:i** – röviden: **t**, **i**) és **speciális megfeleltetések** (pl. **t:c**)
- ❑ **Dzsókerszimbólum** (általában: **@**) az adott ábécé tetszőleges karaktere helyett állhat, pl. **t:c** $\Rightarrow$ **__ i:@**
- ❑ **Nullszimbólum** (általában: **0** vagy **ε**): beszúrásnál és törlésnél a kétszintes rendszer egyik szalagján üres szimbólumnak kell állnia, mert a két szalag csak egyenlő számú szimbólum esetén használható megfeleltetésre, pl.

LR: **0 t a t + i**

SR: **' t a c 0 i**

- ❑ **Határszimbólum** (általában: **#**) jelzi a szó elejét vagy végét (kizárólag egy másik határszimbólummal párban: **##**)
- ❑ **Részhalmaz:** egyszavas nevekkel jelzett karakterhalmazok, pl. **C** a mássalhangzók halmaza, **V** a magánhangzóké, **T** a felpattanó hangzóké, vagy **NAS** a nazálisoké:

SUBSET C **b c d f g p t k b d g m n n g s l r w y**

SUBSET V **i e a o u**

SUBSET T **p t k b d g**

SUBSET NAS **m n n g**

Kétszintes lexikonok

- ☐ **Kétszintes lexikonok:**
a **tőlexikonok**,
az **alternációs minták lexikonjai** és
a **toldaléklexikonok** (inflexiószak és derivációszak)
- ☐ **Lexikon:** egy név és az alábbi formájú lexikális elemek listája
- ☐ **Lexikális elemek:** lexéma, folytatási osztály, kimeneti információ
- ☐ A **P folytatási osztály** definíciója (ami valahol másol van megadva):
(**P** = **PS** **K0** #)
azt állítja, hogy az eredeti lexikális elemet vagy a **PS** allexikon
(birtokos toldalékok), vagy a **K0** allexikon (klitikumok) valamelyik
eleme követheti, vagy egy határszimbólum

Az angol morfológia kétszintes leírása

(Karttunen & Wittenburg 1985)

Alternációk

ALTERNATIONS

```
( /N = N )  
( /IN = C1 )  
( /MN = MN )  
( /V = P3 PS PP PR I AG AB )  
( /IV1 = PR I AG AB )  
( /IV2 = P3 PR I AG AB )  
( /A = PA CA CS Ly )  
( /IP3 = IP3 )  
( /IPS = IPS )  
( /IPF = IPP )  
( C1 = C1 )  
( C2 = C2 )  
( I = I )  
( P3 = P3 )  
( PS = PS )  
( PP = PP )  
( PR = PR )  
( AG = AG )  
( PA = PA )  
( CA = CA )  
( CS = CS )  
( Ly = Ly )  
( Root = Root )  
( AB = AB )  
( # = )
```

END

Az angol morfológia kétszintes leírása

(Karttunen & Wittenburg 1985)

Lexiconok 1.

LEXICON N	0	C1	"N SG";	+s	C2	"N PL"
LEXICON MN	0	C2	"MASS N"			
LEXICON C1	0	#	"";	's	#	" GEN"
LEXICON C2	0	#	"";	'	#	" GEN"
LEXICON P3	+s	#	"V PRES SG 3RD"			
LEXICON IP3	0	#	"V PRES SG 3RD"			
LEXICON PS	+ed	#	"V PAST"			
LEXICON IPS	0	#	"V PAST"			
LEXICON PP	+ed	#	"V PAST PRT"			
LEXICON IPP	0	#	"V PAST PRT"			
LEXICON IP	0	#	"V PAST"			
LEXICON PR	+ing	#	"V PROG"			
LEXICON I	0	#	"V"			
LEXICON IP1	0	#	"V PRES SING 1ST"			
LEXICON AG	+er	/N	"AG "			
LEXICON PA	0	#	"A"			
LEXICON CA	+er	#	"A COMP"			
LEXICON CS	+est	#	"A SUP"			
LEXICON Ly	ly	#	"ADV"			
LEXICON AB	+able	#	"VERB ABL"			

Az angol morfológia kétszintes leírása

(Karttunen & Wittenburg 1985)

Lexiconok 2.

LEXICON Root

am	#	"V PRES SING 1ST";
am	#	"AUX";
are	#	"AUX";
are	#	"V PRES PL";
are	#	"V PRES SING 2ND";
at	#	"PREP";
at`tack	/N	"";
at`tack	/V	"";
a`gree	/V	"";
be	#	"AUX";
be	/IV1	"";
beer	/N	"";
believe	/V	"";
big	/A	"";

... ..

under`stand	/IV2	"";
under`stood	/IPP	"";
under`stood	/IPS	"";
undid	/IPS	"";
undo	/IV1	"";
undoes	/IP3	"";
undone	/IPP	"";
untie	/V	"";
went	/IPS	"";
went	/IPS	"";

END

Párhuzamos szabályalkalmazás és a szótárak

- ☐ A kétszintes rendszer szabályainak alkalmazása önmagában még nem oldaná meg a korábban a *kaNpat* példával illusztrált „túlelemzési” problémát (a két megszorító szabály a *kammat* alakot engedi többféleképpen, *kaNpat*, *kammat* vagy *kammat* alakra is visszavezetni), azonban a kétszintes rendszerben ezt a problémát azért lehet megoldani, mert csak két szint van, és az egyik elemzésekor minden lépésben konzultálni tud **a másik szinten további megszorításokat bevezető lexikonnal**
- ☐ Koskenniemi modelljében tehát a lexikonhoz fordulás és a felszíni alak elemzése együtt, egyfajta **tandem**-működésben valósul meg
- ☐ A lexikonok **szófa-erdők**, melyek szorosan együttműködnek a folytatásiosztály-linkekkel, azaz a lexikonhoz fordulás az egyik szófa leveleit a szófaerdő egy másik fájának (vagy fáinak) gyökeréhez kapcsolva determinisztikusan halad
- ☐ Az a lexikon, mely az aktuális lexikális füzért tartalmazza, egyfajta **folytonos lexikális szűrőként** működik

A 'kanPat' példa kétszintes szabályokkal és lexikonnal

$N \rightarrow m / \_p$; elsewhere n.

$p \rightarrow m / m\_$

Alphabet

a b c d e f g h i j k l m N:m N:n n o p q r s t u v x y w z ;

Rules

"N realized as m"

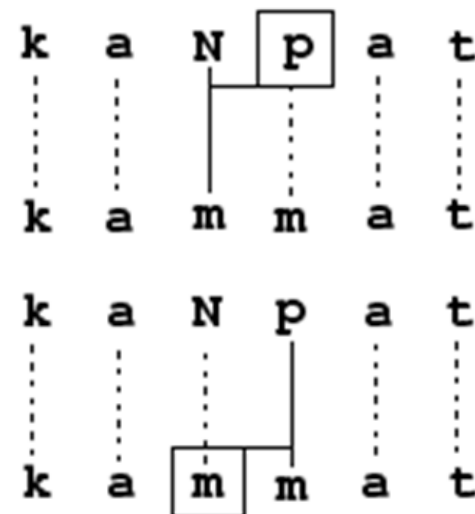
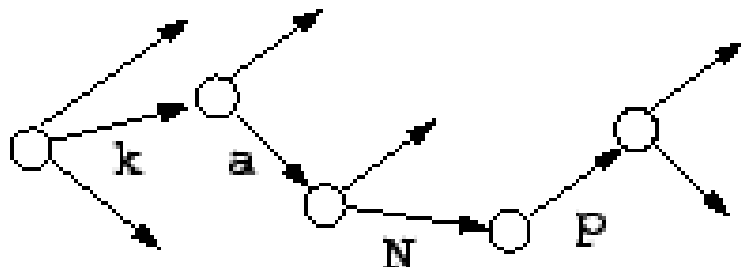
! Before a lexical "p." Always, and only there.

$N:m \Leftrightarrow \_p$;

"p realized as m"

! After a surface "m". Always and only there.

$p:m \Leftrightarrow :m \_$;



Figures from <http://www.ling.helsinki.fi/~koskenni/essli-2001-karttunen/>

Lexikon + lexikális reprezentáció + felszíni alak

Lexikális szint { **f** **o** **x** **+N** **+PL** }

LEXICON-FST

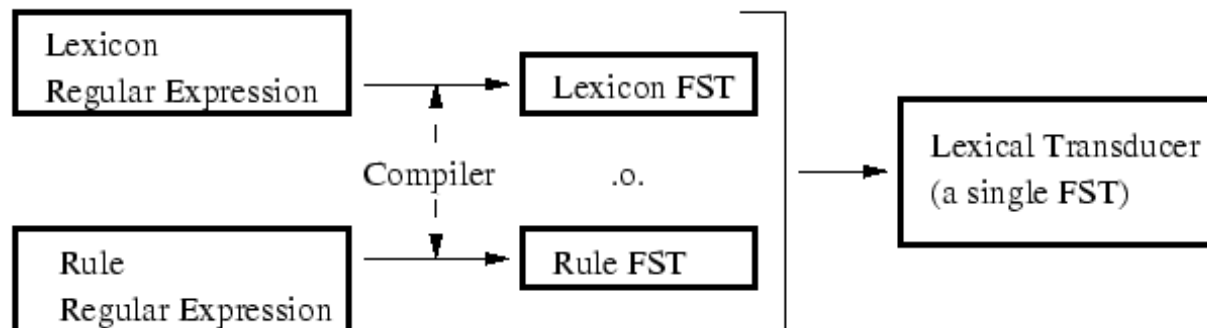
Köztes szint { **f** **o** **x** **^** **s** }

FST₁ *helyesírási szabályok* *FST_n*

Felszíni szint { **f** **o** **x** **e** **s** }

A Xerox lexikonkompozíciója

- ❑ Ha elvégezzük a kompozíció műveletét a kétszintes lexikon és a kétszintes szabályhalmaz átalakítóin, a lexikon által nem megengedett füzérek automatikusan kiesnek
- ❑ Eleinte a nagyméretű lexikonok és a nagyméretű szabályrendszer kompozíciója eredményétől tartottak a kutatók, hogy az használhatatlanul nagy méretű lesz
- ❑ Lauri Karttunen és más Xerox-kutatók kimutatták, hogy a lexikon és a kétszintes rendszer FST-inek kompozíciója soha **nem lesz szignifikánsan nagyobb** a kiinduló lexikon FST-jénél, és **sokkal kisebb** lesz, mint a szabályok véges átalakítóinak metszeteként kapott FST
- ❑ A kapott egyetlen lexikális FST a kiinduló lexikon minden lexikális alakját és ezeknek a szabályok által realizált összes felszíni reprezentációját tartalmazza:



Hogy konvertálunk kétszintes szabályokat FST-be?

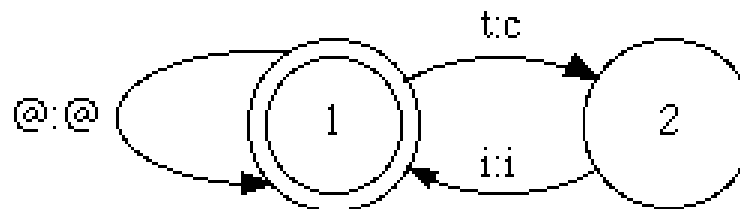
- ☐ A felhasználó környezetfüggő szabályokat használhat
- ☐ Minden nyelvi jelenséget egy önálló kétszintes szabály segítségével fogalmazzunk meg (a többit maga a kétszintes rendszer kezeli)
- ☐ A környezetfüggőnek tűnő kétszintes szabályrendszert egyetlen FST-be lehet fordítani
- ☐ Ez a fordítás a TWOL-rendszer létrehozásától (1983) évekig csak kézzel történt
- ☐ A kézi szabályfordítás az átalakítók részletes ismeretét és az újszerű szabályok szemantikájának mély megértését követelte, amit nem sok kutató tudott az elvárt szinten elsajátítani, hiszen a sokszor egymással komplex interakcióba lépő szabályok működésének megértése sok-sok órás koncentrált munkát követelt mind a létrehozáskor, mind a teszteléskor
- ☐ Az első automatikus szabályfordítót Koskenniemi and Karttunen (1987) hozta létre, Ron Kaplan (Xerox) véges-állapotú kalkulusa első implementációjának segítségével, aminek alapját Kaplan és Kay (1994) véges állapotú nyelvészeti leírása adta

Példa: $a \Rightarrow$ konverziója FST-be

A kétszintes szabály:

$t:c \Rightarrow \_\_ i$

Az ekvivalens FSA:



Az ekvivalens FSA táblázatos alakban:

	t	i	@
c	i	@	
1:	2	1	1
2:	0	1	0

Példák kétszintes szabályok táblázatos alakjára

$$t:c \leftarrow \_\_\_ i$$

	t	t	i	@
	c	@	i	@

1:	1	2	1	1
2:	1	2	0	1

$$t:c \Leftrightarrow \_\_\_ i$$

	t	t	i	@
	c	@	i	@

1:	3	2	1	1
2:	3	2	0	1
3:	0	0	1	0

$$t:c \nleftarrow \_\_\_ i:\hat{e}$$

	t	i	@
	c	$\hat{e}$	@

1:	2	1	1
2:	2	0	1

Kétszintes definíciók és szabályok

ALPHABET a b c d e f g h i j k l m n o p q r s t u v w x y z
+

; + is morpheme boundary

NULL 0

ANY @

BOUNDARY #

SUBSET C b c d f g h j k l m n p q r s t v w x y z

SUBSET V a e i o u

; more subsets

RULE „Defaults” 1 29

a b c d e f g h i j k l m n o p q r s t u v w x y z + @

a b c d e f g h i j k l m n o p q r s t u v w x y z 0 @

1: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

RULE "Voicing s:z <=> V____V" 4 4

V s s @

V z @ @

1: 2 0 1 1

2: 2 4 3 1

3: 0 0 1 1

4: 2 0 0 0

; more rules



Az angol morfológia kétszintes leírása

(Karttunen & Wittenburg 1985)

Automata 1.

ALPHABET

```
a b c d e f g h i j k l m n o p q r s t u v w x y z '
+ `
NULL 0
ANY =
SUBSET V a e i o u
SUBSET C b c d f g h j k l m n p q r s t v w x z
SUBSET S s x z
END
```

"Surface Characters" 1 28

```
a b c d e f g h i j k l m n o p q r s t u v w x y z ' =
a b c d e f g h i j k l m n o p q r s t u v w x y z ' =
1: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

"Epenthesis" 6 8

```
c h s S y + + =
c h s S i e 0 =
1: 2 1 4 3 3 0 1 1
2: 2 3 3 3 3 0 1 1
3: 2 1 3 3 3 5 6 1
4: 2 3 3 3 3 5 6 1
5: 0 0 1 0 0 0 0 0
6: 1 1 0 1 1 1 1 1
```

"Gemination" 16 26

```
V b d f g l m n p r s t + + + + + + + + + + + ` =
V b d f g l m n p r s t b d f g l m n p r s t 0 0 =
1: 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
2: 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
3: 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
4: 16 5 6 7 8 9 10 11 12 13 14 15 0 0 0 0 0 0 0 0 0 0 0 0 0 1 116
5: 16 16 16 16 16 16 16 16 16 16 16 16 16 2 0 0 0 0 0 0 0 0 0 0 0 3 1 1
6: 16 16 16 16 16 16 16 16 16 16 16 16 16 16 0 2 0 0 0 0 0 0 0 0 0 3 1 1
7: 16 16 16 16 16 16 16 16 16 16 16 16 16 16 0 0 2 0 0 0 0 0 0 0 0 3 1 1
8: 16 16 16 16 16 16 16 16 16 16 16 16 16 16 0 0 0 2 0 0 0 0 0 0 0 3 1 1
9: 16 16 16 16 16 16 16 16 16 16 16 16 16 16 0 0 0 0 2 0 0 0 0 0 0 3 1 1
10: 16 16 16 16 16 16 16 16 16 16 16 16 16 16 0 0 0 0 0 2 0 0 0 0 0 3 1 1
11: 16 16 16 16 16 16 16 16 16 16 16 16 16 16 0 0 0 0 0 0 2 0 0 0 0 3 1 1
12: 16 16 16 16 16 16 16 16 16 16 16 16 16 16 0 0 0 0 0 0 0 2 0 0 0 3 1 1
13: 16 16 16 16 16 16 16 16 16 16 16 16 16 16 0 0 0 0 0 0 0 0 2 0 0 3 1 1
14: 16 16 16 16 16 16 16 16 16 16 16 16 16 16 0 0 0 0 0 0 0 0 0 2 0 3 1 1
15: 16 16 16 16 16 16 16 16 16 16 16 16 16 16 0 0 0 0 0 0 0 0 0 0 2 3 1 1
16: 16 16 16 16 16 16 16 16 16 16 16 16 16 16 0 0 0 0 0 0 0 0 0 0 0 16 116
```


Az angol morfológia kétszintes leírása

(Karttunen & Wittenburg 1985)

Automata 2.

"Y-Spelling" 6 7

C y y + i a =
C y i = i a =
1: 2 1 0 1 1 1 1
2: 2 5 3 1 1 1 1
3: 0 0 0 4 0 0 0
4: 1 1 0 1 0 0 1
5: 1 1 0 6 1 1 1
6: 0 0 0 0 1 1 0

"Elision" 15 8

V i e e + g c =
V i e 0 0 g c =
1: 2 2 3 4 1 11 11 1
2: 1 1 5 6 1 11 11 1
3: 1 1 5 6 9 11 11 1
4: 0 0 0 0 8 0 0 0
5: 1 1 1 4 7 11 11 1
6: 0 0 0 0 10 0 0 0
7: 1 1 0 0 0 1 1 1
8: 1 1 1 0 0 0 0 0
9: 0 0 0 0 0 1 1 1
10: 0 0 1 0 0 0 0 0
11: 1 1 14 12 1 1 1 1
12: 0 0 0 0 13 0 0 0
13: 0 1 1 0 0 0 0 0
14: 1 1 1 0 15 11 11 1
15: 1 0 0 0 1 11 11 1

"I-Spelling" 7 6

i e + i e =
y 0 0 i e =
1: 2 1 1 5 1 1
2: 0 3 0 0 0 0
3: 0 0 4 0 0 0
4: 0 0 0 1 0 0
5: 1 1 1 1 6 1
6: 0 1 7 0 0 1
7: 0 0 0 0 0 1

END

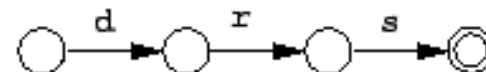
Nemkonkatenatív morfológiák FST-ben

- ❑ **Az összefésülő algoritmus:** egy olyan mintakitöltő művelet, mely két reguláris nyelvet kombinál: a mintát (template) és a kitöltőt (filler) egyetlen reguláris alakká
- ❑ A minta kezdőállapotából kiindulva az algoritmus megpróbálja megtalálni az összes megfelelő illesztést a mintaélek és a kitöltőélek között
- ❑ Az illesztés akkor sikeres, ha a kitöltőél címkéje benne van abban az osztályban, amit a mintaél címkéje határoz meg ($d \in C, r \in C, s \in C, u \in V, i \in V$)

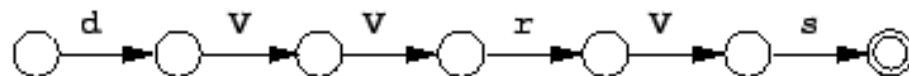
Minta:



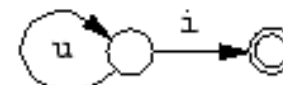
Mássalhangzókitöltő:



Köztes eredmény:



Magánhangzókitöltő:



Végeredmény:



A kétszintes leírás néhány nehézsége

- ☐ A felszín és a lexikális alak kötelezően azonos hosszúsága
- ☐ A szuppletív (lexikalizálódott) alakok és a nem produktív toldalékolás kezelése
- ☐ Ha a szótár „mindent kibír”, miért kell a „nehéz” alakokat is levezetésekkel kezelni?
- ☐ Pl. *jöv+ök, jösz+sz, jön+0, jöt+tök, jö+het, ...*
- ☐ *..., jó, ..., gyere, gyerünk, gyertek*
- ☐ Mit ér a reguláris rendszer, ha vannak reguláris nyelvvel nem leírható morfológiai jelenségek?
- ☐ Pl. *nagy, nagy+obb, leg+nagy+obb*



A nyelvtechnológia alapjai 4.

Unifikációs morfológia,
szófaji egyértelműsítés,
alkalmazások

2019/2020. tanév, I. félév

Gráfok, fák, DAG-ok

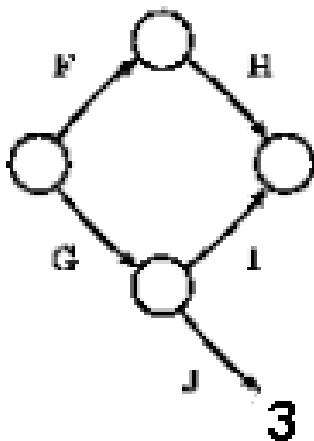
- **Gráf:** egy olyan objektumhalmaz reprezentációja, ahol egyes objektumpárokat kapcsolatok kötnek össze; az objektumokat csomópontoknak, az összekötéseket éleknek mondjuk
- **Ösvény:** egy adott u_0 csomópontból kiinduló, u_n -be vezető olyan élsorozat, melyben minden él u_{i-1} -ből u_i -be ($i=1, \dots, n-1$) vezető él végpontjában levő csomópontból kiindul egy u_{i+1} -be vezető él
- u és v csomópontok **kapcsoltak**, ha a gráfnak van u -ból v -be menő ösvénye, és a gráf **összefüggő** gráf, ha minden különböző csomópontja kapcsolt
- **Írányított gráf:** csomópontok halmaza élek rendezett párjaival
- **Gyökér:** speciális csomópont, amelyből bármely csomóponthoz vezet ösvény
- **Levél:** kimenő él nélküli csomópont
- **Írányított körmentes gráf (DAG):** irányított gráf, kör nélkül

Az unifikáció

- A jegyszerkezetek **irányított körmentes gráfokkal** reprezentálhatók, ahol a csomópontok felelnek meg a változóknak és az ösvények a változóneveknek
- A jegyszerkezeteket leggyakrabban jegy-érték **mátrixokként** írják le, melyeknek két oszlopuk van (az egyik a jegyek nevének, a másik az értékeinek)
- Az **érték** vagy atomi érték vagy egy másik jegyszerkezet
- **Jegyszerkezet-unifikáció:** két jegyszerkezeten működő művelet, mely csak akkor hiúsul meg, ha van olyan ösvény mindkét szerkezetben, melyeknek eltérő atomi értékei vannak; minden más esetben a létrejövő szerkezet az eredeti két szerkezet összes ösvényét tartalmazza (az azonosakat egybeejtve)
- Az unifikáció absztrakt művelet, és **NEM** a morfológiához tartozik, csak a legegyszerűbb formáját itt tárgyaljuk először

Jegyszerkezetek mint DAG-ok

A jegyszerkezet egy gráf:



A jegyszerkezet mint attribútum-érték mátrix:

$$\left[\begin{array}{c|c} F & H & \boxed{1} \\ \hline G & \begin{bmatrix} I & \boxed{1} \\ J & 3 \end{bmatrix} \end{array} \right]$$

•Source: Frederik Fouvry - <http://www.dfki.de/compling/pdfs/ubg-slides.pdf>

Az unifikáció művelete

Az unifikáció olyan művelet,
mely egyedül akkor nem hajtható végre,
ha a tagokban egyazon jegy
különböző értékekkel szerepel:

[NUMBER SG] U [NUMBER PL] = *fail!*

de:

[NUMBER SG] U [NUMBER SG] = [NUMBER SG]

[NUMBER SG] U [NUMBER []] = [NUMBER SG]

[NUMBER SG] U [PERSON 3] = [[NUMBER SG][PERSON 3]]

Példák unifikációra

$$\begin{bmatrix} \text{CATEGORY} & \textit{noun} \end{bmatrix} \sqcup \begin{bmatrix} \text{NUMBER} & \textit{singular} \end{bmatrix} = \begin{bmatrix} \text{CATEGORY} & \textit{noun} \\ \text{NUMBER} & \textit{singular} \end{bmatrix}$$

$$\begin{bmatrix} \text{CAT} & [] \end{bmatrix} \sqcup \begin{bmatrix} \text{CAT} | \text{CASE} & \textit{accusative} \end{bmatrix} = \begin{bmatrix} \text{CAT} | \text{CASE} & \textit{accusative} \end{bmatrix}$$

$$\begin{bmatrix} \text{F} & [1] \\ \text{H} & [1] \end{bmatrix} \sqcup \begin{bmatrix} \text{F} & [] \\ \text{H} | \text{G} & [] \end{bmatrix} = \begin{bmatrix} \text{F} & [1] \\ \text{H} & [1] \end{bmatrix} \begin{bmatrix} \text{G} & [] \end{bmatrix}$$

$$\begin{bmatrix} \text{CATEGORY} & \textit{noun} \end{bmatrix} \sqcup \begin{bmatrix} \text{CATEGORY} & \textit{verb} \end{bmatrix} = \text{fail}$$

$$\begin{bmatrix} \text{AGR} & [1] \\ \text{SUBJ} & \begin{bmatrix} \text{AGR:} & [1] \end{bmatrix} \end{bmatrix} \begin{bmatrix} \text{NUM} & \textit{sg} \end{bmatrix} \sqcup \begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{AGR} & \begin{bmatrix} \text{PERS} & \textit{third} \end{bmatrix} \end{bmatrix} \end{bmatrix} = \begin{bmatrix} \text{AGR} & [1] \\ \text{SUBJ} & \begin{bmatrix} \text{AGR} & [1] \\ \text{NUM} & \textit{sg} \\ \text{PERS} & \textit{third} \end{bmatrix} \end{bmatrix}$$

•Shieber, S.. An introduction to unification-based approaches to grammar . CSLI Lecture Notes 15. Stanford (1986)

Az unifikáció implementációkban

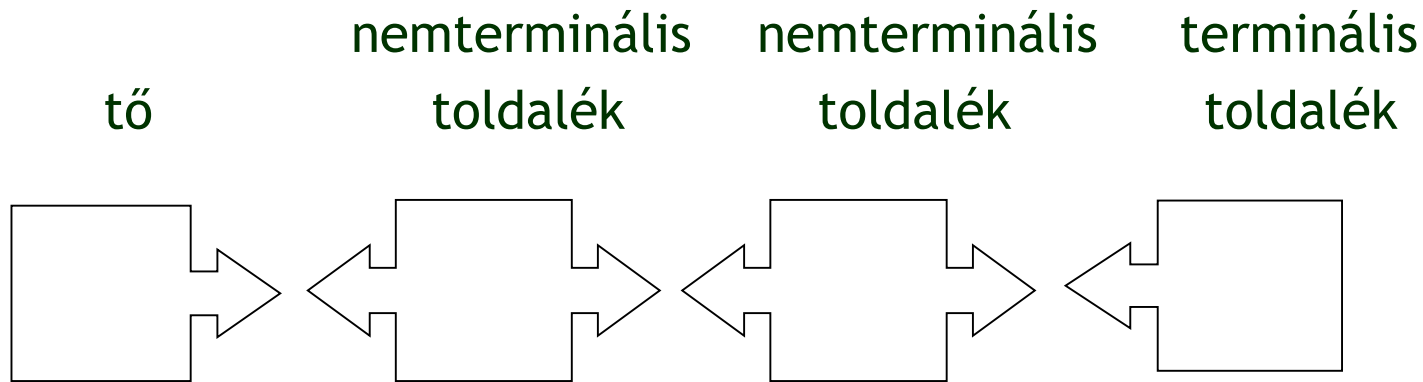
Destruktív unifikáció:

- Két szerkezet unifikációjakor az egyik megszűnik, és helyén létrejön az eredményszerkezet
- Gyors
- Bizonyos esetekben nem kívánt hatása is lehet (pl. ha egy szabályt alkalmazunk, annak formáját nem szeretnénk a művelet következtében megváltoztatni)

Nem-destruktív unifikáció (unifikálhatóság-ellenőrzés):

- A kiinduló szerkezetek nem változnak, az eredmény egy új, harmadik szerkezet
- Az implementációkban igen gyakori
- Könnyű kezelni, viszont másolási műveletet igényel
- Az unifikálhatóság-ellenőrzés egy reláció, igen/nem eredménnyel, más kimenet nélkül

Szóalaktani alapséma



(relatív) tő / relatív todalék

relatív tő / relatív todalék

relatív tő / (relatív) todalék

A HUMOR morfológiai rendszer

- ❑ High-speed Unification Morphology
- ❑ Az unifikáció mint művelet NEM a morfológia sajátja (lesz ilyen a többi nyelvi szinteken is), csak itt tárgyaljuk először
- ❑ Unifikáció van a logikai programozásban is, csak ott az argumentumok sorrendje is rögzített (Prolog)
- ❑ A lényeg: jegy-érték párok
- ❑ Az unifikáció destruktív művelet, de a HUMOR esetében csak unifikálhatósági relációról beszélünk
- ❑ Minden tulajdonság jegyként jelenik meg (az is, ami literálisan is jól látszik!)
- ❑ A folytatási osztályok találkozási pontjain működik
- ❑ Nincs más aktív művelet, csak az unifikálhatóság-ellenőrzés

Bináris kérdések a magyar morfo-fonológiáról

		$\alpha = +$	$\alpha = -$
1	α névszó	névszó	ige
2	α fn	főnév	melléknév, számnév
3	α szótári	szótári alapalak	nem szótári alapalak
4	α elöl	elöl képzett	hátsó képzett
5	α kerek	ajakkerekítéses	nem ajakkerekítéses
6	α PL	többes szám	nem állhat többes számban
7	α PLkötő	PL kötőhanggal	PL nem kötőhanggal
8	α PERS	birt. szem.ragos	nem kap birt. szem.ragot
9	α ACC	van tárgyesete	nem tárgyesetes
10	α ACCkötő	ACC kötőhanggal	ACC nem kötőhanggal
11	α DAT	van részesesete	nincs részesesete
12	α INS: β	van eszk.h.esete	nincs eszk.h.esete
13	α ÁS	-ás/-és képzős	nem kap -ás/-és képzőt
14	...		

Szótövek tára

hó []
[+névszó +fn +szótári -elől -kerek -PL
-PERS -ACC -ACCKötő +DAT +INS:V]

hav []
[+névszó +fn -szótári -elől -kerek +PL
+PLkötő +PERS +ACC -DAT -INS]

képez []
[-névszó +szótári +elől -kerek -ÁS]

képz []
[-névszó -szótári +elől -kerek +ÁS]

Toldalékok tára

<i>ás</i>	$[-\text{névszó} -\text{elöl} +\text{ÁS}]$ $[+\text{névszó} +\text{fn} +\text{szótári} -\text{elöl} -\text{kerek} +\text{PL}$ $+ \text{PLkötő} +\text{ACC} -\text{ACCkötő} +\text{DAT} +\text{INS:S}]$
<i>és</i>	$[-\text{névszó} +\text{elöl} +\text{ÁS}]$ $[+\text{névszó} +\text{fn} +\text{szótári} +\text{elöl} -\text{kerek} +\text{PL}$ $+ \text{PLkötő} +\text{ACC} -\text{ACCkötő} +\text{DAT} +\text{INS:S}]$
<i>ak</i>	$[+\text{névszó} -\text{elöl} -\text{kerek} +\text{PL} +\text{PLkötő}]$ $[+\text{névszó} -\text{elöl} -\text{kerek} -\text{PL} -\text{PERS} +\text{ACC}$ $+ \text{ACCkötő} +\text{DAT} +\text{INS:K}]$
<i>ek</i>	$[+\text{névszó} +\text{elöl} -\text{kerek} +\text{PL} +\text{PLkötő}]$ $[+\text{névszó} +\text{elöl} -\text{kerek} -\text{PL} -\text{PERS} +\text{ACC}$ $+ \text{ACCkötő} +\text{DAT} +\text{INS:K}]$
<i>nak</i>	$[+\text{névszó} -\text{elöl} +\text{DAT}]$ $[]$
<i>nek</i>	$[+\text{névszó} +\text{elöl} +\text{DAT}]$ $[]$

Az unifikáció pseudokódos definíciója

function UNIFY(*f1*, *f2*) **returns** *fstructure* or failure

f1-real $\leftarrow$ Real contents of *f1*

f2-real $\leftarrow$ Real contents of *f2*

if *f1-real* is null **then**

f1.pointer $\leftarrow$ *f2*

return *f2*

else if *f2-real* is null **then**

f2.pointer $\leftarrow$ *f1*

return *f1*

else if *f1-real* and *f2-real* are identical **then**

f1.pointer $\leftarrow$ *f2*

return *f2*

else if both *f1-real* and *f2-real* are complex feature structures **then**

f2.pointer $\leftarrow$ *f1*

for each *feature* **in** *f2-real* **do**

other-feature $\leftarrow$ Find or create

 a feature corresponding to *feature* in *f1-real*

if UNIFY(*feature.value*, *other-feature.value*) **returns** failure **then**

return failure

return *f1*

else return failure



Az unifikációs morfológia működése

<i>hó</i>	[+névszó +fn +szótári -elöl -kerek -PL -PERS -ACC -ACCKötő +DAT +INS:V]
<i>*hav</i>	[+névszó +fn <u>-szótári</u> -elöl -kerek +PL +PLkötő +PERS +ACC +DAT -INS]
<i>hó+nak</i>	[+névszó +fn +szótári -elöl -kerek -PL -PERS -ACC -ACCKötő +DAT +INS:V] [+névszó -elöl +DAT]
<i>*hav+nak</i>	[+névszó +fn -szótári -elöl -kerek +PL +PLkötő +PERS +ACC <u>-DAT</u>] [+névszó -elöl <u>+DAT</u>]
<i>hav+at</i>	[+névszó +fn -szótári -elöl -kerek +PL +PLkötő +PERS +ACC -DAT] [+névszó -elöl +ACC]
<i>*hó+vel</i>	[+névszó +fn +szótári <u>-elöl</u> -kerek -PL -PERS -ACC -ACCKötő +DAT +INS:V] [+névszó <u>+elöl</u> +INS:V]
<i>képz+és+nek</i>	[-névszó -szótári +elöl -kerek +ÁS] [-névszó +elöl +ÁS] [+névszó +fn +szótári +elöl -kerek +PL +PLkötő +ACC -ACCKötő +DAT +INS:S] [+névszó +elöl +DAT]

Az unifikációról általában

- ☐ Az unifikáció mint művelet **NEM** a morfológiához tartozik, mindössze egy egyszerűbb változatát itt tárgyaljuk először
- ☐ Az unifikációt elsősorban irányított ciklusmentes gráfok (directed acyclic graph, DAG) esetében szokás használni
- ☐ A morfológiai leírás esetében egyszerűbb struktúrákon működtetjük az unifikációt
- ☐ A DAG több mint fa, és épp ezáltal használhatóbb nyelvi szerkezetek leírására, hogy egy csomópontnak a nyelvi szerkezetekben sokszor kell, hogy több szülője lehessen
- ☐ A közös csomópontok elsősorban a szintaktikai szerkezetek kezelésénél válnak fontossá

Morfo-fonológiai „guesser”

kacsónak	✓ N
kacsóna + k	* N + PL
kacsón + ak	* N + PL
kacsó + nak	* V + PL3
kacsó + nak	✓ N + DAT
kacs + ó + nak	* V + PART + DAT
ka csónak	* N N

Innentől a szóalak-gyakoriság dönt:

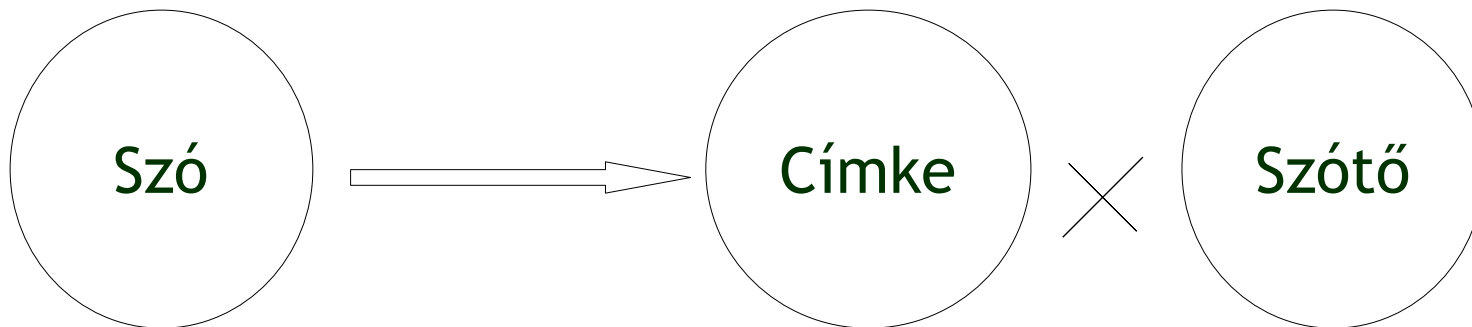
- kacsónak, *kacsónakot, *kacsónakra, ... → {}
- kacsó, kacsót, kacsóra, kacsói, ... → {kacsó}



Szófaji egyértelműsítés

A szófaji egyértelműsítési feladat

Egyértelműsítő:



{tokenizált mondat, címkekészlet}
→ **szófaji egyértelműsítő** →
{címkézett (és szótövezett) mondat}

Markov-modellek

- ☐ **Markov-lánc:** diszkrét sztochasztikus folyamatot ír le
- ☐ **Markov-folyamat:** adott jelenbeli állapot mellett a rendszer jövőbeni állapota nem függ a múltbeliektől
- ☐ **Rejtett Markov-modell (HMM):** csak a modell működésének az eredményét (a kimenetet, azaz a generált szekvenciát) ismerhetjük, a modell maga és a paraméterei számunkra ismeretlenek
- ☐ Következésképp csak a kimenetből következtethetünk a modell felépítésére és a működését leíró paraméterekre (az átmeneti és a kibocsátási valószínűségekre)
- ☐ A modell karakterisztikus tulajdonsága tehát a rejtett állapot, azaz a (sztochasztikus) állapotfejlődés és annak passzív (sztochasztikus) megfigyelése

Az alapvető szófaji egyértelműsítő módszerek

- ❑ Szabály-alapú módszerek
 - ❑ Megszorításos modellek
 - ❑ Transzformációs modellek
- ❑ Valószínűségi módszerek (W: szósorozat, T: címkesorozat)
 - ❑ Rejtett Markov-modellen alapuló módszerek → a legjobb címkesorozat, ami maximalizálja a $P(T|W)$ -t:

$$\begin{aligned}\hat{T} &= \operatorname{argmax}_T P(T|W) \\ &= \operatorname{argmax}_T P(W|T)P(T) \\ &= \operatorname{argmax}_T \prod_i P(\text{word}_i | \text{tag}_i) \prod_i P(\text{tag}_i | \text{tag}_{i-1})\end{aligned}$$

- ❑ Maximum entrópián alapuló módszerek → a $P(T|W)$ kiszámítása a lehetséges címkesorozatok közül választva:

$$\begin{aligned}\hat{T} &= \operatorname{argmax}_T P(T|W) \\ &= \operatorname{argmax}_T \prod_i P(t_i | w_i, t_{i-1})\end{aligned}$$

Egy megszorításos egyértelműsítő: EngCG

- ❑ Kb. 3600 szabály (az angolra)
- ❑ Pl. a (*That round*) *table might collapse*. mondat (vége):
 - "table" N NOM SG
 - "table" <SVO> V SUBJUNCTIVE VFIN @+FMAINV
 - "table" <SVO> V IMP VFIN @+FMAINV
 - "table" <SVO> V INF
 - "table" <SVO> V PRES -SG3 VFIN @+FMAINV
 - "might" <-Indef> N NOM SG
 - "might" V AUXMOD VFIN @+FAUXV
 - "collapse" N NOM SG
 - "collapse" <SV> <SVO> V IMP VFIN @+FMAINV
 - "collapse" <SV> <SVO> V INF
 - "collapse" <SV> <SVO> V PRES -SG3 VFIN @+FMAINV
 - "<\$.>"

Egy transzformációs módszer: a Brill-tagger

❑ Transzformáció-alapú tanulás:

{címkézett korpusz}

→ Brill →

{transzformációs szabályok sorozata}

❑ A szabályok formája:

ÍRD ÁT a z címkét y-ra, HA az előző címke x

ÍRD ÁT a z címkét y-ra, HA az előző szó w,

ÉS a következő szó címkéje z

❑ A rendszer működése:

- 1) Inicializálás: a leggyakoribb címkék választása
- 2) Szabályválasztás: a „leghasznosabb” szabály választása
- 3) Szabály alkalmazása a tanító korpuszra, majd 2

A TnT egyértelműsítő

- ❑ Thorsten Brants (2002)
- ❑ Másodrendű, simított Markov-modell:

$$\operatorname{argmax}_{t_1 \dots t_T} \left[\prod_{i=1}^T P(t_i | t_{i-1}, t_{i-2}) P(w_i | t_i) \right] P(t_{T+1} | t_T)$$

$$P(t_3 | t_1, t_2) = \lambda_1 \hat{P}(t_3) + \lambda_2 \hat{P}(t_3 | t_2) + \lambda_3 \hat{P}(t_3 | t_1, t_2)$$

- ❑ Fontosabb tulajdonságai:
 - ismeretlen szavak kezelése
 - végződések levágása
 - nagy kezdőbetűk figyelembevétele

Magyar egyértelműsítők

❑ hunpos (2007)

- BME MOKK
- a TnT OCaml-ben történt nyílt forráskódú újraimplementálása
- MSD kódok

❑ MagyarLánc (2010)

- SZTE
- Stanford POS-tagger + morfológiai elemző
- MSD kódok

❑ PurePOS (2013)

- PPKE ITK
- a TnT alapján Javában (azóta Pythonban is)
- integrált morfológiai elemzővel (Humor)

Mi egyértelmű?

Az utca végén levő legelőre mentem legeltetni.

Az = **az[DET]=Az**

az[NM]=Az

utca = **utca[FN]+[NOM]**

végén = **vég[FN]+é[POS]+n[SUP]**

vég[FN]+é[PSe3]+n[SUP]

levő = **lesz[IGE]=lev+ő[MIF]+[NOM]**

legelőre = **legelőre[HA]**

legelő[FN]+re[SUB]

legel[IGE]+ő[MIF]+re[SUB]

mentem = **ment[IGE]+em[Te1]**

ment[MN]+em[PSe1]

megy[IGE]=men+tem[Me1]

megy[IGE]=men+t[MIB]+em[PSe1]

legeltetni. = **legeltet[IGE]+ni[INF]+.[STOP]**

legel[IGE]+tet[MUV]+ni[INF]+.[STOP]



Számítógépes morfológiai alkalmazások

A hibák és normák kezelése („proofing tools”)

- ☐ A helyesírási normák, a szabályok és a gépi helyesírás-ellenőrzés viszonya
- ☐ A helyesírási szabályzatok viszonya a gépi helyesírás-ellenőrzéshez
- ☐ Formai és jelentéstani szempontok gépi kezelése
- ☐ A gép a „nyelvhelyességi tanácsadó” szerepében
- ☐ A nehezen értelmezhető elemzések okozta problémák
- ☐ Szószinten javítható hibák: betűhibák és elütések, hibás egybeírások, hibás kötőjelezés, szóismétlések (!)
- ☐ Szószinten nem javítható hibák: hibás különírások, központoszási hibák, egyeztetési hibák, szórendhibák, vonatkozó névmások helytelen használata, stílushibák, helytelen szóhasználat, szóközök többszörözése vagy elhagyása, „értelmes” szóhibák

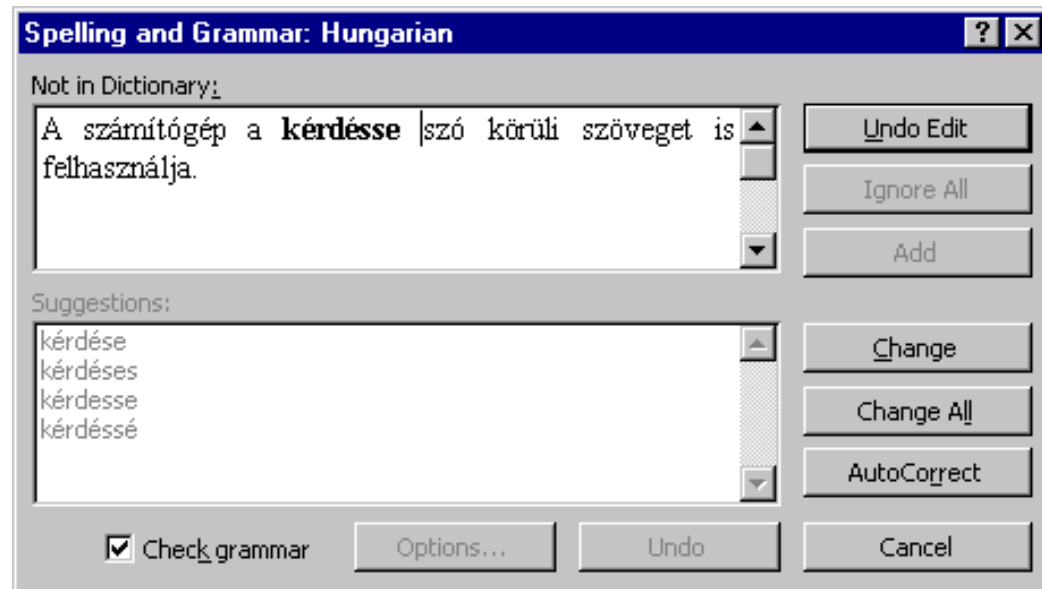
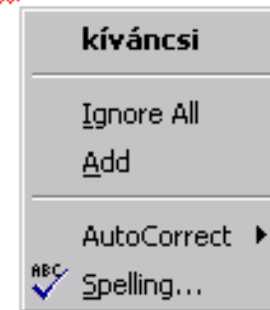
Tipikus hibák a számítógéppel létrehozott dokumentumokban

- ☐ Karakterhibák, elütések
- ☐ Valódi helyesírási hibák (általában szószintűek)
- ☐ Nyelvhelyességi hibák (általában nem szószintűek)
- ☐ Tipográfiai hibák (az átlag felhasználó nem is ismeri őket)
- ☐ A helyesírás-ellenőrzés lehetőségei és korlátai a szavak szintjén
- ☐ A szóellenőrzés és az ún. nyelvhelyesség-ellenőrzés viszonya
- ☐ A nyelvi programrendszer lehetséges hibái: a túlgenerálás (kör/kőr, alak/ház/tan + it/ít)

A szóellenőrző moduljai

- ☐ Alapszótár ill. morfológia
- ☐ Ajánlómodul és adatbázisai
- ☐ Időleges saját szótárak
- ☐ Kiegészítő szótár
- ☐ Kizáró szótár
- ☐ A ragozó nyelvek esetén hasznos toldalékoló saját szótárak lehetőségei és nehézségei

kíváncsi



Két füzér Levenshtein-távolsága

- ❑ A Levenshtein-távolság a Hamming-távolság egyfajta általánosítása
- ❑ Nemcsak egyforma hosszú füzérekre, és nemcsak betűcsere van az átalakításban, hanem beszúrás és törlés is

- ❑ Pszeudokódja:

```
int LevenshteinTavolsag(char str1[1..lenStr1], char str2[1..lenStr2])  
  // d: egy lenStr1+1 sorból és lenStr2+1 oszlopból álló mátrix  
  declare int d[0..lenStr1, 0..lenStr2]  
  // i, j: a str1 és str2 feletti iterációhoz  
  declare int i, j, cost  
  for i from 0 to lenStr1  
    d[i, 0] := i  
  for j from 0 to lenStr2  
    d[0, j] := j  
  for i from 1 to lenStr1  
    for j from 1 to lenStr2  
      if str1[i] = str2[j] then cost := 0  
      else cost := 1  
      d[i, j] := minimum(d[i-1, j] + 1,           // törlés  
                          d[i, j-1] + 1,          // beszúrás  
                          d[i-1, j-1] + cost)     // csere  
  return d[lenStr1, lenStr2]
```


Damerau-Levenshtein-távolság

- ❑ A Hamming-távolság egy másik általánosítása
- ❑ Minimális számú beszúrás, törlés, betűcsere mellett helycsere is van
- ❑ A Damerau-Levenshtein-távolság két füzér között a Levenshtein-számításból úgy kapható, hogy a fő ciklust kiegészítjük ezzel:
if($i > 2$ and
 $j > 2$ and
 $str1[i-1] = str2[j-2]$ and
 $str1[i-2] = str2[j-1]$)
 then $d[i,j] := \text{minimum}(d[i,j], d[i-2,j-2] + \text{cost})$ // *transzpozíció*
- ❑ Damerau annak idején (1964) azt is állította, hogy az emberi elütések 80%-a ezekkel korrigálható
- ❑ Ugyanakkor Damerau csak egyetlen karakteres elütésekkel számolt, szemben a Levenshtein-távolsággal

A szóellenőrzés menete

(1) Morfológiai elemzés

kérdesse ☞ <nincs ilyen szó a magyarban>

(2) Ajánlás

törlés:

érdesse, krdesse, kérésse, kédesse, kérdése, kérdéss

helycsere:

ékrdesse, krédesse, kérédsse, ..., kérdéses

nyelvspecifikus csere:

kérdéssé, kérdésse, ...

...

(3) Ellenőrzés a generátumok morfológiai elemzésével

kérdése, kérdéses, kérdésse, kérdéssé

Szóellenőrzés morfológiával

kérdése

kérdés[FN]+e[PSe3]

főnévi

kérd[IGE]+és[IF]+e[PSe3]

főnévi

kérdéses

kérdéses[MN]

melléknévi

kérdés[FN]+es[SKEP]

melléknévi

kérd[IGE]+és[IF]+es[SKEP]

melléknévi

kérdesse

kérd[IGE]+es[MUV]+se[TPe3]

igei

kérdéssé

kérdés[FN]+sé[FAC]

főnévi

kérd[IGE]+és[IF]+sé[FAC]

főnévi

Nyelvhelyesség-ellenőrzés a szóhatáron túl

- ☐ Lehetséges-e mondat szintű helyesírás-ellenőrzés?
- ☐ „Grammar checker” ?
- ☐ Parciális elemzések
- ☐ Hiba-nyelvtan vs. „szokásos” nyelvtan
- ☐ Hibaelemzések, a hibák súlyozása
- ☐ A hiba és a nem-hiba határának elmosódása
- ☐ A nyelvi vagy a formai természetű hibák szűrésének preferálása
- ☐ Stílusellenőrzés számítógéppel
- ☐ Környezetfüggő szóellenőrzés (pl. **a ház mellett**)

A mondat szintjén felismerhető hibajelenségek

- ☐ Névelő-egyeztetés: *Az kutya ugat.*
- ☐ Vesszőhiány: *Nem látok vak vagyok.*
- ☐ Algoritmikus szóösszetétel: *hat lábú*
- ☐ Szemantikus szóösszetétel: *kecske béka*
- ☐ Hibás szóösszetétel: *azután a nő után*
- ☐ Hiányos szerkezetek: *Megy hasú ebédelni.*
- ☐ Téves szóhasználat: *Egyenlőre nincs mit tenni.*
- ☐ Idegen szavak szűrése: *Elromlott a printer.*
- ☐ Terjengős kifejezések: *büntetést eszközöl*
- ☐ Trágár szavak szűrése: *Le van ***va.*
- ☐ Szóközhiány, -felesleg: *Nem ,én nem akarom.*

Mondatellenőrzés a gyakorlatban

Aki másnak vermet és maga esik bele.

Elment mellettem anélkül hogy köszönt volna.

A **hog**y kötőszó elé vesszőt kell írni!
Javaslat: **mellettem**„

Ignore Sentence

 Grammar...

Ha ez egy összetett mondat, a jelzett helyen
vagy környékén egy vessző hiányzik.
Ellenőrizze!
Javaslat: **maga**„

Ignore Sentence

 Grammar...

Spelling and Grammar: Hungarian

Grammatical error:

Elszeretném adni az autómat.

Suggestions:

Az igekező nem a segédigéhez tartozik.
Javaslat: **El szeretném**.



☒ Check grammar

Options...

Undo

Helyesírási Szabályzat

File Edit Bookmark Options Help

Contents Search Back Print

131.c A magyar helyesírás szabályai 11. kiadás

Az igekező külön szó marad, ha közte és igeje (vagy igeeneve) között más szó is van: *el ne késs, le nem tenné, újjá is építjük, közre kell bocsátani, haza szabad menni, föl sem véve* stb. Hasonlóképpen: *be nem avatkozás, meg nem értés* stb.

Az igekezők különírása vagy egybeírása gyakran értelmi különbségre is utal: annak jelölője, hogy az igekező a közvetlenül utána álló igehez tartozik-e, vagy az ezt követőhöz. Más és más tehát: *megvan húszéves*, de: *meg van töltve*; *megfogom a lepkét*, de: *meg fogom nézni*; *leszokott a dohányzásról*, de: *le szokott utazni vidékre*; *beleszeretett a lányba*, de: *bele szeretett volna szólni*; *elleszek nélküle*, de: *el leszek fáradva*; *megtudták a hírt*, de: *meg tudták javítani*; stb.

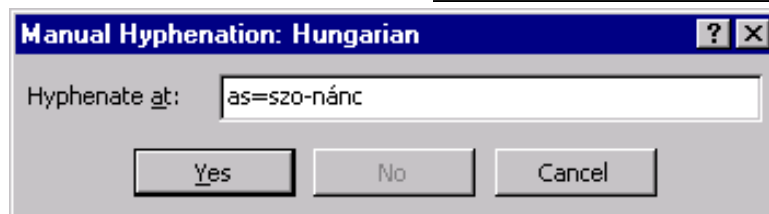
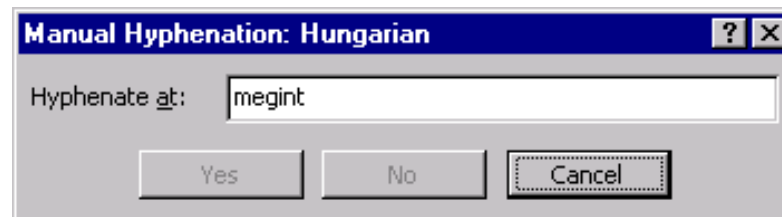
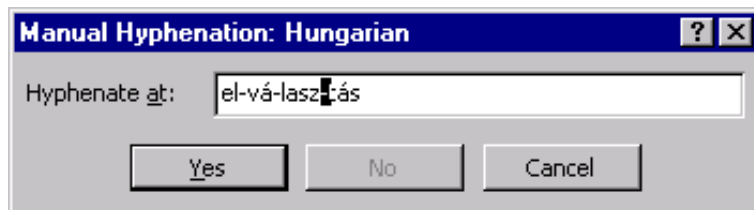
Automatikus szövegelválasztás

- ☐ Az elválasztás alkalmazása
- ☐ Automatikus és interaktív elválasztó módszerek
- ☐ Az elválasztás alapszabályai
- ☐ A morfológiai felülbíráltás kérdése
- ☐ Alternatív elválasztások kezelése
(a többértelműségek, illetve a szabályok „engedékenysége” miatt)
- ☐ Szótagolás és elválasztás: a tipográfiai szokások hatása az elválasztásra
- ☐ Különleges elválasztások (hosszú kettős mássalhangzók, mássalhangzó-háromszorozódás) helyes kezelése

A magyar elválasztás alapszabályai

Alap	Elválasztva	Példa
#VV	#VV	<u>autó</u>
#VC	#VC	<u>al</u> aki
VV#	VV#	haz <u>a</u> i
VV	V-V	<u>ba</u> -uxit
VC_1C_2V	VC_1-C_2V	<u>er</u> -kély
VC_iC_iV	VC_i-C_iV	<u>vet</u> -tem
VCc_1c_2V	$VC-c_1c_2V$	<u>mor</u> -zsa
Vc_1c_2CV	Vc_1c_2-CV	<u>asz</u> -tal
$Vc_{11}c_{12}c_{21}c_{22}V$	$Vc_{11}c_{12}-c_{21}c_{22}V$	<u>tarisz</u> -nya
$Vc_1c_1c_2V$	$Vc_1c_2-c_1c_2V$	össze/össz-sze

Elválasztási problémák



Más nyelven nehezebb a dolog (pl. angol):

- Oxford Advanced Learners' Dictionary (1989)

hyp-not-ize, hy-po-thesis, pro-vi-sional

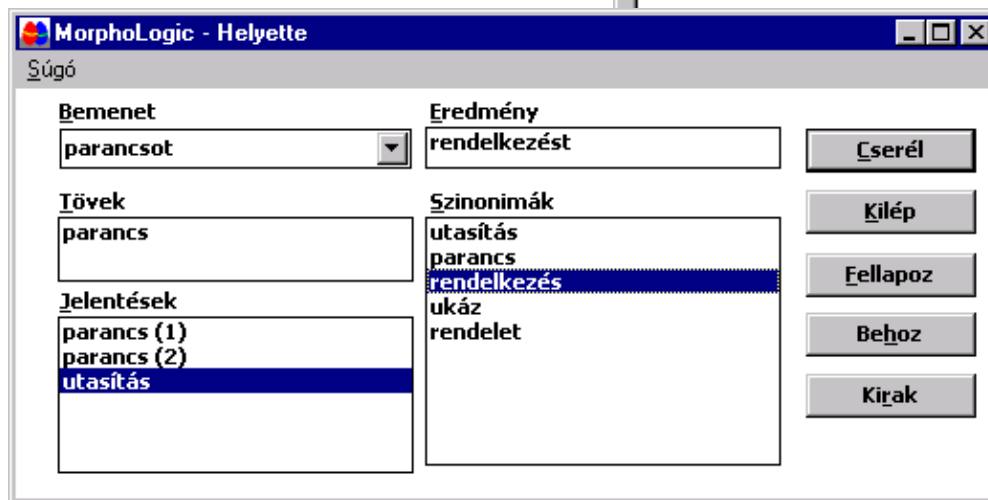
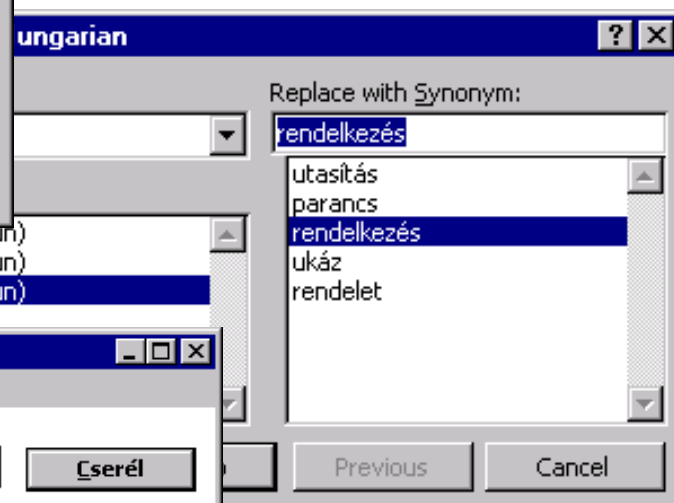
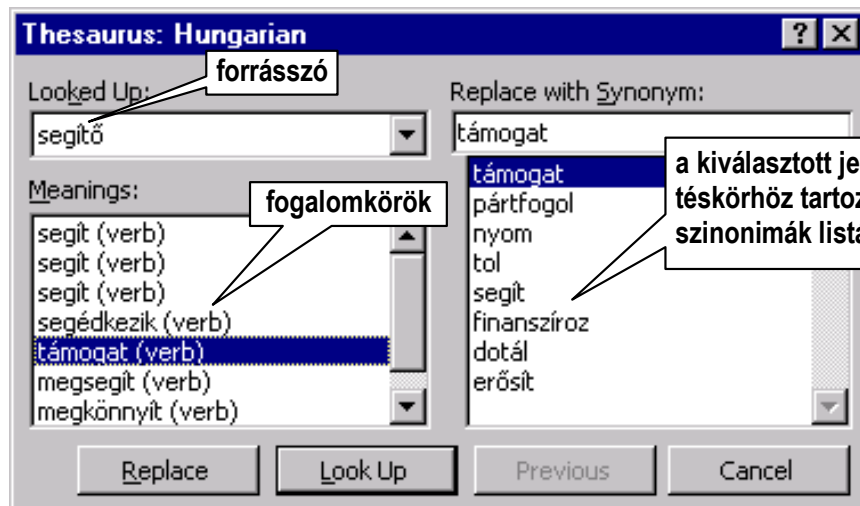
- Longman Dictionary of Contemporary English (1995)

hyp-no-tize, hy-poth-e-sis, pro-vi-sion-al

Számítógépes szinonimaszótárak és tezauruszok

- ☐ A szinonimákról
- ☐ Szinonimaszótár vagy tezaurusz?
- ☐ Tárolási és keresési problémák
- ☐ A rokon értelműség definíciója
- ☐ Az automatikus csere problémái
- ☐ Tő-visszaállítás
- ☐ Többértelműségek kezelése
- ☐ A lexikai és a szintaktikai szó különbségéből adódó nehézségek
- ☐ Az összetett szavak szinonimáinak problémája
- ☐ Morfológiai generálás minta alapján

Szinonimaszótárak





A nyelvtechnológia alapjai

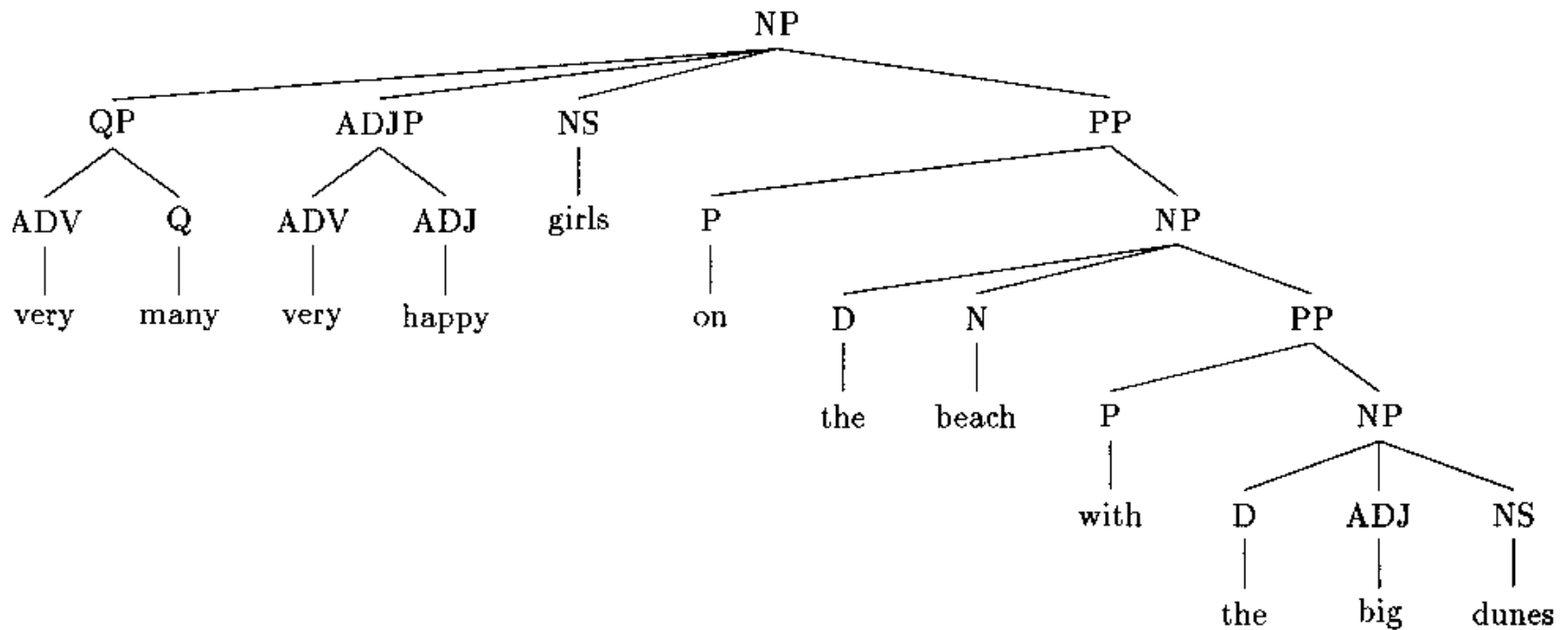
5.

Mondatszintaxis, mondatelemzés



Szintaktikai reprezentációk

Szintaktikai szerkezetek reprezentációja fákkal





A fareprezentáció kettős szerepe

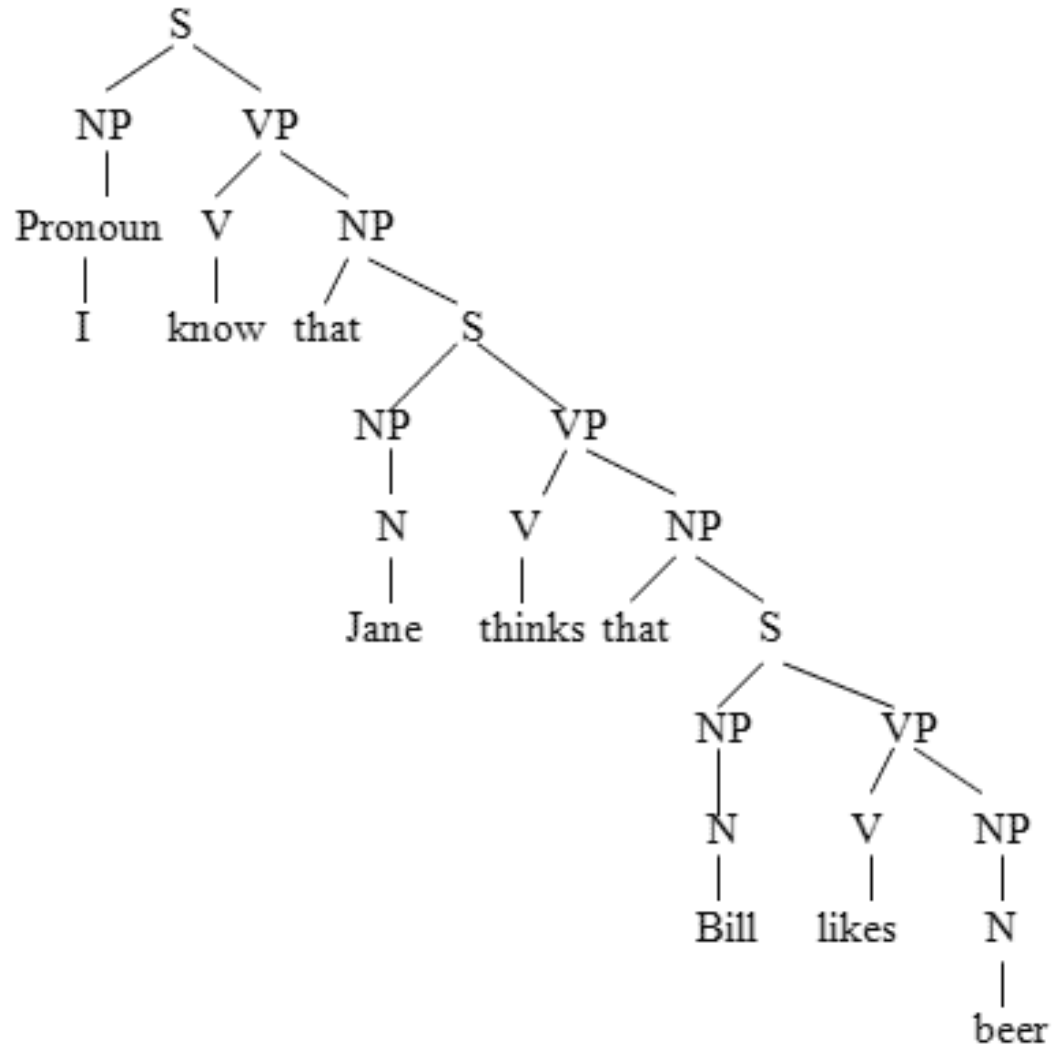
ID: 0 dominance

= mi függ mitől az adott szerkezetben?

LP: linear precedence

= mi mit előzhet meg?
(az egész nyelvtenban)

Közvetlen összetevők





Morfológia és szintaxis határán

Megbántottam őt.

$S \rightarrow VP \ PP$

$VP \rightarrow V$

$V \rightarrow \text{megbántottam}$

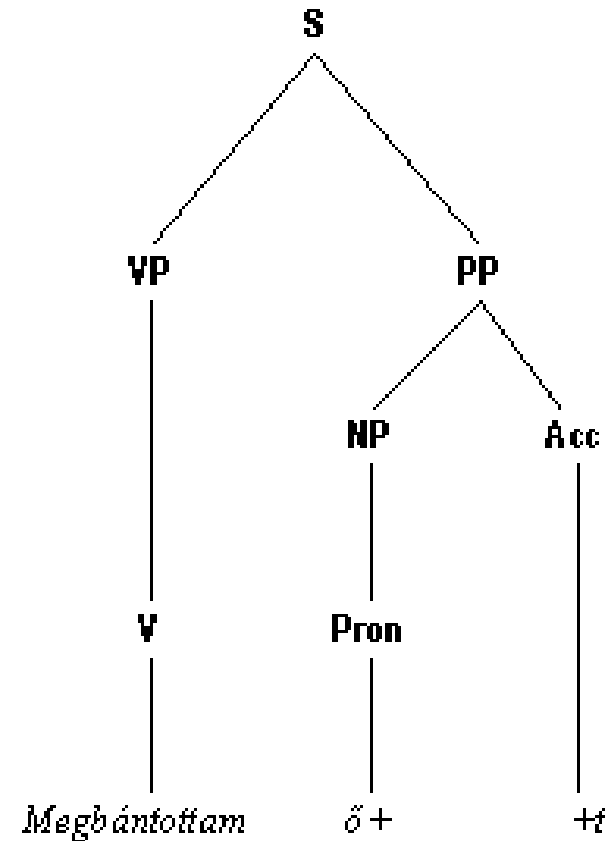
$PP \rightarrow NP \text{ Acc}$

$NP \rightarrow \text{Pron}$

ahol:

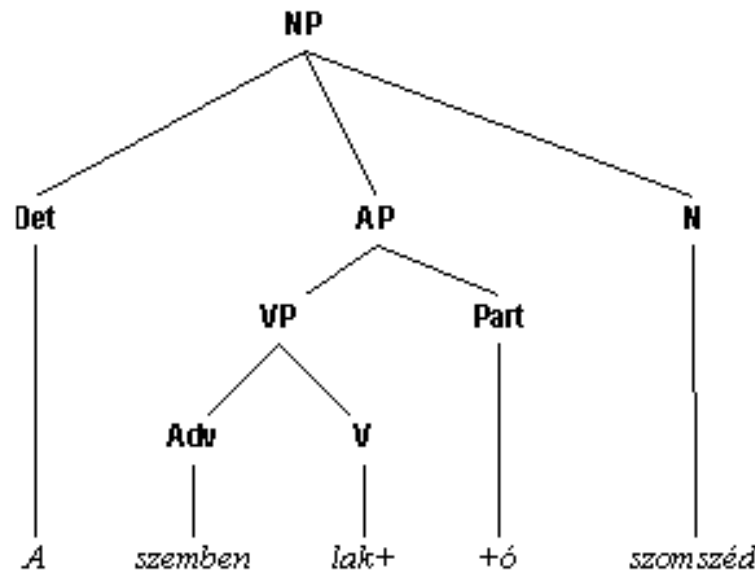
$\text{Pron} \rightarrow \text{ő}$

$\text{ACC} \rightarrow t \mid at \mid et \mid ot \mid öt$



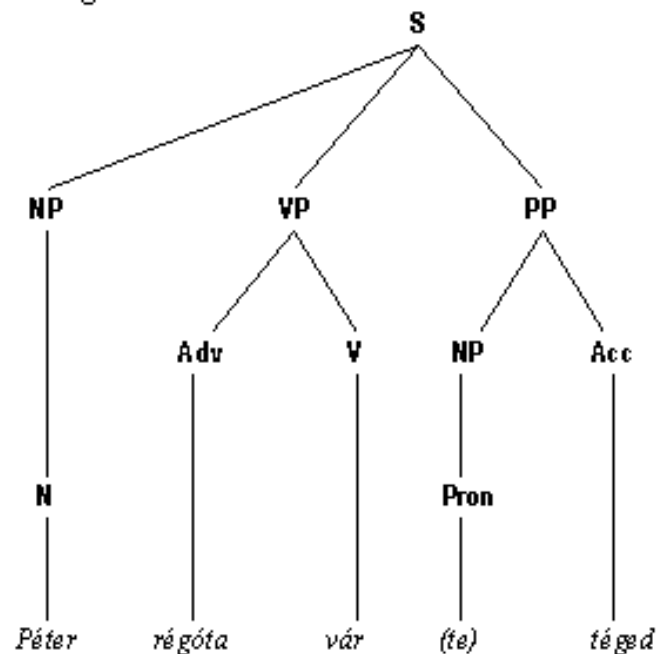
Projektív fa = címkézett zárójelezés

[NP [Det **A**] [AP [VP [Adv **szemben**] [V **lak**]] [Part **ó**]] [N **szomszéd**]]

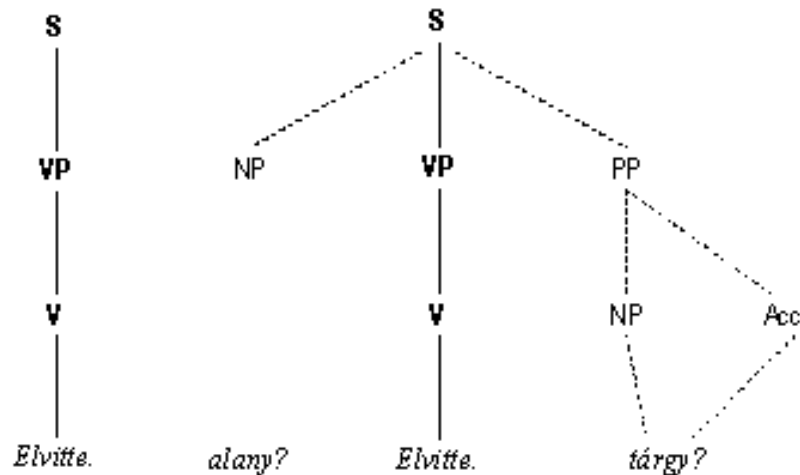


Problémás esetek: üres csomópontok

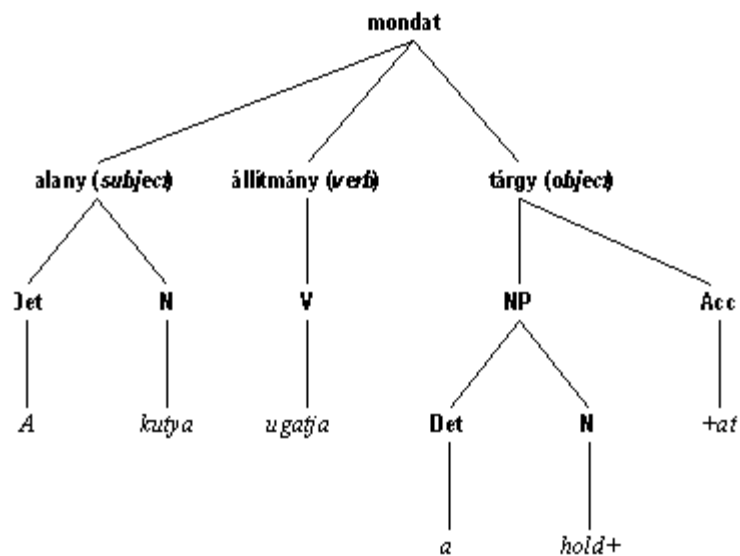
Péter régóta vár téged.



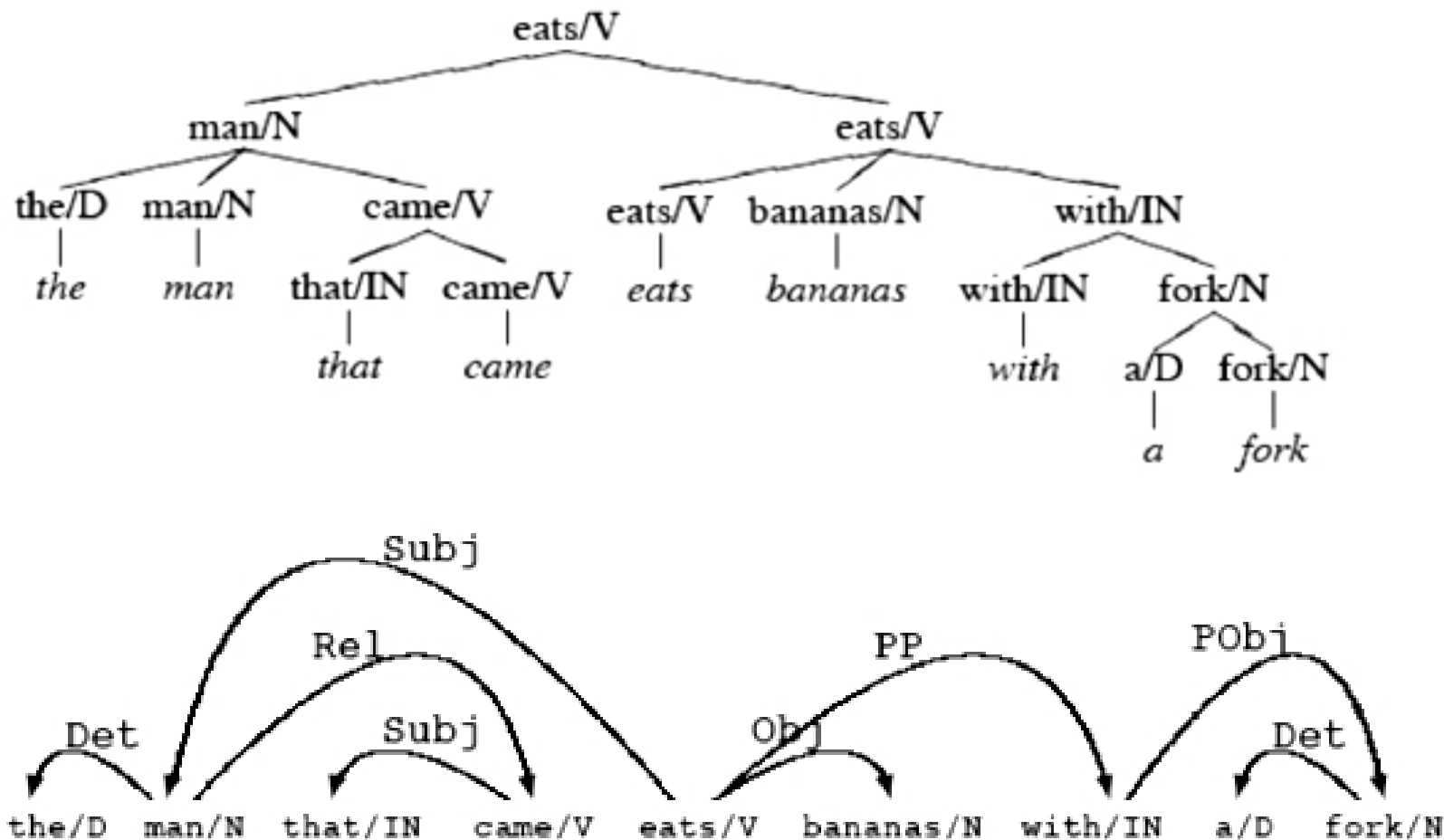
Elvitte.



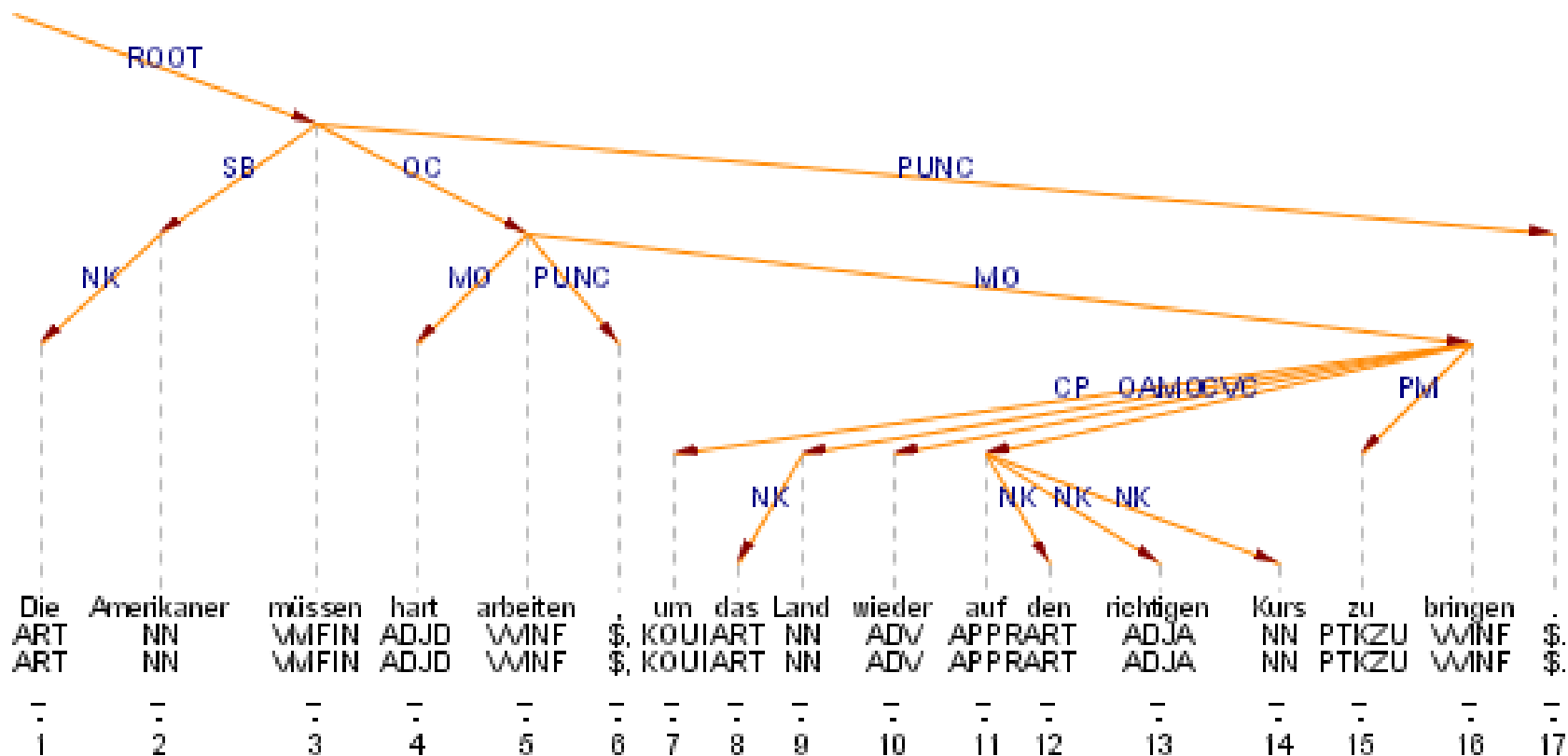
Funkcionális szerkezet



Összetevős és függőségi reprezentációk



Projektív függőségi fa



A mondat szerkezet-leírások összehasonlítása

- ❑ Közvetlen összetevős nyelvtanok:
előnyük a magasabb szintű kategóriák bevezetésének lehetősége,
hátrányuk a szintaktikai viszonyok egy részének „kifejezhetetlensége”
- ❑ Függőségi szerkezet:
előnyük a szintaktikai függőség kifejezésének lehetősége,
hátrányuk a magasabb szintű kategóriák kezelhetetlensége
- ❑ Egy elegáns közös megoldás: az X-vonás nyelvtanok

X-vonás: összetevők és függőség

- ❑ $S \rightarrow NP VP$
- ❑ Az összetevős szerkezetben az NP és a VP „testvérek”, azaz mindketten az S „gyermekai”, de ezt nem fejezi ki a függőségi leírás
- ❑ Azt viszont a közvetlen összetevős leírás nem fejezi ki, hogy testvérek bár, de nem egyforma súllyal, ui. a VP a szerkezet feje
- ❑ X-vonás szabályként: $V'' \rightarrow N' V'$
- ❑ Azaz: a V'' a V maximális projekciója, tehát a mondat feje az ige!
- ❑ Csak endocentrikus szerkezetekre!
(v.ö. exocentrikus)

Az X-vonás nyelvtanok tulajdonságai

- ❑ Megőrzi mind az összetevős, mind a függőségi leírás előnyeit
- ❑ Szófajok feletti általánosítás:
$$X = V / N / P / Adj / Adv / Det / Q / \dots$$
- ❑ Az X-nél magasabb szerkezeteket is egységesen lehet vele leírni
- ❑ Megkülönböztethetők a különféle jellegű bővítmények (pl. előtte, mögötte)
- ❑ Szintjei:
X (= fej, v. X-nullvonás) a lexikális szerkezet
X' (= X-vonás) a köztes szerkezet
X'' (= X-kétvonás v. X-max) a maximális szerkezet, a tulajdonképpeni frázis

Az X-vonás szabályok

- Az X-vonás szerkezet egy X fejből és egy opcionális komplementumból áll:

$$X' \rightarrow X \text{ (COMPLEMENT)}$$

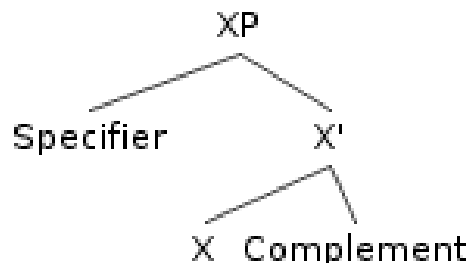
- Az X-vonás ezen kívül opcionálisan állhat egy X-vonás elemből és egy szabad bővítményből (ADJUNCT):

$$(X' \rightarrow X' \text{ ADJUNCT})$$

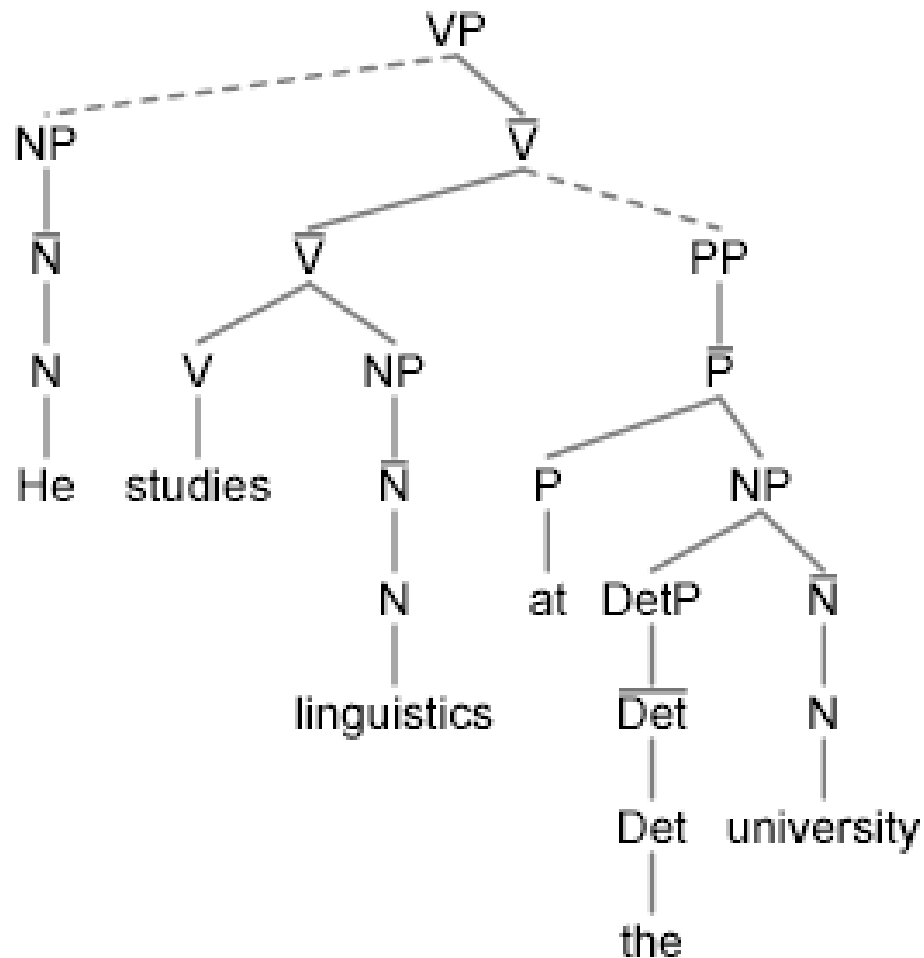
- Egy X-kétvonás (más néven XP)-szerkezet egy opcionális SPECIFIER elemből és egy X-vonásból áll:

$$XP \rightarrow (\text{SPECIFIER}) X'$$

- Az alap XP, azaz az X-vonás frázis tehát:



Példa egy X-vonás szerkezetre





Alapvető szintaktikai elemzési módszerek

Nyelvtan, környezetfüggetlenség, elemzés

- ☐ $G = (N, T, S, P)$, ahol
 - T: a szintaxisfa levelei a nyelvtan terminális szimbólumai
 - N: a szintaxisfa többi pontja, a nemterminális szimbólumok
 - S: a gyökérelem, a nyelvtan kezdőszimbóluma
 - P: a szabályok
- ☐ Két nyelvtan gyengén ekvivalens, ha ugyanazt a nyelvet ismeri fel, és erősen ekvivalens, ha ugyanazokat a szerkezeteket építi fel
- ☐ Az emberi nyelvekhez a
 - reguláris: kevés
 - környezetfüggő: sok! (?)
- ☐ Környezetfüggetlen (Chomsky 2-es típusú) nyelvtan:
$$A \rightarrow \alpha, \quad \text{ahol } A \in N, \alpha \in (N \cup T)^*$$
- ☐ A szintaxisfa belső részének felépítésére több módszer létezik
- ☐ Ha az S szimbólumból kiindulva építjük fel a szintaxisfát: felülről-lefelé (TD) haladó elemzés
- ☐ Ha a szintaxisfa építése a levelekből kiindulva halad az S szimbólum felé: alulról-felfelé (BU) elemzés

Egy „hagyományos” mininyelvten angolra

$S \rightarrow NP VP$

$S \rightarrow Aux NP VP$

$S \rightarrow VP$

$NP \rightarrow Det Nominal$

$NP \rightarrow Proper-Noun$

$VP \rightarrow Verb$

$VP \rightarrow Verb NP$

$Nominal \rightarrow Noun$

$Nominal \rightarrow Noun Nominal$

$Nominal \rightarrow Nominal PP$

$PP \rightarrow Prep NP$

$Det \rightarrow that \mid this \mid a$

$Noun \rightarrow book \mid flight \mid meal \mid money$

$Verb \rightarrow book \mid include \mid prefer$

$Aux \rightarrow does$

$Prep \rightarrow from \mid to \mid on$

$Proper-Noun \rightarrow Houston \mid TWA$



Alulról-felfelé (BU) elemzés

(Book that flight)

 $S \rightarrow NP VP$ $S \rightarrow Aux NP VP$ $S \rightarrow VP$ $NP \rightarrow Det Nominal$ $Nominal \rightarrow Noun$ $Nominal \rightarrow Noun Nominal$ $NP \rightarrow Proper-Noun$ $VP \rightarrow Verb$ $VP \rightarrow Verb NP$

Book that flight

Noun Det Noun
| | |
Book that flight

Verb Det Noun
| | |
Book that flight

NOM NOM
| |
Noun Det Noun
| | |
Book that flight

NOM
|
Verb Det Noun
| | |
Book that flight

NP
| |
NOM NOM
| |
Noun Det Noun
| | |
Book that flight

VP NOM
| |
Verb Det Noun
| | |
Book that flight

NP
| |
Verb Det Noun
| | |
Book that flight

VP NP
| |
Verb Det NOM
| | |
Book that flight

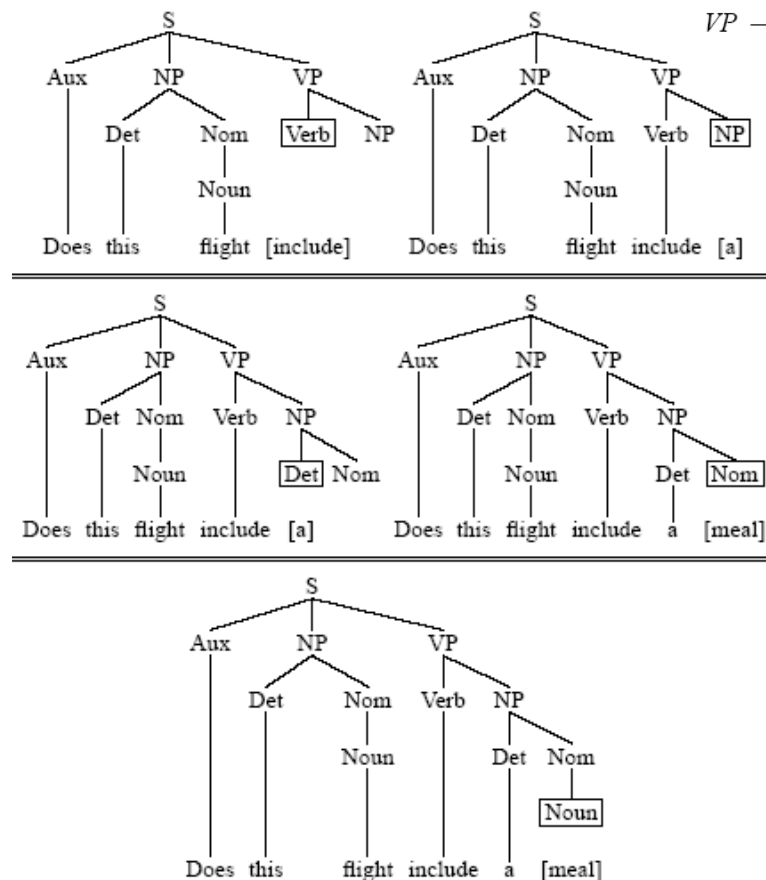
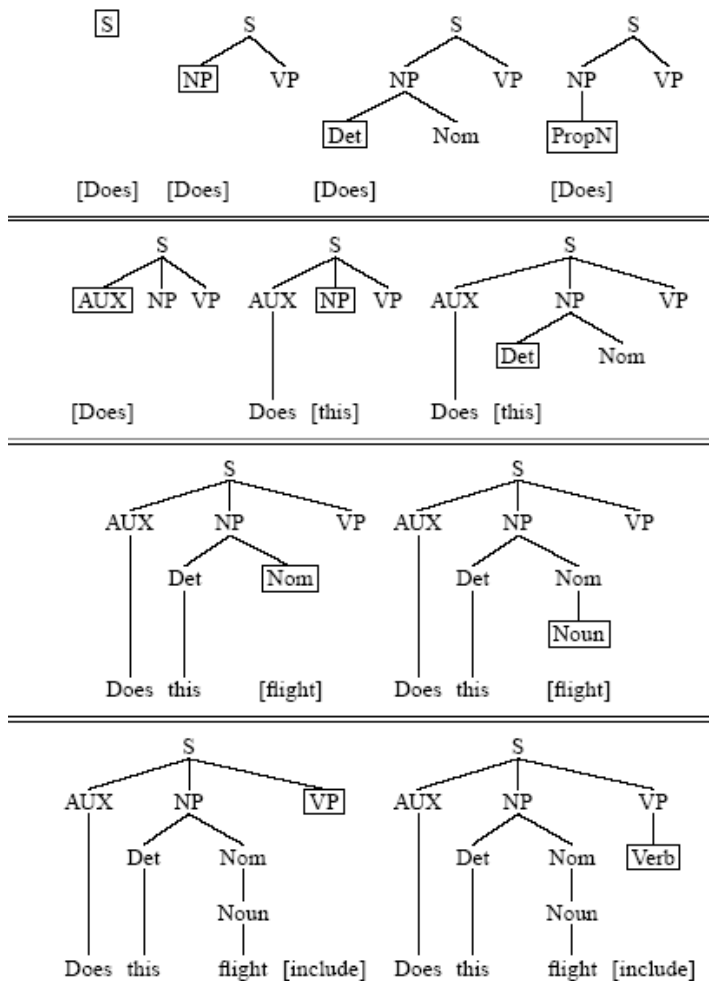
VP NP
| |
Verb Det NOM
| | |
Book that flight

S
|
VP NP
| |
Verb Det Nom
| | |
Book that flight

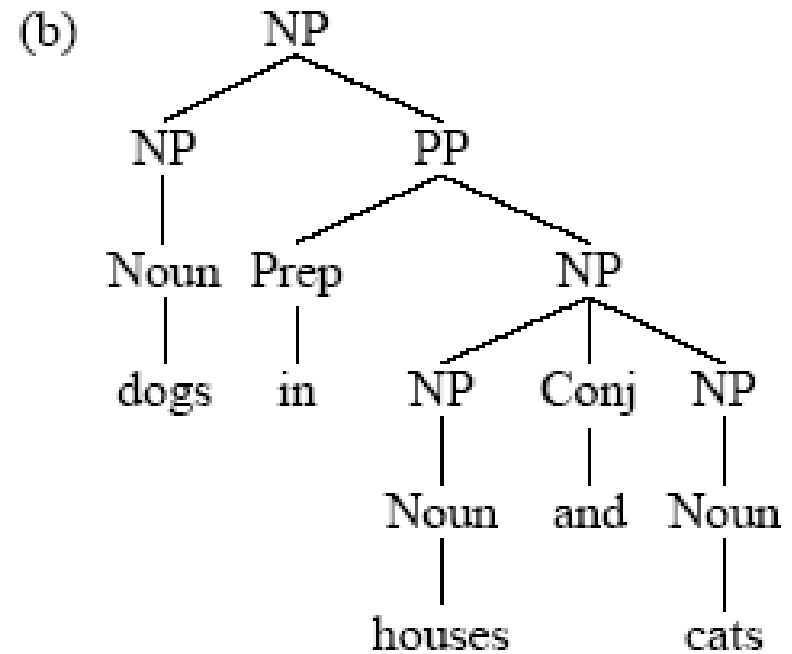
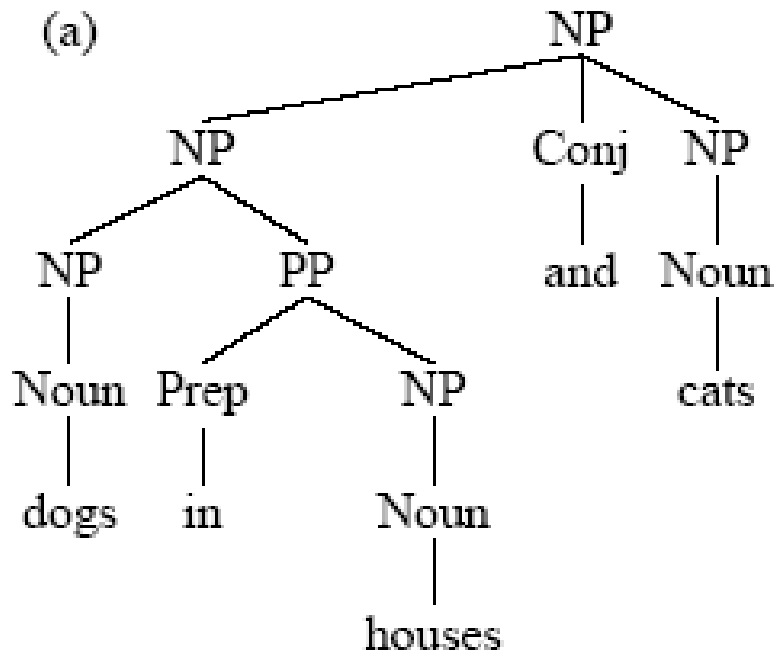
Felülről-lefelé (TD) elemzés

(Does this flight include a meal?)

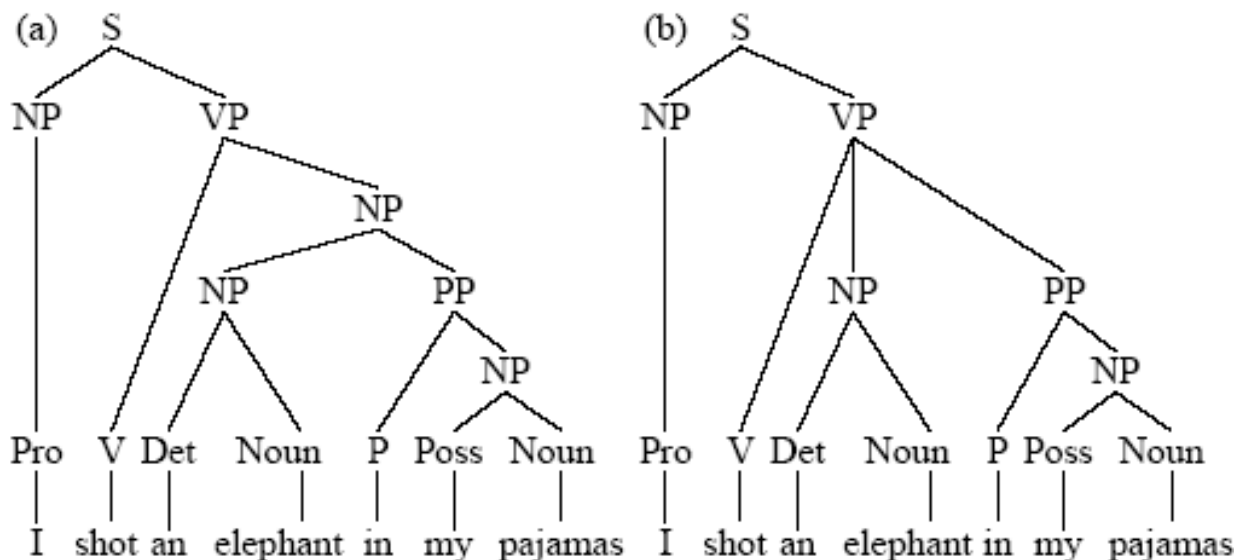
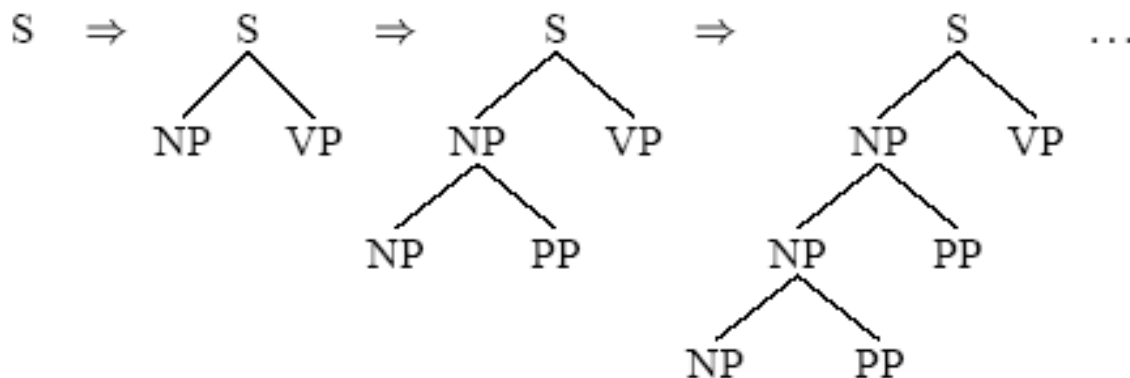
$S \rightarrow NP VP$
 $S \rightarrow Aux NP VP$
 $S \rightarrow VP$
 $NP \rightarrow Det Nominal$
 $Nominal \rightarrow Noun$
 $Nominal \rightarrow Noun Nominal$
 $NP \rightarrow Proper-Noun$
 $VP \rightarrow Verb$
 $VP \rightarrow Verb NP$



Egy örök nyelvi probléma: hová tartozik a PP?



A balrekurzió okozta problémák a TD-elemzésben



Egy példa a balrekurzióra

$n > 1$ esetén a balrekurziót természetes nyelvekben meglehetősen nehéz megérteni:

$$\begin{aligned} S &\rightarrow NP VP \\ NP &\rightarrow RelPron S \end{aligned}$$

0: *A fiú elment.*

1: *A fiú,
akit a barátom meghívott,
elment.*

2: *A fiú,
akit a barátom,
akiről a kollégám mesélt,
meghívott,
elment.*

3: *A fiú,
akit a barátom,
akiről a kollégám,
akivel egy iskolába jártam,
mesélt,
meghívott,
elment.*

Egy példa a jobbrekurzióra

A nyelvekben a jobbrekurzió ($VP \rightarrow PP VP$) mélysége nem jelent problémát az (emberi) feldolgozásban:

*„az agyag ölelő karjai közül kibontakozni akaró
kocsikerék rettentő nyikorgásától megriadt juhászkutya
bundájába kapaszkodó kullancs kidülledt félszeméből
alácseppenő könnycseppben visszatükröződő holdvilág
fényétől illuminált rablólovagvár felvonóhídjából kiálló
vasszegek kohéziós erejének hatása”*

(Fehér G.)

A jobbrekurziós alárendelés természete

1: A karok ölelnek.

→ ölelő karok

2: A kocsikerék ki akar bontakozni (vhonnan).

→ (vhonnan) kibontakozni akaró kocsikerék

3: A juhászkutya megriadt (vmitől).

→ (vmitől) megriadt juhászkutya

4: A kullancs kapaszkodik (vmibe).

→ (vmibe) kapaszkodó kullancs

5: A könnyecsepp alácseppen (vhonnan).

→ (vhonnan) alácseppenő könnyecsepp

6: A holdvilág visszatükröződik (vmiben).

→ (vmiben) visszatükröződő holdvilág

7: A rablólovagvár illuminált (vmitől).

→ (vmitől) illuminált rablólovagvár

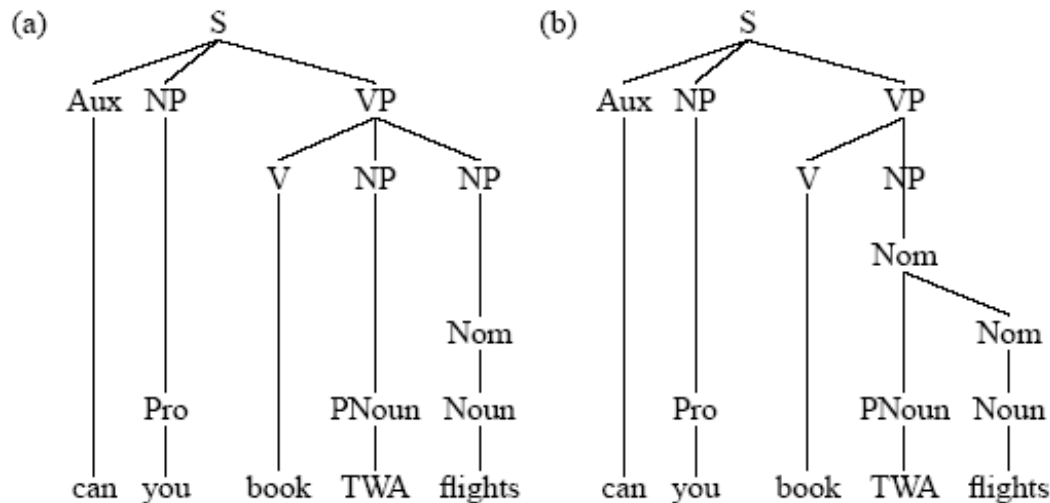
8: A vasszegek kiállnak (vhonnan).

→ (vhonnan) kiálló vasszegek

A TD- és a BU-stratégiák összehasonlítása

- ☐ TD
nem fordulhat elő, hogy ne mondat-szimbóllummal végződő elemzéssel töltsük az időt, és
nincs olyan részszerkezet, melynek ne lenne helye a végső fában
- ☐ BU:
nem fordulhat elő, hogy a bemenettel inkonzisztens szabályokkal kellene töltenie az időt, és
nincs olyan részszerkezet, mely ne lenne kompatibilis a bemenet valamely részével
- ☐ Optimális megoldás: valamiféle kombináció, pl.
BU-elemzés előfordulási valószínűségekkel;
TD-elemzés „BU-jóslással”

A valószínűség bevonása az elemzésbe



Egy megoldás valószínűségek használatával:

Rules	P	Rules	P
S → Aux NP VP	.15	S → Aux NP VP	.15
NP → Pro	.40	NP → Pro	.40
VP → V NP NP	.05	VP → V NP	.40
NP → Nom	.05	NP → Nom	.05
NP → PNoun	.35	Nom → PNoun Nom	.05
Nom → Noun	.75	Nom → Noun	.75
Aux → Can	.40	Aux → Can	.40
NP → Pro	.40	NP → Pro	.40
Pro → you	.40	Pro → you	.40
Verb → book	.30	Verb → book	.30
PNoun → TWA	.40	Pnoun → TWA	.40
Noun → flights	.50	Noun → flights	.50

Balsarok-elemzés

B-t A balsarkának nevezzük, ha van A-nak B-vel kezdődő levezetése

$$S \rightarrow NP VP$$
$$S \rightarrow Aux NP VP$$
$$S \rightarrow VP$$
$$NP \rightarrow Det Nominal$$
$$Nominal \rightarrow Noun$$
$$Nominal \rightarrow Noun Nominal$$
$$NP \rightarrow Proper-Noun$$
$$VP \rightarrow Verb$$
$$VP \rightarrow Verb NP$$

Ennek a nyelvtannak a balsarok-táblája az alábbi:

Szófaj	Balsarok
<i>S</i>	<i>Det, Proper-Noun, Aux, Verb</i>
<i>NP</i>	<i>Det, Proper-Noun</i>
<i>Nominal</i>	<i>Noun</i>
<i>VP</i>	<i>Verb</i>

A balsarok-elemzés működése

Szófaj	Balsarok
<i>S</i>	<i>Det, Proper-Noun, Aux, Verb</i>
<i>NP</i>	<i>Det, Proper-Noun</i>
<i>Nominal</i>	<i>Noun</i>
<i>VP</i>	<i>Verb</i>

Tehát:

a „*Does this flight include a meal?*” esetében a

$$S \rightarrow NP VP$$

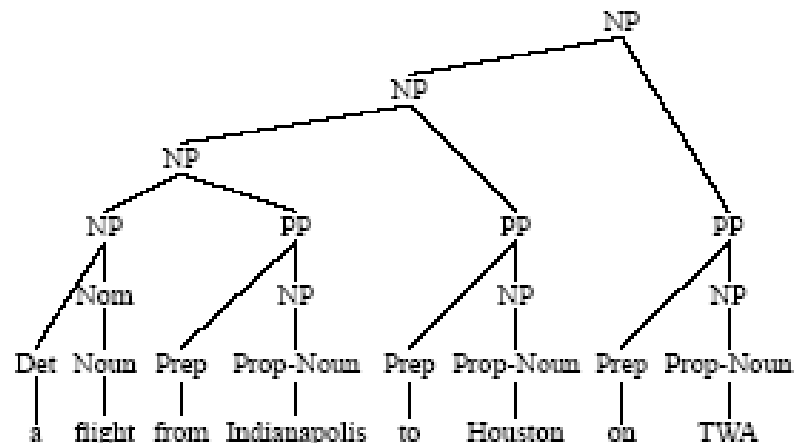
$$S \rightarrow Aux NP VP$$

$$S \rightarrow VP$$

három szabályból TD csak a középső jöhet szóba

a *does* Aux volta miatt!

(visszalépéses TD-elemzés)



1: on TWA, a flight from Indianapolis to Houston on TWA

Mondatelemzés Earley-algoritmussal

- ☐ Earley-algoritmus: párhuzamos TD-elemzés
- ☐ BU-párja a Cocke – Young – Kasami-algoritmus (CYK)
- ☐ Exponenciális idejű problémát polinomiálisra redukál
- ☐ Azaz: dinamikus programozáson alapul
- ☐ Lényege:
a repetitív részszerkezet-kezelés eliminálása
- ☐ Magja:
egy $n+1$ elemű táblázat és az ún. pontozott szabály

A pontozott szabály

Ha az elemzéssel a bemenet j -edik betűjéig jutottunk el, miközben megállapítottuk, hogy léteznek olyan

$$p, q, r, t \in (N \cup T)^*$$

szavak, melyekre

$$S \Rightarrow^* pAq$$

$$p \Rightarrow^* a_1 \dots a_i$$

$$q \Rightarrow^* a_{i+1} \dots a_j$$

$$A \rightarrow rt \in P$$

akkor ezt a körülményt fejezzük ki azzal, hogy az

$$A \rightarrow r \bullet t$$

az $\langle i, j \rangle$ indexű pontozott szabály

A pontozós elemzési lépések fajtái

$S \rightarrow \bullet VP$

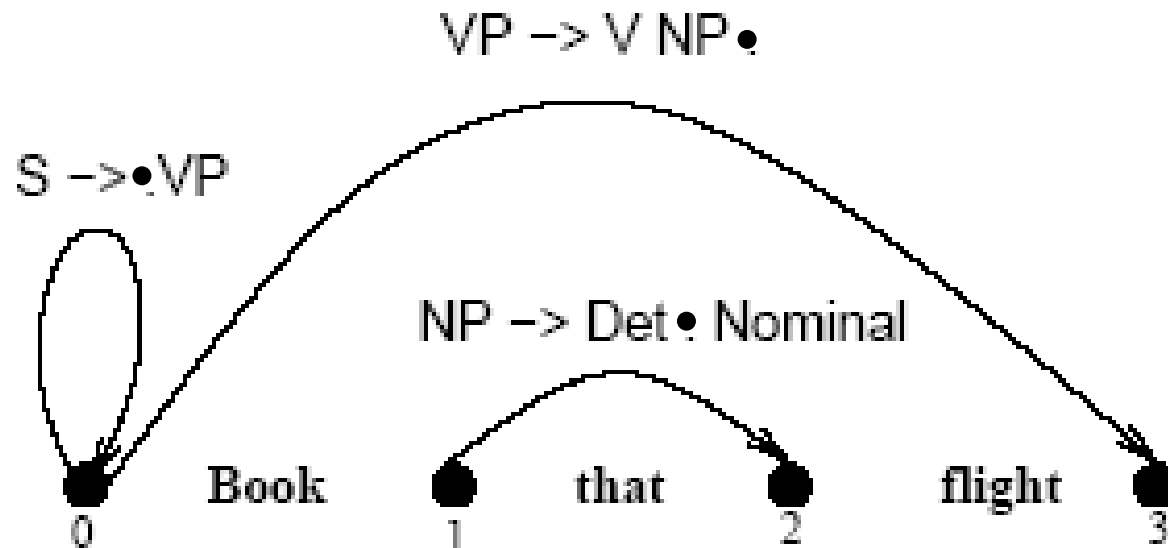
$NP \rightarrow Det \bullet Nominal$

$VP \rightarrow V NP \bullet$

jósló (predictor)

elemző (scanner)

befejező (completer)



Az Earley-algoritmus

```
function EARLEY-PARSE(words, grammar) returns chart
  ENQUEUE( $(\gamma \rightarrow \bullet S, [0, 0])$ , chart[0])
  for  $i \leftarrow$  from 0 to LENGTH(words) do
    for each state in chart[i] do
      if INCOMPLETE?(state) and
        NEXT-CAT(state) is not a part of speech then
        PREDICTOR(state)
      elseif INCOMPLETE?(state) and
        NEXT-CAT(state) is a part of speech then
        SCANNER(state)
      else
        COMPLETER(state)
    end
  end
  return(chart)

procedure PREDICTOR( $(A \rightarrow \alpha \bullet B \beta, [i, j])$ )
  for each  $(B \rightarrow \gamma)$  in GRAMMAR-RULES-FOR(B, grammar) do
    ENQUEUE( $(B \rightarrow \bullet \gamma, [j, j])$ , chart[j])
  end

procedure SCANNER( $(A \rightarrow \alpha \bullet B \beta, [i, j])$ )
  if  $B \subset$  PARTS-OF-SPEECH(word[j]) then
    ENQUEUE( $(B \rightarrow \text{word}[j], [j, j + 1])$ , chart[j+1])

procedure COMPLETER( $(B \rightarrow \gamma \bullet, [j, k])$ )
  for each  $(A \rightarrow \alpha \bullet B \beta, [i, j])$  in chart[j] do
    ENQUEUE( $(A \rightarrow \alpha B \bullet \beta, [i, k])$ , chart[k])
  end

procedure ENQUEUE(state, chart-entry)
  if state is not already in chart-entry then
    PUSH(state, chart-entry)
  end
```

Konkrét implementációk:

https://en.wikipedia.org/wiki/Earley_parser#Implementations

Példa az Earley-elemzésre

(a circle touches a square)

S → NP VP
NP → Det N
VP → VT NP
VP → VI PP
PP → P NP

Det → a
N → circle|square|triangle
VT → touches
VI → is
P → above|below

	a	circle	touches	a	square
□ → .S	scanned	scanned	scanned	scanned	scanned
predicted	□ Det → a.	1 N → circle.	2 VT → touches.	3 Det → a.	4 N → triangle.
□ S → .NP VP	completed	completed	completed	completed	completed
□ NP → .Det N	□ NP → Det.N	□ NP → Det N.	2 VP → VT.NP	3 NP → Det.N	4 NP → Det N.
□ Det → .a	predicted	□ S → NP.VP	predicted	predicted	3 VP → VT NP.
	1 N → .circle	predicted	3 NP → .Det N	5 N → .circle	□ S → NP VP.
	1 N → .square	2 VP → .VT NP	3 Det → .a	4 N → .square	□ → S.
	1 N → .triangle	2 VP → .VI PP		4 N → .triangle	
		2 VT → .touches			
		2 VI → .is			
State set 0	1	2	3	4	5



Az unifikációs szintaxis

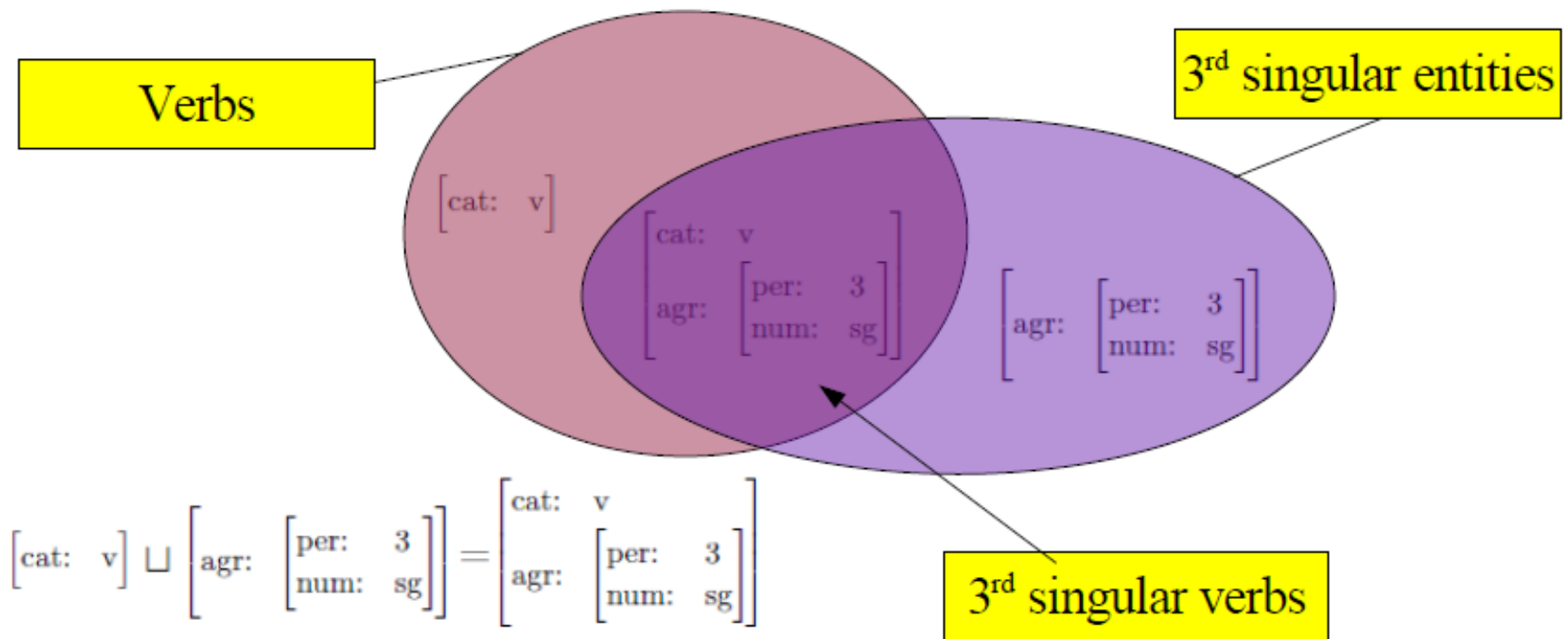


Jegyszerkezetek

- ☐ A kategóriák (szófajok) nem atomiak, tovább alosztályozhatók: *VPto*, *Sthat*, *Non3sgAux*, *3sgNP*, *tV*, *NPmass*, ...
- ☐ Az efféle jegyek hierarchiába rendezhetők
- ☐ A szabályok így szabályosztályokká válnak
- ☐ Jegy-érték párok
- ☐ Jegyek: nyelvtani tulajdonságok, értékek: atomi szimbólumok és jegyszerkezetek
- ☐ Jegy-ösvény: nem egyetlen jegynek van értéke, hanem az ösvénynek
- ☐ Például: *[egyeztetés[szám]] = Plur*
- ☐ Fej-jegyek és láb-jegyek

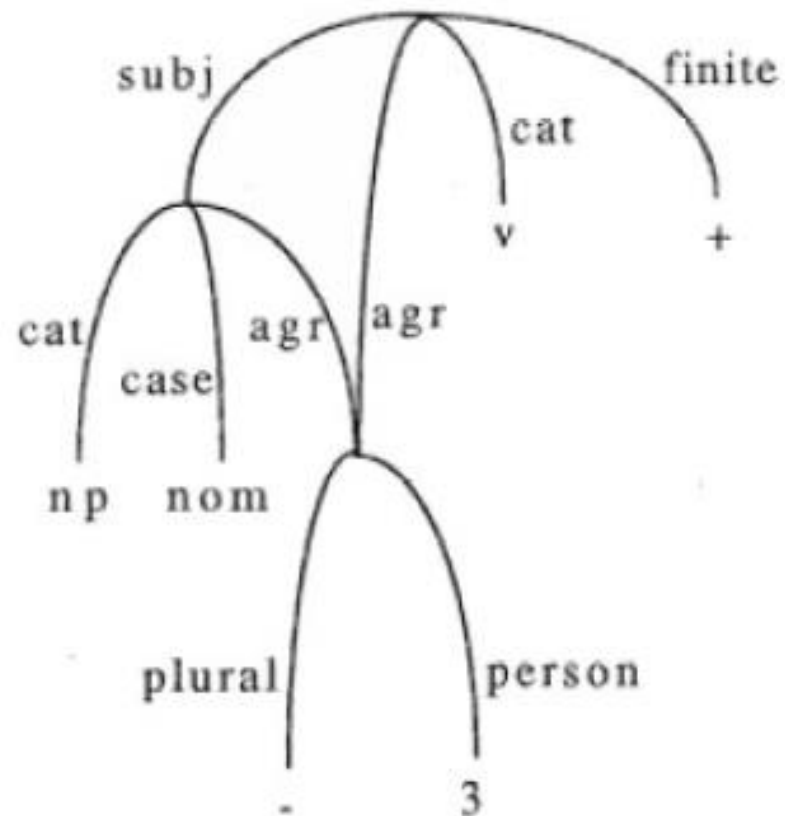
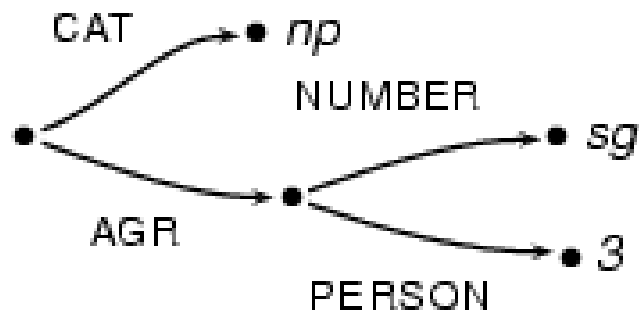
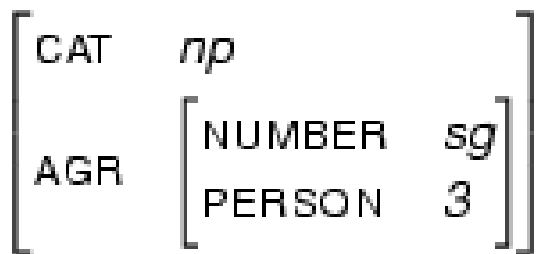
Diszjunkt jegy-érték párok unifikációja

Az információk unifikációja a denotációk metszete:



Berthold Crysmann tananyagából (http://www.coli.uni-saarland.de/courses/msc-prep-06/slides/VL_PrepSyntaxUBG.pdf)

Jegyszerkezetek DAG-ként



A jegyszerkezetek reprezentációja

Az unifikációs rész bemutatásához Berthold Crysmann tananyagát használtam fel (http://www.coli.uni-saarland.de/courses/msc-prep-06/slides/VL_PrepSyntaxUBG.pdf)

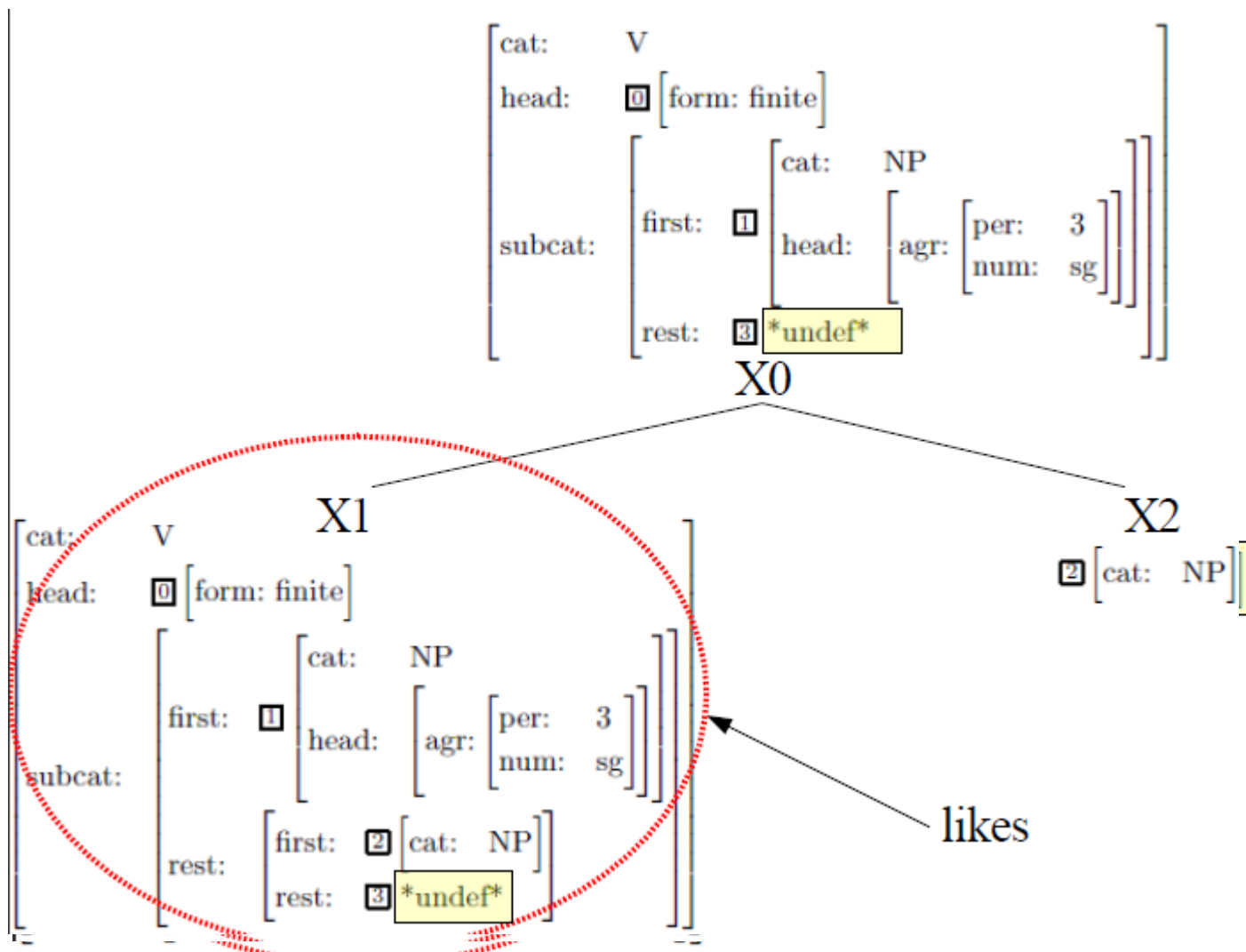
- ❑ Ösvényekre vonatkozó egyenletekkel:
 - $\langle \text{cat} \rangle = v$
 - $\langle \text{finite} \rangle = +$
 - $\langle \text{agr:plural} \rangle = -$
 - $\langle \text{agr:person} \rangle = 3$
 - $\langle \text{subj:cat} \rangle = np$
 - $\langle \text{subj:case} \rangle = \text{nom}$
 - $\langle \text{subj:agr} \rangle = \langle \text{agr} \rangle$

- ❑ Jegy-érték mátrixokkal:

$$\left[\begin{array}{l} \text{cat:} \\ \text{finite:} \\ \text{agr:} \\ \text{subj:} \end{array} \begin{array}{l} v \\ + \\ \boxed{1} \left[\begin{array}{l} \text{plural:} \\ \text{per:} \end{array} \begin{array}{l} - \\ 3 \end{array} \right] \\ \left[\begin{array}{l} \text{cat:} \\ \text{case:} \\ \text{agr:} \end{array} \begin{array}{l} np \\ \text{nom} \\ \boxed{1} \left[\begin{array}{l} \end{array} \right] \end{array} \right] \end{array} \right]$$

A jegyszerkezet mint szabály

Berthold Crysmann tananyagából (http://www.coli.uni-saarland.de/courses/msc-prep-06/slides/VL_PrepSyntaxUBG.pdf)



Az unifikációs Earley-algoritmus

```
function EARLEY-PARSE(words, grammar) returns chart
  ENQUEUE( $(\gamma \rightarrow \bullet S, [0, 0], dag_\gamma, chart[0])$ )
  for  $i \leftarrow$  from 0 to LENGTH(words) do
    for each state in  $chart[i]$  do
      if INCOMPLETE?(state) and
        NEXT-CAT(state) is not a part of speech then
        PREDICTOR(state)
      elseif INCOMPLETE?(state) and
        NEXT-CAT(state) is a part of speech then
        SCANNER(state)
      else
        COMPLETER(state)
    end
  end
  return(chart)

procedure PREDICTOR( $(A \rightarrow \alpha \bullet B \beta, [i, j], dag_A)$ )
  for each  $(B \rightarrow \gamma)$  in GRAMMAR-RULES-FOR( $B, grammar$ ) do
    ENQUEUE( $(B \rightarrow \bullet \gamma, [j, j], dag_B, chart[j])$ )
  end

procedure SCANNER( $(A \rightarrow \alpha \bullet B \beta, [i, j], dag_A)$ )
  if  $B \in$  PARTS-OF-SPEECH(word[j]) then
    ENQUEUE( $(B \rightarrow word[j], [j, j+1], dag_B, chart[j+1])$ )

procedure COMPLETER( $(B \rightarrow \gamma \bullet, [j, k], dag_B)$ )
  for each  $(A \rightarrow \alpha \bullet B \beta, [i, j], dag_A)$  in  $chart[j]$  do
    if  $new\_dag \leftarrow$  UNIFY-STATES( $dag_B, dag_A, B$ )  $\neq$  Fails!
      ENQUEUE( $(A \rightarrow \alpha B \bullet \beta, [i, k], new\_dag, chart[k])$ )
  end

procedure UNIFY-STATES( $dag1, dag2, cat$ )
   $dag1\_cp \leftarrow$  COPYDAG( $dag1$ )
   $dag2\_cp \leftarrow$  COPYDAG( $dag2$ )
  UNIFY(FOLLOW-PATH( $cat, dag1\_cp$ ), FOLLOW-PATH( $cat, dag2\_cp$ ))

procedure ENQUEUE(state, chart-entry)
  if state is not subsumed by a state in chart-entry then
    PUSH(state, chart-entry)
  end
```

További kiterjesztések

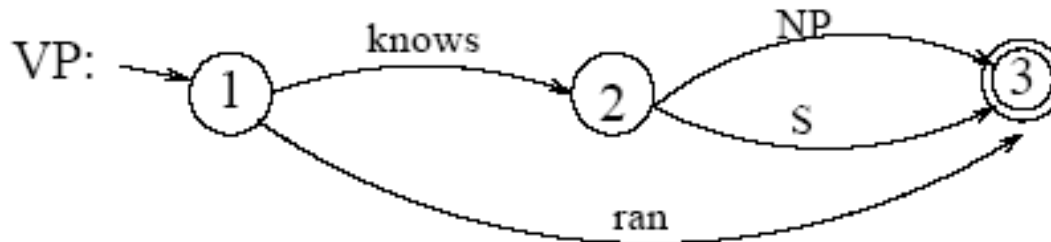
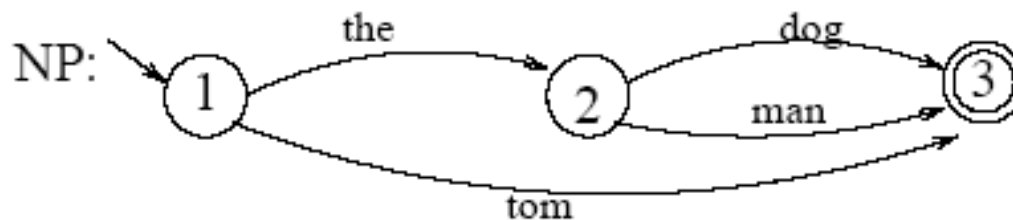
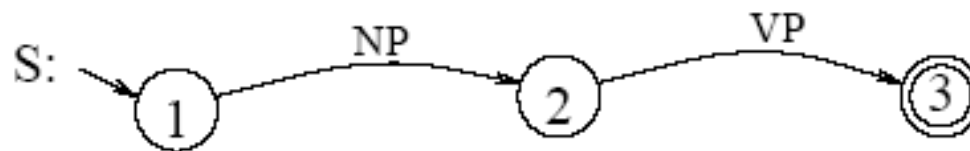
- ☐ Negáció
- ☐ Diszjunkció
- ☐ Ösvény-egyenlőtlenség
- ☐ Halmaz-értékű jegyek
- ☐ Jegyszerkezet-leíró metanyelv



A véges állapotú eszközök kiterjesztései szintaktikai elemzéshez

RTN

(Recursive Transition Network)

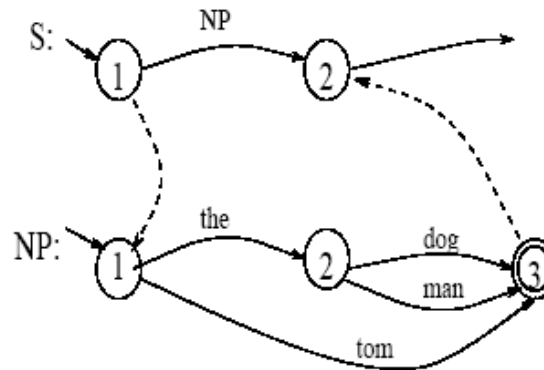


RTN

(kiegészítések a FSA-hoz)

A szokásos FSA működtetésén túl figyelni kell:

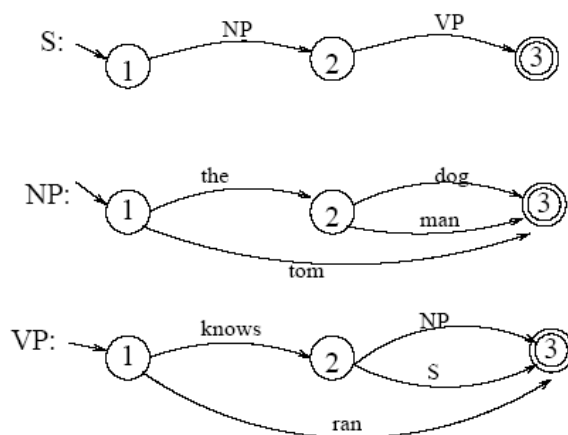
- az aktuális bemeneti pozíciót,
- az aktuális állapotot és
- hogy hova kell visszamenni:



- összegezve:
veremkezelés kell

RTN

(a *Thew man ran* mondat elemzésének menete)



Input	State	Return
<i>the man ran</i>	S:1	—
<i>the man ran</i>	NP:1	S:2
<i>man ran</i>	NP:2	S:2
<i>ran</i>	NP:3	S:2
<i>ran</i>	S:2	—
<i>ran</i>	VP:1	S:3
	VP:3	S:3
	S:3	—

Az RTN összefoglalása

- ☐ Az RTN egymást hívó FSA-k hálózata: az élek címkéi közt megjelenik más FSA-k „neve”
- ☐ A FSA (a reguláris nyelvek) $O(n)$ idő alatt elemezhetők
- ☐ Az RTN valójában egy veremautomata, azaz környezet-független nyelvek elemzésére is alkalmas
- ☐ Az RTN-nel tehát $O(n^3)$ elemzési idő garantálható



ATN (az RTN bővítése)

ÉLCÍMKÉK:

WRD *, CAT *, PUSH *, POP, JUMP *

ÉRTÉKEK:

GETR, *, QUOTE, GETF, BUILDQ *, APPEND

TESZTEK:

T, EQ, AND, OR, NOT

AKCIÓK:

SETR, TO

Az ATN-formalizmus

(hagyományosan és XML-ben)

```
(
  (s  (push np t (setr np *) (to s_np)))
  (s_np (push vp t (setr vp *) (to s_vp)))
  (s_vp (pop (buildq (s + +) np vp) t))
  (np  (cat det t (setr det (buildq (det *))) (to np_det))
      (cat npr t (setr np (buildq (np (npr *)))) (to np_np)))
  (np_det(cat n t (setr np (buildq (np (det +) (noun *))) det)
          np_np))
  (np_np (jump np_np1 t (setr pps ())))
  (np_np1 (push pp t (setr pps (append (getr pps) *))(to
          np_np1))
          (jump np_end t))
  (np_end (pop (buildq (np + (pps +)) np pps) t))
  (vp  (cat verb t (setr verb *) (setr pps ())(to vp_v)))
  (vp_v (push np t (setr obj *) (to vp_np)))
  (vp_np (push pp t (setr pps (append (getr pps) *))(to vp_
          (jump vp_np1 t))
  (vp_np1 (pop (buildq (vp (verb +) + +) verb obj pps) t))
  (pp  (cat prep t (setr prep *) (to pp_p)))
  (pp_p (push np t (setr np *) (to pp_pp)))
  (pp_pp (pop (buildq (pp (p +) +) prep np) t))
)
```

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE grammar SYSTEM "gram.dtd">

<grammar>

  <network name="S">
    <state name="S.START">
      <PUSH dest="NP">
        <SETR name="NP" value="*"/>
        <JUMP dest="S.2"/>
      </PUSH>
    </state>
    <state name="S.2">
      <CAT name="v" dest="S.END"/>
    </state>
    <state name="S.END">
      <POP>
        <BUILDQ>(S(NP(+)(VP(v(+))))), NP, *</BUILDQ>
      </POP>
    </state>
  </network>

  <network name="NP">
    <state name="NP.START">
      <CAT name="det" dest="NP.2">
        <SETR name="DET" value="*"/>
      </CAT>
      <JUMP dest="NP.END"/>
    </state>
    <state name="NP.2">
      <CAT name="n" dest="NP.END"/>
    </state>
    <state name="NP.END">
      <POP>
        <BUILDQ>(det(+))(n(+)), DET, *</BUILDQ>
      </POP>
    </state>
  </network>

</grammar>
```

1



A számítógépes szemantika felé

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

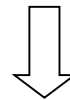
A book was sold by Jane to Bill.

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))

A szintaktikai szerepek és a jelentés viszonya

(sell (agent SUBJECT)
(object DIRECT OBJECT)
(counter-agent
INDIRECT OBJECT
or
PREP. OBJECT "to"))

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))



Többértelműség

A többértelműség megjelenései

A szavak szintjén

- ☐ Homonímia (azonos alakúság): *vár¹, vár²*
- ☐ Poliszémia (többértelműség): *vírus¹, vírus²*
- ☐ Szinonímia (rokon értelműség): *humoros, vicces, tréfás, mókás, poénos, mulatságos, kacagtató, mulattató, ...*

A szóalaktan szintjén

- ☐ Homonímia (azonos alakúság): *nemzet+é¹+t, nemzet+é²+t*
- ☐ Poliszémia (többértelműség): *úr+nak¹ (szolgál), úr+nak² (születik)*

A mondat szintjén

- ☐ Szerkezeti többértelműségek: *Láttalak a teraszon ülve.*

A szöveg szintjén

- ☐ A közlés szándéka, a kommunikációs helyzet, a beszélő és a hallgató viszonya befolyásolja az üzenet jelentését:
De szeretlek!

Mire jó a jelentéstani elemzés?

- ❑ A többértelműség egyik fontos forrása a jelentések nem egyértelmű volta
- ❑ A mondszerkezet ismerete ennek eldöntéséhez még nem nyújt elég információt
- ❑ Az üzenet kibocsátója szempontjából egyértelmű üzenet lehet jelentésbeli többértelműségekkel, melyeket az üzenet vevőjének kell feloldani
- ❑ Ilyen viszont nincs:
*Miért fejt több ember keresztretjvényt, mint szenet?
(... és tehenet?)*
- ❑ A legfontosabb célok:
(1) feloldani a jelentéstani többértelműségeket, és
(2) reprezentálni a mögöttes tartalmat

A nem egyértelmű jelentések kezelése

- ❑ Az agyunk optimalizál, egyensúlyt tart, ui.
 - ha sok a többértelműség, akkor kisebb lesz a lexikon, ezért kevesebb memória kell, de bonyolultabb lesz a feldolgozás
 - ha kevesebb a többértelműség, akkor nagyobb lesz a lexikon, ezért lassabb lesz a tanulás, de könnyebb lesz az értelmezés
- ❑ Az emberi nyelvekben gyakori a többértelműség, mégis ritkán okoz gondot egy-egy fogadott jelsorozat értelmezése
- ❑ A számítógép számára szinte teljesen kezelhetetlen a többértelműség - ezt kell megoldani!



Jelentés-egyértelműsítés

Jelentés-egyértelműsítés (WSD)

Word sense disambiguation:

egy szóalak jelentésének kiválasztása, egy előre megadott (lehetséges jelentés-) halmazból

- a lehetséges jelentések szótárból
- osztályozási modellek alkalmazhatók
- előre egyértelműsített példák szükségesek

Word sense discrimination:

egy adott szóalak különböző használati eseteit (=jelentéseit) elkülöníteni, anélkül, hogy a lehetséges jelentéseket kívülről megadnánk (és így címkézett példáink sem lehetnek)

- felügyelet nélküli statisztikai modellek

A WSD két fő iránya

Minden szövegbeli szóra:

- a cél az összes szóalak egyértelműsítése, folyó szövegben
- nagyon kevés erőforrás áll rendelkezésre
- olyan mintákat kell tanulni, melyek függetlenek az adott szóalaktól
- nincs kielégítő megoldás még
- gyakorlatban ez lenne jól használható

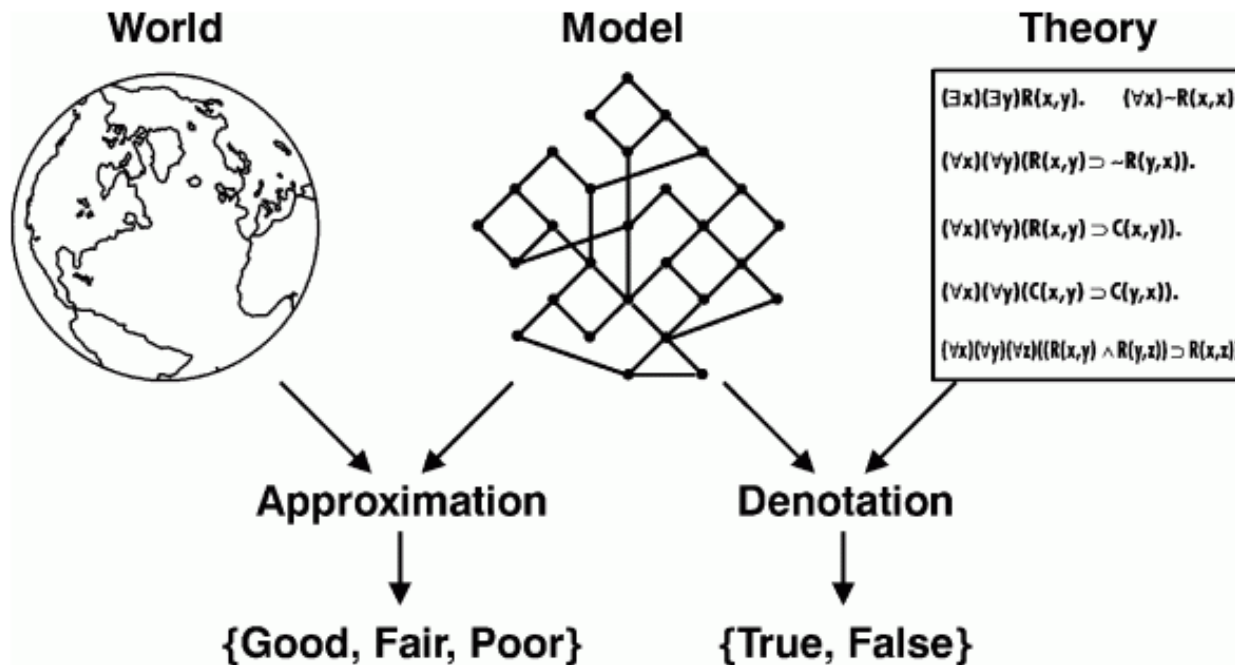
A szöveg egyes szavaira:

- a cél bizonyos szóalakok egyértelműsítése, minden szóalakra önálló modell
- több erőforrás van, és olcsóbban előállítható 1-1 új szóalakra
- az adott szóra jellemző mintákat tanulhatunk, könnyebb feladat
- tűrhető (nem jó!) megoldások vannak
- gyakorlatban ritkán használható



Jelentésábrázolás

Világ - modell - elmélet





Jelentés vagy világismeret?

- (1) *Péter megvette a könyvet.*
- (2) *Péter megvette Az ember tragédiáját.*
- (3) *Vett könyvet Péter?*

A világismeretet tárolni kell:
mi micsoda, és
milyen viszonyban van a többi ismert dologgal?

Világismeret-reprezentáció

Péter megvette Az ember tragédiáját.

(elad (agent ember17)

(object AET3791)

(recipient ember35)

(tense múlt))

AET3791: „Az ember tragédiája” könyv egy példánya

ember35 neve: Péter

ember17: (jelenleg) nem ismerjük



A nyelvtechnológia alapjai

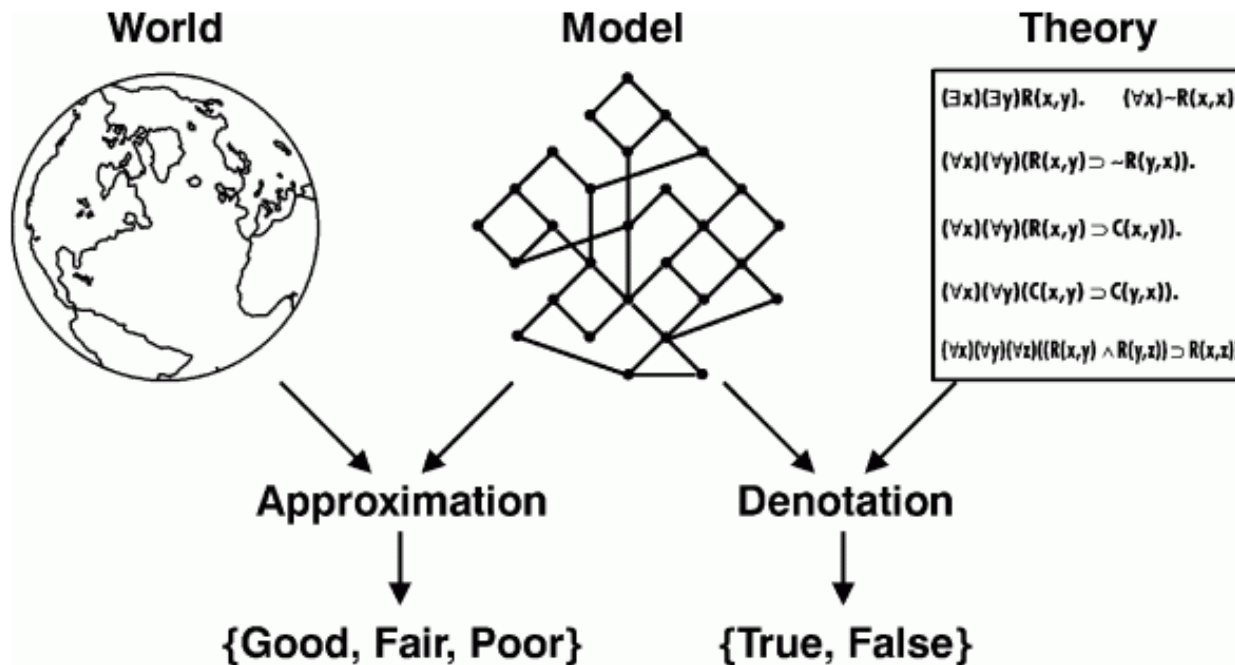
6.

Számítógépes szemantika



Jelentésábrázolás

Világ - modell - elmélet



Jelentés vagy világismeret?

- (1) *Péter megvette a könyvet.*
- (2) *Péter megvette Az ember tragédiáját.*
- (3) *Vett könyvet Péter?*

A világismeretet tárolni kell:
mi micsoda, és
milyen viszonyban van a többi ismert dologgal?

Világismeret-reprezentáció

Péter megvette Az ember tragédiáját.

(elad (agent ember17)

(object AET3791)

(recipient ember35)

(tense múlt))

AET3791: „Az ember tragédiája” könyv egy példánya

ember35 neve: Péter

ember17: (jelenleg) nem ismerjük

Irányzatok a számítógépes jelentésábrázolásban

- ❑ Matematikai logikai reprezentációk
- ❑ Konceptuális reprezentációk:
szemantikus hálók,
fogalmi gráfok,
fogalmi függőség
- ❑ Lexikális szemantikai reprezentációk:
lexikális szemantikus hálók,
ontológiák
- ❑ Mélytanulás, neurális hálók



A logikák szerepe



Logikák a számítógépes jelentésábrázolásban

- ☐ Elvi problémák és a számítógépes nyelvészet igényei
- ☐ Elsőrendű logika
- ☐ Magasabb rendű logikák
- ☐ Modális logikák
- ☐ Intenzionális logikák
- ☐ Montague elmélete (1970)

Montague-nyelvtanok

- ☐ A Montague-nyelvtan alapfeltételezése: a mondatok jelentése igazságfeltételekkel megadható
- ☐ A *Péter olvas egy könyvet* acsa igaz, ha Péter olvas egy könyvet.
- ☐ Ezeket az igazságfeltételeket logikai formulákkal reprezentálhatjuk:
Péter olvas egy könyvet. $\rightarrow \exists x(\text{könyv}(x) \wedge \text{olvas}(p^*, x))$
- ☐ Indirekt interpretáció: TNY $\rightarrow$ logika $\rightarrow$ modellek
- ☐ A kompozicionalitás elve: egy komplex kifejezés jelentése a részei jelentéseinek és az őket leíró szintaktikai szerkezetnek a függvénye
- ☐ E. Bach: „rule-to-rule” hipotézis
- ☐ Az elsőrendű logika nem elég:
John is an intelligent student $\Rightarrow \text{intelligent}(j^*) \wedge \text{student}(j^*)$
John is a good student $\Rightarrow \text{good}(j^*) \wedge \text{student}(j^*)$??
John is a former student $\Rightarrow \text{former}(j^*) \wedge \text{student}(j^*)$???

Montague-nyelvtanok (2)

❑ Alapja a kategoriális nyelvtan (v.ö. X-vonás nyelvtan!)

❑ Alapkategóriák:

Mondat: S

Intranszitiv igék: V

Főnevek: N

❑ Ha A, B kategóriák, akkor A/B is kategória

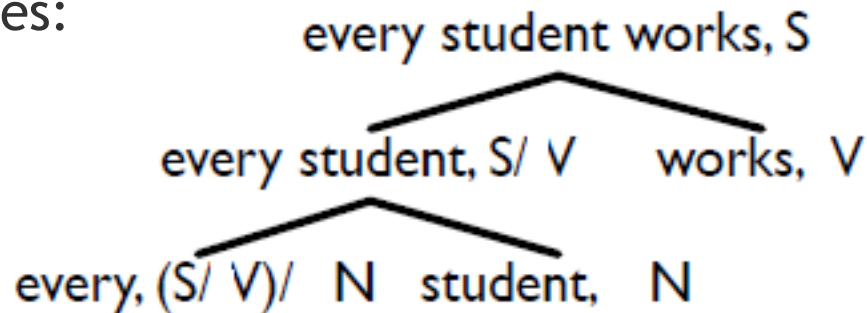
❑ Levezetett kategóriák:

Főnévi csoportok: S/V

Tranzitív igék: V/(S/V)

Determinánsok: (S/V)/N

❑ Egy levezetés:



A logikai szemantikai levezetésekről

- ❑ Egy komplex kifejezés jelentése a részei jelentésének és azoknak a szintaktikai szabályoknak a függvénye, melyek őket összekapcsolják.
- ❑ Néhány lexikális kategória „szemantikai fordítása”:
 $every \rightarrow \lambda P \lambda Q \forall x (P(x) \Rightarrow Q(x))$
 $student \rightarrow student$
 $works \rightarrow work$
- ❑ $every\ student$
 $\rightarrow \lambda P \lambda Q \forall x (P(x) \Rightarrow Q(x))(student)$
 $= \lambda Q \forall x (student(x) \Rightarrow Q(x))$
- ❑ $Every\ student\ works.$
 $\rightarrow \lambda Q \forall x (student(x) \Rightarrow Q(x))(work)$
 $= \forall x (student(x) \Rightarrow work(x))$

A logikai szemantikai levezetésekről 2.

John works. $\rightarrow \text{work}(j^*)$

A student works. $\rightarrow \exists x(\text{student}(x) \wedge \text{work}(x))$

Every student works. $\rightarrow \forall x(\text{student}(x) \Rightarrow \text{work}(x))$

John and Mary work. $\rightarrow \text{work}(j^*) \wedge \text{work}(m^*)$

Lambda-absztrakcióval kifejezve:

John $\rightarrow \lambda P.P(j^*)$

a student $\rightarrow \lambda P \exists x(\text{student}(x) \wedge P(x))$

every student $\rightarrow \lambda P \forall x(\text{student}(x) \Rightarrow P(x))$

John and Mary $\rightarrow \lambda P.P(j^*) \wedge P(m^*)$

A logikai szemantikai levezetésekéről 3.

Például a tranzitív igék fordítása:

$read \rightarrow \lambda Q \lambda x. Q(\lambda y. read^*(y)(x))$

(1) *John reads a book* $\rightarrow \exists y(book(y) \wedge read(y)(j^*))$

(2) *Every student reads a book.* $\rightarrow$

a book $\rightarrow \lambda P \exists z(book(z) \wedge P(z))$

reads $\rightarrow \lambda Q \lambda x. Q(\lambda y. read^*(y)(x))$

reads a book $\rightarrow$

$\rightarrow \lambda Q \lambda x. Q(\lambda y. read^*(y)(x))(\lambda P \exists z(book(z) \wedge P(z)))$

$\rightarrow \lambda x. \lambda P \exists z(book(z) \wedge P(z))(\lambda y. read^*(y)(x))$

$\rightarrow \lambda x. \exists z(book(z) \wedge (\lambda y. read^*(y)(x))(z))$

$\rightarrow \lambda x. \exists z(book(z) \wedge read^*(z)(x))$

every student reads a book $\rightarrow$

$\rightarrow \lambda P \forall w(student(w) \Rightarrow P(w))(\lambda x. \exists z(book(z) \wedge read^*(z)(x)))$

$\rightarrow \forall w(student(w) \Rightarrow \exists z(book(z) \wedge read^*(z)(w)))$



Fogalmi gráfok

Sowa fogalmi gráfjai

- ❑ John F. Sowa (IBM, 1976)
- ❑ Grafikus interfész az elsőrendű logikához (Display Form, DF, illetve Linear Format, LF)
- ❑ Gráf-alapú tudásreprezentáció és következtetési modell
- ❑ Lineáris notációja a Conceptual Graph Interchange Format (CGIF) - ISO/IEC 24707:2007
- ❑ Knowledge Interchange Format (KIF)
- ❑ Grafikus megjelenítő eszközök:
 - CoGUI:** Java-alapú grafikus eszköz COGXML formátumú gráfok építésére (<http://www2.lirmm.fr/cogui/>)
 - Cogitant:** C++-csomag(<http://cogitant.sourceforge.net/>)

Példák fogalmi gráfokra 1.

A cat is on mat.

DF:



LF:

$[Cat]^\circ(On)^\circ[Mat].$

CGIF:

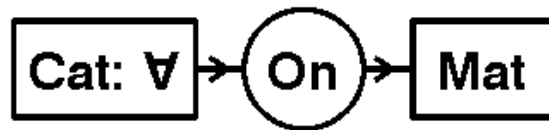
$[Cat: *x] [Mat: *y] (On ?x ?y)$

KIF:

$(exists ((?x Cat) (?y Mat)) (On ?x ?y))$

Példák fogalmi gráfokra 2.

Every cat is on a mat.



LF:

$[Cat: @every]^\circ(On)^\circ[Mat].$

CGIF:

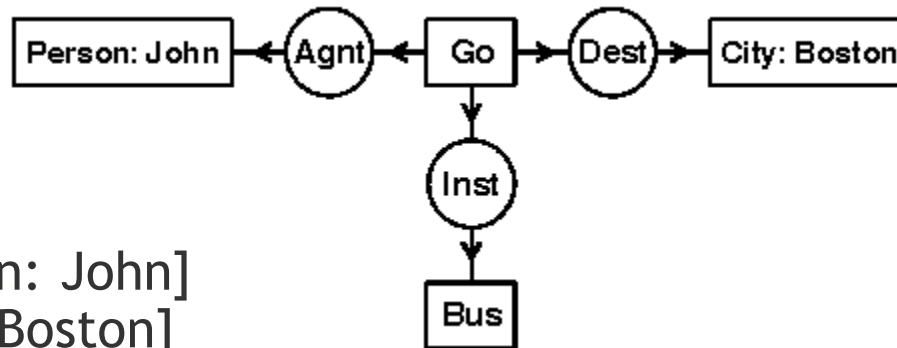
$[Cat: @every*x] [Mat: *y] (On ?x ?y)$

KIF:

$(\text{forall } ((?x \text{ Cat})) (\text{exists } ((?y \text{ Mat})) (On ?x ?y)))$

Példák fogalmi gráfokra 3.

John is going to Boston by bus.



LF:

[Go]-

(Agnt)®[Person: John]

(Dest)®[City: Boston]

(Inst)®[Bus].

CGIF:

[Go: *x] [Person: John *y] [City: Boston *z] [Bus: *w]

(Agnt ?x ?y) (Dest ?x ?z) (Inst ?x ?z)

KIF:

(exists ((?x Go) (?y Person) (?z City) (?w Bus)))

(and (Name ?y John) (Name ?z Boston) (Agnt ?x ?y)

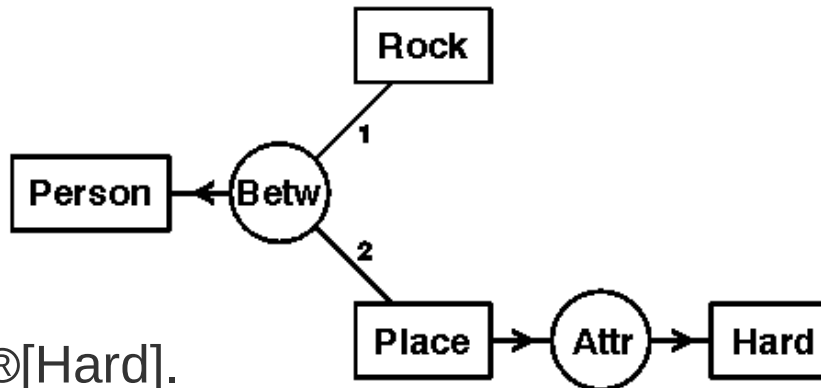
(Dest ?x ?z) (Inst ?x ?w)))

Példák fogalmi gráfokra 4.

A person is between a rock and a hard place.

LF:

[Person]¬(Betw)-
 ¬1-[Rock]
 ¬2-[Place]®(Attr)®[Hard].



CGIF:

(Betw [Rock] [Place *x] [Person]) (Attr ?x [Hard])

KIF:

(exists ((?x person) (?y rock) (?z place) (?w hard)))
 (and (betw ?y ?z ?x) (attr ?z ?w)))

Példák fogalmi gráfokra 5.

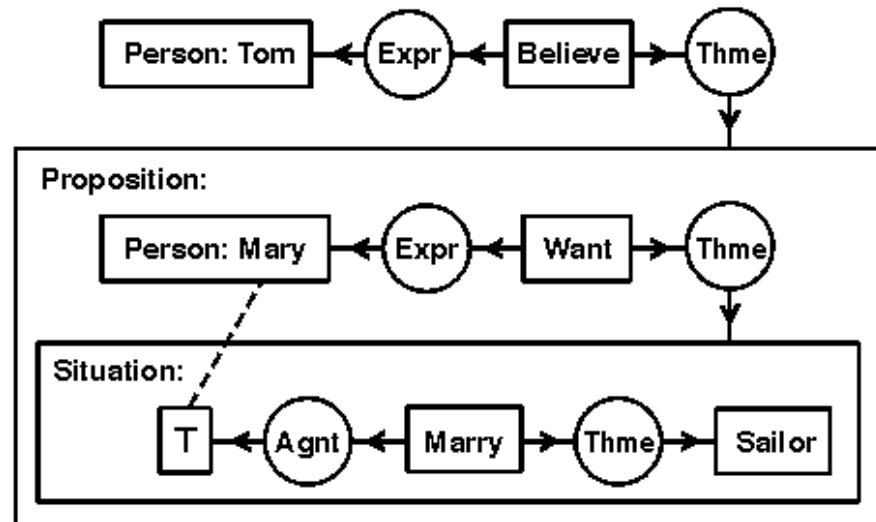
Tom believes that Mary wants to marry a sailor.

LF:

[Person: Tom]-(Expr)-[Believe]®(Thme)-
[Proposition:
[Person: Mary *x]-(Expr)-[Want]®(Thme)-
[Situation:
[?x]-(Agnt)-[Marry]®(Thme)®[Sailor]]].

CGIF:

[Person: *x1 'Tom'] [Believe *x2]
(Expr ?x2 ?x1) (Thme ?x2
[Proposition: [Person: *x3 'Mary'] [Want *x4]
(Expr ?x4 ?x3) (Thme ?x4
[Situation: [Marry *x5] (Agnt ?x5 ?x3)
(Thme ?x5 [Sailor])]])



KIF:

(exists ((?x1 person) (?x2 believe))
(and (expr ?x2 ?x1) (thme ?x2
(exists ((?x3 person) (?x4 want) (?x8 situation))
(and (name ?x3 'Mary) (expr ?x4 ?x3) (thme ?x4
?x8) (dscr ?x8 (exists ((?x5 marry) (?x6 sailor))
(and (Agnt ?x5 ?x3) (Thme ?x5 ?x6))))))))))



Fogalmi függőség

Conceptual dependency = fogalmi függőség

Mary took a book from John.

((actor Mary)
 (action MTRANS)
 (object book)
 (direction (to Mary)
 (from John)))

A fogalmi függőség igeosztályai

Primitive	Definition
ATRANS	The abstract transfer of possession or control from one entity to another.
PTRANS	The physical transfer of an object from one location to another
MTRANS	The transfer of mental concepts between entities or within an entity.
MBUILD	The creation of new information within an entity.
PROPEL	The application of physical force to move an object.
MOVE	The integral movement of a body part by an animal.
INGEST	The taking in of a substance by an animal.
EXPEL	The expulsion of something from an animal.
SPEAK	The action of producing a sound.
ATTEND	The action of focusing a sense organ.



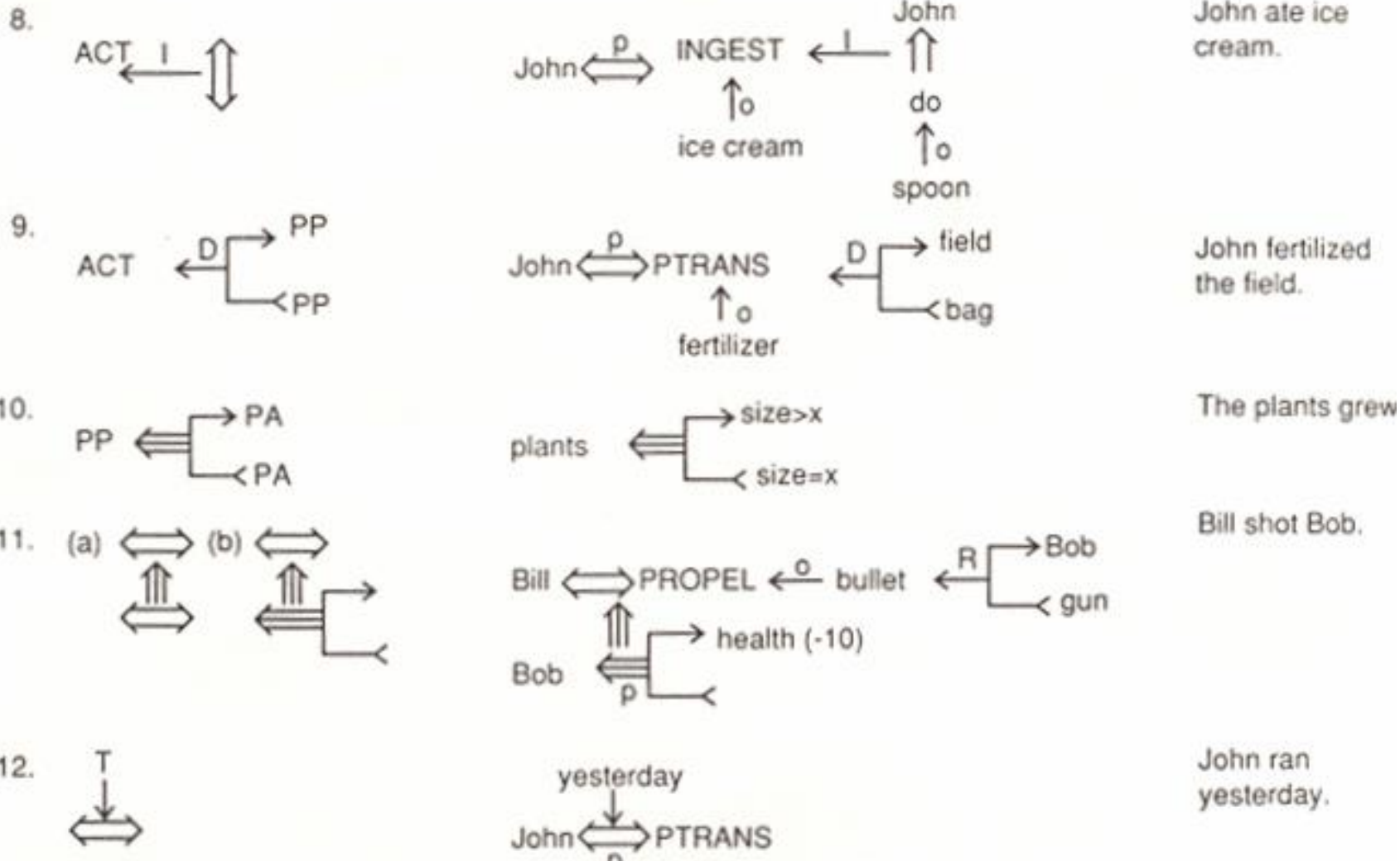
A fogalmi függőség állapotosztályai

<i>state</i>	<i>scale</i>	<i>example</i>
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 eseményábrázolása (1)

1.	$PP \longleftrightarrow ACT$	John $\xrightarrow{P}$ PTRANS	John ran.
2.	$PP \longleftrightarrow PA$	John $\longleftrightarrow$ height (>average)	John is tall.
3.	$PP \longleftrightarrow PP$	John $\longleftrightarrow$ doctor	John is a doctor.
4.	PP ↑ PA	boy ↑ nice	A nice boy
5.	PP ↑↑ PP	dog ↑↑ POSS-BY John	John's dog
6.	$ACT \xleftarrow{O} PP$	John $\xrightarrow{P}$ PROPEL $\xleftarrow{O}$ cart	John pushed the cart.
7.	ACT $\xleftarrow{R}$ $\begin{cases} \rightarrow PP \\ \leftarrow PP \end{cases}$	John $\xrightarrow{P}$ ATRANS $\xleftarrow{R}$ $\begin{cases} \rightarrow John \\ \leftarrow Mary \end{cases}$ ↑ book	John took the book from Mary.

Schank eseményábrázolása (2)



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	<i>constraints</i>	<i>defaults</i>
PATRON	*HUMAN*	*ADULT*
WAITER	*HUMAN*	*ADULT*

COOK	...
MANAGER	...

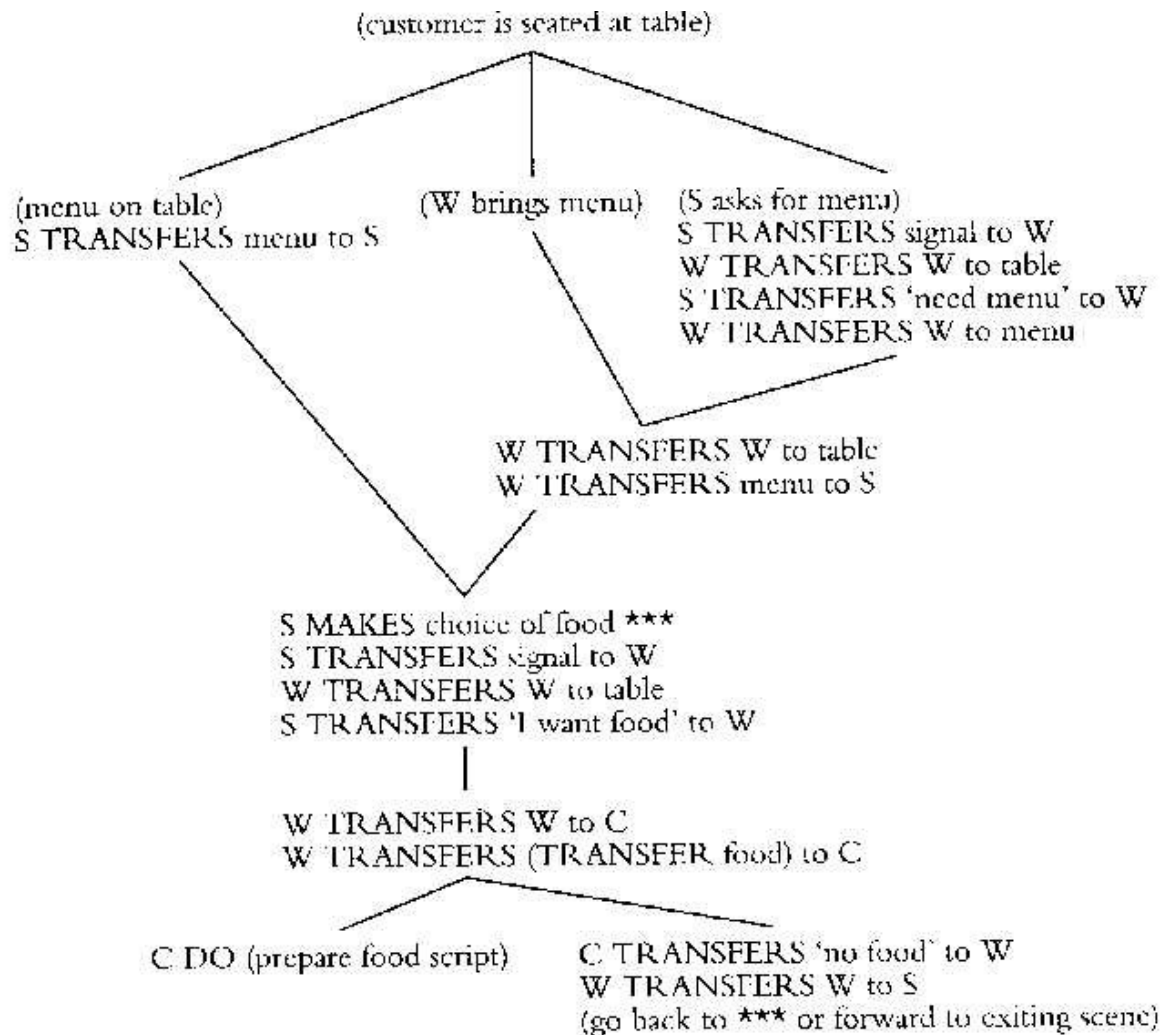
HEADER

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

BODY

EVENT_CHAIN

Az „étterem” teljes forgatókönyve





Lexikális szemantikai reprezentációk

Szemantikus hálók

- ❑ Címkézett gráf
- ❑ Csúcsai: objektumok (fogalmak)
- ❑ Élek: a csúcsok közötti különféle lehetséges relációk
- ❑ A csúcsokhoz rendelhetők attribútumok
- ❑ Az attribútumokat egyes relációk közvetíthetik a relációk mentén

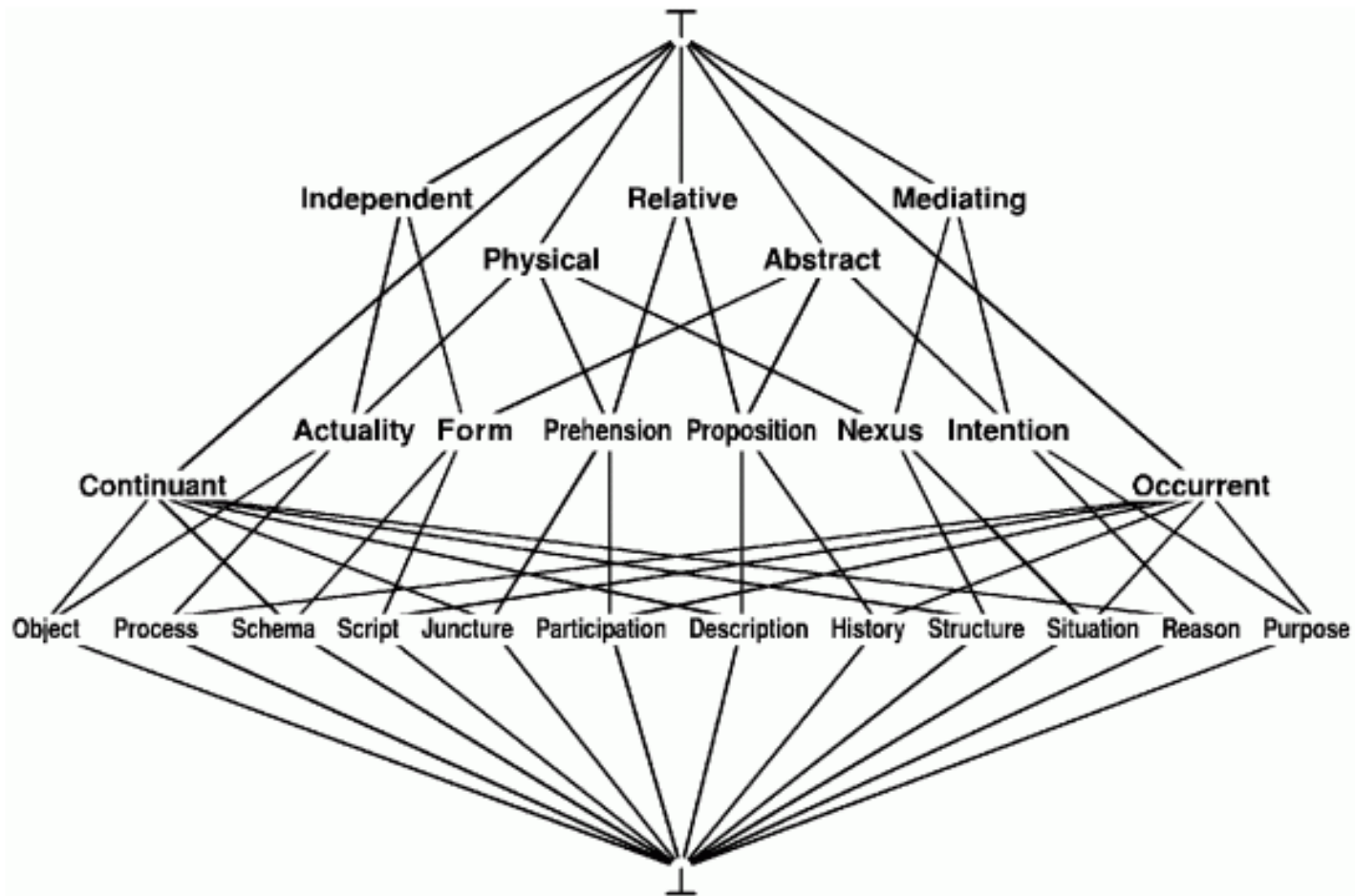
Lexikális szemantikus hálók

- ❑ A modern nyelvészet és a szójelentés: Katz & Fodor, Fillmore, ...
- ❑ Korai pszicholingvista irány, amikor először jelennek meg a hierarchiák (IS-A, HAS-A, PART-OF, ...): Quillian, Minsky, Charniak, ...
- ❑ Információtechnológiai irány:
CyC, MindNet (Microsoft), FrameNet (Fillmore), ...
- ❑ Későbbi pszicholingvista irány:
WordNet, EuroWordNet, SUMO, ...

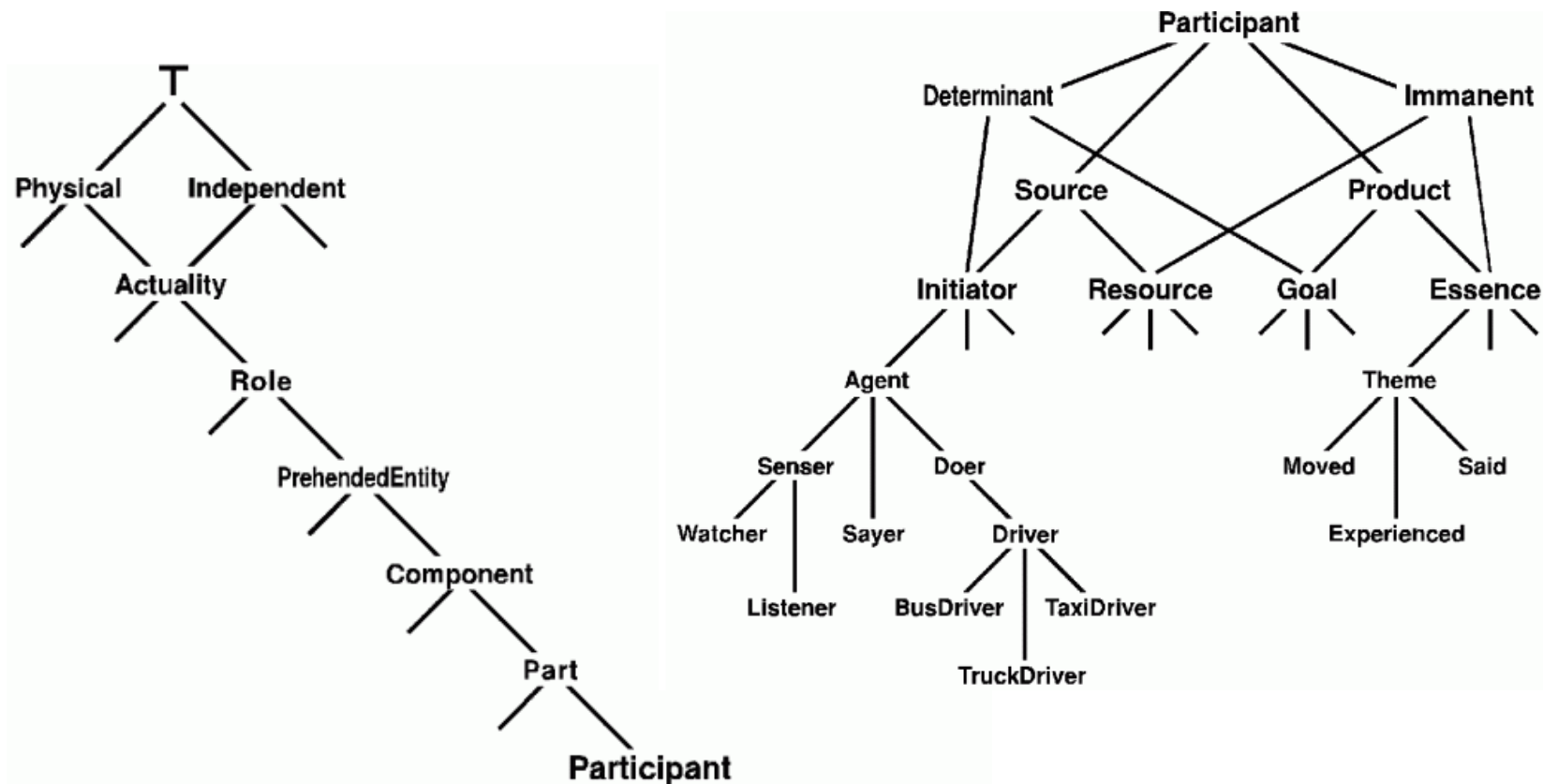
Ontológia

- ❑ Ontológia = „a felfogás leírása”
- ❑ Filozófiai tudományág, a létezést, létező dolgokat vizsgálja, megadja a vizsgált univerzumban létező fogalmak kategóriáit, metafizikai kereteket
- ❑ A cél: a világnak (legtöbbször egy alkalmazás szemszögéből történő) formális leírása
- ❑ Az ontológia értelmezésével érvényes logikai következtetések végezhetők
- ❑ Az informatika több területén népszerű (nem csak a nyelvtechnológiában)

Felső ontológia (felsőszintű kategóriák)

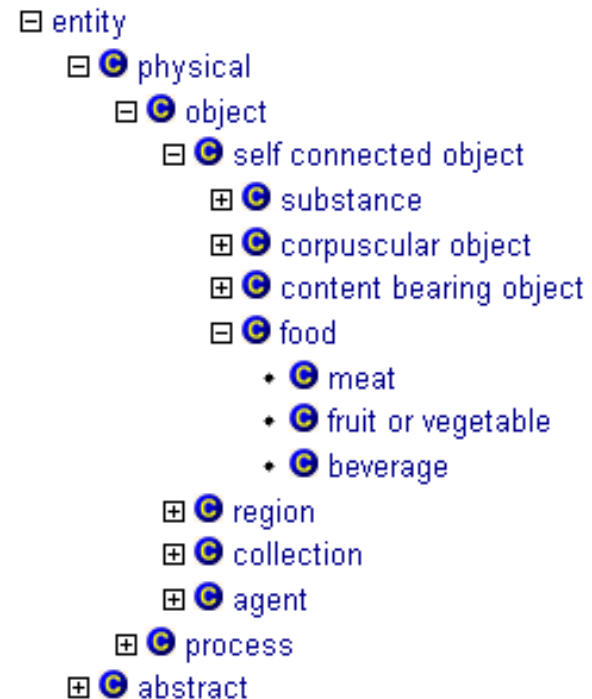


A felső és az alsó szintű ontológiák egy találkozása



SUMO

- ❑ Suggested Upper Merged Ontology
- ❑ 2000 óta
- ❑ Általános magas szintű ontológia (1000 fogalom)
- ❑ Specifikus magas szintű ontológiák (pl. pénzügyi tranzakciókra 20.000 fogalom)
- ❑ Alacsonyabb szintű ontológiákra van belőle leképezés
- ❑ IEEE-tulajdon, de public domain
- ❑ www.ontologyportal.org



WordNet

- ☐ A legelső WordNet (1990): a Princeton WordNet
- ☐ Eredetileg az emberi agy nyelvi tudásreprezentációjának modellje
- ☐ A legnagyobb ingyenes egységes gépileg feldolgozható lexikai adatbázis
- ☐ Ma: WordNet 3.0
- ☐ A WordNet legfőbb jellemzői:
 - szemantikus háló
 - a csúcsok címkéi: a „jelentések” mint szinonimahalmazok (synset)
 - az élek címkéi: a leggyakoribb lexikai relációk
 - az élcímkék szófajfüggőek: főnéviek, igeiek, melléknévi

PWN 2.0

Szófaj	Szavak	Synsetek	Szó-jelentés párok
Főnevek	114 648	79 689	141 690
Igék	11 306	13 508	24 632
Melléknevek	21 436	18 563	31 015
Határozószók	4 669	3 664	5 808
Összesen	152 059	115 424	203 145

Poliszémia a PWN 2.0-ban

Szófaj	Átlag-poliszémia	
	Az egyértelmű szavakkal	Az egyértelmű szavak kizárásával
Főnevek	1.23	2.79
Igék	2.17	3.66
Melléknevek	1.44	2.80
Határozószók	1.24	2.49

A WordNet relációi

Főnév

- *hipernima*: *Y* hipernimája *X*-nek, ha minden *X* (egyfajta) *Y*
(pl. A szuka hipernimája a *kutyá*-nak)
- *hiponima* : *Y* hiponimája *X*-nek, ha minden *Y* (egyfajta) *X*
(pl. A *kutya* hiponimája a *szuká*-nak)
- *rokon fogalom*: *Y* rokon fogalma *X*-nek, ha *X*-nek és *Y*-nak van közös hipernimája
(pl. *farkas* és *kutya*)
- *holonima* : *Y* holonimája *X*-nek, ha *X* része *Y*-nak
(pl. Az *épület* holonimája az *ablak*-nak)
- *meronima* : *Y* is a meronym of *X* if *Y* is a part of *X*
(pl. Az *ablak* holonimája az *épület*-nek)

Ige

- *hipernima*: az *Y* ige hipernimája az *X* igének, ha az *X* aktivitás (egyfajta) *Y*
(pl. *to perceive* is an hypernym of *to listen*)
- *troponima*: az *Y* ige troponimája az *X* igének, ha *Y* valamilyen módon végrehajtott *X*
(pl. A *csacsogás* troponimája a *beszéd*-nek)
- *velejáró*: *Y* ige velejárója *X*-nek, ha *X*-et csinálva *Y*-t is kell csinálni
(pl. *Horkolni* csak úgy lehet, ha *alszunk*)
- *rokon ige*: azok az igék, amelyeknek közös hipernimájuk van

Melléknév

- *rokon főnév*
- *hasonlít*
- *igenév*

Határozószó

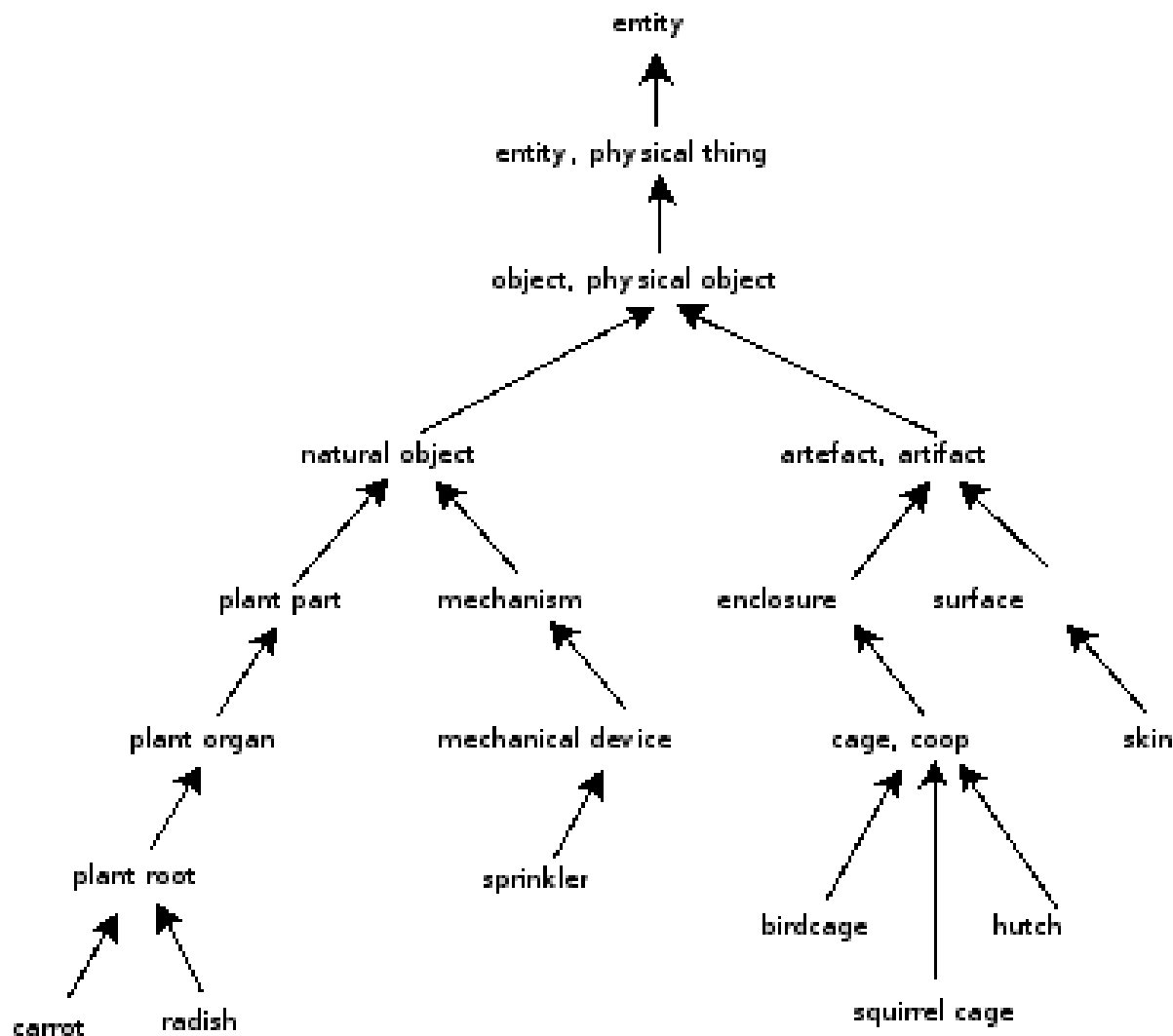
- *melléknévgyök*



A WordNet tematikus szerepei példákkal

Thematic Role	Example
AGENT	<i>The waiter spilled the soup.</i>
EXPERIENCER	<i>John has a headache.</i>
FORCE	<i>The wind blows debris from the mall into our yards.</i>
THEME	<i>Only after Benjamin Franklin broke the ice...</i>
RESULT	<i>The French government has built a regulation-size baseball diamond...</i>
CONTENT	<i>Mona asked "You met Mary Ann at a supermarket"?</i>
INSTRUMENT	<i>He turned to poaching catfish, stunning them with a shocking device...</i>
BENEFICIARY	<i>Whenever Ann Callahan makes hotel reservations for her boss...</i>
SOURCE	<i>I flew in from Boston.</i>
GOAL	<i>I drove to Portland.</i>

Egy „szelet” a Princeton WordNetből



WordNet-jelentések: „bass”

The noun “bass” has 8 senses in WordNet.

1. bass - (the lowest part of the musical range)
2. bass, bass part - (the lowest part in polyphonic music)
3. bass, basso - (an adult male singer with the lowest voice)
4. sea bass, bass - (flesh of lean-fleshed saltwater fish of the family Serranidae)
5. freshwater bass, bass - (any of various North American lean-fleshed freshwater fishes especially of the genus *Micropterus*)
6. bass, bass voice, basso - (the lowest adult male singing voice)
7. bass - (the member with the lowest range of a family of musical instruments)
8. bass - (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

WordNet-öröklődés: „bass”

Sense 3

bass, basso --

(an adult male singer with the lowest voice)

=> singer, vocalist

=> musician, instrumentalist, player

=> performer, performing artist

=> entertainer

=> person, individual, someone...

=> life form, organism, being...

=> entity, something

=> causal agent, cause, causal agency

=> entity, something

Sense 7

bass --

(the member with the lowest range of a family of musical instruments)

=> musical instrument

=> instrument

=> device

=> instrumentality, instrumentation

=> artifact, artefact

=> object, physical object

=> entity, something



WordNet a weben

<http://wordnetweb.princeton.edu>

WordNet Search - 3.0 - [WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

Display Options:

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

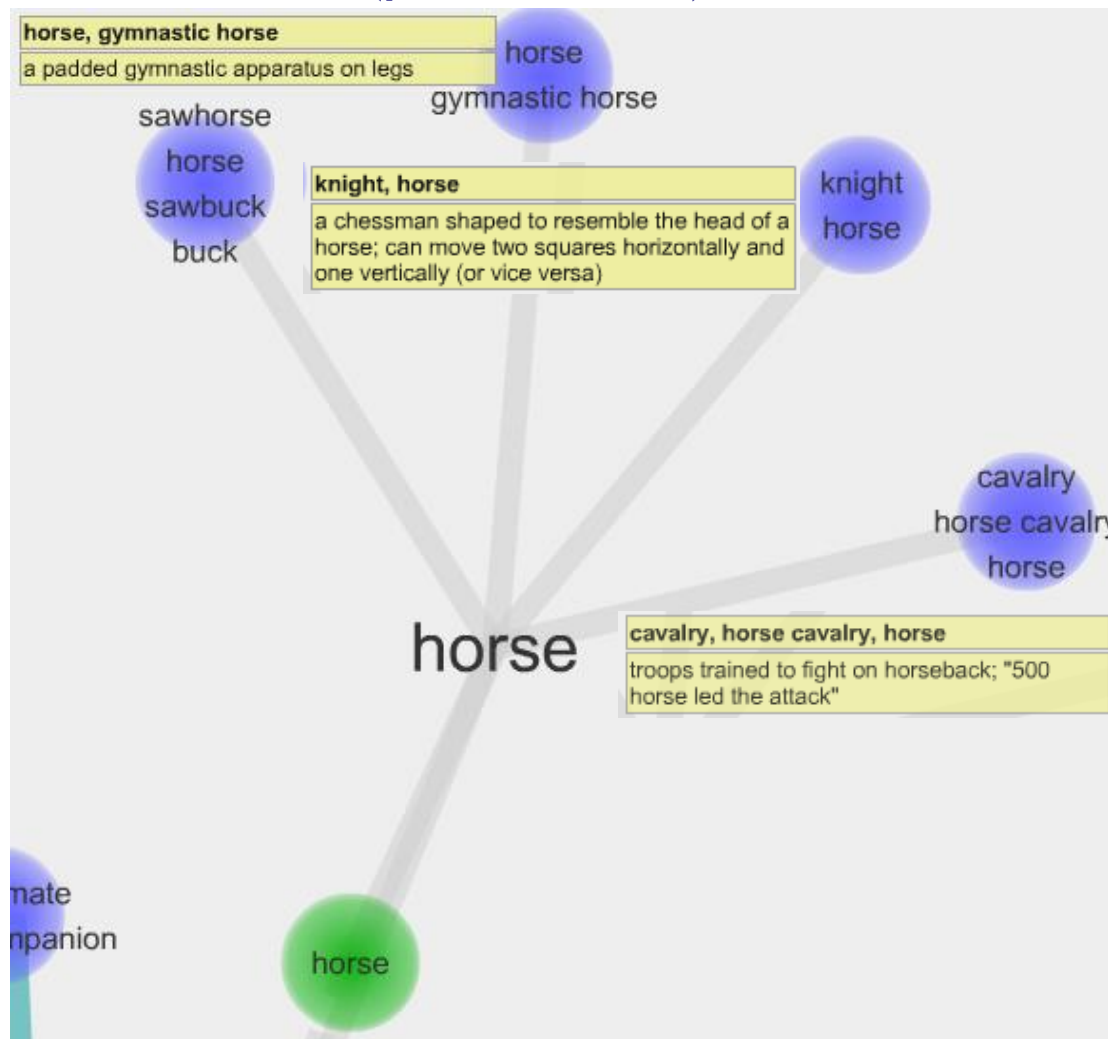
Noun

- [S:](#) [\(n\)](#) **horse**, [Equus caballus](#) (solid-hoofed herbivorous quadruped domesticated since prehistoric times)
- [S:](#) [\(n\)](#) **horse**, [gymnastic horse](#) (a padded gymnastic apparatus on legs)
- [S:](#) [\(n\)](#) **cavalry**, [horse cavalry](#), **horse** (troops trained to fight on horseback)
"500 horse led the attack"
- [S:](#) [\(n\)](#) **sawhorse**, **horse**, [sawbuck](#), [buck](#) (a framework for holding wood that is being sawed)
- [S:](#) [\(n\)](#) **knight**, **horse** (a chessman shaped to resemble the head of a horse; can move two squares horizontally and one vertically (or vice versa))

Verb

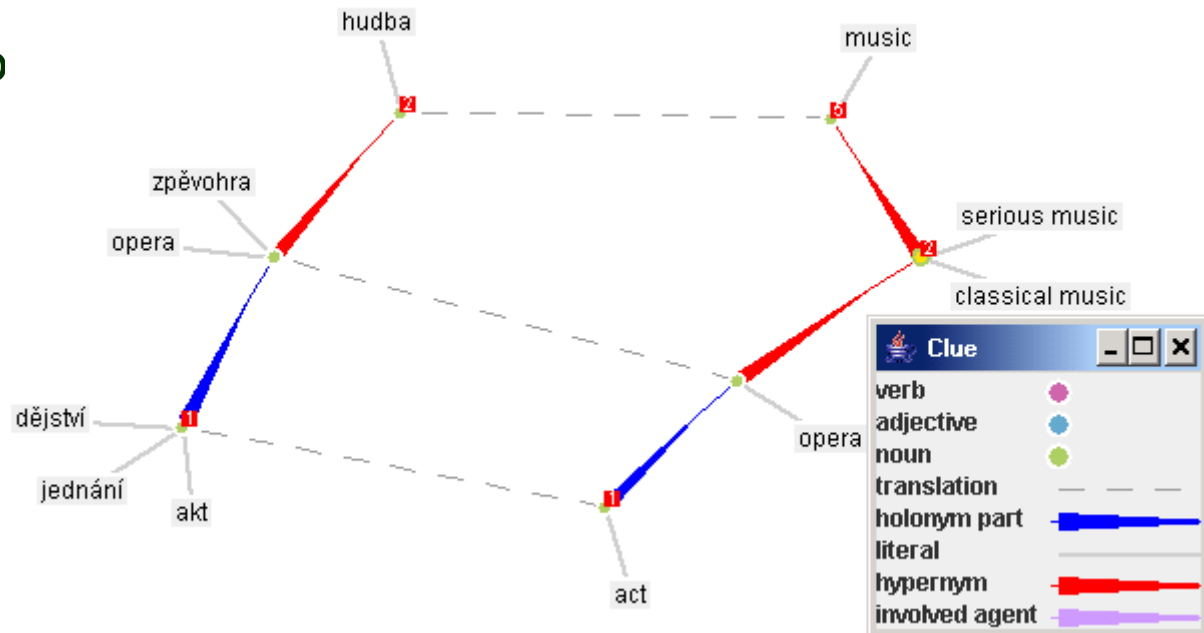
- [S:](#) [\(v\)](#) **horse** (provide with a horse or horses)

Különféle vizuális megjelenítés (pl. VisuWords)



Az EuroWordNet

- ❑ Többnyelvű ontológia a WordNet modellje alapján
- ❑ A nyelvek közti kapcsolódást egy nyelvközi indexszel oldják meg: Inter Lingual Index, ILI
- ❑ EuroWordNet Core: PWN 1.5
- ❑ <http://www.hum.uva.nl/~ewn>
- ❑ Két különböző nyelvi synsetet nyelvközi (ekvivalencia) reláció köt össze, ha mindkettő ugyanahhoz a PWN synsethez van csatolva
- ❑ Jelentések közti ekvivalenciáról van szó
- ❑ Már informatikai motivációval indult el a fejlesztése

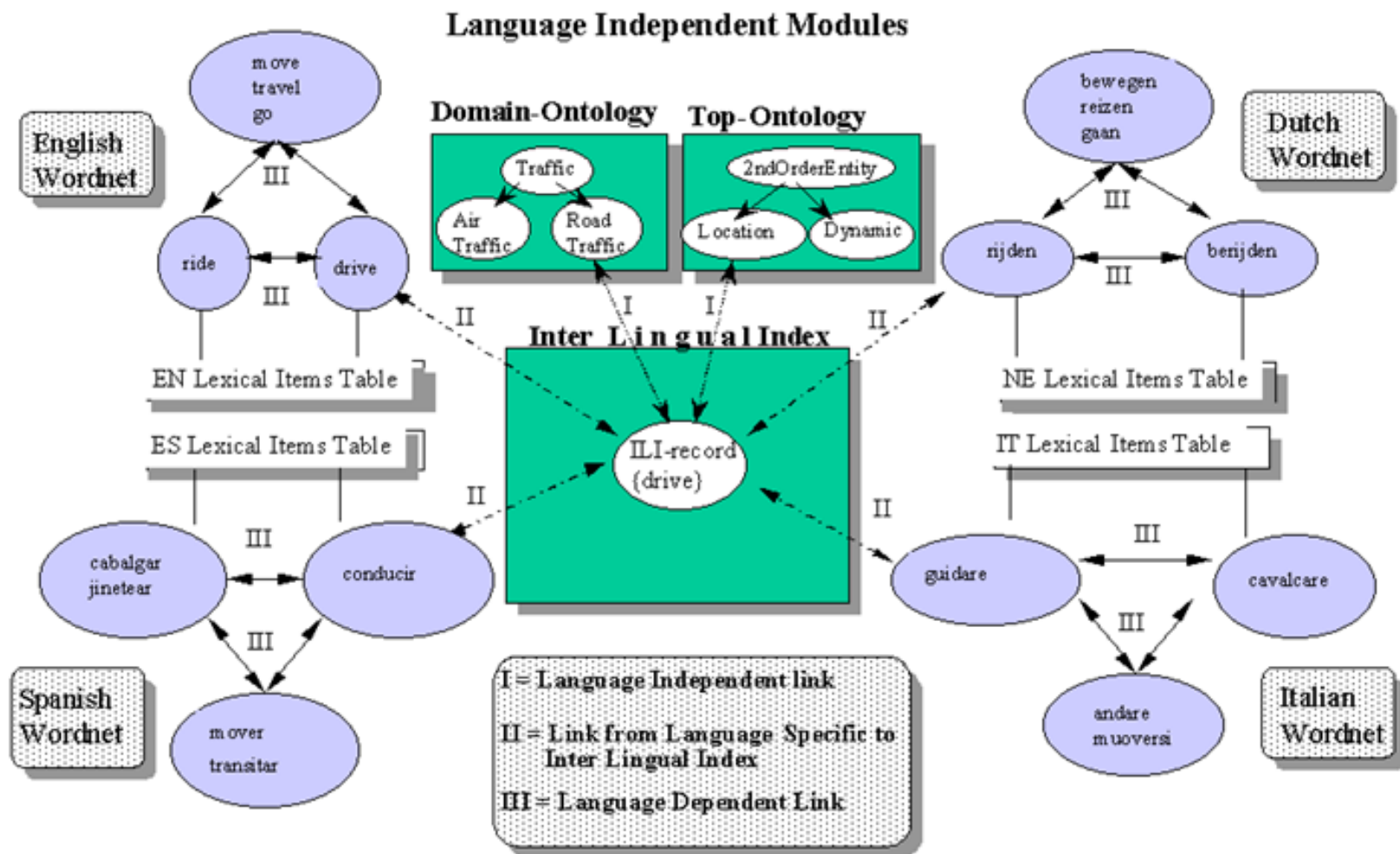




EWN Base Concepts & Top Ontology

- ❑ EWN Base Concepts: maximalizálni akarták az átfedést a különböző nyelvi hálózatok topológiája közt
- ❑ Az EWN BC-t minden EWN tag kötelező jelleggel létrehozza a saját WordNetjében
- ❑ Minimális átjárhatóságot biztosít a nyelvek közt
- ❑ BC-k kiválasztása: ha BC-ként két nyelvben is be akartak egy adott fogalmat vezetni, akkor a többire kötelezővé vált (1059 db synset, amiből 796 N, 263 V)
- ❑ EWN Top Ontology: a különböző WN-ek hasonlóságának maximalizálására, a BC-k fölé, illetve azokból
- ❑ Az EWN ToP Ontology nyelvfüggetlen: minden célnyelvre azonos, és a legabsztraktabb fogalomosztályok közösek
- ❑ A BC-k és a TOP ontológia kialakításával értelemszerűen az angol WN-t is bővíteni, ill. módosítani kellett → levált a princetoni vonalról a fejlődése

Az EWN felépítése



BalkaNet

- ❑ Balkáni nyelvek csatlakozása a EuroWordNet rendszerbe
- ❑ <http://www.ceid.upatras.gr/Balkanet/>
- ❑ Az EWN követelményein túlmutató fejlesztések
 - bővített BC halmaz: vették az egész hipernim és holonim lezártját
 - módosított ILI: 8516 fogalom a BILI-ben
 - minőségbiztosítással konzisztensebb adatbázis
 - XML adatformátum
 - validálták a nyelvek közti összefüggéseket a hálóban

Magyar WordNet

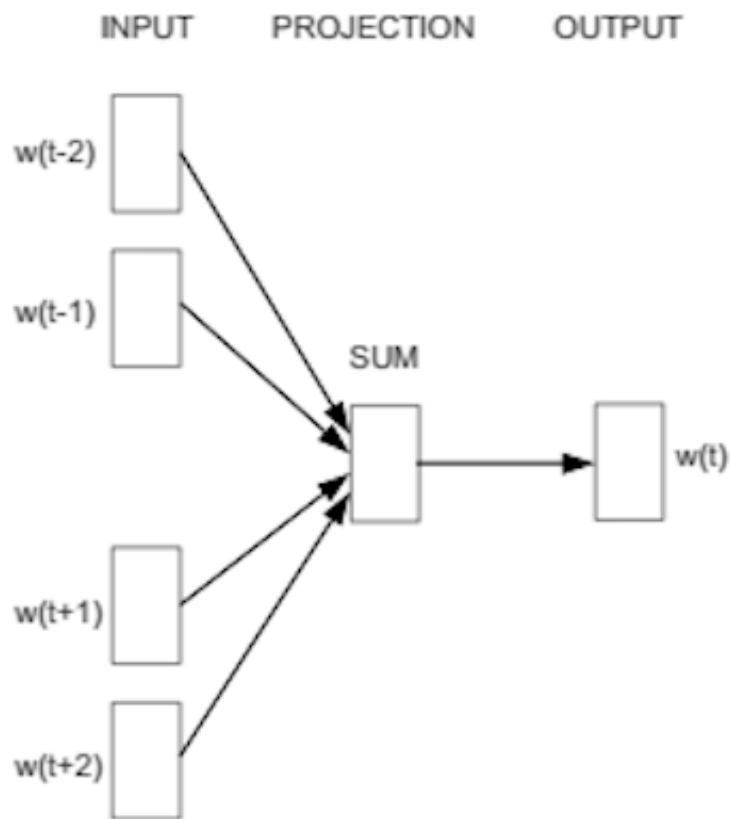
- ❑ MorphoLogic Kft, MTA NyTI, SzTE: 2007-ben fejeződött be a fejlesztése
- ❑ Kutatási célokra hozzáférhető: <https://github.com/dlt-rilmta/huwn.rdf>
- ❑ **Forrásai:**
 - idegennyelvűek (pl. más WordNetek)
 - kétnyelvűek (pl. elektronikus szótárak)
 - egynyelvűek (pl. ÉKSZ, szinonimaszótárak)
- ❑ 42.288 fogalom, a BalkaNet specifikációját (ILI, BCS, stb.) követve, kézi validálással, javítással

Szófaj	Synsetek	Jelentések	Szavak
Főnevek	33530	45508	39079
Igék	3607	6947	4905
Melléknévek	4112	6215	4900
Határozószók	1039	1793	1354

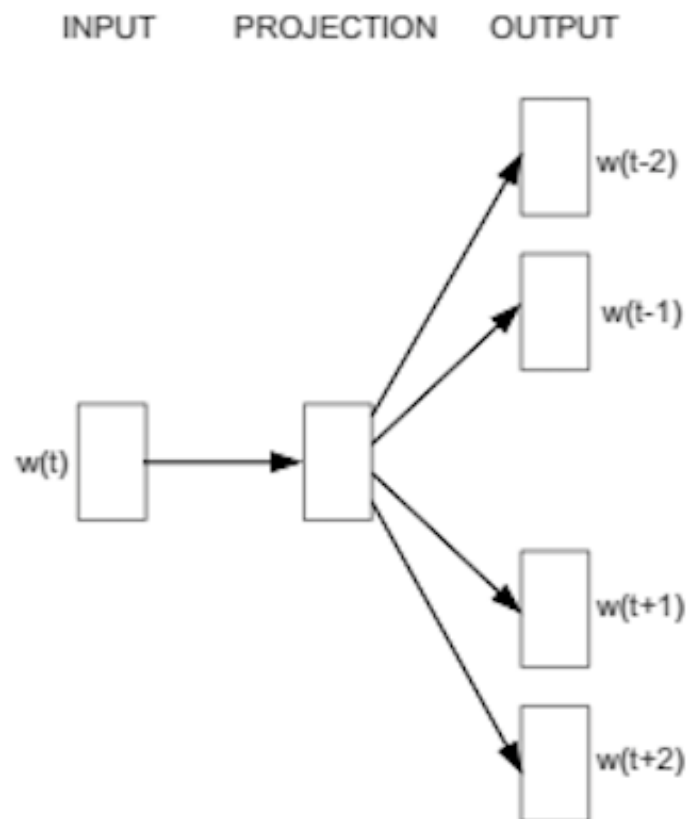
Mélytanulás és szójelentés

- Szóbeágyazás: nyelvmodellezési és jegytanulási technikák gyűjtőneve, ahol a szavakat egy n -dimenziós valós vektor reprezentálja
- A leképezés neurális hálókkal, a szókörnyezet reprezentációjával történik
- Mikolov (2013): word2vec egy kétrétegű neurális háló szövegreprezentációs célokra, ahol a bemenet a szövegkorpusz és a kimenet a korpusz szavait reprezentáló jegyvektorok halmaza
- A word2vec maga nem mélyneurális háló, de a kimenetét mélyneurális hálók is tudják kezelni
- Két lehetséges alapmodell:
 - (1) környezet→szó: folytonos szózsák (Continuous Bag Of Words, CBOW)
 - (2) szó→környezet: skip-gram

Környezet→szó vagy szó→környezet



CBOW



Skip-gram

Forrás: <https://deeplearning4j.org/word2vec>

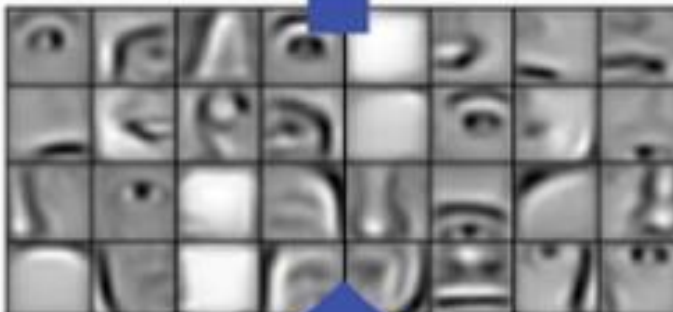
Mélytanulás

- A hagyományos tanulás sekély hálókon alapul, egy bemeneti és egy kimeneti réteggel, maximum egy rejtett réteggel közöttük
- A mélytanulásban ennél több, tehát több mint három réteg van, és minden csomópont-réteg az előző réteg kimenetének diszjunkt jegyhalmazán tanul

Forrás: <https://deeplearning4j.org/neuralnet-overview>



Harmadik réteg

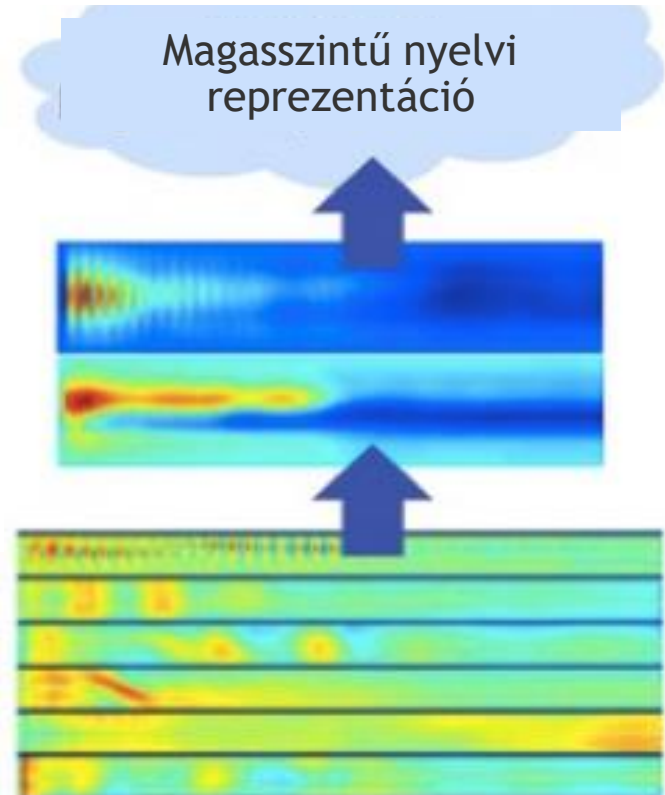


A részek összekapcsolása objektumokká

Második réteg

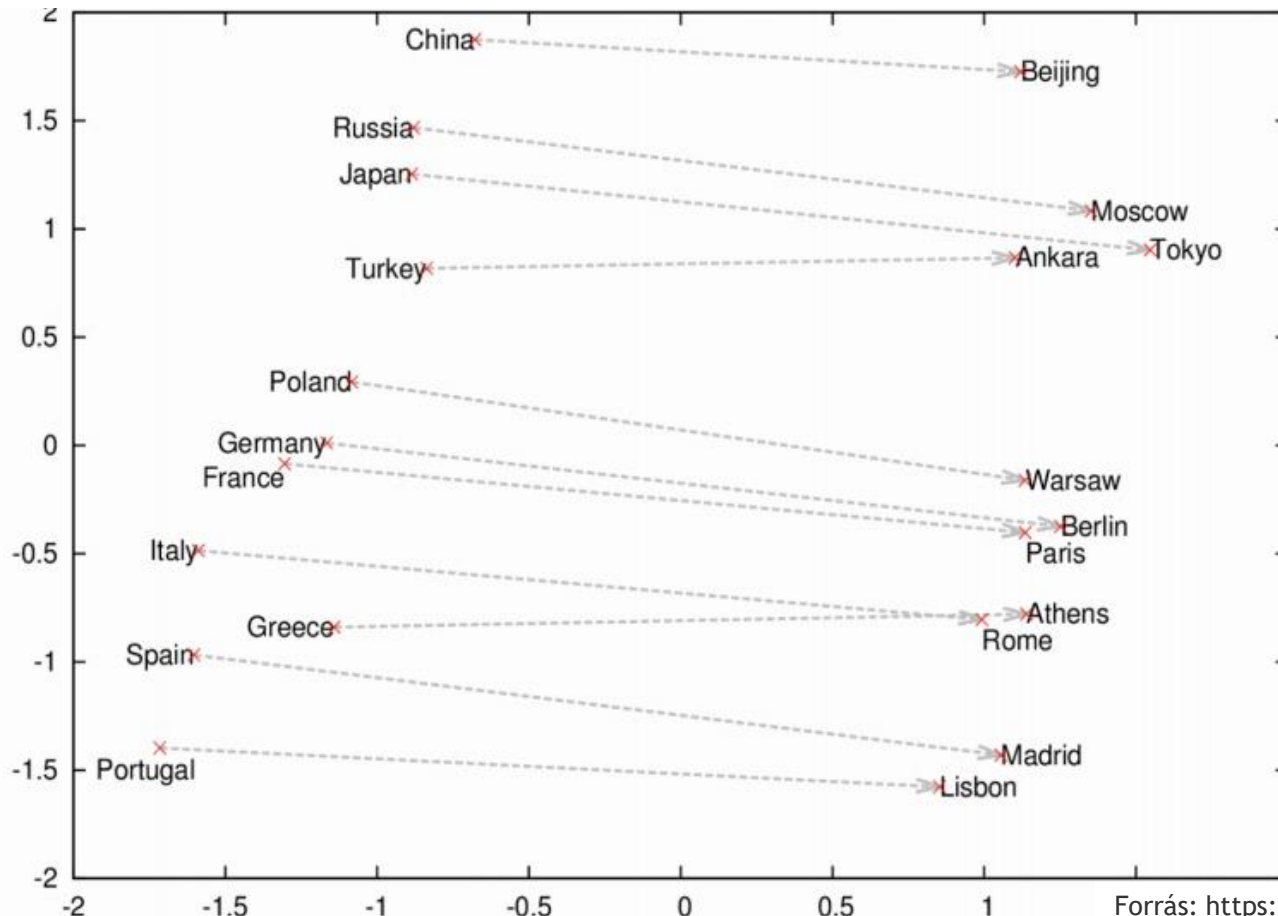


Első réteg



Megjelenik az analógia

Ha v a w szót az n -dimenziós vektorreprezentációba képező függvény,
 akkor pl. $v(\text{főváros}_i) - v(\text{ország}_i) + v(\text{ország}_k) \approx v(\text{főváros}_k)$
 ahol $\approx$ jobboldala a baloldal értékének legközelebbi szomszédja





Az első kísérletek magyarra: főnevek, igék

0	kenyerek	1	2270	0	sógorom	1	2176	0	iszunk	1	6737
1	zsemlék	0.8105	283	1	sógornőm	0.8524	1855	1	megiszunk	0.8151	907
2	péksütemények	0.8048	997	2	nagynéném	0.8374	3088	2	igyunk	0.8046	5814
3	kekszek	0.7972	1046	3	nagybátyám	0.8314	4394	3	ihatunk	0.7878	875
4	pékáruk	0.7957	771	4	apósom	0.8235	2538	4	eszünk	0.7638	18243
5	tészták	0.7881	2466	5	nővérem	0.8208	10266	5	innánk	0.7497	432
6	lepények	0.7849	202	6	keresztanyám	0.8170	859	6	fogyasztunk	0.7429	7196
7	kiflik	0.7843	349	7	vejem	0.8121	526	7	ittunk	0.7342	8195
8	kalácsok	0.7841	277	8	keresztapám	0.8105	931	8	megisszuk	0.7339	694



Tulajdonnevek, új szavak

0	Trabantok	1	277	0	Princeton	1	1433	0	migráns	1	1229
1	Wartburgok	0.8822	142	1	Yale	0.8927	2809	1	bevándorló	0.7223	7427
2	Skodák	0.8569	237	2	Stanford	0.8637	3154	2	migránsok	0.7033	1199
3	Zsigulik	0.8537	111	3	Harvard	0.8556	6293	3	bevándorlók	0.6566	13222
4	Ladák	0.8511	410	4	Rutgers	0.8502	563	4	roma	0.6542	76409
5	Moszkvicsok	0.8506	91	5	Georgetown	0.8371	454	5	menedékkérő	0.6204	246
6	Volgák	0.8189	90	6	Northwestern	0.8350	445	6	menedékkérők	0.5934	644
7	Daciák	0.7917	115	7	Cambridge-i	0.8299	1307	7	migránsoknak	0.5908	72
8	Szukik	0.7893	272	8	Harward	0.8296	165	8	marginalizált	0.5860	444



Fogalmi rokonságok tövesített korpusszal

0	franciakulcs	1	255	0	csavargó	1	6123	0	migráns	1	3264
1	feszítővas	0.8590	846	1	koldus	0.8026	15793	1	bevándorló	0.8162	23954
2	csavarkulcs	0.8445	473	2	zsvány	0.7691	3497	2	menedékkérő	0.8043	1309
3	csípőfogó	0.8242	345	3	haramia	0.7624	2024	3	vendégmunkás	0.7241	4808
4	pajszer	0.8219	567	4	csirkefogó	0.7570	2019	4	menekült	0.7218	27430
5	hidegvágó	0.8054	156	5	vadember	0.7531	2497	5	kivándorló	0.7083	1893
6	csavarhúzó	0.7984	4369	6	útonálló	0.7472	1942	6	szexmunkás	0.6668	803
7	csőfogó	0.7890	111	7	utcagyerek	0.7401	1653	7	roma	0.6571	162822
8	villáskulcs	0.7890	764	8	gazfickó	0.7366	1651	8	EU-állampolgár	0.6269	316



Antonimák, szlengek, elütések

0	lerombol	1	18374
1	szétrombol	0.9202	3158
2	szétzúz	0.8700	3662
3	elpusztít	0.8554	38350
4	újjaépít	0.8423	8664
5	szétver	0.8360	15517
6	újraépít	0.8344	3063
7	feléget	0.8329	4243
8	rombol	0.8175	28932

0	bealszik	1	4325
1	elszundít	0.8061	781
2	elálmosodik	0.7889	2118
3	elbóbiskol	0.7833	1507
4	visszaalszik	0.7742	5858
5	elalszik	0.7646	62217
6	kipurcan	0.7399	1187
7	szunyál	0.7315	974
8	elbambul	0.7285	1953

0	rövidnac	1	43
1	pizs	0.7731	180
2	napszemcs	0.7584	37
3	sap	0.7460	374
4	zacs	0.7259	170
5	szemcs	0.7209	37
6	pih	0.7198	149
7	suzuk	0.6943	131
8	nemtomm	0.6795	47



... meg érzi a stílust is?

0	Katalin	1	88546	0	Eufrozina	1	254	0	Kincső	1	1242
1	Zsuzsanna	0.8893	30461	1	Jolánta	0.7732	307	1	Csenge	0.8689	4680
2	Ilona	0.8783	33342	2	Konstancia	0.7679	275	2	Evelin	0.8662	3497
3	Ágnes	0.8750	69813	3	Gertrúd	0.7469	1530	3	Bianka	0.8620	4242
4	Gabriella	0.8735	27494	4	Eugénia	0.7418	342	4	Fanni	0.8465	10955
5	Judit	0.8730	74435	5	Adelhaid	0.7410	185	5	Kitti	0.8452	6544
6	Szilvia	0.8483	18932	6	Amália	0.7187	1748	6	Cintia	0.8387	1194
7	Ildikó	0.8465	55454	7	Sarolt	0.7168	795	7	Villő	0.8358	542
8	Klára	0.8442	22176	8	Gertrud	0.7093	802	8	Lilla	0.8296	15016



... és érti a tulajdonneveket?

0	Smith	1	30236	0	McCartney	1	4494	0	Luther	1	16466
1	Wilson	0.8491	19408	1	Elton	0.7574	4107	1	Zwingli	0.7782	853
2	Thompson	0.8476	6506	2	Lennon	0.7094	6597	2	Melanchton	0.7558	873
3	Harris	0.8447	8168	3	Sting	0.6988	3622	3	Augustinus	0.7487	1749
4	Walker	0.8429	6358	4	Clapton	0.6888	2106	4	Kálvin	0.7467	19884
5	Fisher	0.8418	4496	5	Oakenfold	0.6880	358	5	Melanchthon	0.7313	626
6	Adams	0.8413	7335	6	Taylor	0.6796	20785	6	Bucer	0.7218	274
7	Taylor	0.8407	20785	7	Richards	0.6776	3616	7	Servet	0.7128	1202
8	Wright	0.8360	6385	8	Perry	0.6765	8923	8	reformátor	0.7068	4649



Bevezetés a nyelvtechnológiába 7. Fordítástámogatás: szótárak

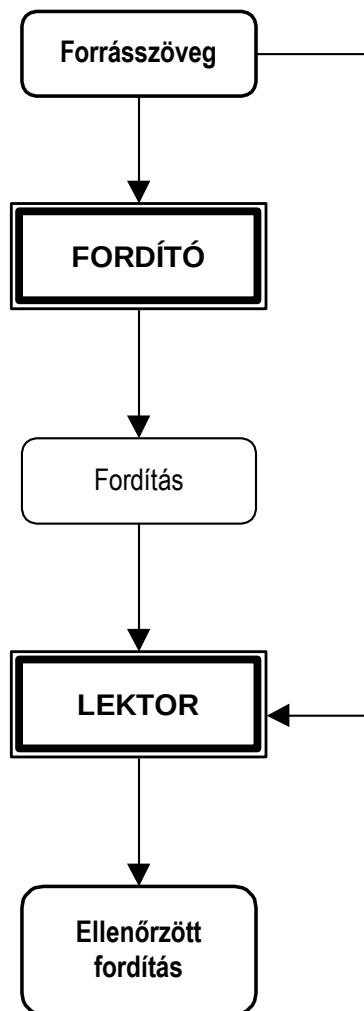
A nyelvtechnológia szerepe a lexikográfiában

- ❑ A szótárak célja ma
 - emberek számára készülnek
 - gépek számára készülnek
- ❑ Miből hozunk létre ma szótárakat?
 - semmiből, technikai támogatással
 - szövegekből, gépi támogatással
 - meglevő szótár(ak)ból, gépi támogatással
- ❑ A géppel támogatott szótárkészítés típusai
 - embernek írt forrásokból, ember számára
 - embernek írt forrásokból, gép számára
 - gépi forrásokból, ember számára
 - gépi forrásokból, a gép számára

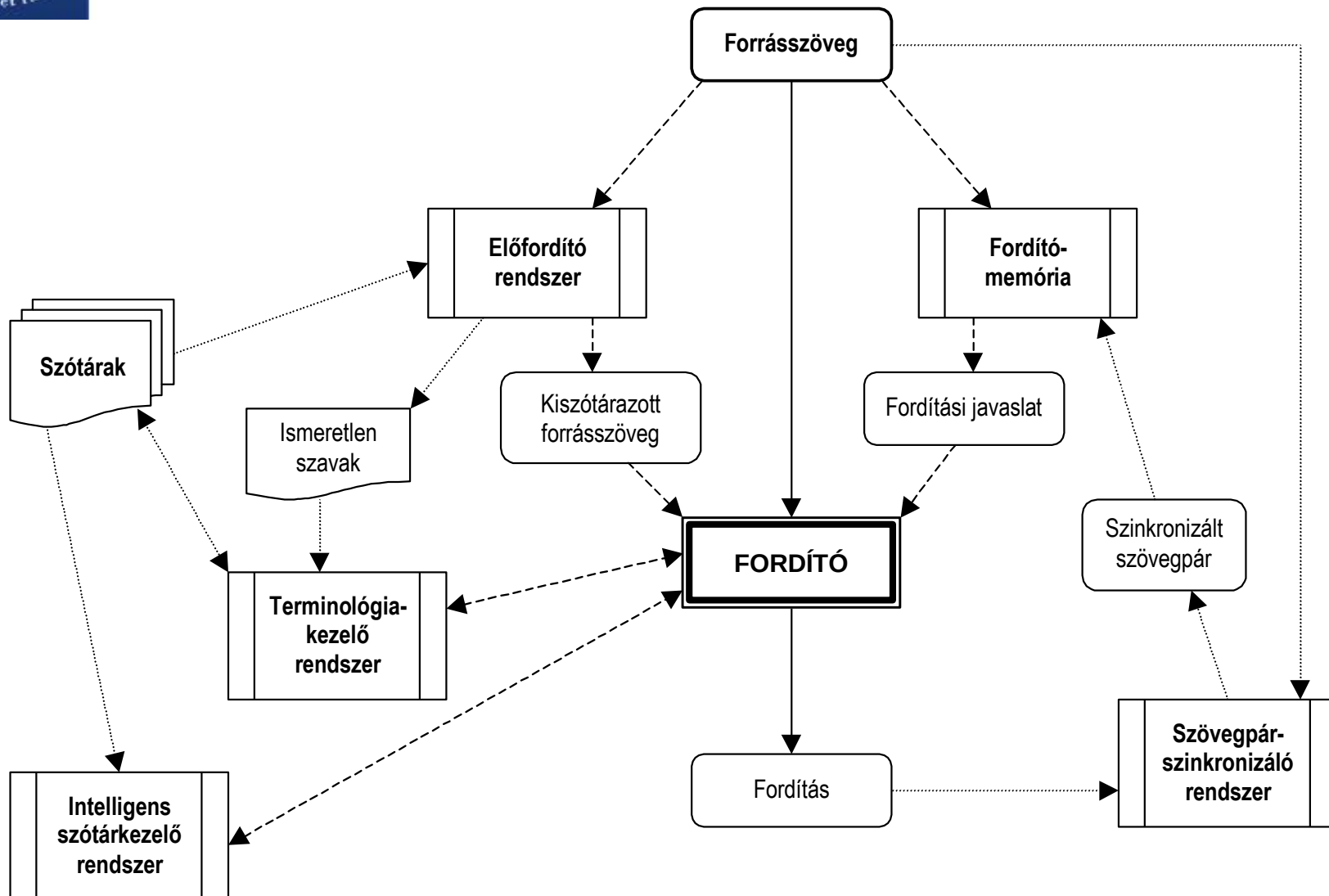


A szótár helye a fordítási folyamatban

Az emberi fordítás gépi támogatása



A fordító számítógépes segédeszközei

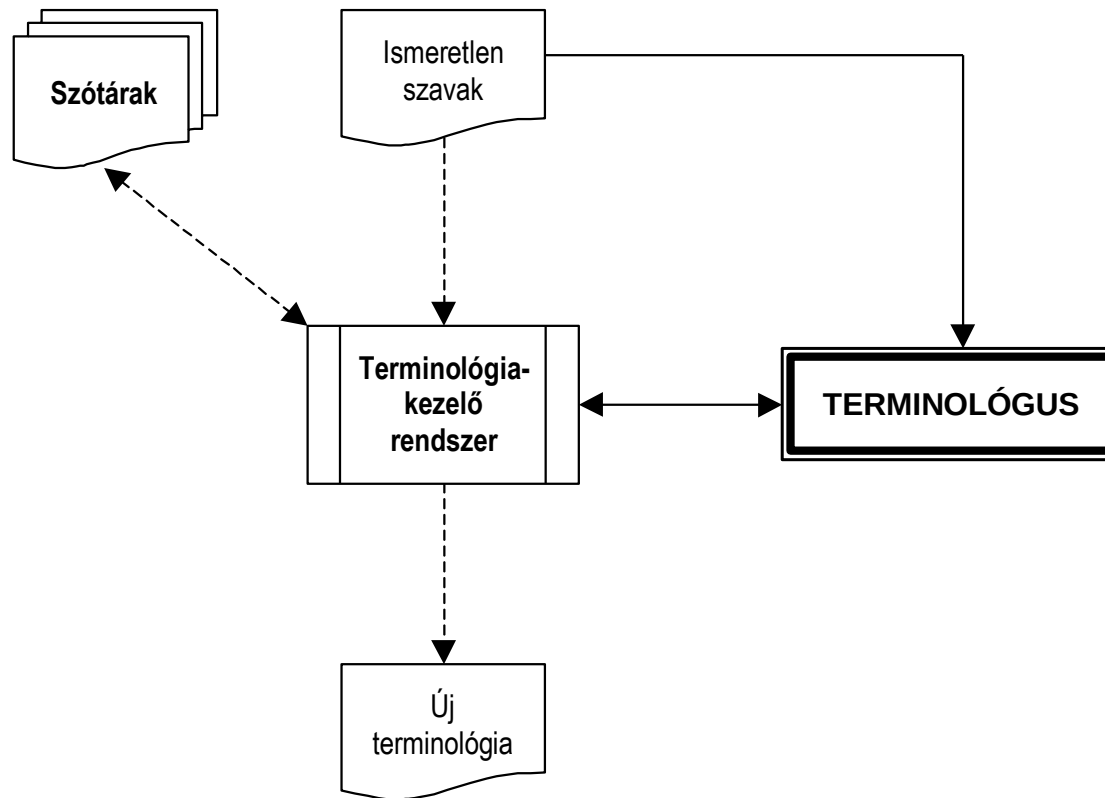


```

graph TD
    FSz[Forrásszöveg] --> L[LEKTOR]
    Ft[Fordítás] --> L
    Ut[Új terminológia] --> L
    L --> EF[Ellenőrzött fordítás]
    L -.-> SzSz[Szövegpár-szinkronizáló rendszer]
    SzSz -.-> SzSzS[Szinkronizált szövegpár]
    SzSzS -.-> KEr[Konzisztencia-ellenőrző rendszer]
    KEr -.-> IK[Inkonzisztenciák]
    IK -.-> L
    L -.-> TK[Terminológia-kezelő rendszer]
    TK -.-> ÚT[Új terminológia]
    TK <--> Sz[Szótárak]
    Sz -.-> ISz[Intelligens szótárkezelő rendszer]
    ISz -.-> L
    Sz -.-> KEr
  
```

The diagram illustrates the architecture of the Lektor system. At the top, 'Forrásszöveg' (Source text) and 'Fordítás' (Translation) feed into the central 'LEKTOR' (Lektor) component. 'Új terminológia' (New terminology) also feeds into 'LEKTOR'. 'LEKTOR' outputs 'Ellenőrzött fordítás' (Checked translation). 'LEKTOR' also interacts with several other components: 'Szövegpár-szinkronizáló rendszer' (Text pair synchronization system), 'Szinkronizált szövegpár' (Synchronized text pair), 'Konzisztencia-ellenőrző rendszer' (Consistency checking system), 'Inkonzisztenciák' (Inconsistencies), 'Terminológia-kezelő rendszer' (Terminology management system), 'Új terminológia' (New terminology), 'Szótárak' (Dictionaries), and 'Intelligens szótárkezelő rendszer' (Intelligent dictionary management system). The 'Terminológia-kezelő rendszer' is linked to 'Új terminológia' and 'Szótárak'. 'Szótárak' is linked to 'Intelligens szótárkezelő rendszer'. 'Intelligens szótárkezelő rendszer' feeds into 'LEKTOR'. 'Szövegpár-szinkronizáló rendszer' feeds into 'Szinkronizált szövegpár', which feeds into 'Konzisztencia-ellenőrző rendszer', which feeds into 'Inkonzisztenciák', which feeds back into 'LEKTOR'. 'LEKTOR' also feeds into 'Terminológia-kezelő rendszer'.

A terminológus helye a fordítási folyamatban



Leegyszerűsített szótártipológia

- ❑ Nyomtatott szótárak
Elektronikus szótárak
- ❑ Szótárak
Terminológiai adatbázisok
- ❑ Közvetlen szótárak
Közvetett (elektronikus) szótárak
- ❑ Kétnyelvű szótárak
Egynyelvű szótárak
Többnyelvű szótárak
- ❑ A forrásnyelv és a célnyelvek aszimmetriája
papír- és elektronikus alapon



Szótártípusok

A nyelvek száma szerint

- ❑ Ezek voltak először: lexikon, enciklopédia
- ❑ Mai egynyelvűek: értelmező, nyelvtörténeti, etimológiai, nyelvjárási, szólás-, szleng-, idegen szavak, helyesírási, gyakorisági, ragozási, a tergo (szóvégmутató) szótárak
- ❑ Kétnyelvűek
- ❑ Többsnyelvűek: valójában párba állított kétnyelvűek
- ❑ Sőt, az egynyelvűek is két különböző nyelvet használnak: a címszavak és a szócikkek „nyelve, nyelvezete” különbözik

A szótárak általános szerkezete

- ❑ Önálló és utaló szócikkek
- ❑ Szócikkfej (=a szótári baloldal): címszó, homonimák, alak- és írásváltozatok, kiejtés, elválasztás, szófaj, főbb toldalékos alakok, nyelvtani megjegyzés, stílusminősítés
- ❑ Jelentéscsoportok (=a szótári jobboldal) alapjelentés, jelentésárnyalatok, értelmezések és példák, továbbá szólások, közmondások, más szavakkal alkotott összetételek, származékszók

A szótárelemek nyelvtechnológiai felhasználása

- ☐ Címszó: kiindulás helyesírási programokhoz
- ☐ Variánsok és toldalékolt alakok: a morfológiai rendszerhez
- ☐ Szótagolás: elválasztó programokhoz
- ☐ Kiejtés: beszédkeltő rendszerekhez
- ☐ Szófaj: egyértelműsítőkhöz
- ☐ Témakód: szövegtípus-azonosításhoz
- ☐ Definíciók: jelentés-egyértelműsítéshez
- ☐ Példák: a címszó körüli többszavas kifejezések azonosításához
- ☐ „Lásd még” szavak: szinonimák, antonimák

Segéd- és szakszótárak

- ☐ Szakmai szótárak (enciklopédiák, lexikonok?)
- ☐ Terminológiai adatbázisok
- ☐ Keveredés az alapszótárakkal: általános nyelvi nagyszótárak (tele szakszócikkekkel!)
- ☐ Az egyidejű használat lehetősége miatt megváltozott szerepük az elektronikus világban
- ☐ Dinamikus szakszótár-kiegészítés alapszótárakhoz

Nyomtatott szótárak vs. elektronikus szótárak

Nyomtatott szótárak	Elektronikus szótárak
Zárt szerkezetű szöveget tartalmaznak.	Zárt vagy nyitott szerkezetű szöveget tartalmaznak.
Van területi korlát a címszójegyzék összeállításában és a jelentésstruktúra felépítésében.	Nincs területi korlát.
A címszavak hierarchikusan, betűrendben követik egymást.	A címszavakat asszociatív útvonalon is elérhetjük.
A szótári rész szócikkekből vagy utaló szócikkekből áll.	A szótári rész szócikkekből áll, az utaló szócikk szerepét a hiperlink veszi át.
A szócikkeket statikus ábrák, képek illusztrálják.	A szócikkeket interaktív, multimediális anyagok is illusztrálják.
Manuális keresésre van lehetőség.	Gépi keresésre van lehetőség.
Csak címszó szerint kereshető.	Többszavas kifejezések, toldalékolt alakok szerint is kereshető.
Több szótár együttes használata manuálisan lehetséges.	Több szótár együttes használata keresőfelület segítségével lehetséges.

A „hagyományos” elektronikus szótárak





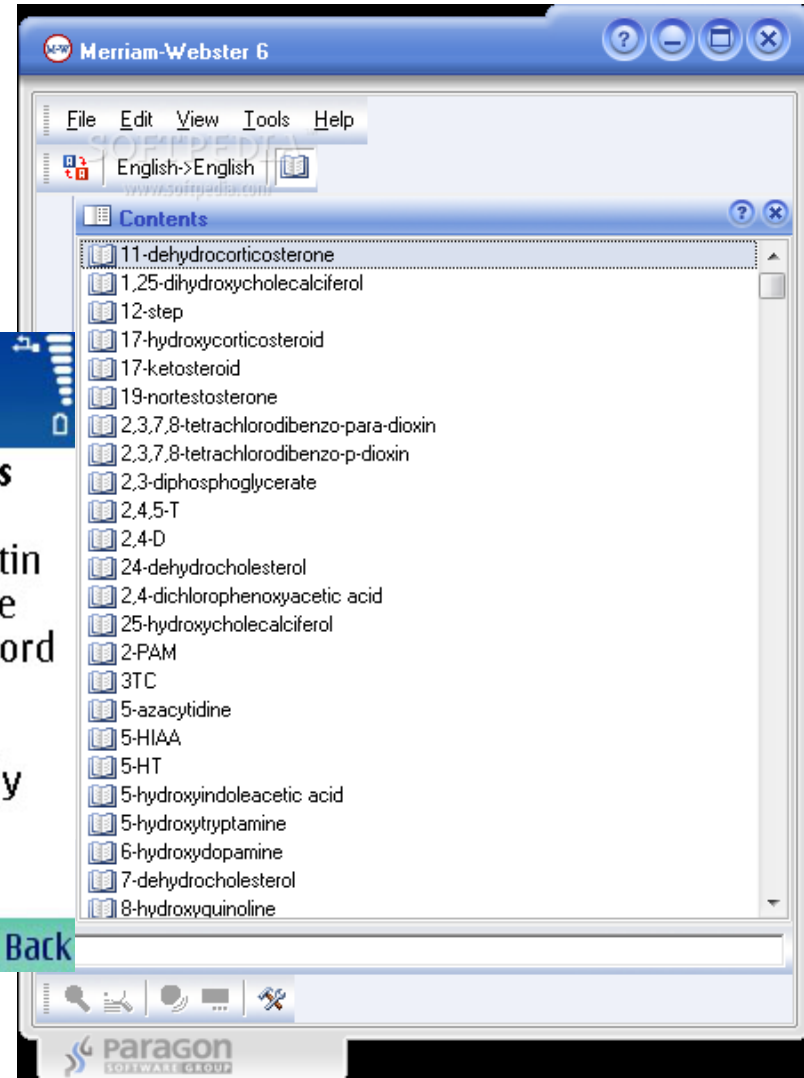
Mindenféle platformon



Merriam-Webster
En-En
Unabridged

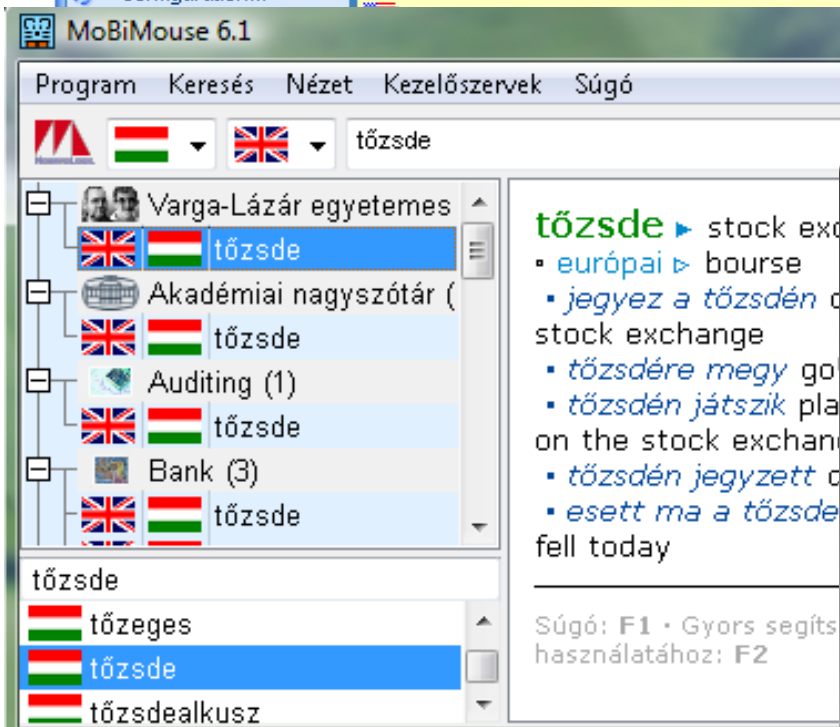
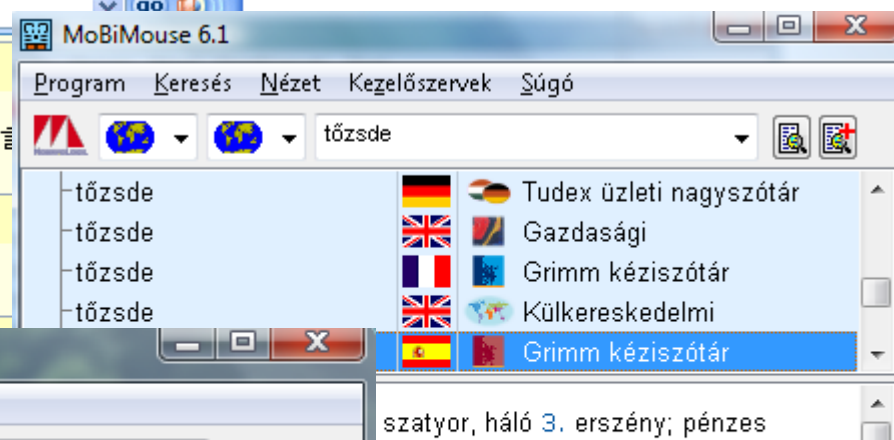
dic-tio-nary, **noun** -es
often attributive
Etymology: Medieval Latin
dictionarium, from Late
Latin **dictio-**, **dictio** word
+ Latin **-arium** -ary
1. : a reference book
containing words usually
alphabetically arranged
along with information
about their forms

Options Back

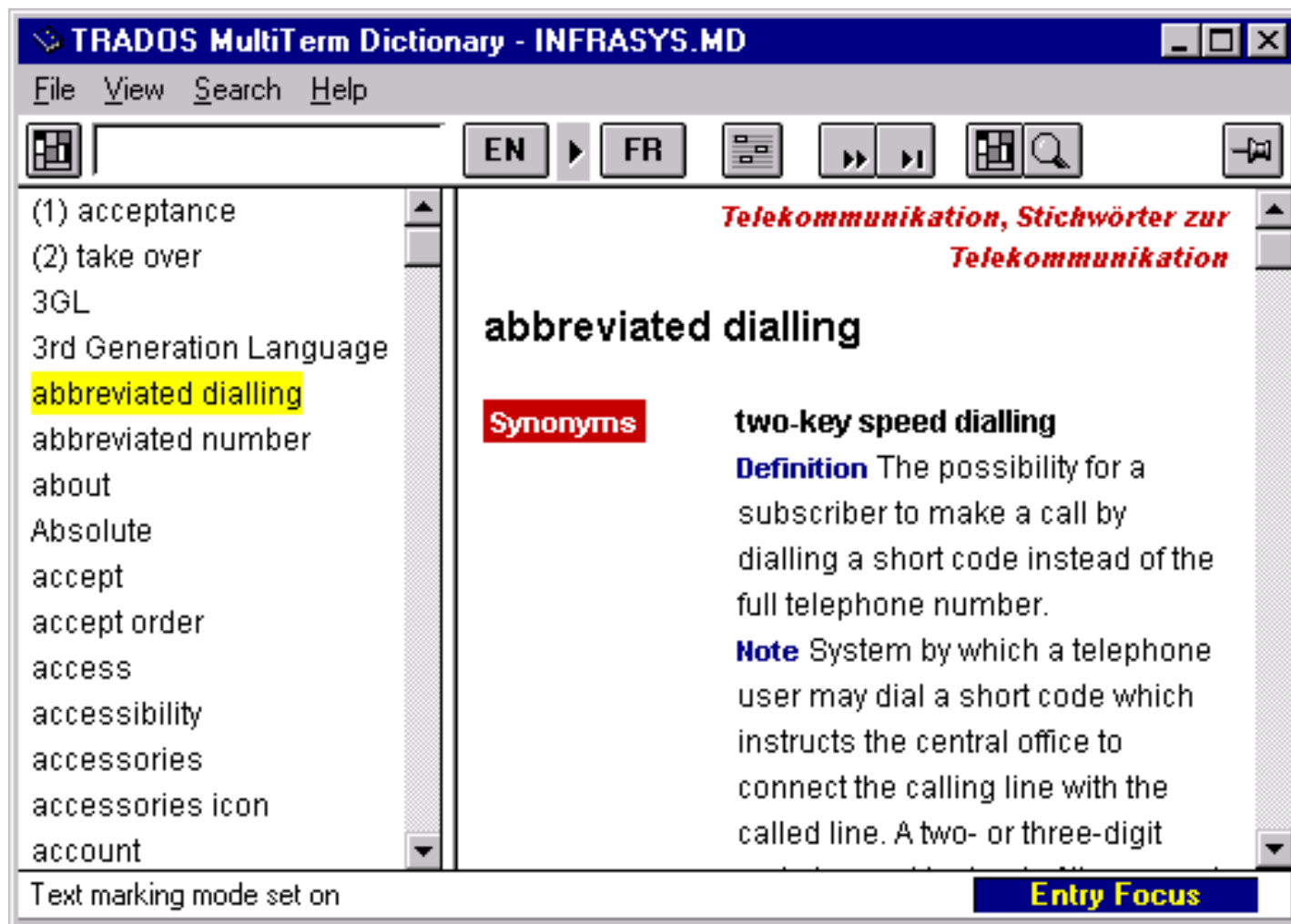




Elektronikus multiszótár



Terminológiakezelők (1)



Terminológiakezelők (2)

iate.europa.eu



InterActive Terminology
for Europe

magyar (hu) ▾

[keresési preferenciáim](#) [keresési preferenciáim törlése](#) [segítség](#)

keresési kritériumok

kifejezés
keresése

keres

forrásnyelv

bg - bolgár ▾

[beállítások](#)
[betöltése](#)

célnyelvek

☐ bg ☐ cs ☐ da ☐ de ☐ el ☐ en ☐ es ☐ et ☐ fi ☐ fr ☐ ga ☐ hu
☐ it ☐ la ☐ lt ☐ lv ☐ mt ☐ nl ☐ pl ☐ pt ☐ ro ☐ sk ☐ sl ☐ sv

☐ bármely
[törlés](#)

* Ez a szimbólum kötelezően megadandó mezőt jelöl.

további választható kritériumok

Válassza ki a keresett témakört!

bármely témakör ▾ ?

keresés típusa:

☐ kifejezés ☐ rövidítés ☒ mind

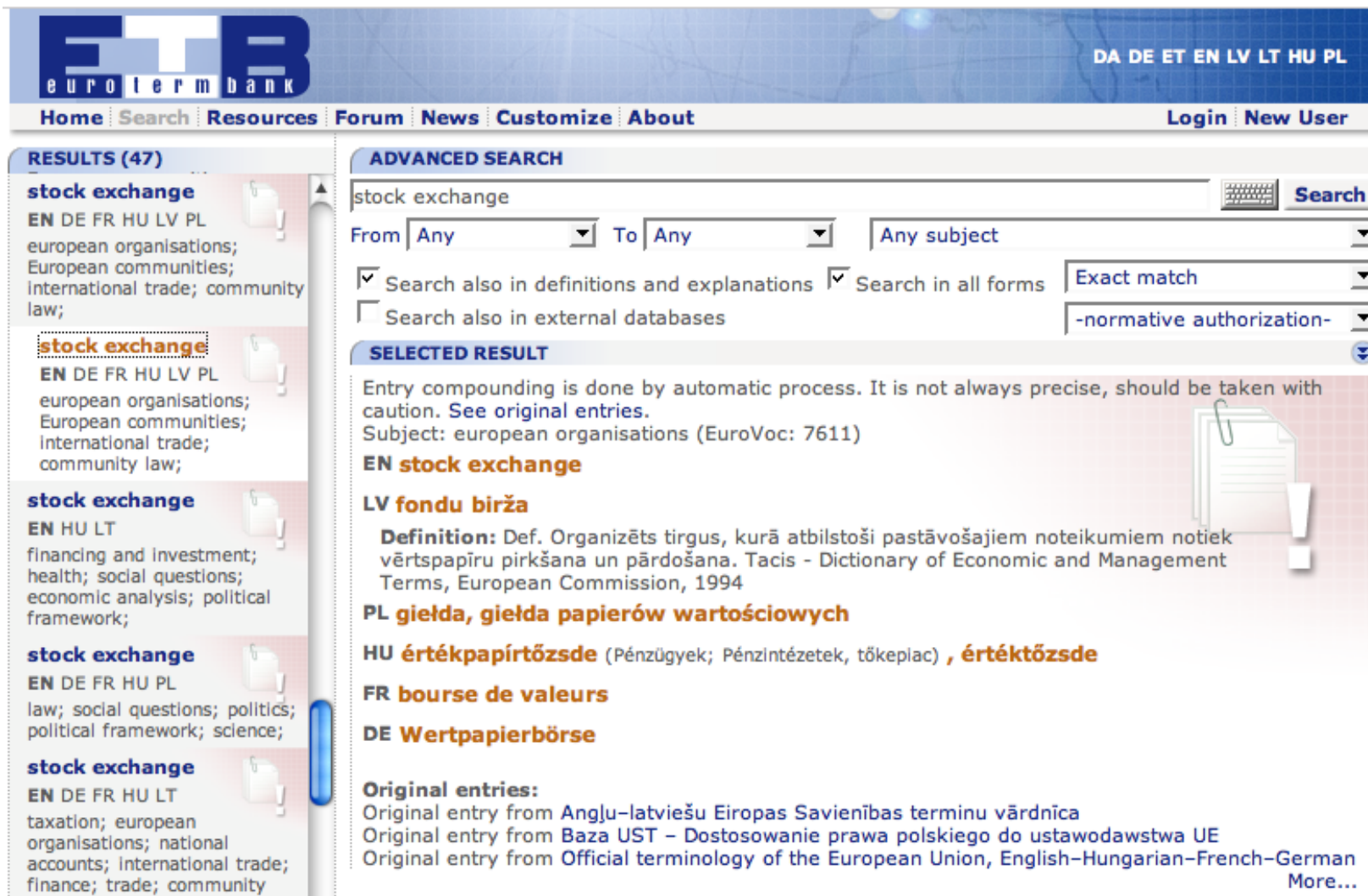
utolsó 10 keresés

választás az elmentett keresések közül ▾

iate diffusion version [1.3.35](#) © [Copyright Disclaimer](#) [About IATE](#) [Contact us](#)

Terminológiakezelők (3)

www.eurotermbank.com



The screenshot shows the EuroTermBank website. At the top, there is a header with the logo and navigation links: Home, Search, Resources, Forum, News, Customize, About, Login, and New User. The main content area is divided into two columns. The left column displays search results for the term 'stock exchange' in various languages (EN, DE, FR, HU, LV, PL). The right column shows the 'ADVANCED SEARCH' interface with a search bar and filters. Below the search bar, there are checkboxes for 'Search also in definitions and explanations' and 'Search also in external databases'. The 'SELECTED RESULT' section shows the definition of 'stock exchange' in English, with links to original entries in various languages.

RESULTS (47)

stock exchange
EN DE FR HU LV PL
european organisations;
European communities;
international trade; community
law;

stock exchange
EN DE FR HU LV PL
european organisations;
European communities;
international trade;
community law;

stock exchange
EN HU LT
financing and investment;
health; social questions;
economic analysis; political
framework;

stock exchange
EN DE FR HU PL
law; social questions; politics;
political framework; science;

stock exchange
EN DE FR HU LT
taxation; european
organisations; national
accounts; international trade;
finance; trade; community

ADVANCED SEARCH

stock exchange

From To Any subject

☒ Search also in definitions and explanations ☒ Search in all forms

☐ Search also in external databases

SELECTED RESULT

Entry compounding is done by automatic process. It is not always precise, should be taken with caution. See original entries.
Subject: european organisations (EuroVoc: 7611)

EN stock exchange

LV fondu birža
Definition: Def. Organizēts tirgus, kurā atbilstoši pastāvošajiem noteikumiem notiek vērtspapīru pirkšana un pārdošana. Tacis - Dictionary of Economic and Management Terms, European Commission, 1994

PL giełda, giełda papierów wartościowych

HU értékpapírtőzsde (Pénzügyek; Pénzüntézetek, tőkepiac) , **értéktőzsde**

FR bourse de valeurs

DE Wertpapierbörse

Original entries:
Original entry from Angļu-latviešu Eiropas Savienības terminu vārdnīca
Original entry from Baza UST – Dostosowanie prawa polskiego do ustawodawstwa UE
Original entry from Official terminology of the European Union, English–Hungarian–French–German
[More...](#)

A korszerű internetes szótárszolgáltatás kritériumai

- ☐ Folyamatosan bővíthető szótárkínálat
- ☐ Sajátszótár-készítési lehetőség
- ☐ Tetszőleges webes tartalom integrált megjelenítése
- ☐ A kifejezések intelligens kezelése
- ☐ Közösségi jelenlét
- ☐ Egymás segítségének és a (jogos) kritikának a fóruma
- ☐ A rendszer szemantikus ismereteinek erősítése a felhasználó keresési szokásainak elemzésével
- ☐ Könnyű keresés-indítási lehetőség
- ☐ Saját menthető beállítások a környezet személyre szabásához

Szótárszolgáltatás + közösségépítés



The screenshot displays a web-based Hungarian dictionary interface. At the top, there is a search bar with the text "ló". Below the search bar, there are two main panels. The left panel, titled "'ló': 14 találat", shows search results from two sources: "Közösségi szótár" and "Szinonimaszótár". The right panel, titled "'ló': 30 kifejezés (Közösségi szótár)", lists 30 related terms. Below these panels, there is a section titled "elérhető szótárak (9)" which allows users to select different dictionaries. At the bottom right, there are buttons for "súgó" (help) and "beállítások" (settings).

keres **minden nyelv** **professzionális elrendezés**

ló

'ló': 14 találat

- ☐   **Közösségi szótár**
- ló** (sakkban) [n] horse
- ló** (sakkfigura) [n] knight
- ló** (tornaszer) [n] horse
- ló** (ügető) [n] trotter
- ló** [n] horse
- ló** [n] pommel horse
- ☐   **Szinonimaszótár**
- csikó** n **ló**
- gebe** n **ló**
- kanca** n **ló**
- ló** n paci, lovacska
- ló** n huszár
- mén** n mén**ló**, csöddör, **ló**
- paripa** n háta**ló**, csata**ló**, herélt, herélt **ló**, **ló**
- táltos** n **ló**

'ló': 30 kifejezés (Közösségi szótár)

- aiándék ló
- állatorvosi ló
- amerikai félvad ló
- döglött ló
- elszabadult ló
- elszökött ló
- herélt ló
- herélt ló
- hóka ló
- kehes ló
- kivénhedt ló

ezt keresik a többiek

- kollégám
- megvizsgálni
- akkor
- escalating
- jegyes

utóbbi ismeretlen szavak

- butones
- pançetta
- MOČÁR
- tőke-törlesztés

elérhető szótárak (9)

szótárak csoportos ki-be kapcsolása

☒   **Közösségi szótár**

  **Angol-magyar közösségi szótár**

súgó **beállítások**



A szótári keresés

Nyelvfüggő szótárproblémák

- ❑ A forrás- és célnyelv karakterkészleteinek ismerete
- ❑ A forrás- és a célnyelv ábécérendjeinek ismerete
- ❑ A fonetikai információ kezelése
- ❑ Egységes jelölés: nyelvi keresésnél a szótár grammatikai információival való kompatibilitás



Keresési technikák elektronikus szótárakban

- ☐ Betű szerint
- ☐ Csonkolt keresés
- ☐ Hasonlósági keresés (fuzzy, soundex, 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



Többszavas kifejezések keresési módjai

- ☐ Csak címszóként
- ☐ Betű szerint
- ☐ Teljes szövegű kereséssel
- ☐ Reguláris kifejezésként
- ☐ Tőindexekkel:
készítéskor vagy elemzési időben (is)

Többszavas kifejezések keresése

MoBiMouse 6.1

Program Keresés Nézet Kezelőszervek Súgó

head tail

I'll toss you for it - heads or tails?		
there's just one cake left, I'll toss y...		
heads I win tails you lose		
heads or tails		
heads or tails		
heads you win tails I lose		
I can make neither head nor tail of		

head

	HE
	He
	he
	he
	head
	head back
	head case
	head clerk
	head cold

reidont vkt; → **ear* 1.1.**; **be hanging over sy's head** fenyeget vkt [*pl. veszély*]; **give orders over sy's head** vk tudta/beleegyezése nélkül (v. vknek a megkerülésével) ad utasításokat/parancsokat [*személyzetnek stb.*]; **be promoted over sy's head** pretereál vkt, vkt átugrik a számléltrán; **sell a house over sy's head** vk tudta nélkül (v. háta mögött) eladja a házat; **be without a roof to one's head** hajléktalan b) (ábrázolt) fej [*pl. műalkotás, érmén*]; arckép, mellszobor; **heads or tails** fej vagy írás; **toss heads or tails** fej vagy írást játszik; **I can't make**

MorphoMouse

heads tails

can't make head or tail of smth nem ért vmit; nem tud semmit se kihozni vmiből

heads I win, tails you lose ha fej, akkor én nyertem, ha írás, te vesztettél

heads or tails fej vagy írás

can't make head or tail of smth nem ért vmiből; nem igazodik el vmin

I'll toss you for it – heads or tails? feldobok egy pénzt, hogy eldöntsük – fej vagy írás?

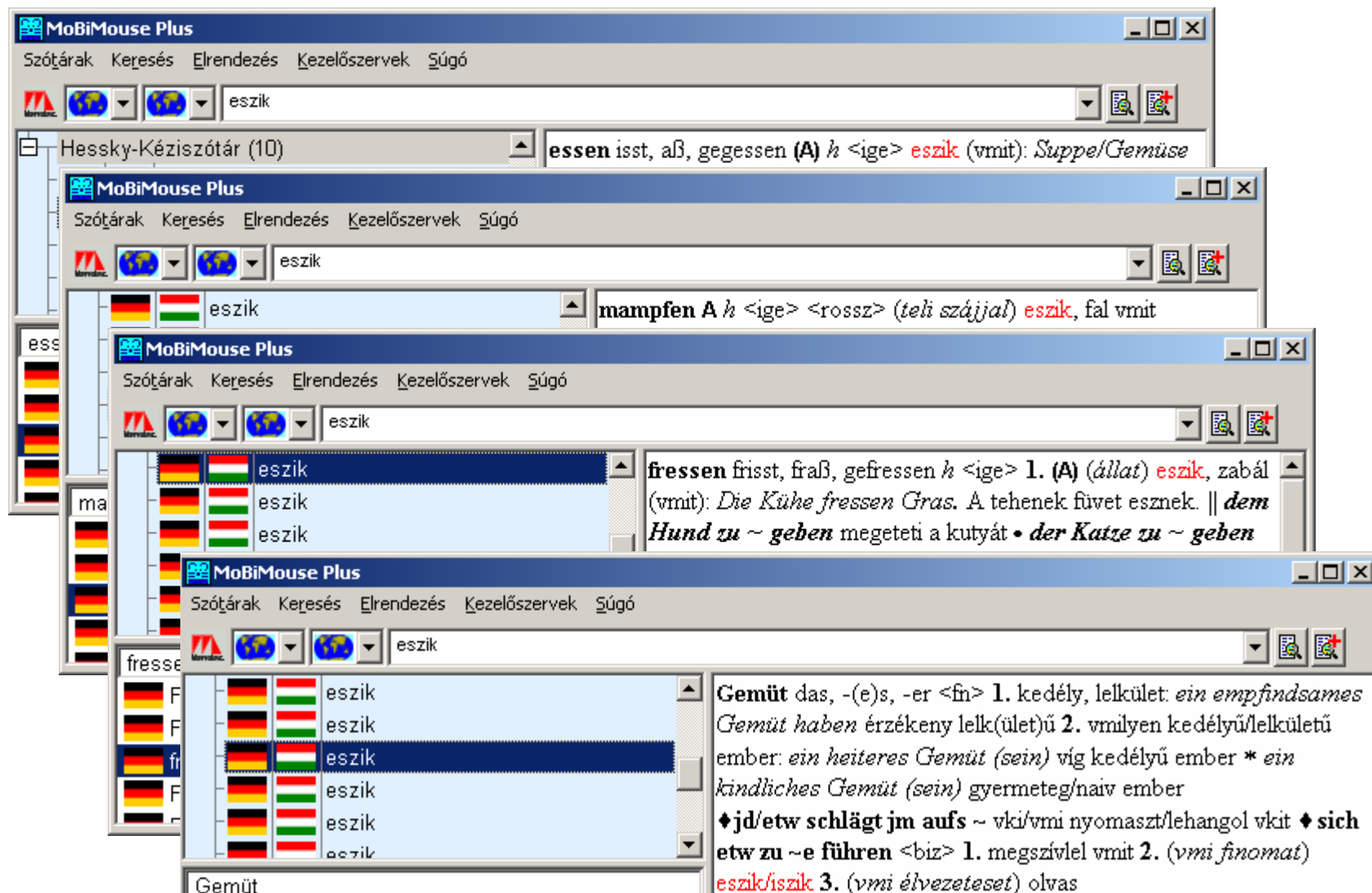
heads I win, tails you lose ha fej, én nyertem, ha írás, te vesztettél

there's just one cake left, I'll toss you for it – heads or tails? csak egy sütemény maradt, feldobok egy pénzt – fej vagy írás?

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 „kifordítható” szótár



MoBiMouse Plus
Szótárak Keresés Elrendezés Kezelőszervek Súgó

eszik

Hessky-Kéziszótár (10) **essen** isst, aß, gegessen (A) *h* <ige> **eszik** (vmit): *Suppe/Gemüse*

MoBiMouse Plus
Szótárak Keresés Elrendezés Kezelőszervek Súgó

eszik

MoBiMouse Plus
Szótárak Keresés Elrendezés Kezelőszervek Súgó

eszik

MoBiMouse Plus
Szótárak Keresés Elrendezés Kezelőszervek Súgó

eszik

fressen frisst, fraß, gefressen *h* <ige> 1. (A) (állat) **eszik**, zabál (vmit): *Die Kühe fressen Gras. A tehenek füvet esznek. || dem Hund zu ~ geben* megeteti a kutyát • *der Katze zu ~ geben*

Gemüt das, -(e)s, -er <fn> 1. kedély, lelkület: *ein empfindsames Gemüt haben* érzékeny lelk(ület)ű 2. vmilyen kedélyű/lelkületű ember: *ein heiteres Gemüt (sein)* vig kedélyű ember • *ein kindliches Gemüt (sein)* gyermeteg/naiv ember ♦ *jd/etw schlägt jm aufs ~* vki/vmi nyomaszt/lehangol vkit ♦ *sich etw zu ~e führen* <biz> 1. megszával vmit 2. (vmi finomat) **eszik/iszik** 3. (vmi élvezetese) olvas



Szótármegjelenítés

LMF

- ☐ Lexical Markup Framework (LMF):
ISO-szabvány (ISO-24613:2008) szótárak leírásához
- ☐ 1. lépés: létező szótárak struktúráinak konzisztens feltérképezése
- ☐ 2. lépés: az összes feldolgozott szótárat lefedő leírás létrehozása
- ☐ 3. lépés: 61 szakértő bevonásával az összes szóba jövő szótárszerkezet megvizsgálása
- ☐ Fontos szempont volt a morfológiai támogatás biztosítása
- ☐ A fejlesztés eredménye (2004-2008): koherens UML-modell



Szótármegjelenítés

Ma már szinte csak XML + XSLT alapú megoldások

kar ☒ kifejezésben

bármilyen nyelv súgó >> szótárak >> visszajelzés >> keres >>

<

Kétféle szótármegjelenítés

(Atkins 2002)

2. It's the economy, stupid

It was a debate about foreign policy, an important subject that plays second fiddle to the No. 1 issue on voter's minds this election: the economy -- OK, in addition to issues such as health and Security.

Funniest tweets from the

There was substantive foreign policy during the the economy received

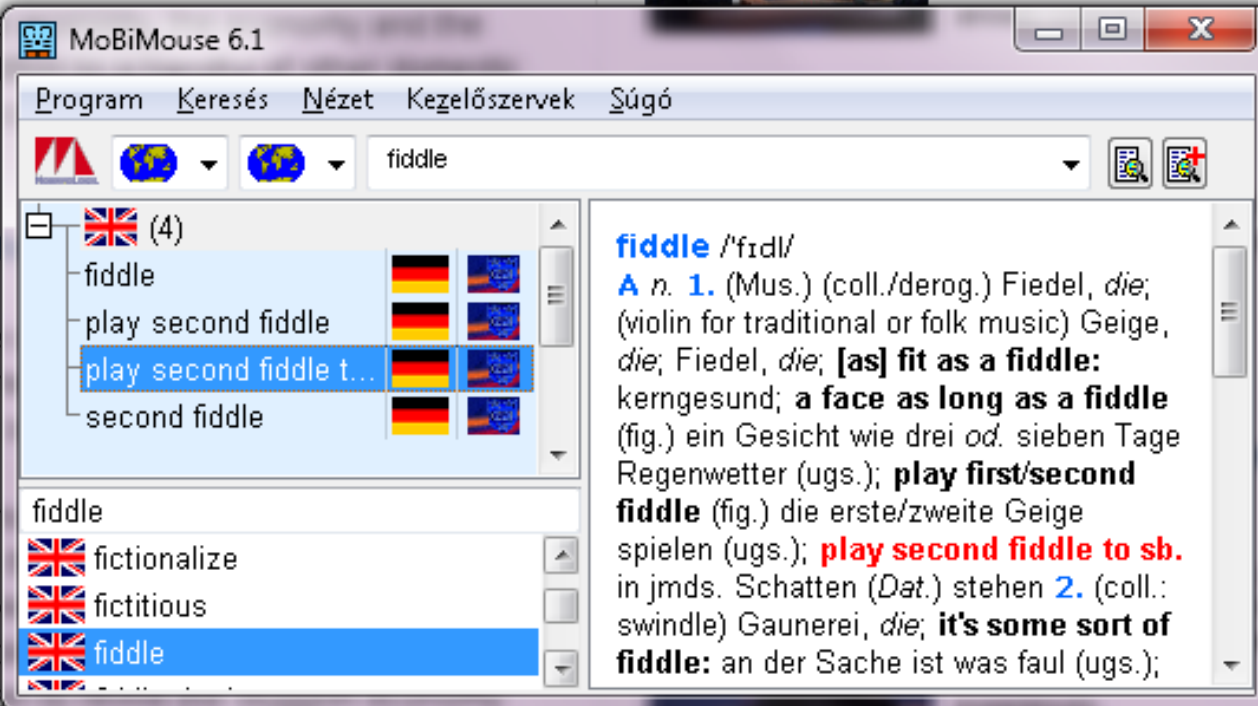
Romney tried to convince national security issues standing in the world. A opportunity, the Republican present his five-point plan

punches

updated 7:48 AM EDT, Tue October 23, 2012



President Barack Obama headed into the final debate of the 2012 campaign with the biggest advantage of all: he's



MoBiMouse 6.1

Program Keresés Nézet Kezelőszervek Súgó

fiddle

(4)

- fiddle
- play second fiddle
- play second fiddle t...
- second fiddle

fiddle

- fictionalize
- fictitious
- fiddle

fiddle /'fɪdl/

A n. **1.** (Mus.) (coll./derog.) Fiedel, *die*; (violin for traditional or folk music) Geige, *die*; Fiedel, *die*; **[as] fit as a fiddle:** kerngesund; **a face as long as a fiddle** (fig.) ein Gesicht wie drei od. sieben Tage Regenwetter (ugs.); **play first/second fiddle** (fig.) die erste/zweite Geige spielen (ugs.); **play second fiddle to sb.** in jmds. Schatten (*Dat.*) stehen **2.** (coll.: swindle) Gaunerei, *die*; **it's some sort of fiddle:** an der Sache ist was faul (ugs.);



Szótárkészítés



Szótárszerkesztő eszközök (LWB)

The screenshot shows the LWB web application interface. The top navigation bar includes the Morphologic logo and various tool icons like save, undo, delete, new, sort, prev, next, locks, keyboard, autolayout, users, upload, export, password, and log out. The main content area is divided into three sections:

- Szótár [!!! László Varga demo !!!]**: A sidebar on the left containing a search bar and a list of words. The word **abandon** is selected.
- Szócikk**: The central area showing the detailed entry for **abandon**. It includes fields for the word form (abandon), its IPA transcription (/əˈbændən/), and its meaning (elhagy). The entry is structured into sections: **szócikk**, **szócikkfej**, **homöindex**, **címszavak**, **változatok**, **címszó**, **kiejtések**, **IPA**, **szófaj**, **szócikktest**, **blokk**, **jelentés**, **ekvivalensek**, **ekvi.blokk**, **ekvivalens**, **példák**, **példa**, **pl. angolul**, **pl. fordításai**, **ekvi**.
- 2abandon**: The rightmost section showing the definition of the word: **2abandon** /əˈbændən/ ige 1. elhagy 2. felad.; letesz vmiről || **abandon ship** elhagyja a süllyedő hajót 3. lemond; átenged.



Automatizált szótárkészítés korpuszelemzéssel

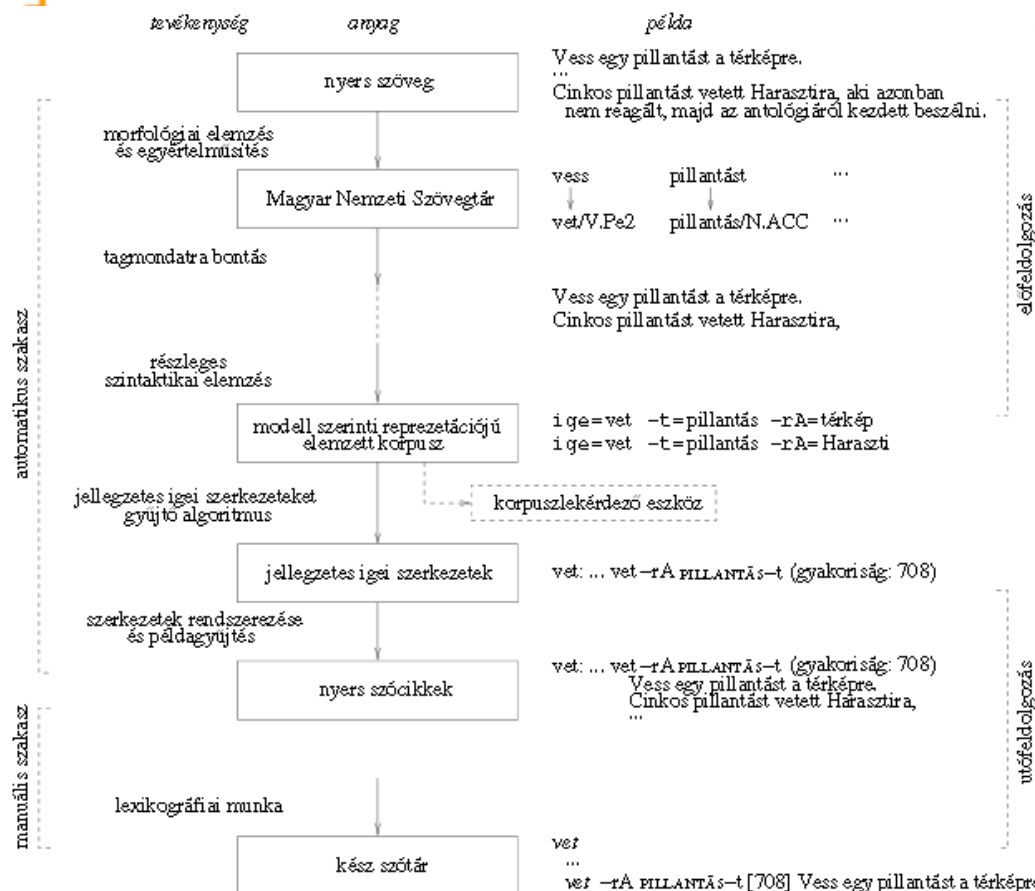
Szótárszerkesztő eszközök (Mazsola)

8674 találat. hang [477] lehetőség [412] válasz [301] otthon [281] tájékoztatás [170] hír [199] ok [188] áldás [85]
esély [114] lendület [74] hely [150] pénz [129] magyarázat [87] szám [128] alkalom [101] tanács [85] koncert [71] munka [116]
utasítás [61] felvilágosítás [52] hangverseny [49] támogatás [93] kép [85] igaz [82] lökés [41] százalék [92] segítség [70]
engedély [59] megbízás [59] forint [86] információ [59] felhatalmazás [39] voks [38] ízeltő [25] alap [89] mód [59] jel [49] nyomaték [27] tér [46]
7649 találat. igaz [20] válasz [275] tanács [172] hang [149] tipp [67] ami [212] lehetőség [155] pénz [151] magyarázat [88]
hála [67] ok [116] név [109] esély [83] ötlet [68] kegyelem [49] mi [119] hi
felmentés [32] több [79] valami [59] fegyver [41] áldás [28] az [136] maga [62] al
parancs [28] megbízás [28] puszi [18] erő [40] hír [35] engedély [33] hit [38] kedve

Mazsola vonzatkereső
(Sass 2007)

Igei szerkezetek
gyakorisági szótára
(Sass 2011)

Prószéky Gábor



Hasonló fogalmak gyűjtése korpuszból, szóbeágyazással

0	PPKE	1	970
1	KRE	0.8159	629
2	BCE	0.8029	1052
3	ÁJK	0.7895	895
4	BMGE	0.7840	276
5	PTE	0.7788	7721
6	JPTE	0.7733	366
7	ÁJTK	0.7728	331
8	BMF	0.7719	762
9	NYME	0.7712	966
10	SZIE	0.7683	2574

0	bionika	1	196
1	sejtbiológia	0.7714	298
2	biofizika	0.7539	501
3	biokémia	0.7464	1993
4	genomika	0.7420	714
5	bioinformatika	0.7371	420
6	rendszerbiológia	0.7292	113
7	mikrobiológia	0.7286	859
8	filogenetika	0.7264	41
9	gyógyszerkémia	0.7132	64
10	nanotechnika	0.7073	166

Szemantikus jegyek automatikus kinyerése korpuszból

"bionika[FN]"	"PPKE[FN]"	"Pázmány[FN]"	"Dzsudzsák[FN]"
bionika	PPKE	Pázmány	Dzsudzsák
atom	educate	@Catholic_Church	team
chemistry	institution	holy	sport
science	science	christian	city_2
method	group_5	book	two_3
knowledge	study	religion	INSTRUMENT_8
HAS_31	student	priest	ball
bionika[FN]	PPKE[FN]	Pázmány[FN]	Dzsudzsák[FN]
CHEMISTRY_22	COLLEGE_12	CHRISTIANITY_5	SPORT_18
BIOLOGY_35	COLLEGE_13	CHRISTIANITY_4	FOOTBALL_5
EDUCATION_10	EDUCATION_9	CHRISTIANITY_16	SPORT_10
MICROBES,_GENETICS,_BIOCHEM_9	COLLEGE_8	CHRISTIANITY_23	THEATRE_12
PHYSICS_1	COLLEGE_11	CHRISTIANITY_7	SPORT_14
HARD_SCIENCE_1	EDUCATION_13	CHRISTIANITY_2	SPORT_7



Kapcsolat más nyelvészeti megoldásokkal



Az „ablakos” kommunikáció nehézségei

- ☐ Kilépni az adott alkalmazásból
- ☐ Elindítani
- ☐ Kinyitni vagy felnagyítani
- ☐ Beírni
- ☐ Klikkelni
- ☐ Átmozgatni
- ☐ Lekicsinyíteni vagy bezárni
- ☐ Visszalépni az eredeti alkalmazásba

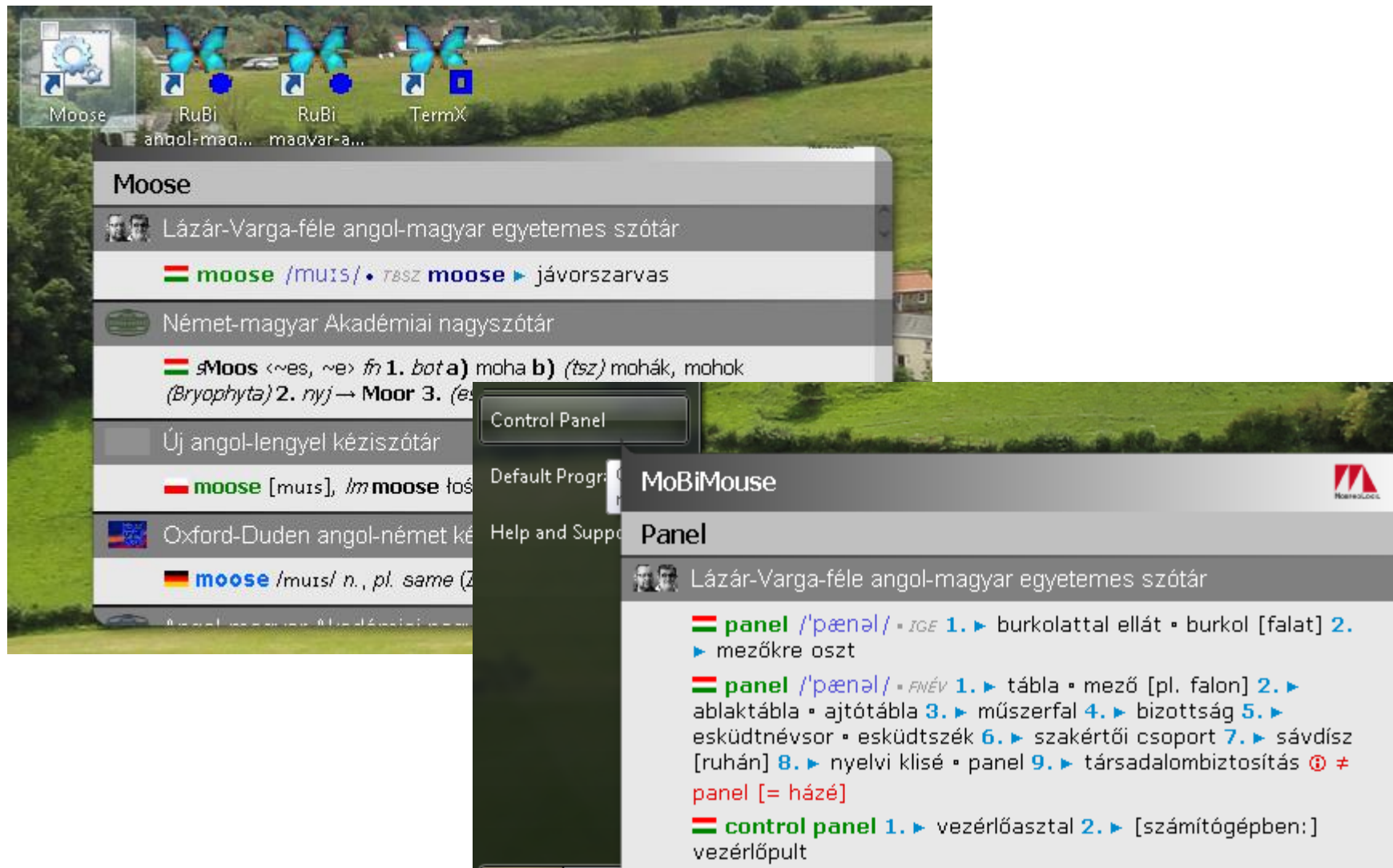
A gyorsfordítók megjelenése

- ❑ 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 „rávetítő” megoldás nyelvtechnológiai elemei

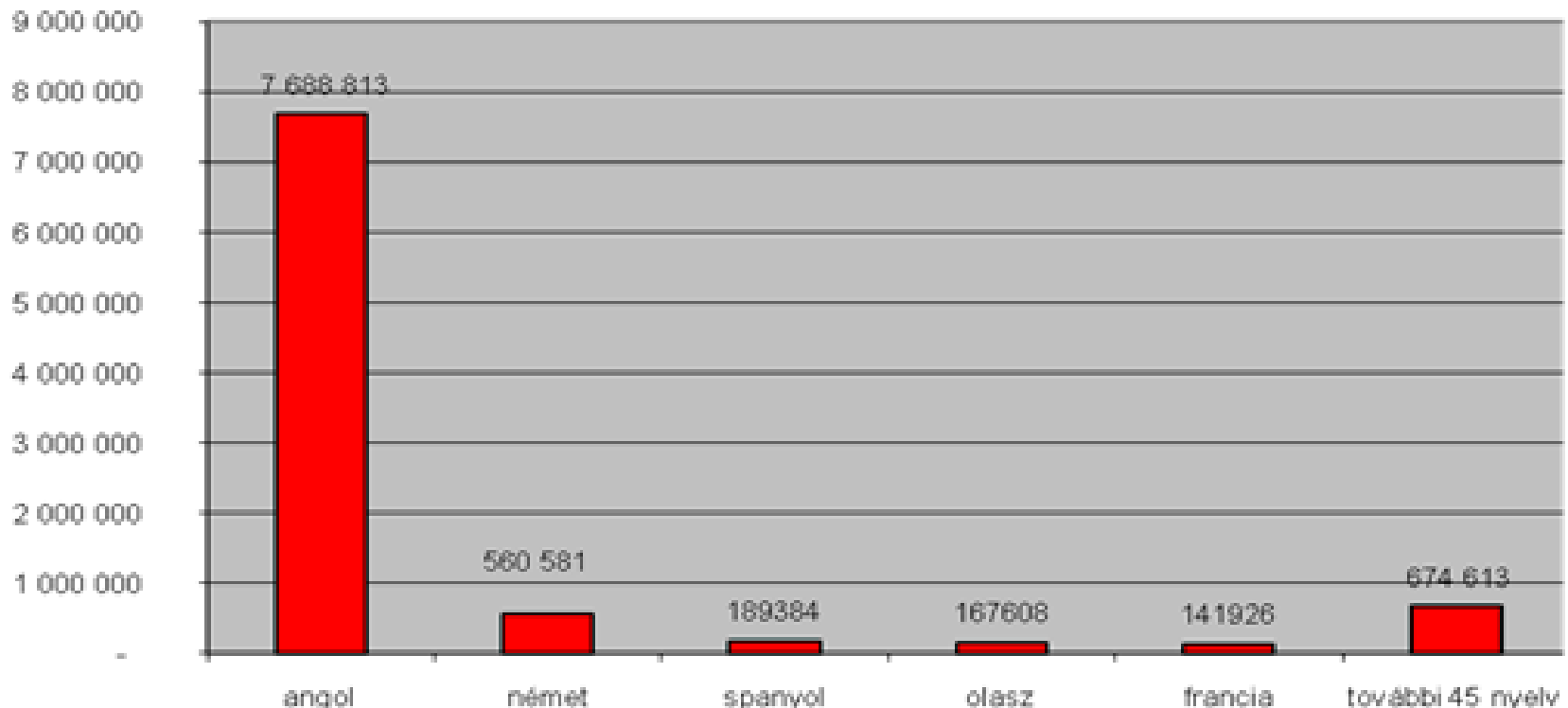
- ☐ Szöveg(rész)-felismerés
- ☐ Nyelvi elemzés: morfológia, lemmák, szókapcsolatok (esetleg környezetelemzés)
- ☐ Szótári keresés: tövesítve vagy csak literálisan
- ☐ Megjelenítés: buborékban vagy fix ablakban
- ☐ Log-fájlok segítségével az automatikus információgyűjtés lehetősége (nemcsak rávetítő szótáraknál!)
- ☐ Megjelennek a felhasználók valódi javaslataival feljavított szótárak

A „rávetítés” a gyakorlatban



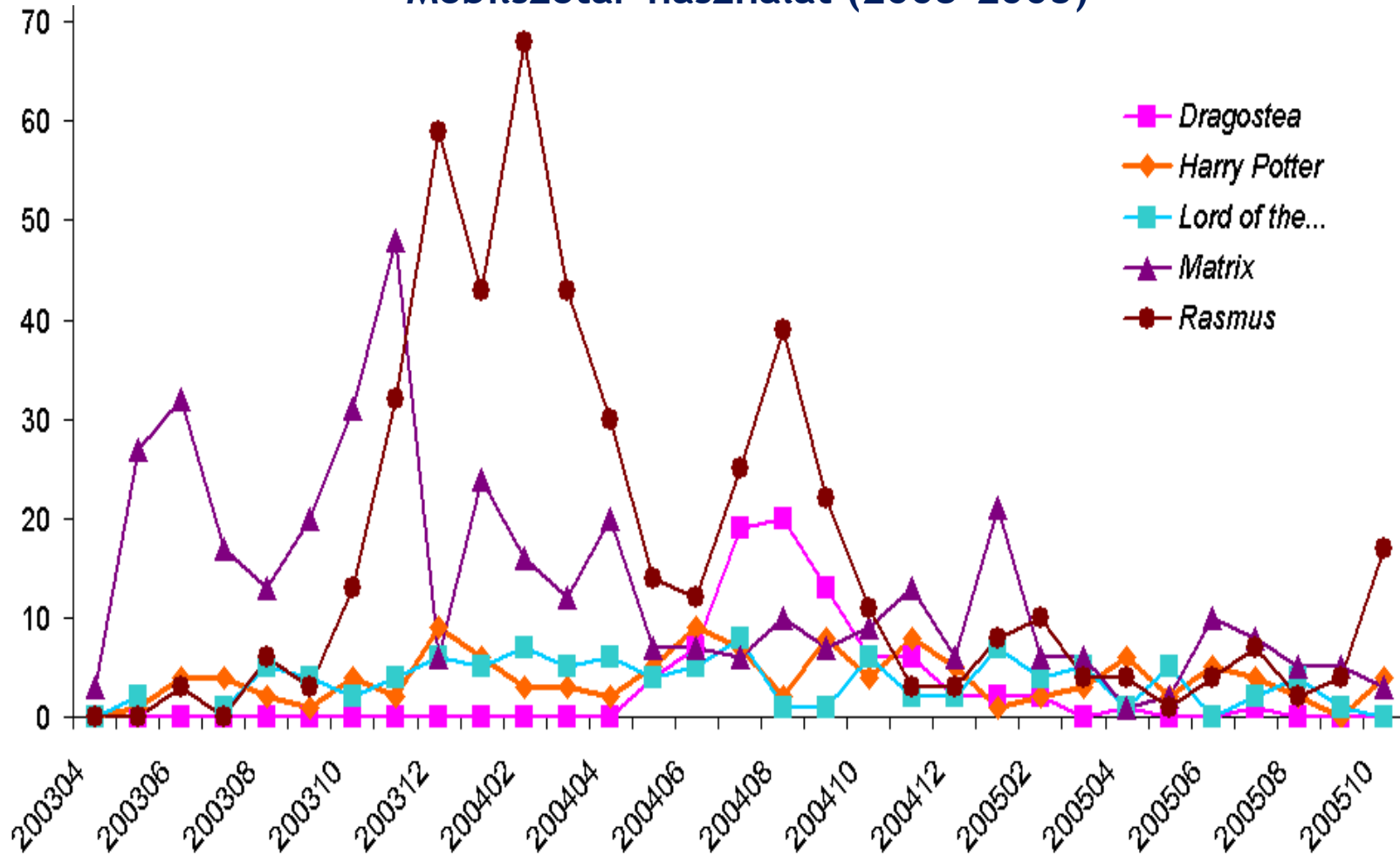
A log-fájlok szociolingvisztikailag elemezhetőek

Szótárlekérdezések a weben



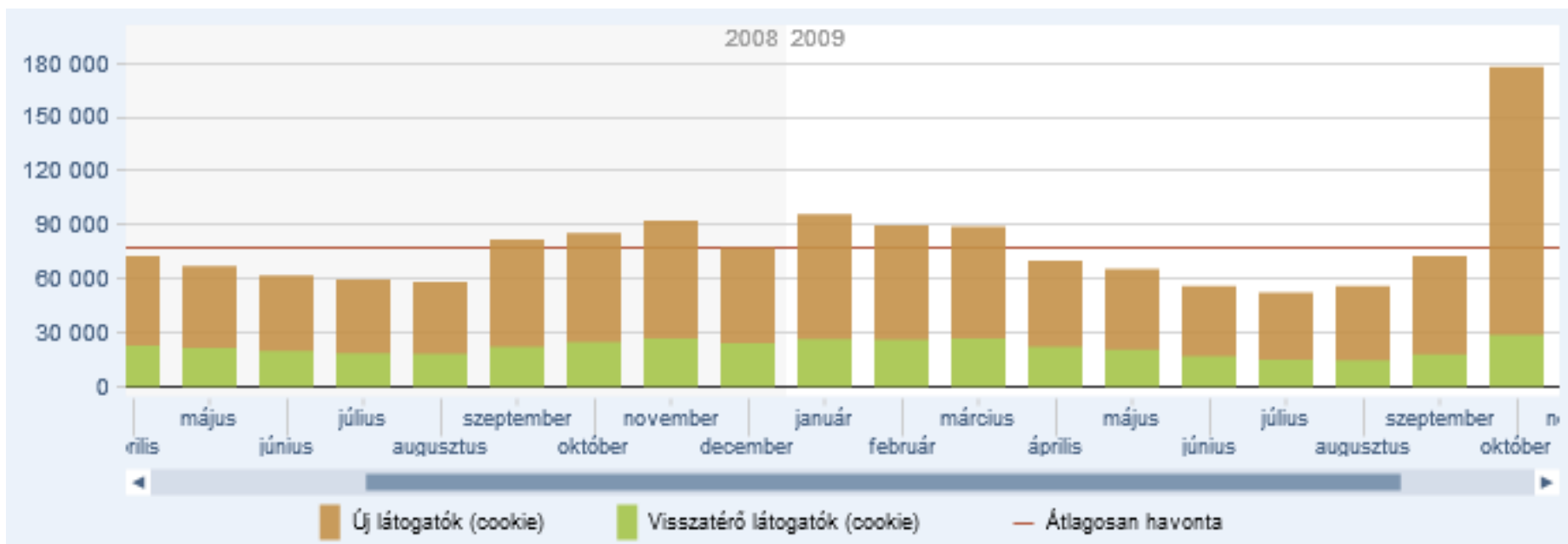
A log-fájlok szociolingvisztikailag elemezhetőek

Mobilszótár-használat (2003-2005)



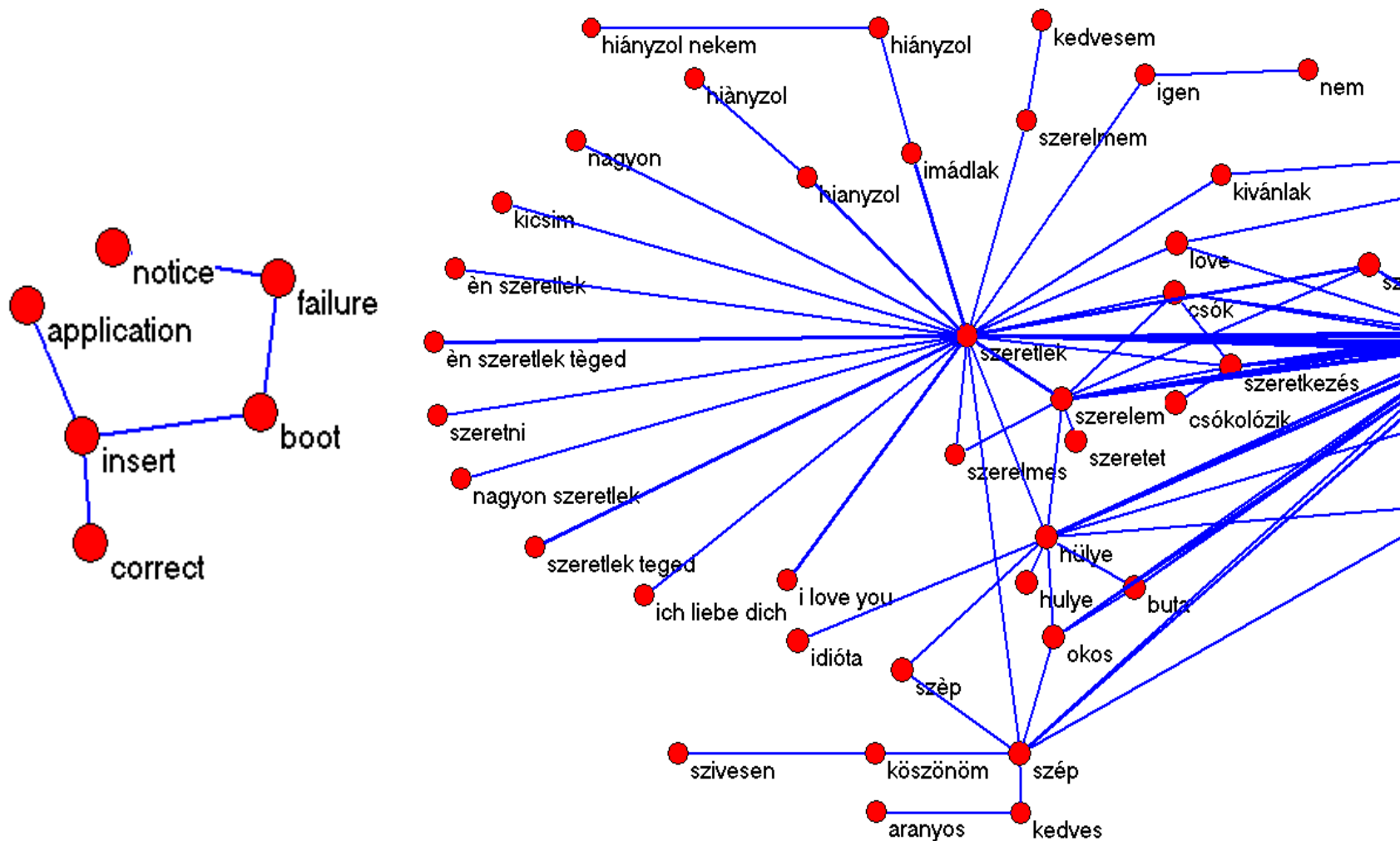
A log-fájlok szociolingvisztikai elemezhetőek

„A szótárazás tanítása”

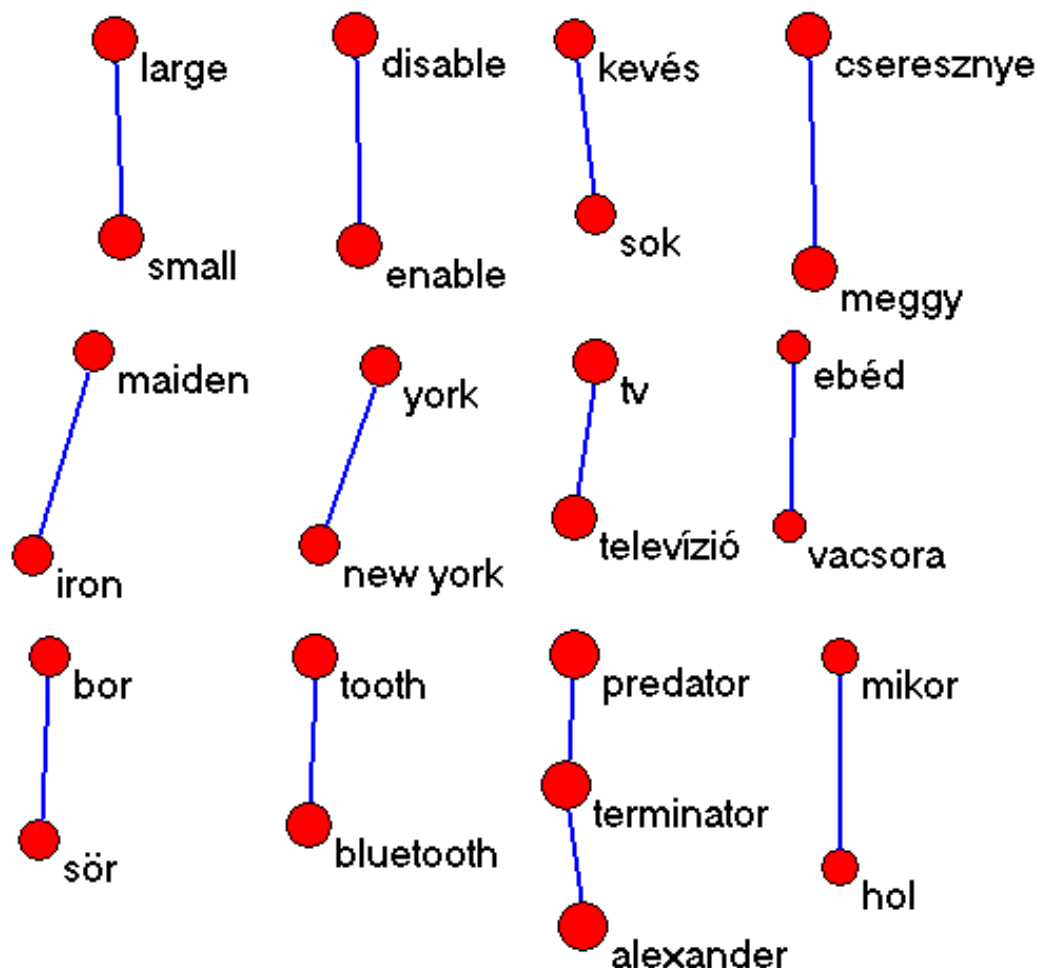


A szótárlekérdezés koincidencia-részgráfjairól

„Aki ezt kérdezte, az még mit kérdezett?”



A szótárlekérdezés koincidencia-részgráfjairól (2)





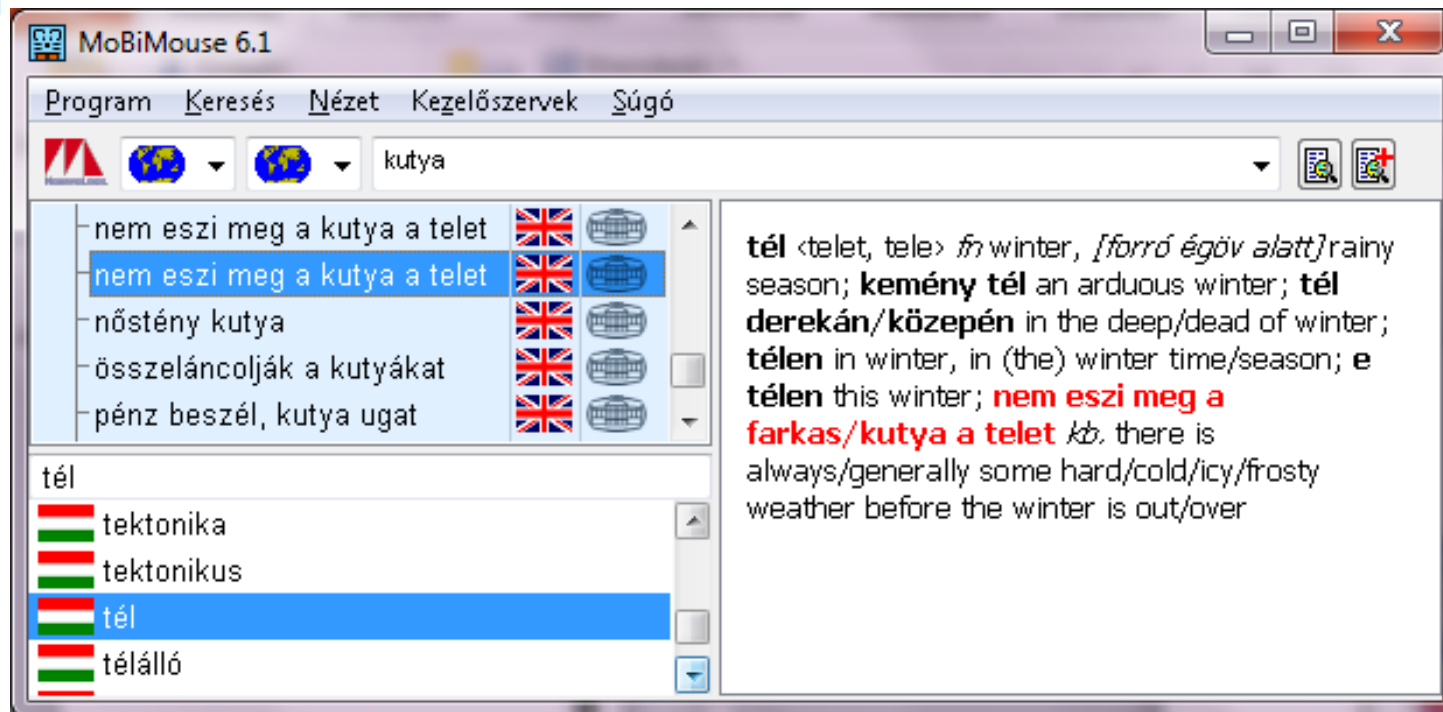
A hagyományos és az „intelligens” szótárak viszonyáról



Az „intelligens” szótárak készítésének problémái

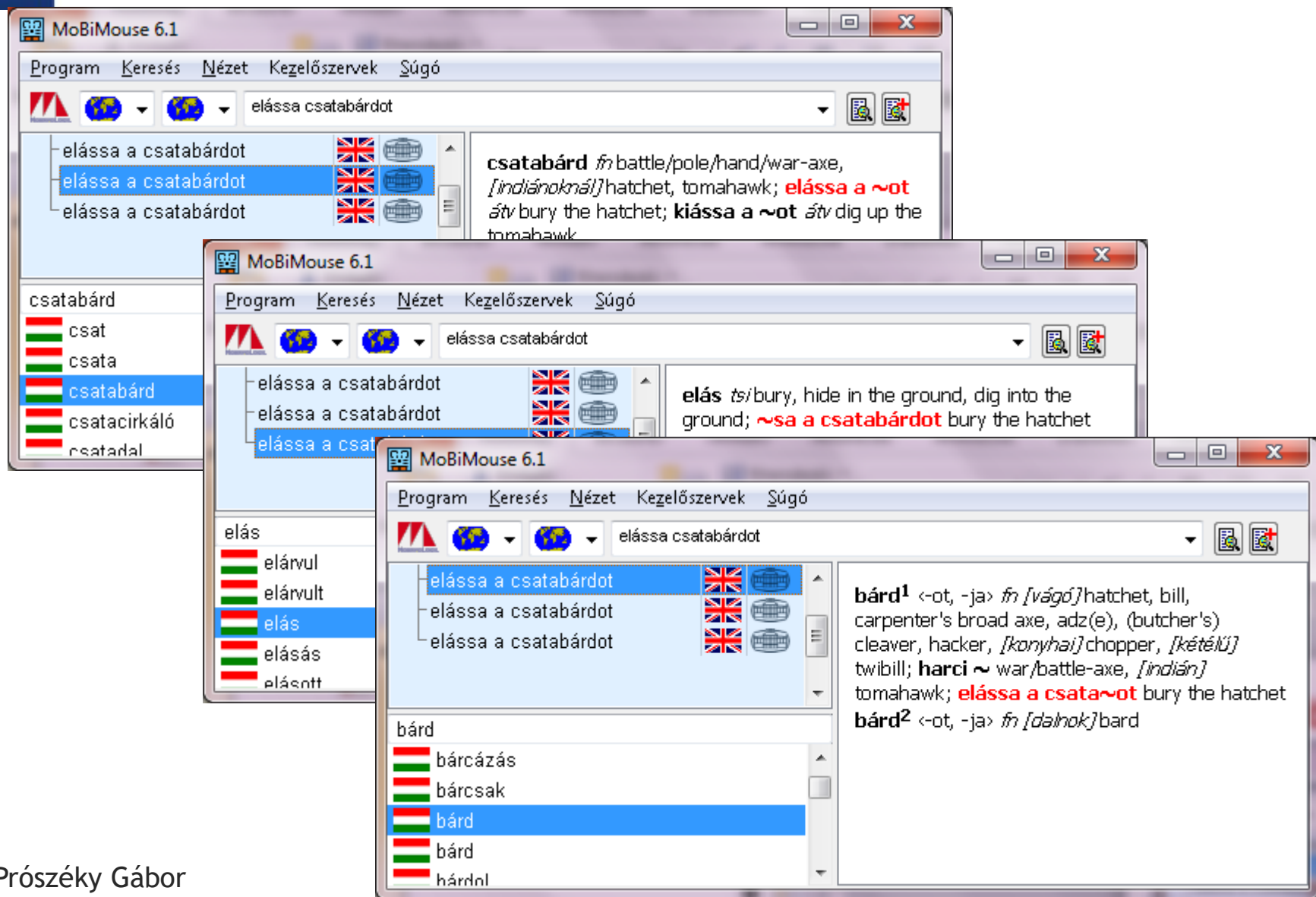
- ❑ A legfőbb baj: a szótárforrások XML-változatainak „amatőr” vagy legalábbis nyomtatás-centrikus megoldásai
- ❑ A második ok: a szótár az embereknek, nem a gépnek készül
- ❑ Egy sor technikai probléma, ami a szótárak „papírszótár” mivoltából ered, ám a gépi változatban át kell ezeket alakítani

A perjel-probléma



- ❑ A szótárak nem a gépnek készülnek (pl. perjelek):
nem szavazó/szavazásra nem jogosító részvény
- ❑ Tudhatja-e egy program összefogni az egybetartozó kifejezést?
{arcára fagy/ráfagy az arcára/lefagy az arcáról} a mosoly

A többszörös előfordulások problémája



The image shows three overlapping windows of the MoBiMouse 6.1 software, demonstrating how the same word can have multiple meanings and grammatical forms, leading to ambiguity in translation.

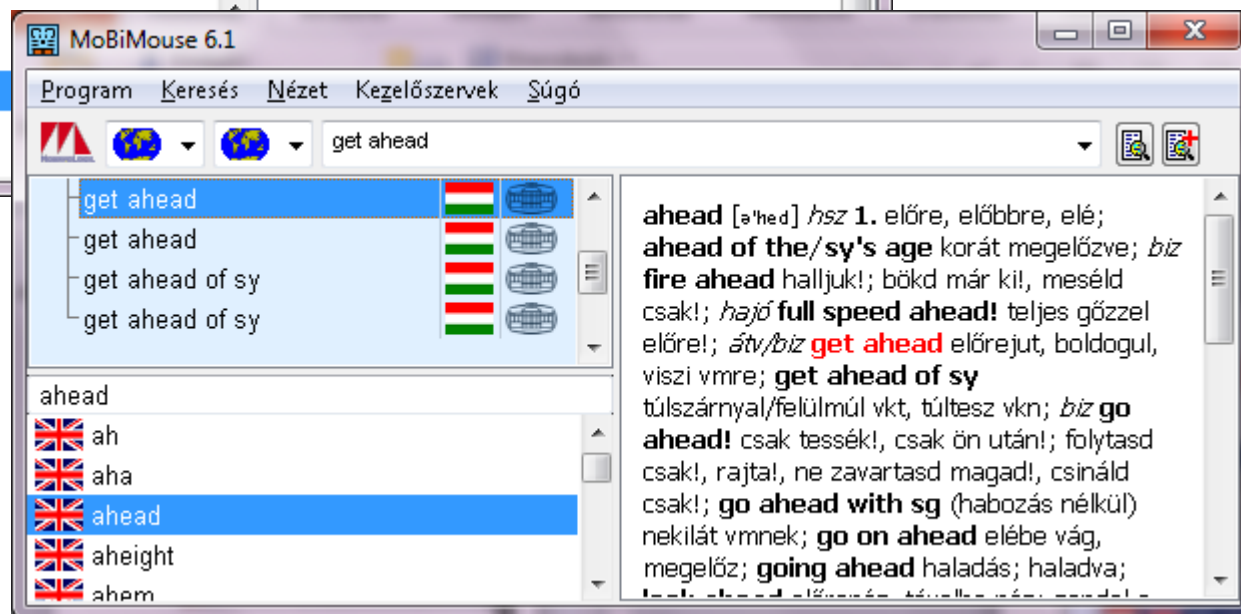
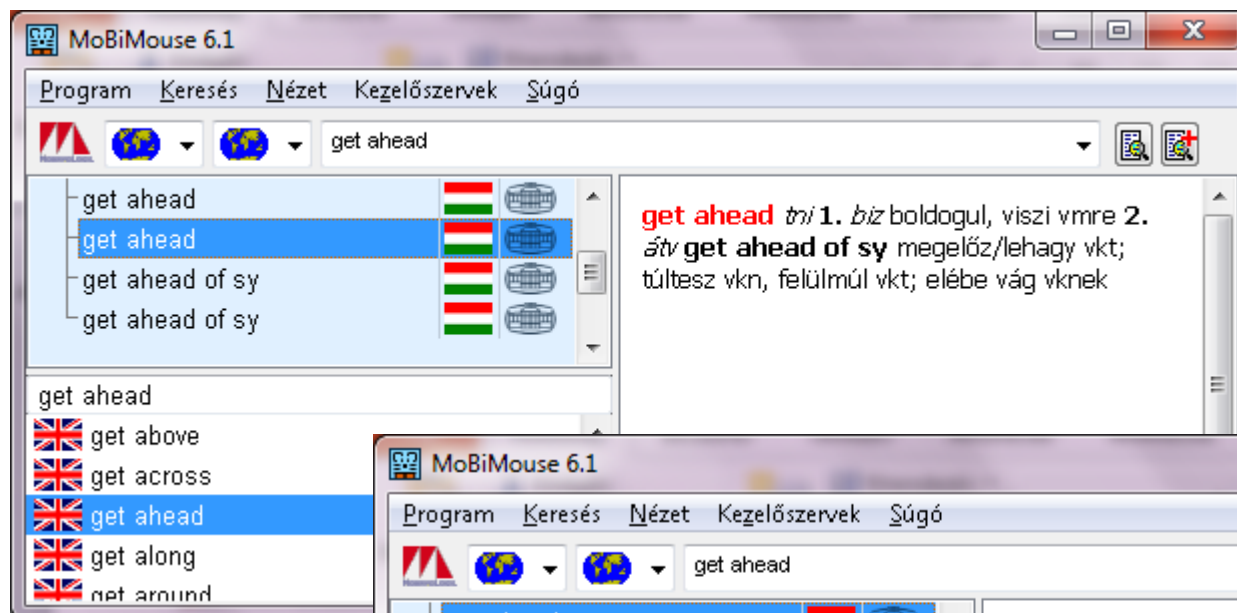
Top Window: The search term is "elássa csatabárdot". The results list three occurrences of "elássa a csatabárdot" with different flags (UK, UK, UK). The definition on the right is: **csatabárd** *fn* battle/pole/hand/war-axe, [indianoknál] hatchet, tomahawk; **elássa a ~ot** *átv* bury the hatchet; **kiássa a ~ot** *átv* dig up the tomahawk.

Middle Window: The search term is "elássa csatabárdot". The results list three occurrences of "elássa a csatabárdot" with different flags (UK, UK, UK). The definition on the right is: **elás** *tsi* bury, hide in the ground, dig into the ground; **~sa a csatabárdot** bury the hatchet.

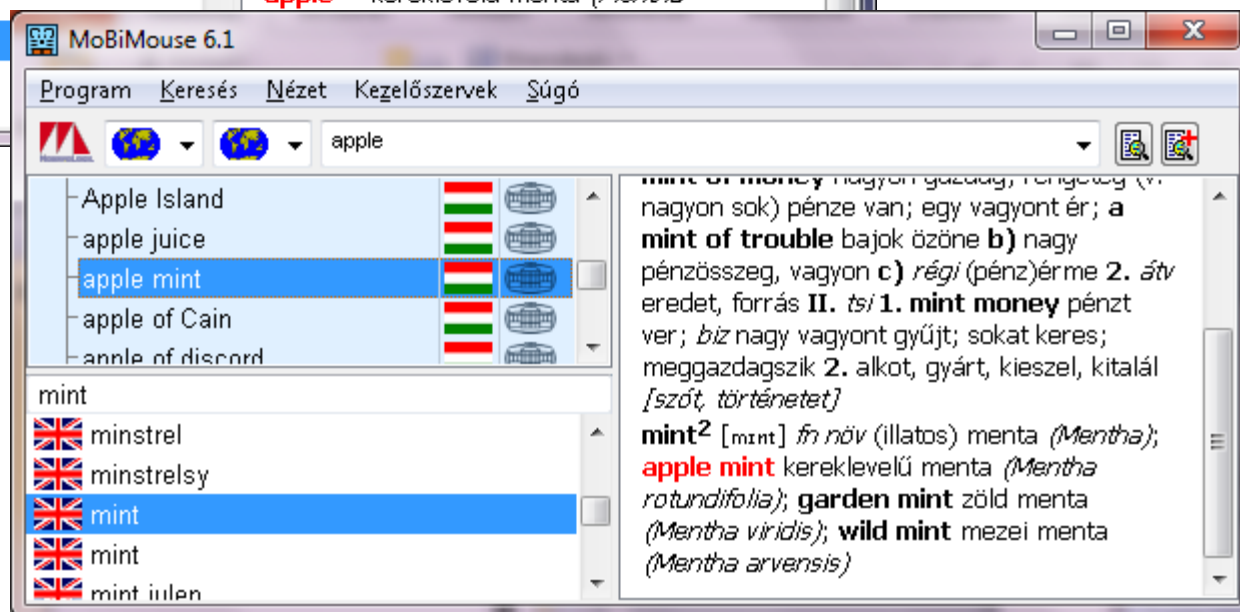
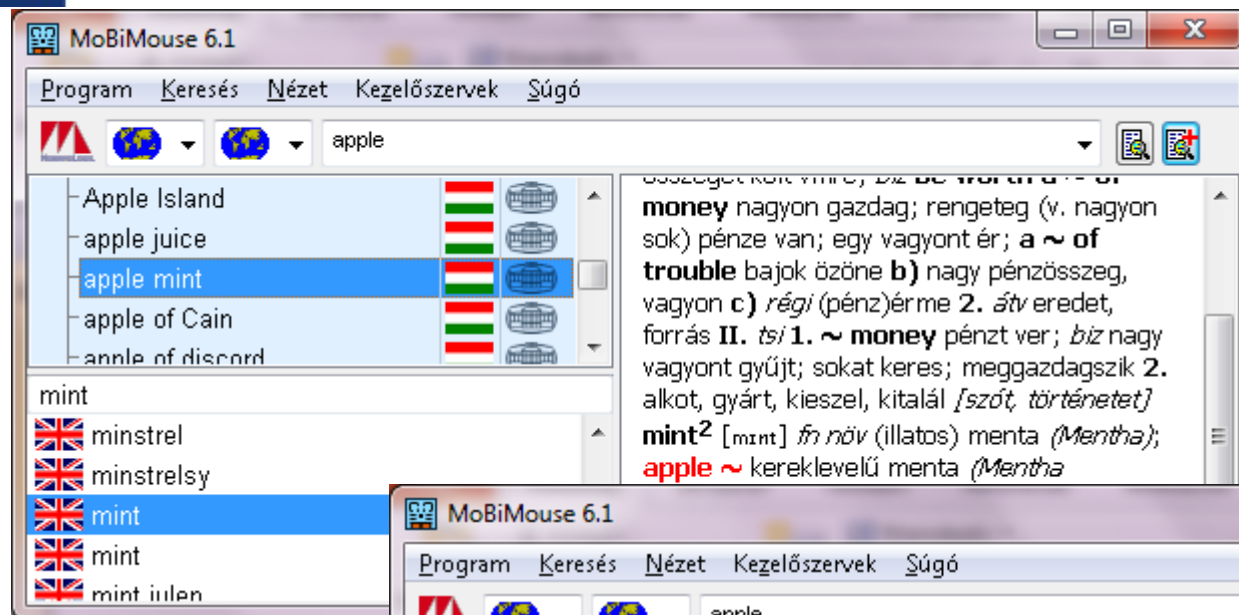
Bottom Window: The search term is "elássa csatabárdot". The results list three occurrences of "elássa a csatabárdot" with different flags (UK, UK, UK). The definition on the right is: **bárd¹** <-ot, -ja> *fn* [vágó] hatchet, bill, carpenter's broad axe, adz(e), (butcher's) cleaver, hacker, [konyhai] chopper, [kétéltű] twibill; **harci ~** war/battle-axe, [indian] tomahawk; **elássa a csata~ot** bury the hatchet; **bárd²** <-ot, -ja> *fn* [dalnok] bard.

The left sidebar of each window shows a list of related words and their grammatical forms, such as "csatabárd", "csat", "csata", "csatabárd", "csatacirkáló", "csatadal", "elás", "elárvul", "elárvult", "elás", "elásás", "elásntt", "bárd", "bárcázás", "bárcsak", "bárd", "bárd", "hárdol".

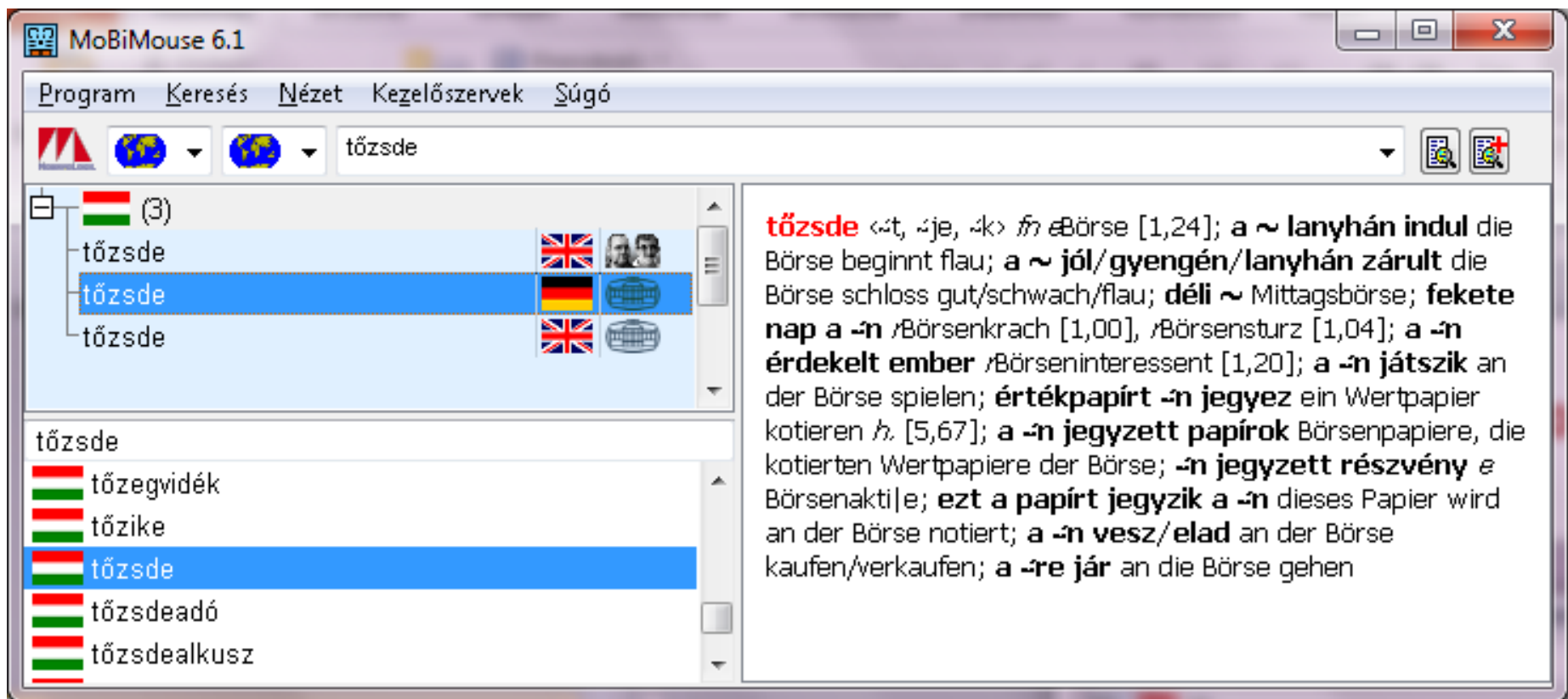
Az ellentmondó előfordulások problémája



A tilde-probléma



A morfológia-probléma

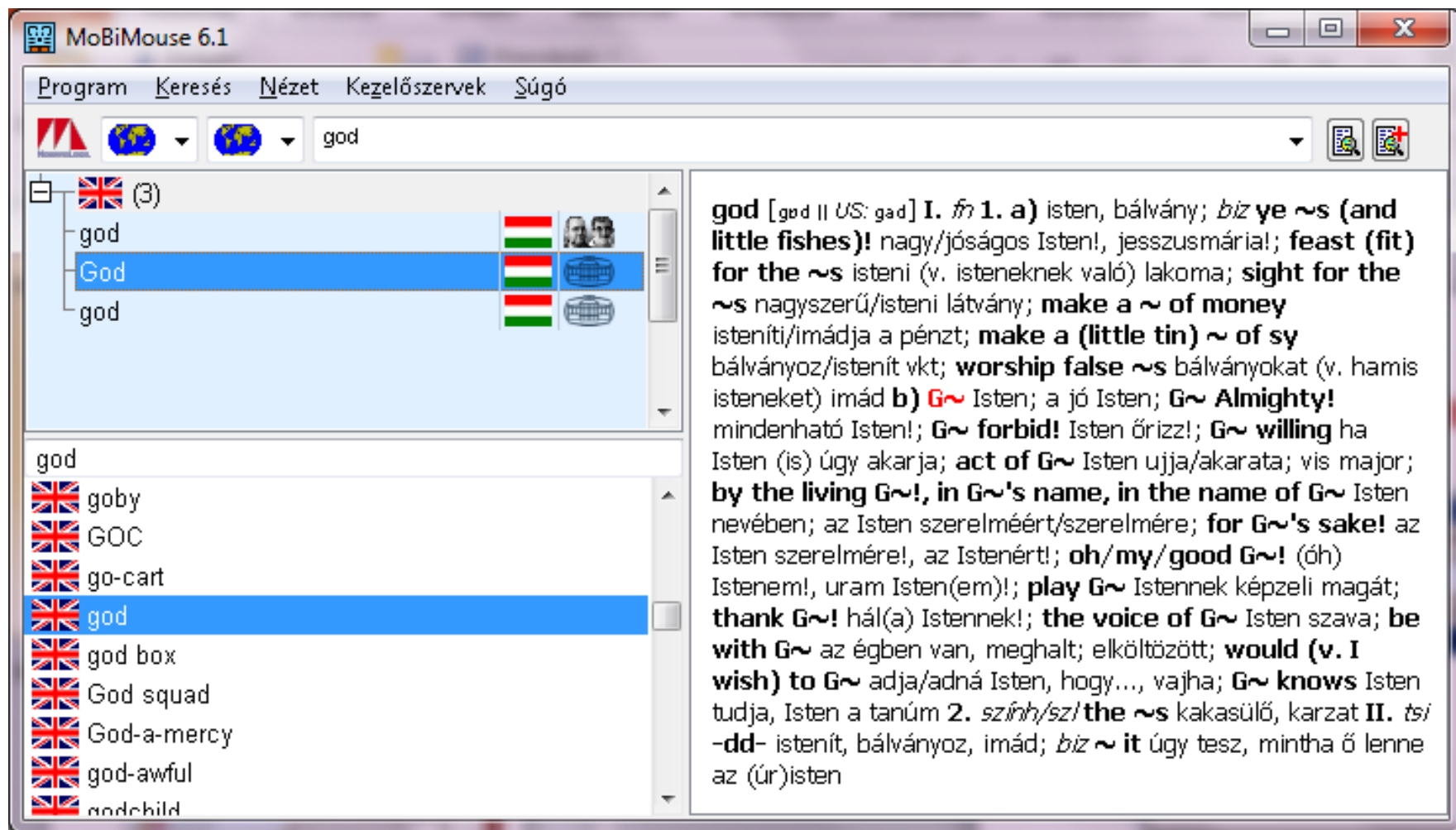


The screenshot shows the MoBiMouse 6.1 software interface. The title bar reads "MoBiMouse 6.1". The menu bar includes "Program", "Keresés", "Nézet", "Kezelőszervek", and "Súgó". Below the menu bar, there are icons for a flag, a globe, and a search icon, followed by a text input field containing "tőzsde".

The main window is divided into two panes. The left pane shows a hierarchical tree structure of the word "tőzsde" with three levels, each associated with a flag (Hungary, Germany, and the UK). The right pane displays a detailed morphological analysis of the word "tőzsde" in Hungarian, including its phonetic transcription, syllable structure, and a list of related words and their meanings.

tőzsde <ʦt, ʦje, ʦk> /ʃ/ Börse [1,24]; a ~ lanyhán indul die Börse beginnt flau; a ~ jól/gyengén/lanyhán zárult die Börse schloss gut/schwach/flau; **déli** ~ Mittagsbörse; **fekete nap** a -n /Börsenkrach [1,00], /Börsensturz [1,04]; a -n **érfdekelt ember** /Börseninteressent [1,20]; a -n **játszik** an der Börse spielen; **értékpapírt** -n **jegyvez** ein Wertpapier kotieren *h.* [5,67]; a -n **jegyzett papírok** Börsenpapiere, die kotierten Wertpapiere der Börse; -n **jegyzett részvény** *e* Börsenakti|e; **ezt a papírt jegyzik** a -n dieses Papier wird an der Börse notiert; a -n **vesz/elad** an der Börse kaufen/verkaufen; a -re **jár** an die Börse gehen

A nagybetű-probléma



MoBiMouse 6.1

Program Keresés Nézet Kezelőszervek Súgó

god

god (3)

god

God

god

god

goby

GOC

go-cart

god

god box

God squad

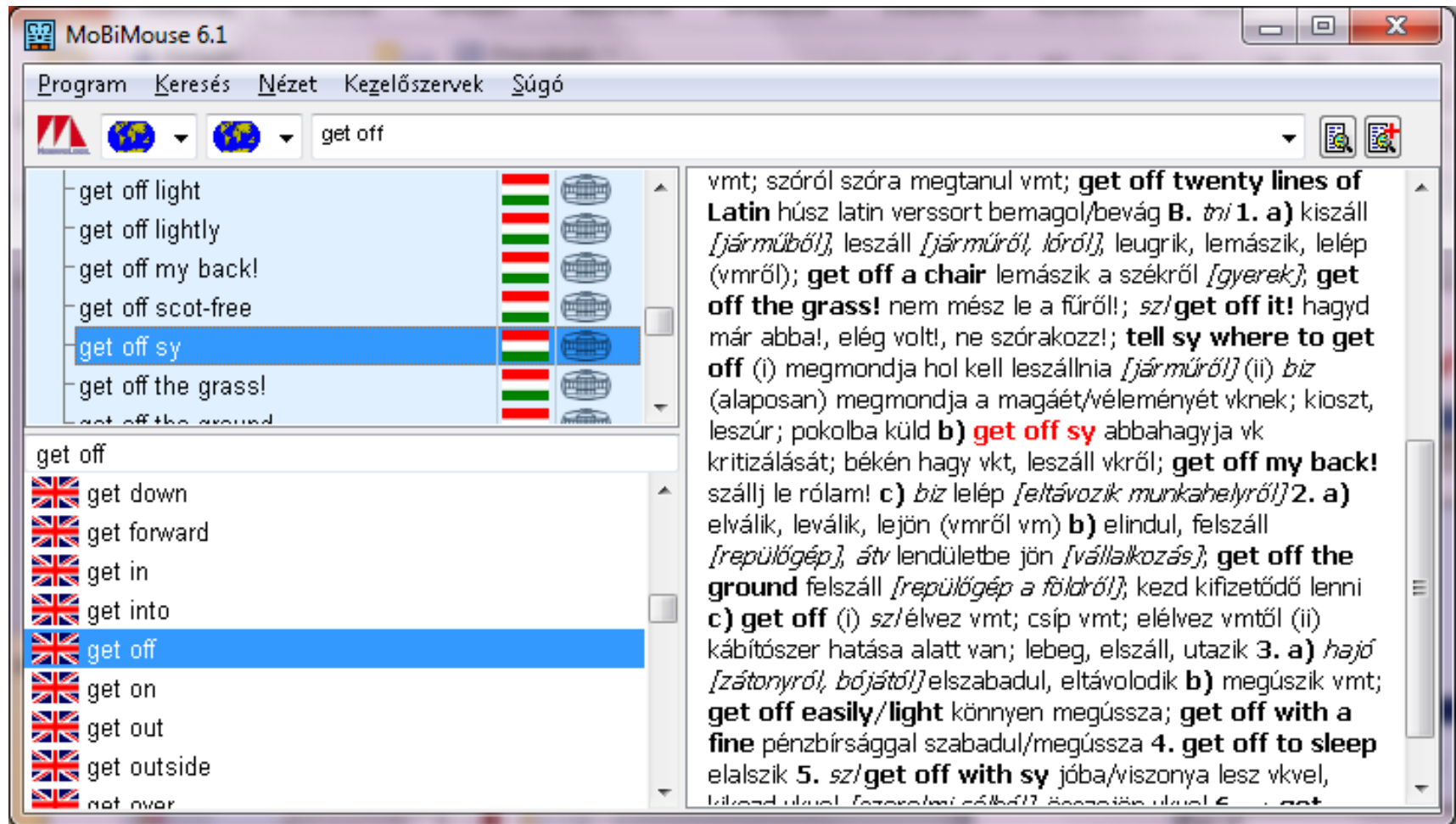
God-a-mercy

god-awful

godchild

god [gɒd II US: gɑd] I. *fn* 1. a) isten, bálvány; *biz ye ~s (and little fishes)!* nagy/jóságos Isten!, jesszusmária!; *feast (fit) for the ~s* isteni (v. isteneknek való) lakoma; *sight for the ~s* nagyszerű/isteni látvány; *make a ~ of money* isteníti/imádja a pénzt; *make a (little tin) ~ of sy* bálványoz/istenít vkt; *worship false ~s* bálványokat (v. hamis isteneket) imád b) G~ Isten; a jó Isten; G~ Almighty! mindenható Isten!; G~ forbid! Isten őrizz!; G~ willing ha Isten (is) úgy akarja; *act of G~* Isten ujja/akarata; vis major; *by the living G~!, in G~'s name, in the name of G~* Isten nevében; az Isten szerelméért/szerelmére; *for G~'s sake!* az Isten szerelmére!, az Istenért!; *oh/my/good G~!* (óh) Istenem!, uram Isten(em)!; *play G~* Istennek képzei magát; *thank G~!* hál(a) Istennek!; *the voice of G~* Isten szava; *be with G~* az égben van, meghalt; elköltözött; *would (v. I wish) to G~* adja/adná Isten, hogy..., vajha; G~ knows Isten tudja, Isten a tanúm 2. *szính/sz!* the ~s kakasülő, karzat II. *tsi -dd-* istenít, bálványoz, imád; *biz ~ it* úgy tesz, mintha ő lenne az (úr)isten

A vonzat-probléma



The screenshot shows the MoBiMouse 6.1 software interface. The main window displays a list of phrases on the left and their translations on the right. The phrase "get off" is highlighted in blue. The translations are in Hungarian and include examples of how to use the phrase in different contexts.

get off

get off light
get off lightly
get off my back!
get off scot-free
get off sy
get off the grass!
get off the ground

get off

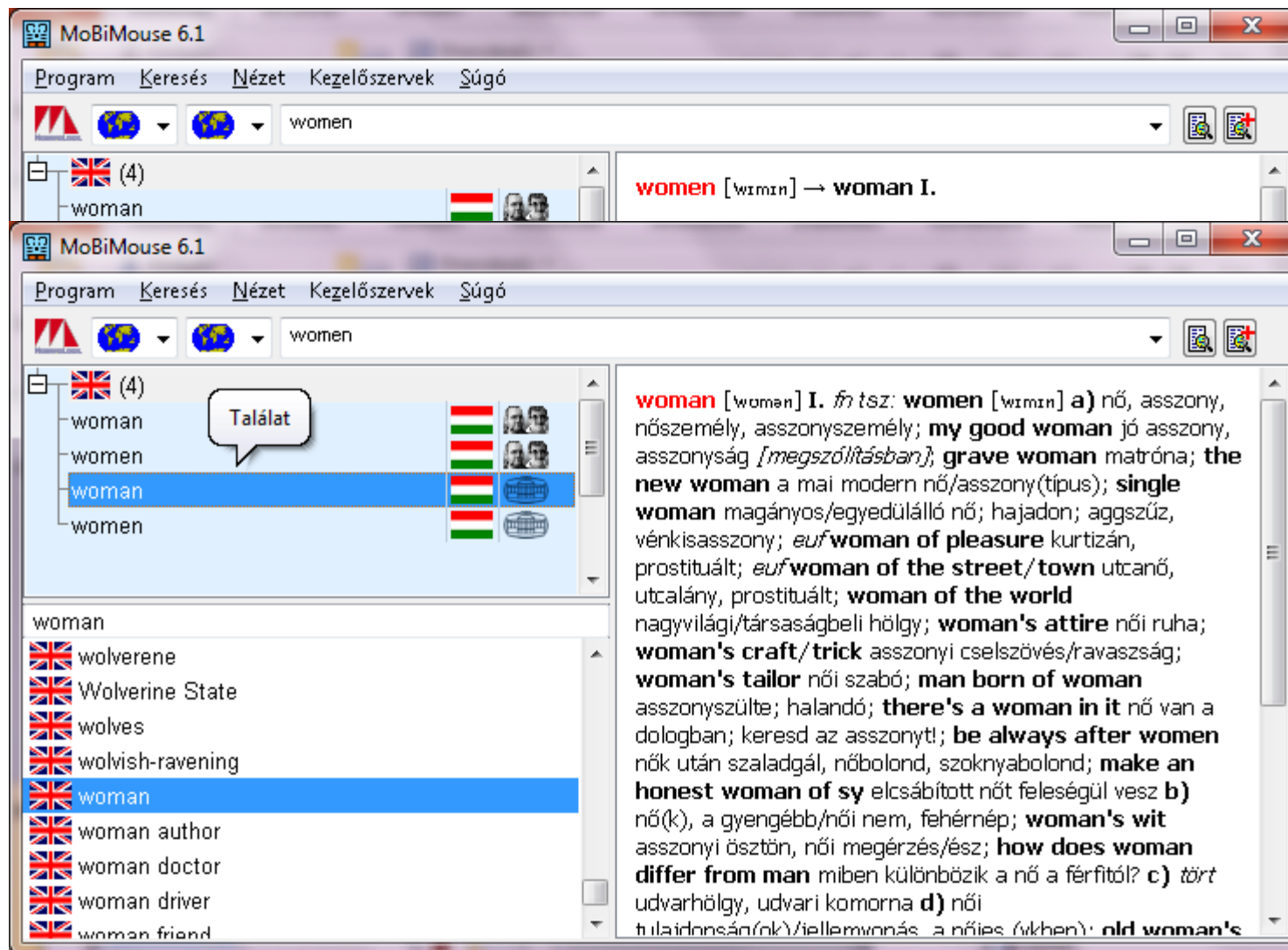
get down
get forward
get in
get into
get off
get on
get out
get outside
get over

vm; szóról szóra megtanul vmt; **get off twenty lines of Latin** húsz latin verssort bemagol/bevág **B. tri 1. a)** kiszáll [járműből], leszáll [járműről, lóról], leugrik, lemászik, lelép (vmről); **get off a chair** lemászik a székről [gyerek]; **get off the grass!** nem mész le a fűről; **sz/get off it!** hagyd már abba!, elég volt!, ne szórakozz!; **tell sy where to get off** (i) megmondja hol kell leszállnia [járműről] (ii) **biz** (alaposan) megmondja a magáét/véleményét vknek; kioszt, leszúr; pokolba küld **b) get off sy** abbahagyja vk kritizálását; békén hagy vkt, leszáll vkről; **get off my back!** szállj le rólam! **c) biz** lelép [eltávozik munkahelyről] **2. a)** elváll, leváll, lejön (vmről vm) **b)** elindul, felszáll [repülőgépre], átv lendületbe jön [vállalkozás]; **get off the ground** felszáll [repülőgép a földről]; kezd kifizetődő lenni **c) get off** (i) **sz/élvez** vmt; csíp vmt; élélvez vmtől (ii) kábítószer hatása alatt van; lebeg, elszáll, utazik **3. a) hajó** [zatonnyról, bőjattól] elszabadul, eltávolodik **b)** megússzik vmt; **get off easily/light** könnyen megússza; **get off with a fine** pénzbírsággal szabadul/megússza **4. get off to sleep** elalszik **5. sz/get off with sy** jóba/viszonya lesz vkvel, kilógat ulat, fozozalmi csikó/2. évezredben ulat **6. sz/get**

A példa-probléma

[hajó] **b) a disease will run its course** egy betegségnek megvan a (szabályos) lefolyása; egy betegségnek nem is nagyon lehet megváltoztatni a lefolyását; **things must run their course** a dolgoknak ki kell futniuk magukat, szabad folyást kell engedni a dolgoknak **c) run a traffic signal** elnéz (v. nem vesz észre) (közúti) jelzőtáblát **2. a)** (meg)futtat, (meg)szalaszt **b)** űz, üldöz, hajsza [vadat], versenyt fut (vkvel); **vad run a scent** nyomon van [kutya]; **átv run sy close/hard** szorongat/megszorít vkt; nyomában van vknek, komoly versenytársa vknek, megközelít vkt; **átv be hard run** szorongatják, szorongatott/szorult helyzetbe kerül; szutyongatják; **biz** (anyagilag) megszorult, le van égve (anyagilag) **3. futtat** [lovat, versenylovat]; **biz run a candidate** jelöltet futtat **4. a)** vezet, irányít, kormányoz, visz [járművet]; **run a ship to London** elvezeti/elkormányozza a hajót Londonba **b) run sy up to town** bevisz vkt a városba [autón] **5. üzemben tart, járat, működtet** [gépet], járat, fenntart [járművet bizonyos útvonalon]; **run a car** kocsit/autót tart **6. a)** vezet [szállót, üzletet], irányít [gazdaságot], üzemben tart [gyárat], kezel, vezet [ügyeket], igazgat [színházat], ellát, (el)vezet [háztartást]; **run a business** üzletet (v. üzleti vállalkozást) irányít/vezet; **biz it is he who runs the show** ő a fejes/góré; **biz he is run by his secretary** a titkára dirigálja, a titkára intéz helyette mindent **b) gazd/biz we are running a cheap line just now that might suit you** jelenleg igen jutányosan beszerezhető árutételt tartunk raktáron, ami talán önt is érdekelné **7. a)** folyvat, ereszt [folyadékot], (ki)önt [formába]; **run sy a bath, run a bath for sy** fürdőt készít vknek, fürdővizet ereszt/enged vknek; **her eyes run tears** szeme könnyet hullat/ont, szeméből könny patakzik; **the streams run blood** a folyók vértől zavarosak/áradnak **b) kiolvaszt; run butter** vajat kisüt **8. (meg)húz, megvon** [vonalat, határt]; **run a parallel too far** túl messzire megy az összehasonlításban **9. biz** csempész [pl. kábítószert]; **run liquor/run** szeszt/alkoholt csempész **10. biz run a temperature** láza/hőemelkedése van **11. behúzó öltéssel varr** [minden 4–5-ik öltés után húzza ki a tűt] **12. a) sp** úgy üti a labdát, hogy leesés után továbbguruljon a lyuk felé [golfsban] **b) sp** átüt [labdát kapun, krikettben] **13. biz run a rig** kirúg a hámból **14. kat/szl run a man** fegyelemsértéssel vádol vkt **15. US run an ad** (állandó) hirdetést ad fel (v. tesz közzé) **16. sp** versenyez **17. elfuvaroz, elvisz** [vkit járművön] **B. tni 1. a) fut, szalad, rohan; run home** hazaszalad; **átv** befut a célba; **run like a hare, run like blazes/hell, run like the devil** fut, mint a nyúl; **run to meet sy** vk elé(ke) fut/szalad; **biz run to meet one's troubles** keresi a bajt; **rules so clear that he who runs may read** olyan világos szabályok, hogy azonnal megért(het)ji mindenki; **he ran a poor third** gyenge harmadiknak futott be; **also ran** futottak még; **an also ran** egy a próbálkozók/résztvevők/futottak közül; **set off running** nekiiramodik, futni kezd, futásnak ered; **biz cut and run** elpucol, elszkol, kereket old **b) biz** kifutóként/küldőnként működik **2. a) fut, megy, halad** [jármű], jár [keréken]; **hajó run free** hátfélsszéllal (v. háromnegyed széllal) megy/halad/vitorlázik **b) gördül** [vers] **c) let one's ideas run freely** gondolatait szabadjára engedi (v. csapongani hagyja) **3. a) jár, működik, megy, dolgozik, üzemben van** [gép, gyár], forog [kerék]; **műsz run hot** túlhevül, hõn fut [csapágy] **b) jár, közlekedik, megy** [jármű menetrendszerûen]; **a bus runs between the two villages** autóbusz jár/közlekedik a két falu között **c) how her tongue runs!** be nem áll a szája!, hogy pereg a

A „lásd”-probléma



MoBiMouse 6.1

Program Keresés Nézet Kezelőszervek Súgó

women

(4)

woman

women [wɪmɪn] → woman I.

MoBiMouse 6.1

Program Keresés Nézet Kezelőszervek Súgó

women

(4)

woman

women

woman

women

Találat

woman

wolverene

Wolverine State

wolves

wolvish-ravering

woman

woman author

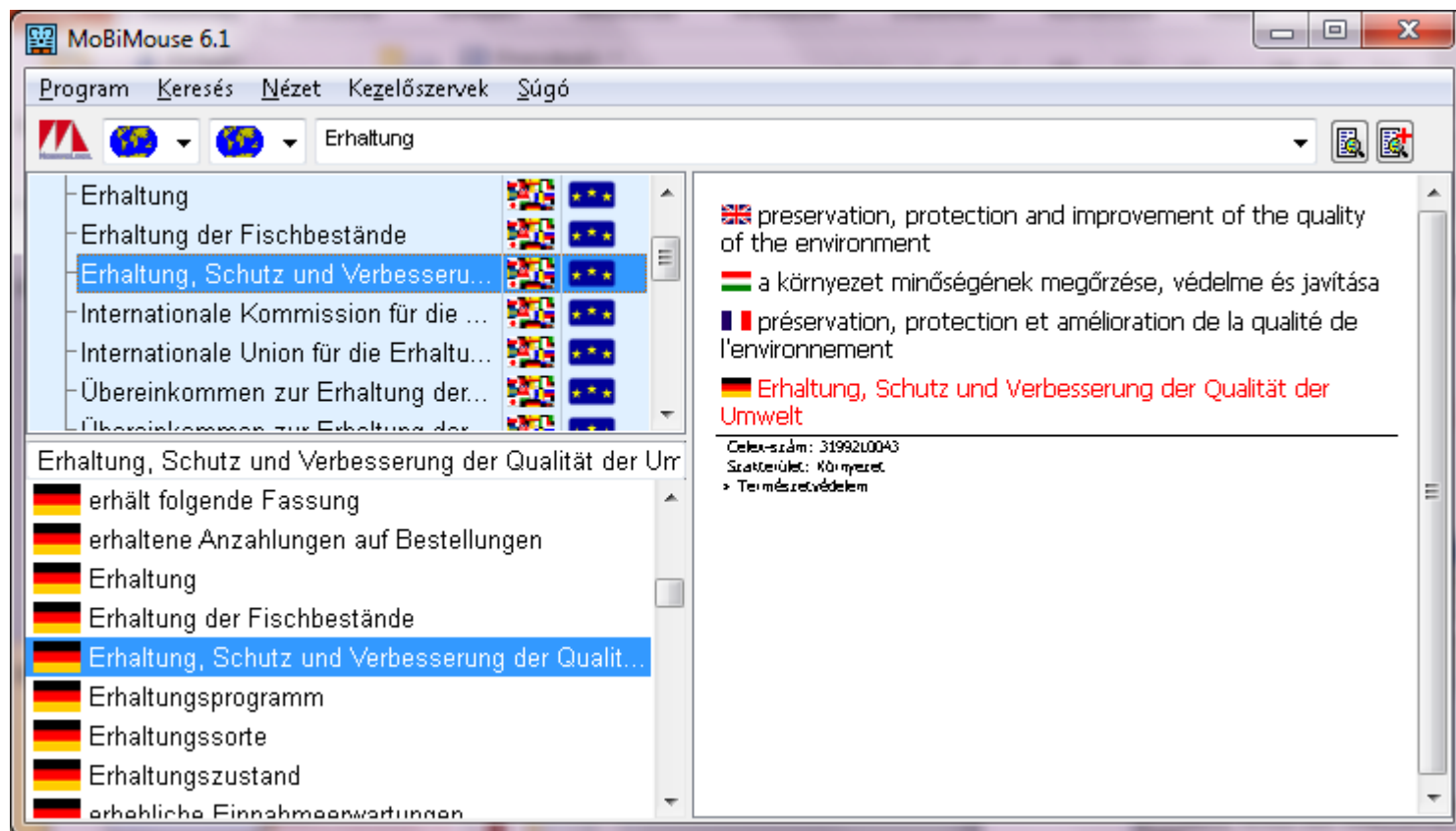
woman doctor

woman driver

woman friend

woman [wʊmən] I. *fr* tsz: **women** [wɪmɪn] a) nő, asszony, nőszemély, asszonyszemély; **my good woman** jó asszony, asszonyság [megszólításban]; **grave woman** matróna; **the new woman** a mai modern nő/asszony(típus); **single woman** magányos/egyedülálló nő; hajadon; aggszűz, vénkisasszony; **euf** **woman of pleasure** kurtizán, prostituált; **euf** **woman of the street/town** utcanő, utcalány, prostituált; **woman of the world** nagyvilági/társaságbeli hölgy; **woman's attire** női ruha; **woman's craft/trick** asszonyi cselszövés/ravaszság; **woman's tailor** női szabó; **man born of woman** asszonyszülte; halandó; **there's a woman in it** nő van a dologban; keresd az asszonyt!; **be always after women** nők után szaladgál, nőbolond, szoknyabolond; **make an honest woman of sy** elcsábított nőt feleségül vesz b) nő(k), a gyengébb/női nem, fehérszínű; **woman's wit** asszonyi ösztön, női megérzés/ész; **how does woman differ from man** miben különbözik a nő a férfitól? c) tört udvarhölgy, udvari komorna d) női tulajdonság(nk)/jellemvonás a nőies (vkhen): **old woman's**

A pontos találatok problémája



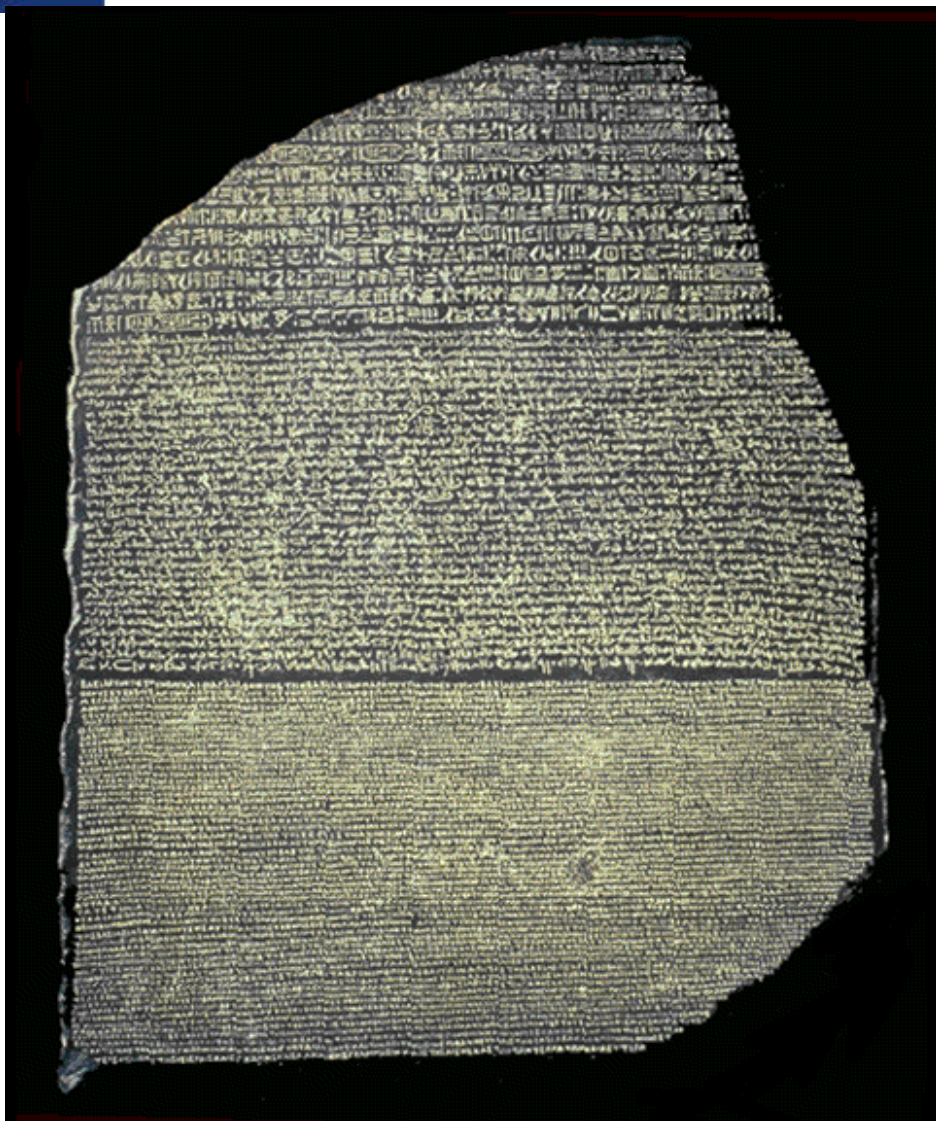


Bevezetés a nyelvtechnológiába

8.

Fordítástámogatás: fordítómemóriák

A Rosetta kő





Párhuzamos korpuszok

A párhuzamos korpuszok

- ☐ Egy könyvben átlagosan néhány százezer szó van
- ☐ Egy művelt ember átlagosan napi tízezer szót olvas
- ☐ Ez összesen évi 3,5 millió és egy életben hozzávetőlegesen néhány százmillió
- ☐ Sok lefordított szöveg akár több százmillió szava elektronikus úton elérhető
- ☐ A gép tehát több szóval tud találkozni, mint egy ember egész életében...
- ☐ Egy lehetséges konklúzió: a fordítási probléma a minták alapján statisztikai alapon megoldható, de legalábbis a fordító munkája másokéval jól támogatható, ha megtaláljuk az épp fordítandóhoz leghasonlóbb mondatot



Kereshető párhuzamos korpusz

A sziklák azonban, amelyekből az Egyetem fölépült, több ezer éven át szívták magukba a mágiát, s annak a rengeteg rendezetlen erőnek kerülnie kellett valahova.	The rocks from which Unseen University was built, however, have been absorbing magic for several thousand years and all that random power has had to go somewhere.
Sajnálattal közlöm, hogy csupáncsak egy ajtó vezet a varázslósághoz, és az a Láthatatlan Egyetem főbejárata, és azon még nő soha nem lépett keresztül.	I am sorry to say there is only one door into wizardry and that is the main gate at Unseen University and no woman has ever passed through it."
Valamikor a nem túl távoli múltban valaki elhatározta valami bizonytalan elképzelés alapján, mely szerint a Tanulás Élvezet Kéne Legyen, hogy az Egyetem ősi folyosóit falfestéssel kell fölvidítani.	At some time in the recent past someone had decided to brighten the ancient corridors of the University by painting them, having some vague notion that Learning Should Be Fun.
Vastag sugárban dőlt az eső a Láthatatlan Egyetem esővízcsatornáiba s a csurgókba folyt, ahol a nyár óta elhagyatott varjúfészkek úgy lebegtek, mint nagyon rosszul megépített csónakok.	Rain gushed across the leads of Unseen University and poured into the gutters where ravens' nests, abandoned since the summer, floated like very badly-built boats.
Mint az egyetem kincstárnoka azt kell mondjam, hogy mi mindig nagy hangsúlyt fektettünk a közösség érdekeit figyelembe vevő jószomszédi politikára motyogta a varázsló, megpróbálva kikerülni Albert szúrós tekintetét.	'As the bursar of this university I must say that we've always encouraged a good neighbour policy with respect to the community,' mumbled the wizard, trying to avoid Albert's gimlet stare.
A másik kérdésről - amely véleményem szerint ugyancsak nehézségeket okozhatott, s az előbbivel együtt nyilván hozzájárult a Párizsi Egyetem válaszának késlekedéséhez - csak futólag beszélnek.	As to the other point which I have thought may have proved a difficulty and cost the University delay, I will touch but a moment upon that, and pass on.



Párhuzamosan is kereshető korpusz

Járt egyetemre , nem?	You went to college , right?
Ezt követően a Bizottság a Cranfield-i Repülésműszaki Egyetem Légi közlekedési Csoportjának a vezetőjét, Doganis professzort bízta meg független szakértői feladatok ellátásával.	Thereafter, the Commission appointed Professor Doganis, Head of the Air Transport Group of the College of Aeronautics at Cranfield University as a new independent expert.
A legtöbben a sebészeti és az általános tanfolyamot párhuzamosan hallgatták, de a törekvőbbek vagy a szorgalmasabbak ezenfelül még további tanulmányokat folytattak, amelyek végeztével elnyerhették a londoni egyetem doktorátusát.	Most of the students took the curriculum of the Conjoint Board of the College of Surgeons and the College of Physicians; but the more ambitious or the more industrious added to this the longer studies which led to a degree from the University of London.
Végig csaltam az egyetemen	I have cheated my way through college
Azt hiszem, a portói az okosabb választás, 1934-ben került az egyetem pincéjébe, ha nem tévedek, ez az egyik legjobb évjárat, ezenfelül brandy nincsen is.	The port I think is the better bet, laid down by the college in 1934, one of the finest vintages I think you'll find, and on the other hand I don't actually have any brandy.
Az egyetemen , addig, amíg végzek.	In college before I dropped out.
Két kölyök egy puccos egyetemen	Two kids in a fancy college



Alkorpuszokra bontott párhuzamos korpusz

Dictionary **egyetem**

college

Subtitles - Ez egy **egyetem**!

- This is **college**!

Subtitles Az **egyetemen** volt.

She was at **college**.

Literature Még **egyetemre** járt.

She was in **college**."

Subtitles Máris mész az **egyetemre**!

Already off to **college**.

Literature Hogy Brett **egyetemre** járhatson.

Toward's Brett's **college**.

Subtitles Járt **egyetemre**, nem?

You went to **college**, right?

Literature Hol járt Claire **egyetemre**?

"Where did Claire go to **college**?"

Subtitles Két kölyök egy puccos **egyetemen**.

Two kids in a fancy **college**.

Subtitles Végig csaltam az **egyetemen**.

I have cheated myway through **college**.

Literature Én akkor már **egyetemre** jártam.

I was in **college** by then.

Literature Ó, **egyetem**... **egyetem**... zihálta Kinderman bánatosan, kimeresztette a szemeit, és bólogatott.

"Ah, **college**, **college**..." breathed Kinderman ruefully, nodding as he stared.

Subtitles Ilyen lányok miatt jöttem az **egyetemre**.

That's the reason I went to **college**.

Subtitles Az **egyetemen** meg hívogatsz állandóan.

Now in **college**, you call me a good time.

Subtitles Istenem, az **egyetem** óta nem.

I haven't had one of these since **college**.

A Hunglish korpusz

- ☐ Magyar–angol párhuzamos korpusz
- ☐ Szoftverdokumentáció,
jogi szövegek,
szépirodalom,
film,
újságok
- ☐ 54,2m token
- ☐ 2,07m mondat
- ☐ BME MOKK, MTA NYTI



Egy tipikus párhuzamos korpusz: Acquis Communautaire

Language ISO code	N° of texts	Text body			Signatures	Annexes	Total No words (text + signatures + annexes):
		Total No words	Total No characters	Average No words	Total No words	Total No words	
bg	11384	16140819	104522671	1417.85	2170075	14114612	32425506
cs	21438	22843279	148972981	1065.55	7225300	16763733	46832312
da	23624	31459627	213468135	1331.68	2629786	16855213	50944626
de	23541	32059892	232748675	1361.87	2542149	16327611	50929652
el	23184	36453749	239583543	1572.37	2973574	16459680	55887003
en	23545	34588383	210692059	1469.03	3198766	17750761	55537910
es	23573	38926161	238016756	1651.3	3490204	19716243	62132608
et	23541	24621625	192700704	1045.9	1336051	14995748	40953424
fi	23284	24883012	212178964	1068.67	2677798	12547171	40107981
fr	23627	39100499	234758290	1654.91	3021013	19978920	62100432
hu	22801	28602380	213804614	1254.44	2529488	15056496	46188364
it	23472	35764670	230677013	1523.72	3120797	18331535	57217002
lt	23379	26937773	199438258	1152.22	2436585	15018484	44392842
lv	22906	27592514	196452051	1204.6	1673124	15437969	44703607
mt	10545	20926909	128906748	1984.53	1336042	15620611	37883562
nl	23564	35265161	231963539	1496.57	3039580	18467115	56771856
pl	23478	29713003	214464026	1265.57	2513141	17027393	49253537
pt	23505	37221668	227499418	1583.56	3034308	19350227	59606203
ro	6573	9186947	60537301	1397.68	514296	11185842	20887085
ro-19211 (readme)	19211	30832212	182631277	1604.92	---	---	30832212
sk	21943	26792637	179920434	1221.01	3227852	16190546	46211035
sl	20642	27702305	178651767	1342.04	3103193	16837717	47643215
sv	20243	29433037	199004401	1453.99	2575771	14965384	46974192
Total	463792	636216050	4288962348	1387.23	60368893	358999011	1055583954



A Biblia mint párhuzamos korpusz

Latin: Biblia Sacra Vulgata

respondens autem lob dixit

Job 23:1 Spanish: La Biblia de las Américas (©1997)

Entonces respondió Job, y dijo:

Hiob 23:1 German: Luther (1912)

Hiob antwortete und sprach:

Job 23:1 French: Louis Segond (1910)

Job prit la parole et dit:

約伯記 23:1 Chinese Bible: Union (Simplified)

约伯回答说:

King James Bible

Then Job answered and said,

American King James Version

Then Job answered and said,

Job 23:1 Danish

Så tog Job til Orde og svarede:

Job 23:1 Dutch Staten Vertaling

Maar Job antwoordde en zeide:

Jób 23:1 Hungarian: Karoli

Felele pedig Jób, és monda:

ljob 23:1 Esperanto

ljob respondis kaj diris:

JOB 23:1 Finnish: Bible (1776)

Job vastasi ja sanoi:

JOB 23:1 Finnish: Pyhä Raamattu (1933/1938)

Job vastasi ja sanoi:

Job 23:1 Greek OT: Septuagint

υπολαβων δε ιωβ λεγει

Jobo knyga 23:1 Lithuanian

Jobas atsakydamas tarė:

Job 23:1 Maori

Katahi ka whakautua e Hopa, i mea ia,

Jobs 23:1 Norwegian: Det Norsk Bibelselskap (1930)

Da tok Job til orde og sa:

Polish: Biblia Gdanska

A odpowiadając ljob rzekł:

Jó 23:1 Portuguese Bible

Então Jó respondeu:

lov 23:1 Romanian: Cornilescu

lov a luat cuvîntul și a zis:

Иов 23:1 Russian: Synodal Translation (1876)

И отвечав Иов и сказал:



Szövegszinkronizáció

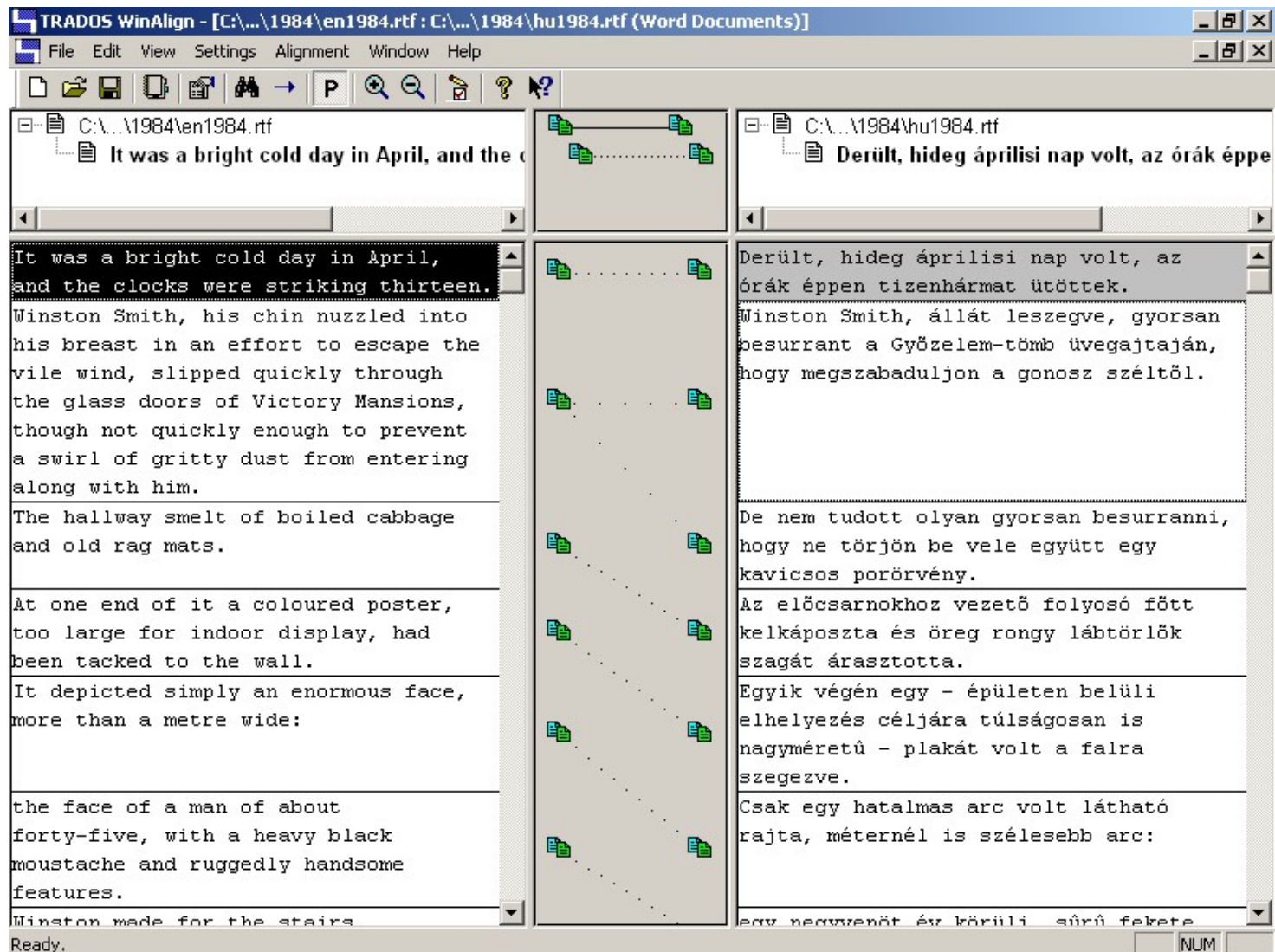
Szövegszinkronizáció-típusok

- ☐ „Alignment”
- ☐ Kétnyelvű szövegpárban az egymás fordításának tekinthető szövegegységek meghatározása
- ☐ Nyelvi egységek szerinti szövegszinkronizáció-típusok:
 - Bekezdésszintű
 - Mondatszintű
 - Kifejezésszintű
 - Szószintű
- ☐ Pontosság/fedés szerinti szövegszinkronizáció-típusok:
 - teljes
 - részleges

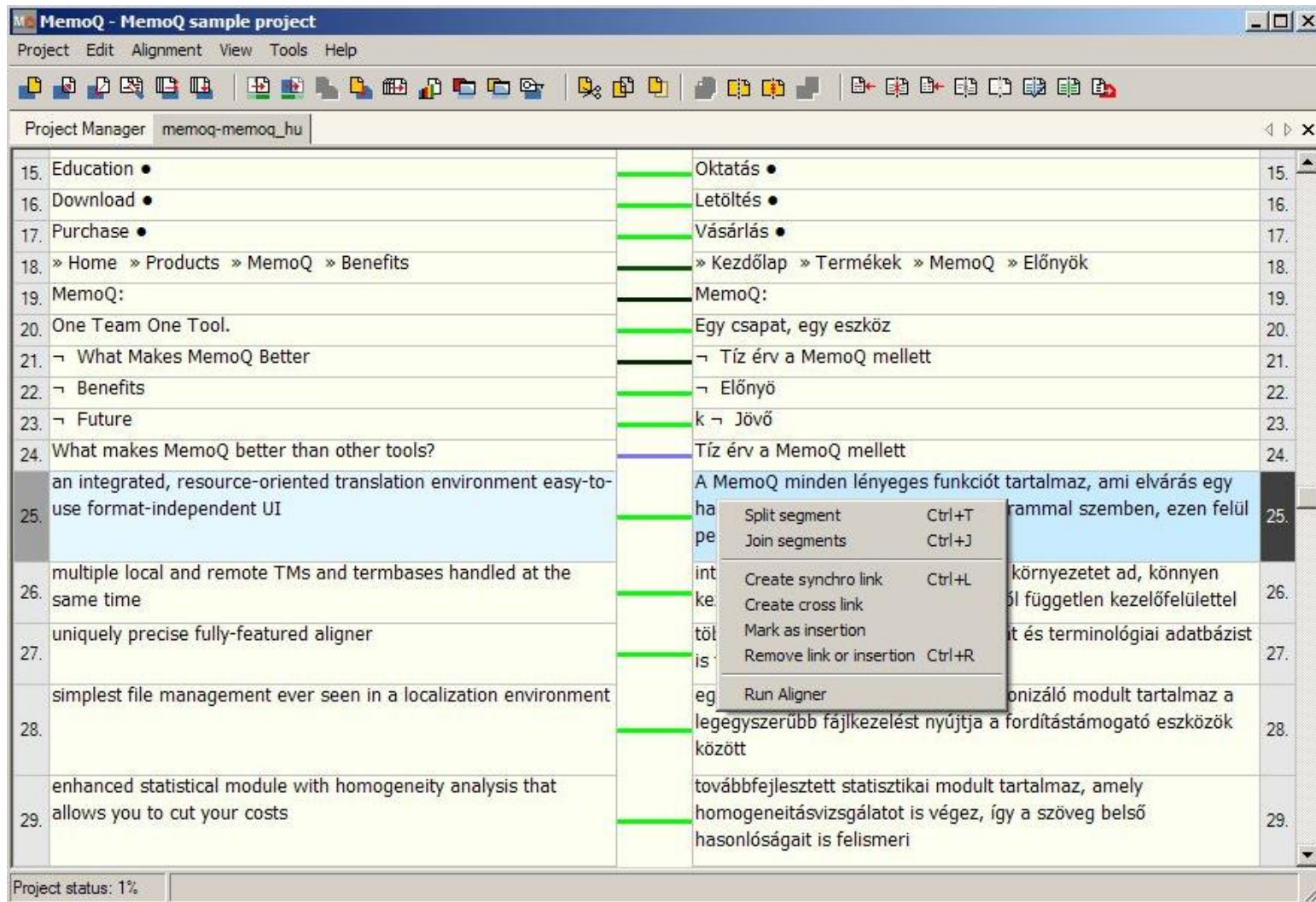
A szövegszinkronizáció felhasználása

- ☐ Párhuzamos korpuszok építése fordítástudományi felhasználásra
 - két- vagy többnyelvű korpuszok szinkronizálása fordítási megfelelések vizsgálatára
- ☐ Fordítások terminológiai konzisztenciájának vizsgálata
 - a forrásszövegben megtalált szakkifejezéseket a fordításban megkeressük
- ☐ Kétnyelvű terminológia gyűjtése
- ☐ Fordítómemória építése
 - a párok kész fordításként felhasználhatók a későbbiekben
- ☐ Statisztikai gépi fordítás

Szövegszinkronizáló eszközök (TRADOS)



Szövegszinkronizáló eszközök (MemoQ)



The screenshot shows the MemoQ software interface. The top menu bar includes Project, Edit, Alignment, View, Tools, and Help. Below the menu is a toolbar with various icons for file operations and alignment. The main window is titled 'MemoQ - MemoQ sample project' and displays a 'Project Manager' pane on the left with a list of segments. The segments are numbered 15 to 29 and contain text in Hungarian. A context menu is open over segment 25, showing options like 'Split segment', 'Join segments', 'Create synchro link', 'Create cross link', 'Mark as insertion', 'Remove link or insertion', and 'Run Aligner'. The status bar at the bottom indicates 'Project status: 1%'.

Segment Number	Segment Content (Hungarian)
15.	Education • Oktatás •
16.	Download • Letöltés •
17.	Purchase • Vásárlás •
18.	» Home » Products » MemoQ » Benefits » Kezdőlap » Termékek » MemoQ » Előnyök
19.	MemoQ: MemoQ:
20.	One Team One Tool. Egy csapat, egy eszköz
21.	↪ What Makes MemoQ Better ↪ Tíz érv a MemoQ mellett
22.	↪ Benefits ↪ Előnyö
23.	↪ Future k ↪ Jövő
24.	What makes MemoQ better than other tools? Tíz érv a MemoQ mellett
25.	an integrated, resource-oriented translation environment easy-to-use format-independent UI A MemoQ minden lényeges funkciót tartalmaz, ami elvárás egy
26.	multiple local and remote TMs and termbases handled at the same time környezetet ad, könnyen
27.	uniquely precise fully-featured aligner től független kezelőfelülettel
28.	simplest file management ever seen in a localization environment és terminológiai adatbázist
29.	enhanced statistical module with homogeneity analysis that allows you to cut your costs szinkronizáló modult tartalmaz a



Mondatszegmentálás



A mondatszegmentálás bonyolult feladat

He can be reached at john.smith@sun.com. The class was written to be part of the java.util.Collections framework. He watched "E!" channel. There is a Jos. A. Banks clothes store on Newbury St. He worked at Smith Corp. "U.S. News reports that he was a model employee," said the anchor. She prefers the title MS. and don't forget it! He saw a 600-lb. gorilla. He flew at 500 m./h. and really high. He had a Ph.D. from UCLA. His name was R. J. Smith, Esq., and don't forget it. I bought fish, fruits, etc. Do you want some? Use a period ("\.") to indicate end of sentence. With \$15.7 billion in annual revenues, Sun can be found in more than 170 countries and on the World Wide Web at <http://sun.com>.

Egy egyszerűbb (?) mondatszegmentálási példa

A Mersey-parti vörösök a hazai vezetés után jobban odatették magukat, aminek eredményeként először Craig Bellamy a 38. percben – először még csak – az oldalhálóba fejelt, a játékrész lefújása előtt nem sokkal viszont már a gólvonal mögé tette a labdát: Steven Gerrard álompasszát a walesi csatár csukafejessel küldte kapura, de Victor Valdes kapus már csak bentről tudta kiemelte a labdát (Dirk Kuyt a biztonság kedvéért azért még a hálóba bombázott).



Szinkronizációs módszerek

A szinkronizációs módszerek csoportosítása

- ☐ Hossz-alapú módszerek
 - teljes szinkronizáció
- ☐ Lexikai információn alapuló teljes szinkronizációt megvalósító módszerek
 - csak részleges szinkronizáció
 - külső adatbázis kell
- ☐ Horgonykereső módszerek
 - csak részleges szinkronizáció
- ☐ Hibrid módszerek
 - több módszert kombinálnak

Hossz-alapú szinkronizáció

- ☐ „Length-based alignment”
- ☐ Gale & Church (1991, 1993)
- ☐ Csak a szövegegységek hosszait vizsgálja
- ☐ Valószínűségi modell
- ☐ Teljes szinkronizáció
- ☐ Dinamikus programozás →
globálisan optimális megoldást keres

A hossz-alapú szinkronizáció előnyei/hátrányai

☐ Pozitívumok

- Globálisan optimális megoldást keres
- Teljes szinkronizációra képes
- Robusztus: minden szövegegységnek van hossza

☐ Negatívumok

- Nem elég a szövegegységek hosszait vizsgálni, ezért könnyen „eltéved”
- A beszúrásokat és az elhagyásokat rosszul kezeli
- A dinamikus programozás költséges ($O(n^2)$), de ha a mátrixnak csak az átló körüli elemeit számítjuk ki, akkor jóval hatékonyabb módszert kapunk

Horgony-alapú szinkronizáció

- ☐ A *horgonyok* a szövegegységek egymásnak nagy pontossággal megfeleltethető pontjai (pl. nevek, számok)
- ☐ Kezdetben két horgony a szövegpár végpontjait köti össze
- ☐ Ezek után olyan szavakat keres, amelyek gyakran fordulnak elő együtt a szövegpár két oldalán nagyjából azonos pozícióban (heurisztika), majd ezekből horgonyokat határoznak meg (de pl. az egymást keresztező horgonyokat elvetik)
- ☐ A horgonyok által kisebb szinkronizálendő szegmensekre bontható a szöveg, ezeken belül újra horgonyokat keresnek
- ☐ Egyes horgonyokkal csak részleges szinkronizáció valósítható meg
- ☐ Az eredmények nem elég jók

A horgonyválasztásról

Latin írást használó nyelvek esetén (kivétel: német)

- ☐ azonos alakú (nagybetűs) szavak
- ☐ azonos szótövű (nagybetűs) szavak
(morfológia kell hozzá, ráadásul olyan,
amelyik az ismeretlen szavakkal is elboldogul)
- ☐ azonosan kezdődő, hasonló szavak („cognates”)
pl. *Parliament, Parlament; Europa, Europe, Európa*
- ☐ görög / cirill átírás??

Nem-latin írást használó nyelvek esetén

- ☐ horgonyjelöltek szótár alapján
- ☐ sok a nem egy-egy típusú megfeleltetés a szótárban

Statisztikai horgonyszűrés

- ❑ A horgonyjelöltek pozíciói (forrásbeli pozíció és a fordításbeli pozíció) alapján kiszámítják a lineáris regressziós egyenest
- ❑ Az egyenestől mért távolságok alapján egy hisztogramot készítenek, és ezzel szűrik a horgonyokat
- ❑ Kiszámítják a lineáris regressziós egyenes egy elfogadható konfidenciasávját, és a sávon kívülre került horgonyjelölteket eldobják



Fordítómemóriák

A fordítómemória

- ☐ A fordítómemória (fordítási memória, translation memory, TM) szegmenspárokat (általában mondatpárokat) tárol
- ☐ A TM forrásnyelvi szegmensek és célnyelvi fordítások párjait tárolja, némi adalékinformációval (rögzítés dátuma, fordító neve, fordítás státusza stb.)
- ☐ Ha később egy hasonló forrásnyelvi szegmenst kell fordítani, a TM program megtalálja és felkínálja a szegmens korábbi fordítását
- ☐ A javasolt mondat
 - (1) teljes egészében megfelelhet: egy az egyben el lehet fogadni
 - (2) jó kiindulópont lehet: némi módosítással fel lehet használni
 - (3) teljesen rossz: elvethető

A fordítómemóriák alapfunkciói

Export/Import

Szövegelemzés

- ☐ központosítás: egyfajta előfeldolgozás
- ☐ markup: lehetséges segítség
- ☐ „kódolt” szövegrészek (dátum, ábra, stb.) azonosítása

Szegmentálás

- ☐ a fordítási egységek megállapítása
- ☐ egynyelvű felszíni elemzéssel
- ☐ a szinkronizálás alapvető feltétele

Szövegszinkronizálás

- ☐ forrásnyelvi–célnyelvi párok létrehozása

Terminológiai kivonatolás

- ☐ induló szótárral, szövegstatistikák segítségével
- ☐ a fordítási munkával kapcsolatos becslések alapjául is szolgál

Szövegelőfeldolgozás (MemoQ)

Statistics

Scope: ☒ Project ☐ Document ☐ Selection ☐ From cursor ☐ Show results for each file

Counts: ☐ Include target counts ☒ Status report

Word counts: ☒ MemoQ ☐ TRADOS-like

Analysis: ☒ Project TM's ☐ Details by TM ☒ Homogeneity ☒ Fragment assembling

Buttons: Calculate, Save as HTML, Save as CSV, Close

Counts

Scope: Project

Type	Segments	Source words	Source chars	Percent
All	61	470	2690	100
Repetition	0	0	0	0
Not started	47	426	2431	91
Pre-translated	12	23	121	5
Fragments	0	0	0	0
Edited	1	19	128	4
Confirmed	1	2	10	0

Analysis

Scope: Project
Resources: Every TM, Homogeneity

Type	Segments	Source words	Source chars	Percent
100%	2	3	18	1
95%-99%	0	0	0	0
85%-94%	1	19	128	4
75%-84%	0	0	0	0
50%-74%	17	39	197	8
Fragments	0	0	0	0
No match	41	409	2347	87

A fordítómemóriák további (ritkább) funkciói

Visszakeresés

- ☐ pontos találat
- ☐ környezet szerint is pontos találat (ICE: In Context Exact)
- ☐ fuzzy találat
- ☐ nyelvileg helyes találat

Módosítás, javítás

- ☐ update-folyamat

Automatikus fordítás

- ☐ automatikus visszakeresés
- ☐ automatikus csere
- ☐ igazi gépi fordítás
- ☐ „intelligens” fordítómemória-funkciók

A fordítómemóriák előnyei és hátrányai

+

- ☐ Konzisztens dokumentumok előállítása: nagy projekteknél elengedhetetlen
- ☐ Gyorsítja a folyamatot: egyszer kell fordítani
- ☐ Pénzt takarít meg: a fordítóiroda tudja, mit tartalmaz az induló memória, illetve meg tudja számolni, hogy mit kellett „igazán” fordítani
- ☐ Egyetlen projekten belül is gazdaságos: menet közben segít, nemcsak a következő projekt indulásakor

–

- ☐ Azon alapul, hogy a mondatok „újrafelhasználhatók”
- ☐ A ma szokásos fordítói munkafolyamat átszervezésével jár
- ☐ Vannak dokumentumformátumok, melyeket még mindig nem támogatnak közvetlenül
- ☐ Idő, míg valaki megtanulja és rászokik
- ☐ Szabadúszó fordítóknak drága és sokszor bonyolult
- ☐ Minden bemenetnek elektronikusnak kell lennie
- ☐ A korábbi fordítások is átalakítandók a fordítómemóriának megfelelő alakúra
- ☐ A hibák is jobban szaporodhatnak a segítségével

Keresés a fordítómemóriában

- ☐ Minden TM-beli mondathoz egy numerikus értéket generálnak (fuzzy kulcs) és egy speciális indexfájlban tárolják egy egyértelmű azonosítóval
- ☐ Kereséskor kiszámítják a keresendő mondat fuzzy-indexét és a legkisebb távolságú fuzzy-indexfájlbeli mondat lesz a találat
- ☐ Indexfájl: egyfajta térkép, ahol egy adott ponttól egy bizonyos sugáron (a megengedett maximális „fuzzyságon”) belül keresünk
- ☐ A fuzzykulcs általában valamilyen n -gram (leggyakrabban a trigram-) technológián alapul
- ☐ Normalizálás: a hosszú mondatok kis helyen való tárolását meg kell oldani (kisbetűsítés, nem alfabetikus karakterek, pl. betűközök eltávolítása stb.): erre a minél kisebb karakterhalmazzal való kalkulálás miatt van szükség

Füzér-összehasonlítás

- ❑ A füzér-összehasonlítás általában a Levenshtein-távolsággal történik (minimális számú törlés, beszúrás és helyettesítés)
- ❑ Dinamikusan épül egy mátrix, melynek minden pozícióján az adott részfüzérék Levenshtein-távolsága áll, pl.

	l	o	v	e	s
h	1	2	3	4	5
a	2	2	3	4	5
t	3	3	3	4	5
e	4	4	4	3	4
s	5	5	5	4	3

- ❑ Értelemszerűen mondatokra is alkalmazható, nemcsak szavakra
- ❑ Az algoritmus optimalizálható azzal, hogy csak az első különböző karaktertől készítjük el (mindkét irányban, tehát hátulról is):

	l	o	v
h	1	2	3
a	2	2	3
t	3	3	3

A fűzér-összehasonlítás eredménye

- ☐ Két hasonló mondat:

Directive 77/93/EEC is hereby amended as indicated in the Annex to the Directive.

Directive 83/107/EEC is hereby amended as indicated in the Annex to the Directive.

- ☐ Az eredeti mátrix $82 \times 83 = 6806$ cellát tartalmazna, míg a redukált mátrix csak $5 \times 6 = 30$ cellát (ami kevesebb, mint fél százalék)
- ☐ Súlyozhatjuk a karaktermanipulációkat is:
 - nagybetű–kisbetű-helyettesítés: 1
 - diakritikus jelek cseréje: 2
 - egyéb csere: 4
 - TM-beli karakter törlési súlya < egy beszúrás súlya

TMX

A TMX-et két szinten definiálják:

1. fordítási egység (Translation Unit: <tu>)
2. egy alacsonyabb szintű meta-markup (Segment: <seg>)

A TMX két implementációs szintje:

1.szint (Plain Text Only):

az adatok a <seg> elemeken belül sima szövegek
Content Markup nélkül, ami szoftverüzenetek
fordításakor elég, formázott dokumentumoknál nem

2. szint (Content Markup):

formázott dokumentumokra (ld. SRX)

Egy TMX-példa

```
<?xml version="1.0"?> <!-- Example of TMX document -->
<tmx version="1.4">
<header creationtool="XYZTool" creationtoolversion="1.01-023" datatype="PlainText" segtype="sentence"
adminlang="en-us" srclang="EN" o-tmf="ABCTransMem" creationdate="20020101T163812Z" creationid="ThomasJ"
changedate="20020413T023401Z" changeid="Amity" o-encoding="iso-8859-1" > <note>This is a note at document
level.</note> <prop type="RTFPreamble">{\rtf1\ansi\tag etc...\fonttbl}</prop> <ude name="MacRoman"
base="Macintosh"> <map unicode="#xF8FF" code="#xF0" ent="Apple_logo" subst="[Apple]"/> </ude>
</header>
<body>
<tu tuid="0001" datatype="Text" usagecount="2" lastusedate="19970314T023401Z" >
<prop type="x-Domain">Computing</prop>
<prop type="x-Project">P&#x00E6;gasus</prop>
<tuv xml:lang="EN" creationdate="19970212T153400Z" creationid="BobW" >
<seg>data (with a non-standard character: &#xF8FF;).</seg>
</tuv>
<tuv xml:lang="FR-CA" creationdate="19970309T021145Z" creationid="BobW" changedate="19970314T023401Z"
changeid="ManonD" >
<prop type="Origin">MT</prop>
<seg>donn&#xE9;es (avec un caract&#xE8;re non standard: &#xF8FF;).</seg>
</tuv>
</tu>
<tu tuid="0002" srclang="*all*" >
<prop type="Domain">Cooking</prop>
<tuv xml:lang="EN"> <seg>menu</seg> </tuv>
<tuv xml:lang="FR-CA"> <seg>menu</seg> </tuv>
<tuv xml:lang="FR-FR"> <seg>menu</seg> </tuv>
</tu>
</body>
</tmx>
```

Fordítómemóriákkal kapcsolatos szabványok

TMX [Translation Memory Exchange format](#). This standard enables the interchange of translation memories between translation suppliers. TMX has been adopted by the translation community as the best way of importing and exporting translation memories. The current version is 1.4b - it allows for the recreation of the original source and target documents from the TMX data.

TBX [Termbase Exchange format](#). This standard allows for the interchange of terminology data including detailed lexical information. The framework for TBX is provided by two ISO 12620, ISO 12200 and ISO Committee Draft 16642, known as TMF or Terminological Markup Framework. ISO 12620 provides an inventory of well-defined “data categories” with standardized names that function as data element types or as predefined values. ISO 12200 (also known as MARTIF) provides the basis for the core structure of TBX. TMF includes a structural metamodel for Terminology Markup Languages in general, regardless of which XML style of representation is used.

SRX [Segmentation Rules Exchange format](#). SRX is intended to enhance the TMX standard so that translation memory data that is exchanged between applications can be used more effectively. The ability to specify the segmentation rules that were used in the previous translation increases the leveraging that can be achieved.

GMX [GILT Metrics](#). GILT stands for (Globalization, Internationalization, Localization, and Translation). The GILT Metrics standard comprises three parts: GMX-V for volume metrics, GMX-C for complexity metrics and GMX-Q for quality metrics. The proposed GILT Metrics standard is tasked with quantifying the workload and quality requirements for any given GILT task.

OLIF [Open Lexicon Interchange Format](#). OLIF is an open, XML-compliant standard for the exchange of terminological and lexical data. Although originally intended as a means for the exchange of lexical data between proprietary machine translation lexicons, it has evolved into a more general standard for terminology exchange.

XLIFF [XML Localization Interchange File Format](#). It is intended to provide a single interchange file format that can be understood by any localization provider. XLIFF is the preferred way of exchanging data in XML format in the translation industry.

TransWS [Translation Web Services](#). TransWS specifies the calls needed to use Web services for the submission and retrieval of files and messages relating to localization projects. It is intended as a detailed framework for the automation of much of the current localization process by the use of Web Services.

xml:tm [xml:tm](#) This is a radical new approach to translation memory based on the concept of text memory which composes author and translation memory. One of the first significant advances in translation memory technology since its inception.



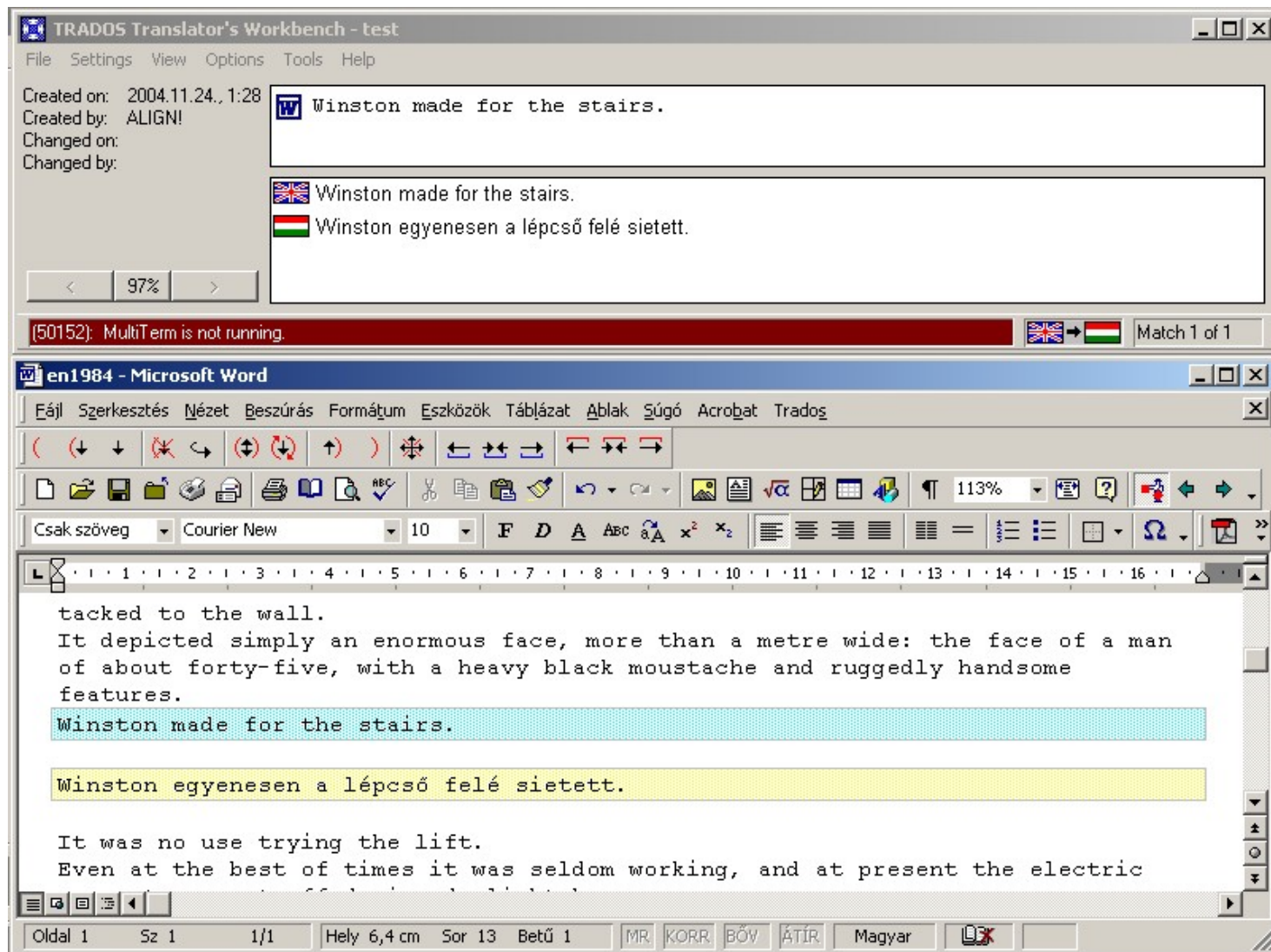
Fordítómemóriák a gyakorlatban



Fontosabb fordítómemóriák

Across Language Server	Windows	http://www.across.net/en/
Déjà Vu	Windows	http://www.atril.com/content/last-update
GlobalSight	Linux, Win	http://www.globalsight.com/
memoQ	Windows	https://www.memoq.com/en/
Memsource	Web, Win/Mac/Linux	https://www.memsource.com/en
MetaTaxis	MS Word add-in	http://www.metataxis.com/
OmegaT	Java	http://omegat.org/
Open Language Tools	Java	https://open-language-tools.java.net/
Poedit	Cross-platform	https://poedit.net/
Pootle	Cross-platform	http://www.translate.org.za/
SDL Trados	Windows	http://www.translationzone.com/
SmartCAT	Cross-platform	https://ru.smartcat.ai/
Star Transit	Windows	https://www.star-group.net/en/products/transit.html
Virtaal	Cross-platform	http://virtaal.translatehouse.org/
Wordfast Classic	MS Word (Win/Mac)	http://www.wordfast.com/
Wordfast Pro	Java	http://www.wordfast.com/

Integrált fordítási környezet (TRADOS)



Önálló fordítási környezet (MemoQ)

MemoQ - MemoQ sample project

Project Edit Translation Format View Tools Help

Project Manager memoq

	Source	Target	Progress	Status
37.	Benefits	Előnyök	100%	✓
38.	MemoQ, the world's most advanced <i>integrated localization environment</i> optimizes the productivity of translation and provides the quickest ROI ever.	A MemoQ, a világ legfejlettebb <i>integrált fordítási környezete</i> , optimalizálja a fordítás és lokalizáció hatékonyságát és grantálja a befektetés lehető legyorsabb megtérülését.	0%	✗
39.	In the MemoQ ILE, all translation resources are only a mouse-click away resulting in an ergonomomy never seen before.		0%	✗
40.	During the design of MemoQ we had only one thing in mind: the optimal environment for translators and project managers.		0%	✗
41.	Kilgray worked hard to avoid all useless keystrokes, file operations, and the result is a one-stop environment with every tool integrated in the most transparent manner.		0%	✗
42.	MemoQ is an easy-to-use concept-oriented system.		0%	✗
43.	An average translator can learn to use 90% of the program's functionality within just two hours.		0%	✗
44.	The remaining 10% is the domain of project managers and clients: you can achieve enormous savings by customizing		0%	✗

Translation Results

Source	Target
MemoQ, the world's most advanc...	1 A MemoQ, a világ legfejlettebb integrá...
MemoQ	2 MemoQ
integrated localization environment	3 integrált fordítási környezet

MemoQ, the world's most advanced integrated localization environment optimizes the productivity of translation and provides the quickest ROI ever.

MemoQ, the world's most advanced integrated localization environment optimizes the productivity of translation *and localization* and provides the quickest ROI ever.

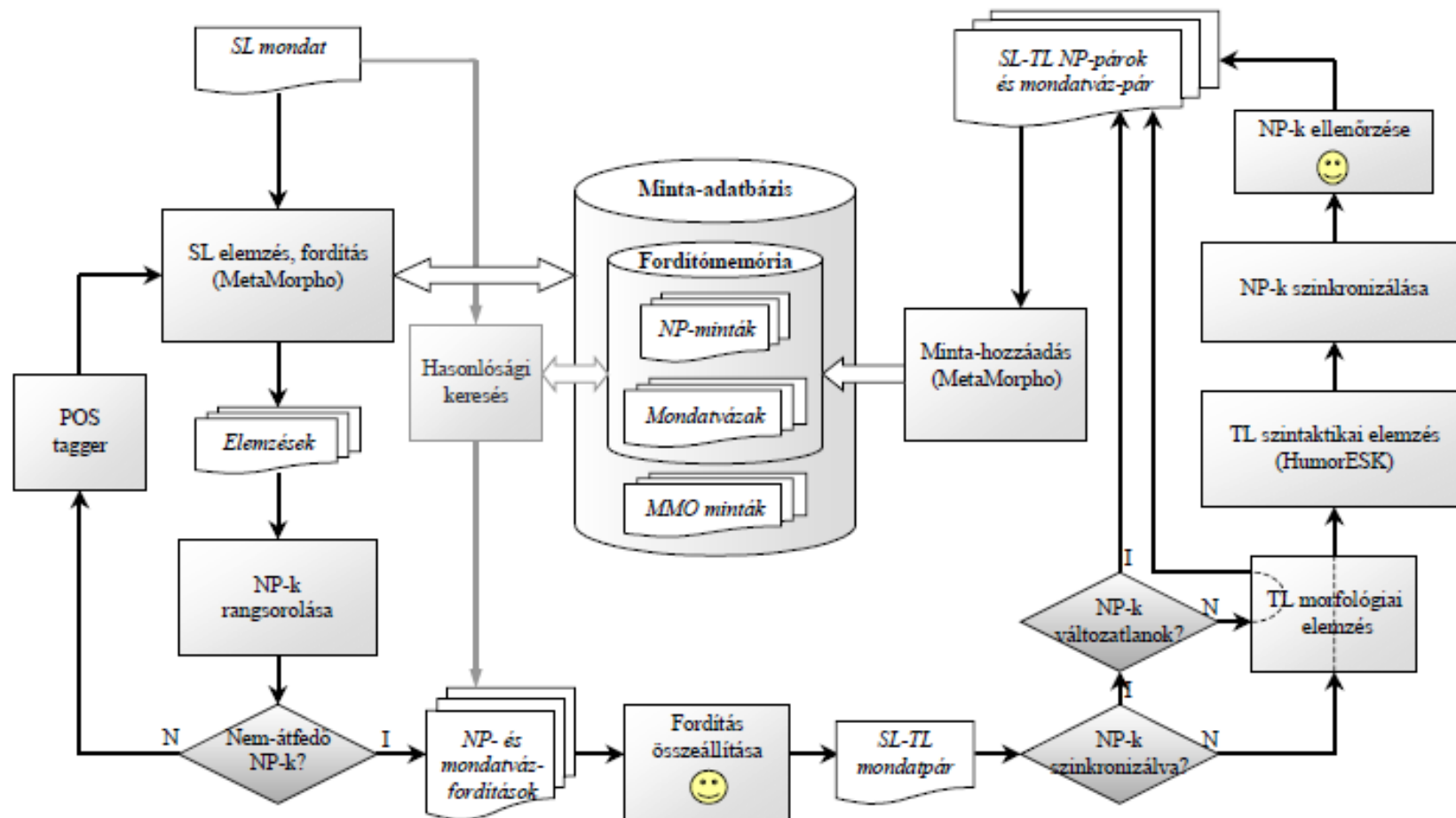
A MemoQ, a világ legfejlettebb integrált fordítási környezete, optimalizálja a fordítás és lokalizáció hatékonyságát és grantálja a befektetés lehető legyorsabb megtérülését.

Pro 101/123 **CLI Famous Client**
Dom IT **Sub CAT**

Aligned Corrected
 MemoQ gabor
 5/3/2006 1:04:18 PM **89%**

Project status: 1% Document status: 1% Conf: 1 Ed: 1 Empty: 47 Pre: 12 Frag: 0 ✗ Errors: 11 Ready

Az intelligens fordítómemória felé





Automatikus kiértékelés

A fordítások hasonlósága: az automatikus kiértékelés alapjai

- ☐ A kézi kiértékelés lassú, és szubjektív
- ☐ Vannak statisztikai módszereink...
- ☐ Nagy tesztkorpuszon közelít az „objektívhez”
- ☐ 2002, IBM: Bilingual Evaluation Understudy (BLEU)
- ☐ A lényeg:
átfedés a referenciafordításokkal
- ☐ Módosítások kellenének:
morfológia, szinonimák, ...

A fordítási minőség mérésének alapjai

Referenciafordítás:

The gunman was shot to death by the police .

BLEU-módszer:

The gunman was shot kill .

Wounded police jaya of

The gunman was shot dead by the police .

The gunman arrested by police kill .

The gunmen were killed .

The gunman was shot to death by the police .

The ringer is killed by the police .

Police killed the gunman .

zöld = 4-gram illeszkedik (jó!)

piros = nem-illeszkedő szó (rossz!)



Bevezetés a nyelvtechnológiába

9.

Gépi fordítás

Ezt olvassuk a gépi fordításról...

- ❑ A gépi fordítással foglalkozni kidobott idő, mert egy gép sosem fog Shakespeare-t fordítani.
- ❑ Állítólag volt egyszer egy gépi fordító rendszer, amelyik arra az angol mondatra, hogy *The spirit is willing, but the flesh is weak*, azt fordította oroszra, hogy: *A vodka jó, de a hús gyenge*. Ebből jól látszik, hogy a gépi fordítás haszontalan dolog.
- ❑ Általánosságban, a gépi fordítás minősége annyira gyenge, hogy semmire sem használható a gyakorlatban.
- ❑ A gépi fordítás elveszi a fordítók munkáját.
- ❑ A japánok kifejlesztettek egy rendszert, amihez csak bele kell beszélni a telefonba, és lefordítja, amit mondunk, és ezt a másik angolul hallja.
- ❑ Van egy dél-amerikai indián nyelv, amelyiknek a szerkezete logikailag annyira tökéletes, hogy nagyon könnyen lehet gépi fordító rendszert csinálni a segítségével.

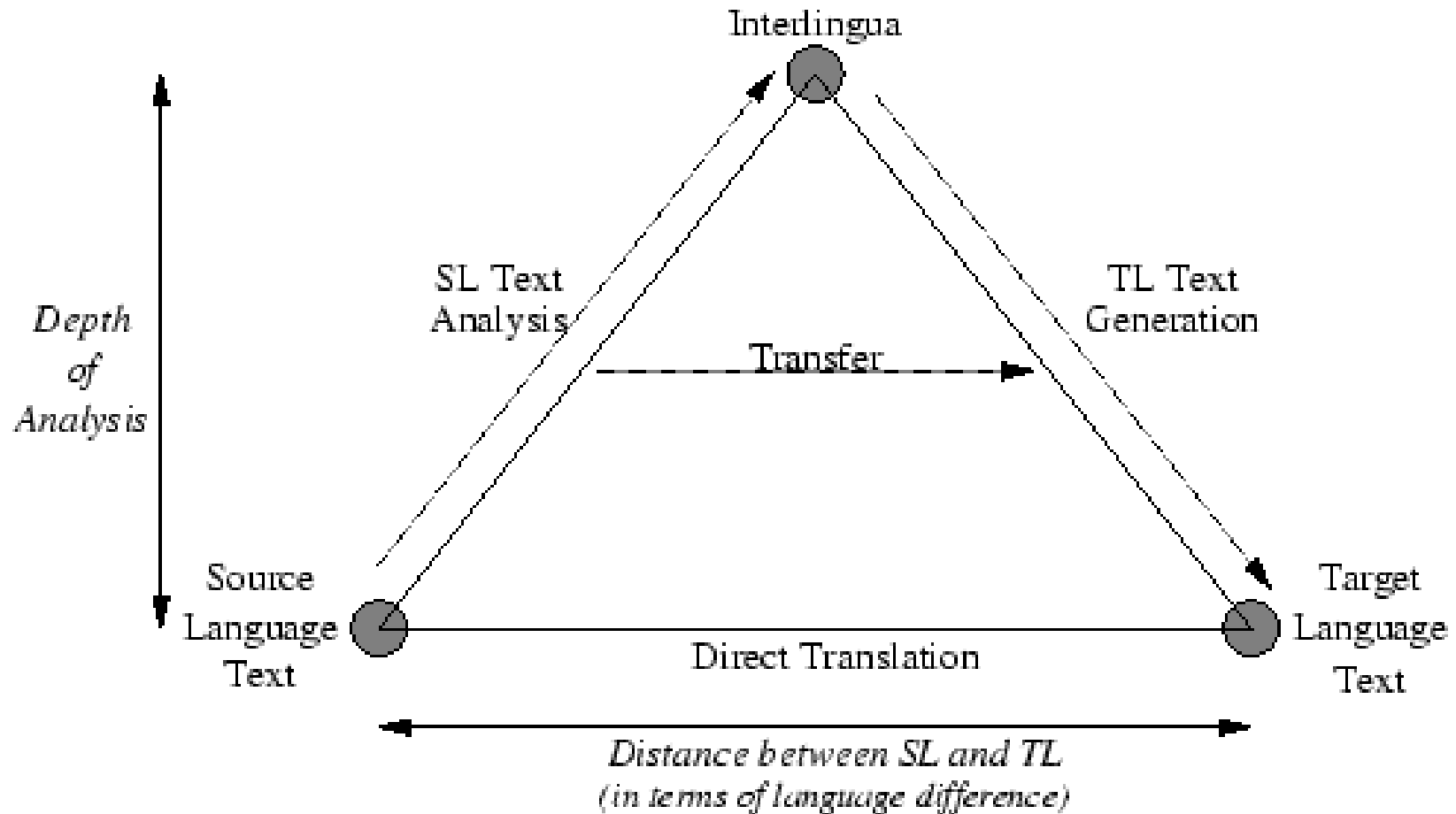
Ez igaz a gépi fordításról...

- ☐ Igaz, hogy a minőség még sokszor hagy kívánnivalót, de a sebesség igen értékes tulajdonsága a GF-rendszereknek.
- ☐ Bizonyos körülmények közt a GF minőségi fordítást is képes produkálni: a METEO 4%-nál kevesebb kimenete igényel emberi korrekciót.
- ☐ A GF hasznos dolog. A fenti METEO rendszer például napi használatban van 1977 óta. 1990 óta 45.000 szót fordít gyakorlatilag minden nap.
- ☐ A GF nem fenyegeti a fordítók munkáját.
- ☐ A beszéd-beszéd GF továbbra is kutatási téma. Általánosságban is igaz: sok kutatási téma van még a GF területén.
- ☐ A GF-rendszerek fejlesztése időigényes munka. A gyakorlatban egy GF-rendszer csak akkor tud megvalósulni, ha a felhasználó komoly munkát fektet bele az egyedi felépítésébe.
- ☐ A statisztikai és a neurális rendszerek készítéséhez nem kell nyelvtudás (csak a teszteléséhez)



Szabályalapú fordítás

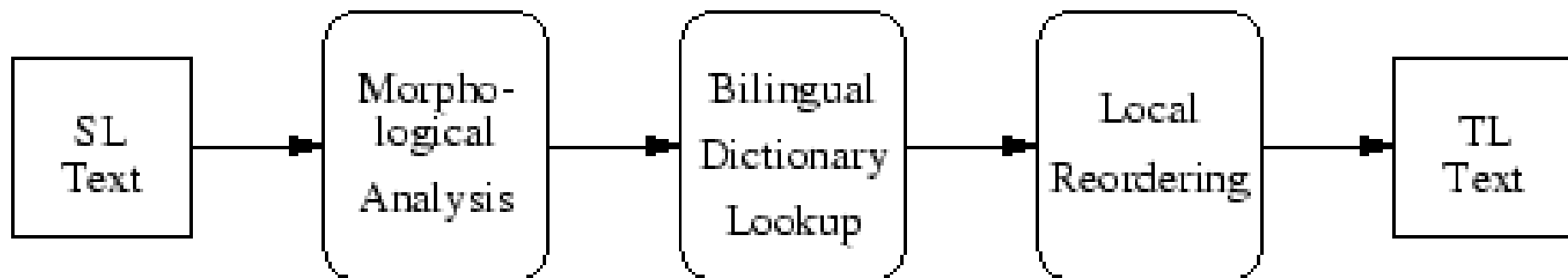
Szabályalapú fordítási stratégiák





Közvetlen fordítás

Közvetlen fordítás



A közvetlen fordítás lépései

Stage	Action
1.	morphological analysis
2.	lexical transfer of content words
3.	various work relating to prepositions
4.	SVO rearrangements
5.	miscellany
6.	morphological generation

Input:	watashihatsukuenouenopenwojonniageta.
After stage 1:	watashi ha tsukue no ue no pen wo jon ni ageru PAST.
After stage 2:	I ha desk no ue no pen wo John ni give PAST.
After stage 3:	I ha pen on desk wo John to give PAST.
After stage 4:	I give PAST pen on desk John to.
After stage 5:	I give PAST the pen on the desk to John.
After stage 6:	I gave the pen on the desk to John.



METEO

METRO TORONTO.

LE GRAND TORONTO.

TODAY:

MAINLY CLOUDY AND COLD WITH
OCCASIONAL FLURRIES.
BRISK WESTERLY WINDS TO 50 KM/H.
HIGH NEAR MINUS 7.

AUJOURD HUI:

GENERALEMENT NUAGEUX ET FROID
AVEC QUELQUES AVERSES DE NIEGE.
VENTS VIFS D'OUEST A 50 KM/H.
MAXIMUM D'ENVIRON MOINS 7.

TONIGHT:

VARIABLE CLOUDINESS.
ISOLATED FLURRIES.
DIMINISHING WINDS.
LOW NEAR MINUS 15.

CETTE NUIT:

CIEL VARIABLE.
AVERSES DE NIEGE EPARSEES.
AFFAIBLISSEMENT DES VENTS.
MINIMUM D'ENVIRON MOINS 15.

FRIDAY:

VARIABLE CLOUDINESS.
HIGH NEAR MINUS 6.

VENDREDI:

CIEL VARIABLE.
MAXIMUM D'ENVIRON MOINS 6.

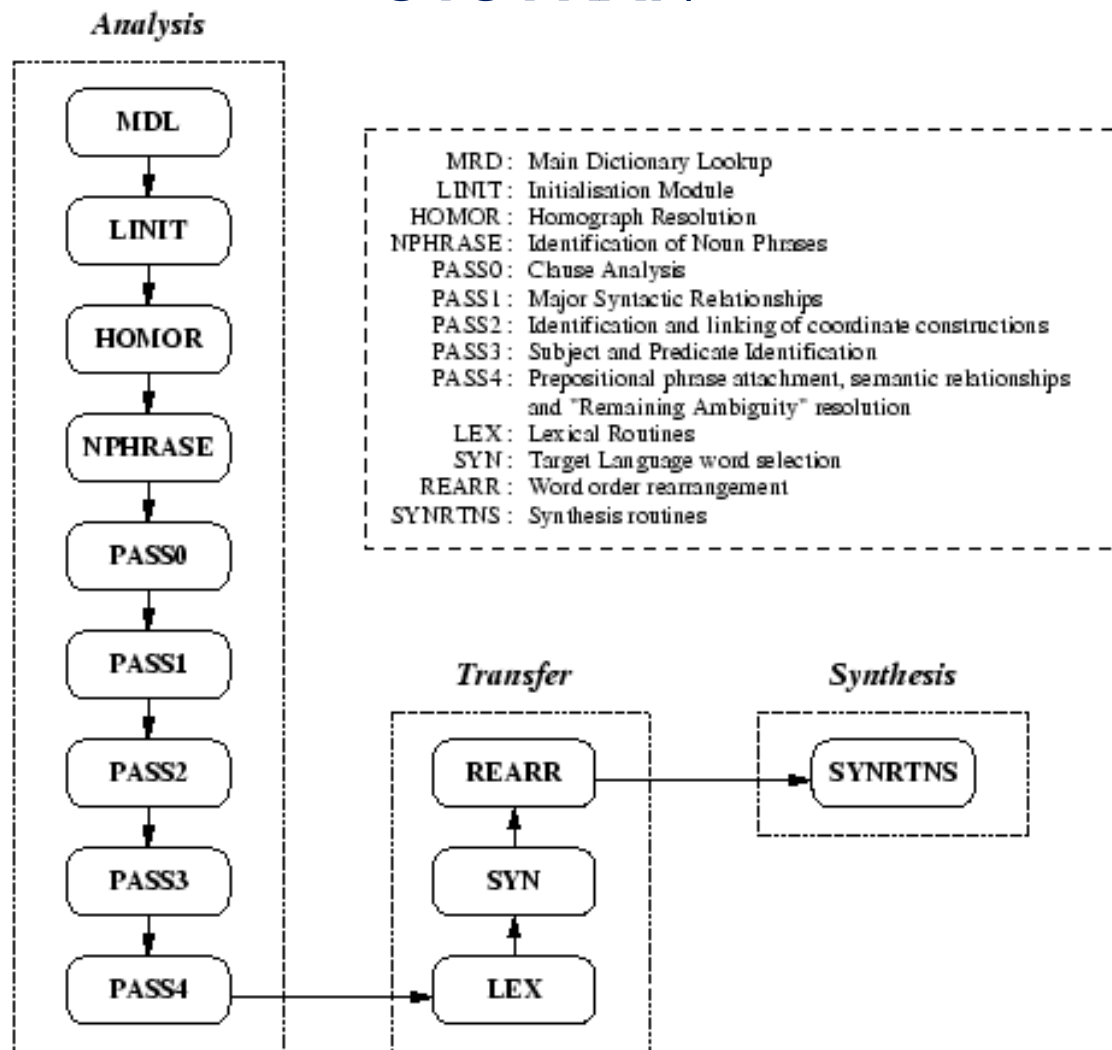
A közvetlen fordítás jellemzői

- ☐ Soha nincs kész semmilyen állapot sem, csak a végén
- ☐ Rengeteg lépésből áll
- ☐ Minden lépés egy adott nyelvi jelenséget dolgoz fel
- ☐ A legfontosabb eszköz a kétnyelvű szótár
- ☐ Komoly probléma a helyes kimeneti szórend és a megfelelő toldalékolás előállítása
- ☐ Ezt a feladatot átrendezésnek hívják

A közvetlen fordítás modulárisává válása

- ❑ Átrendezés: vonzatjellegű információk és általános szórendi kérdések „keveréke”
- ❑ Az átrendezésre jellemző a modularitás teljes hiánya
- ❑ Generálás: a célnyelv lexikális elemeinek morfológiailag megfelelő szósorozata mint kimenet
- ❑ A Systran és a Meteo ilyen átrendezéseit írták átkésőbb transzferrendszerre
- ❑ Az átrendezési műveletek tehát a később megjelenő transzfer és a generálás kombinációjának is tekinthetők

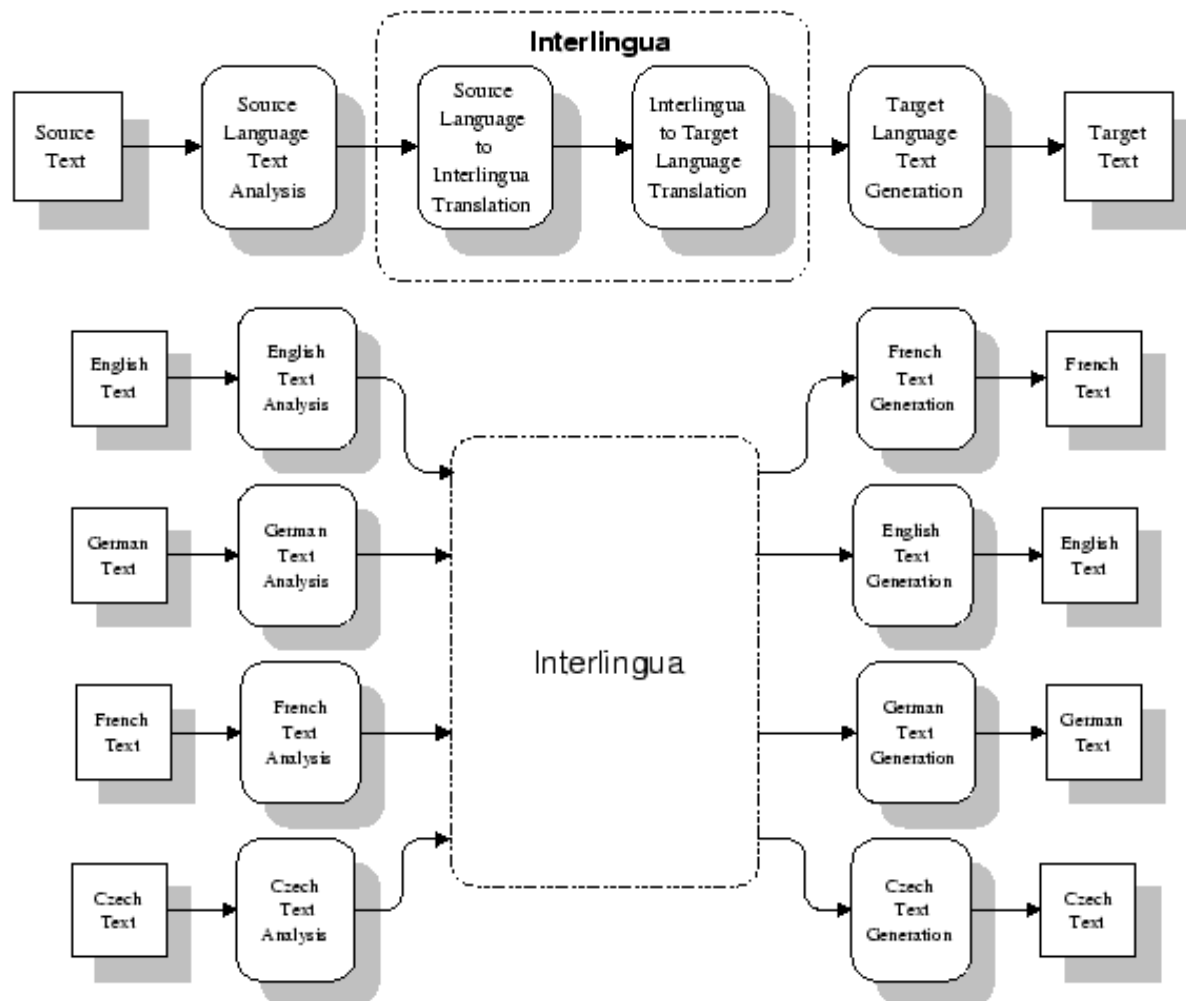
SYSTRAN





Közvetítőnyelves fordítás

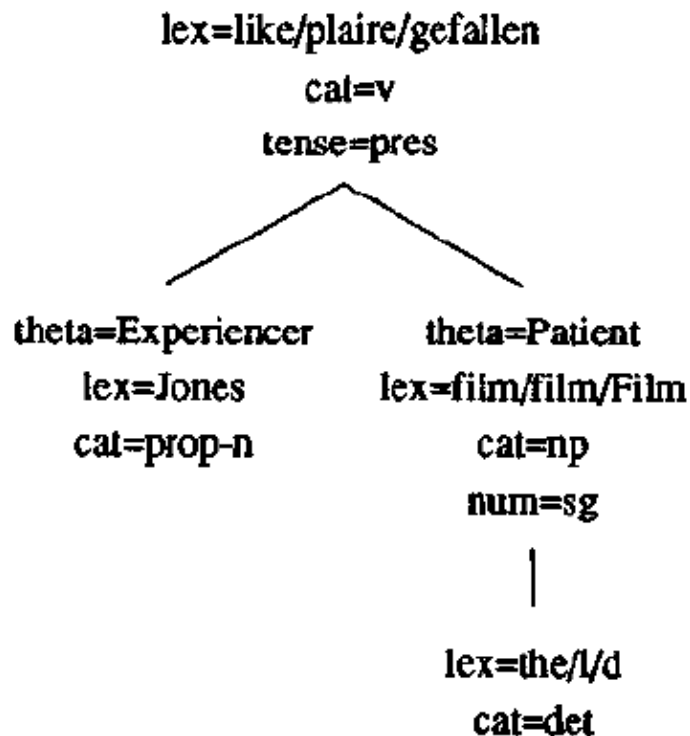
A közvetítőnyelves fordítás



A közvetítőnyelves fordítás jellemzői

- ☐ Nem létezik mindent kielégítő közvetítőnyelv
- ☐ Kísérletek: a logikai formalizmusoktól a formalizált eszperantóig
- ☐ A soknyelvű rendszerek esetén nagyon „vonzó”
- ☐ $n > 3$ esetén: $n \cdot (n-1) > 2 \cdot n$
- ☐ Végül is a két fordítási lépés nem más, mint egy-egy közvetlen fordítás

Példa a közvetítőnyelves fordításra

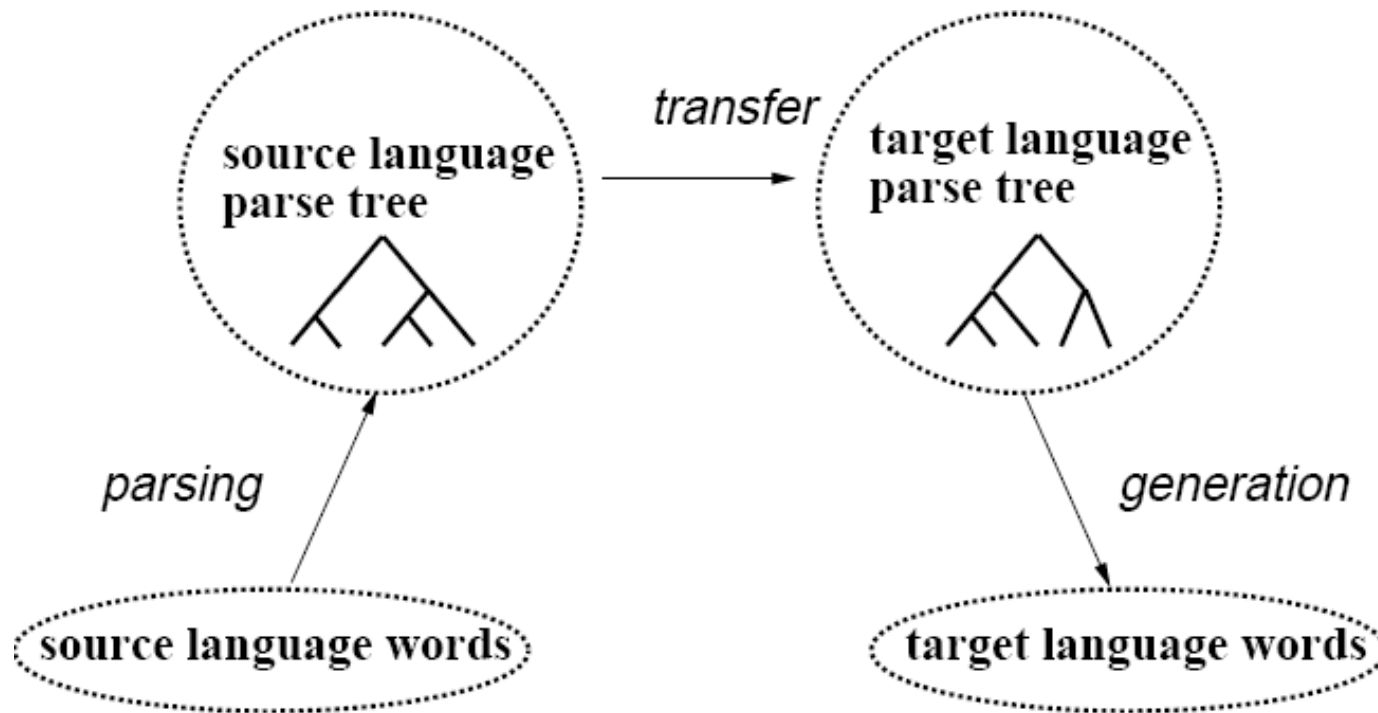


Jones likes the film. Le film plaît à Jones. Das Film gefällt Jones.



Transzfer fordítás

A transzfer fordítás menete

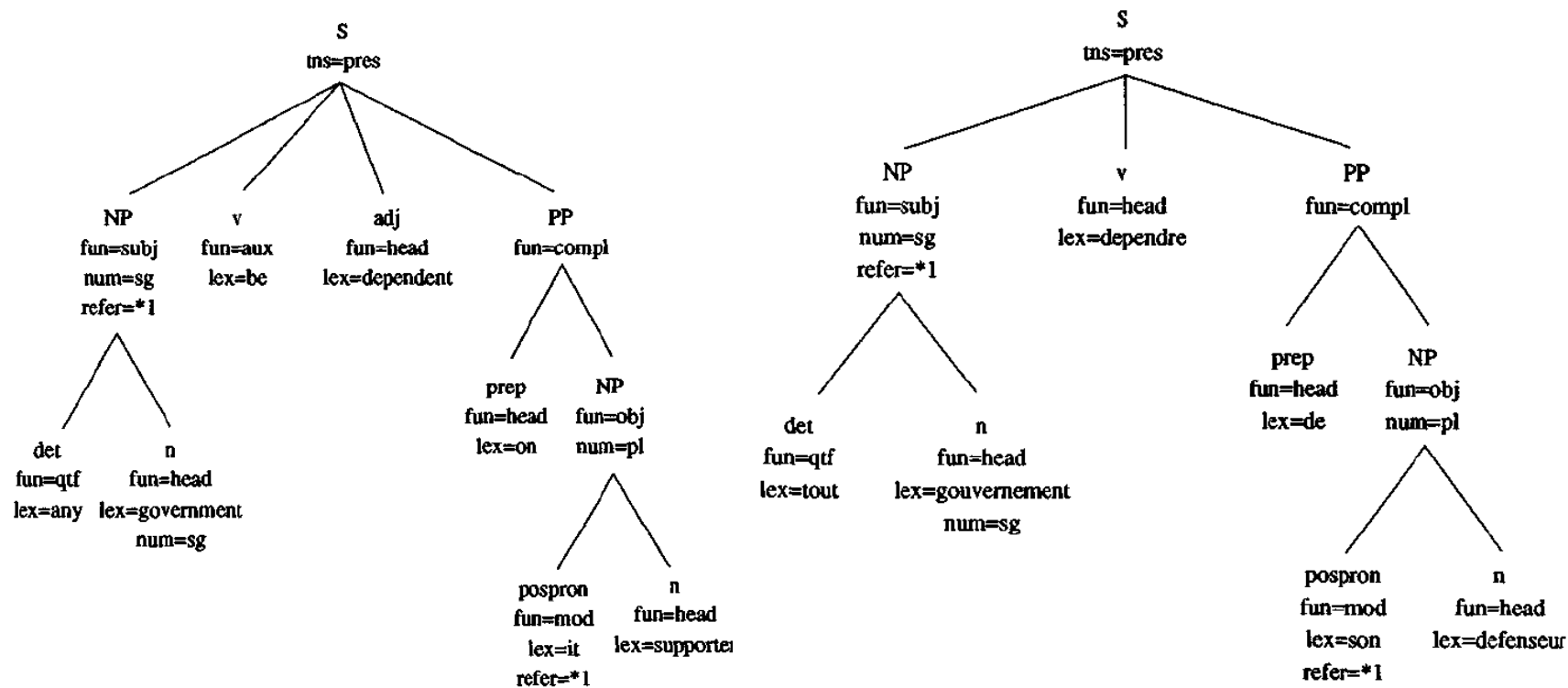


A transzfer fordítás menete és típusai

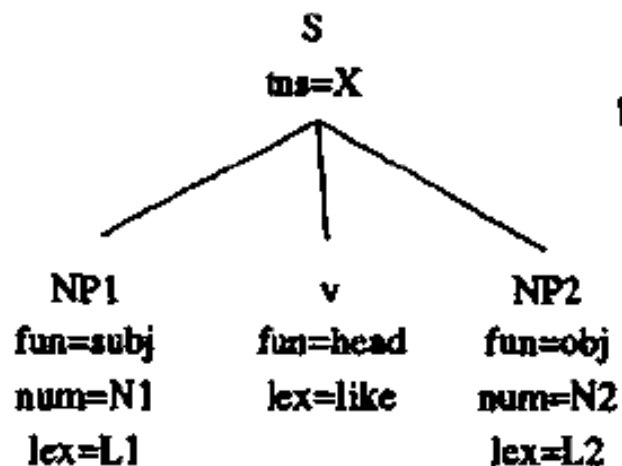
- ☐ Morfológiai elemzés
- ☐ Egyértelműsítés (POS, WSD)
- ☐ Lexikális transzfer: kétnyelvű szótárakkal
- ☐ Szerkezeti transzfer: frázisok, csonkok
- ☐ Morfológiai generálás

- ☐ Szintaktikai/felszíni transzfer: hasonló struktúrájú, különösen rokon nyelvek között
- ☐ Szemantikus/mély transzfer: távolabbi rokonságban lévő nyelvek között

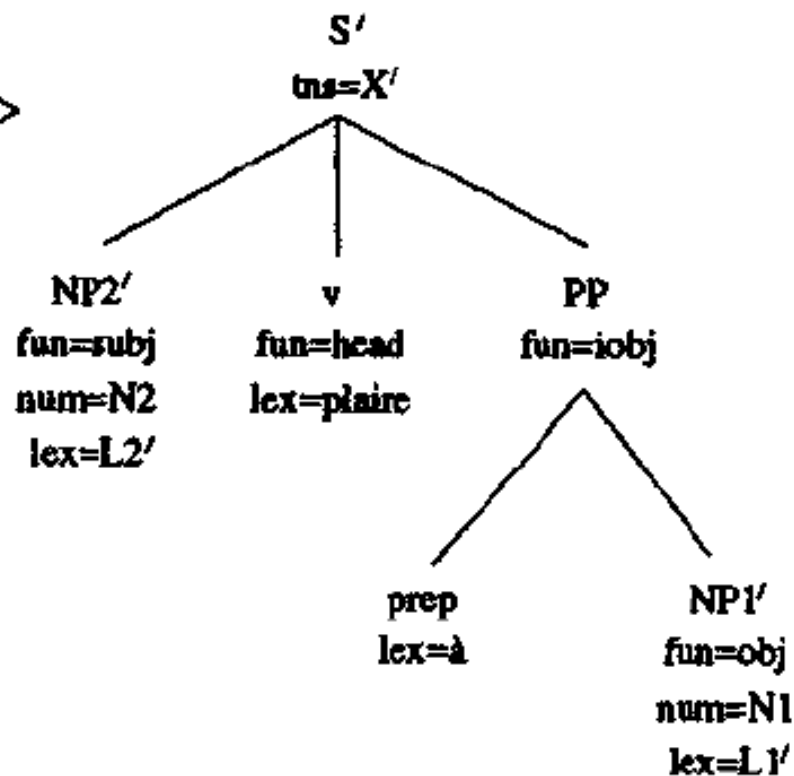
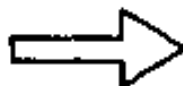
Példa a lexikális transzferre



Példa a szerkezeti transzferre



Jones likes the film.

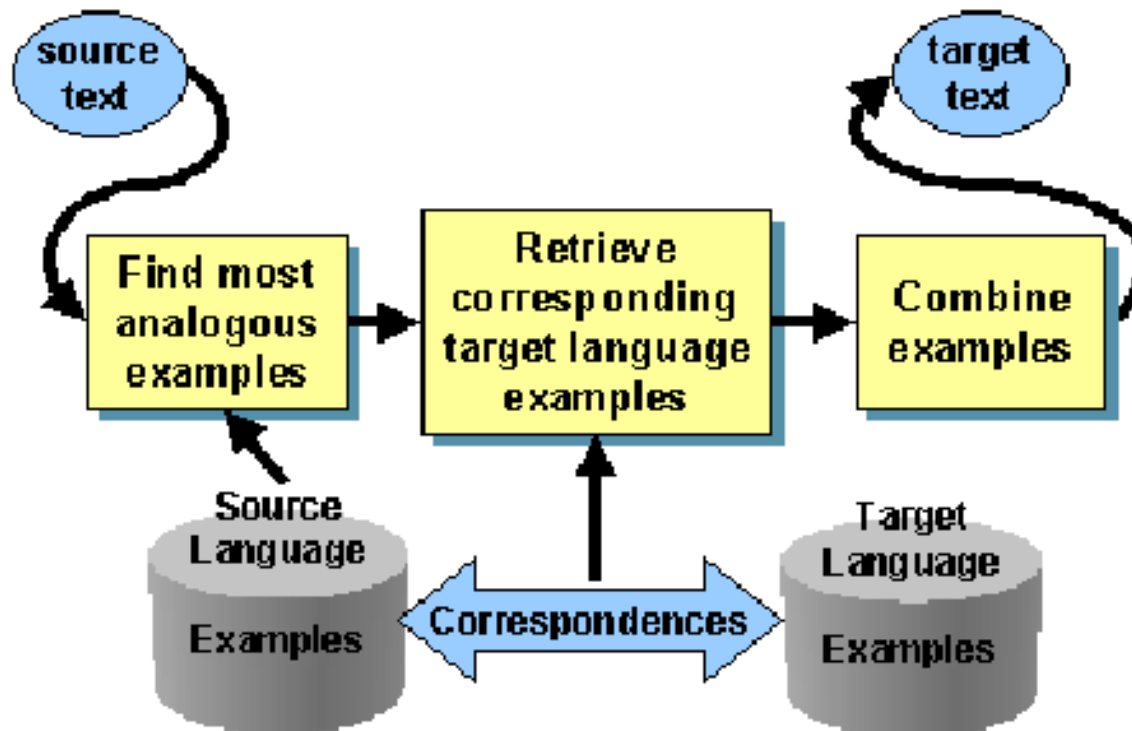


Le film plaît à Jones.



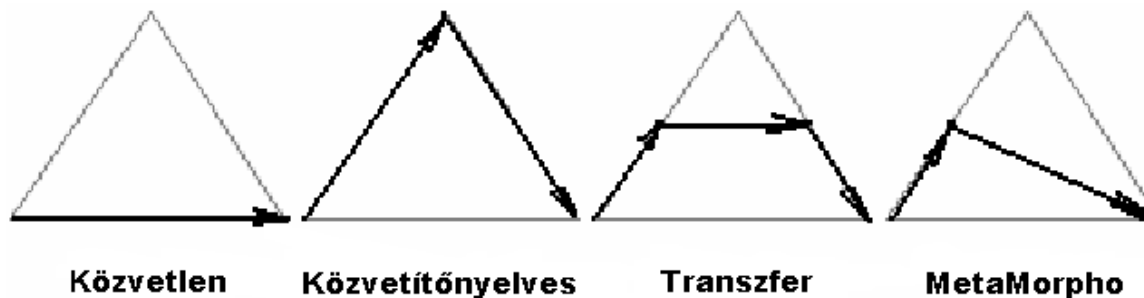
Más - de nem statisztikai - megoldások

Példa-alapú fordítás



MetaMorpho: egy hibrid szabály/példa-alapú fordító

- ❑ 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:





A gépi fordítás néhány tipikus nehézségéről



Lexikai többértelműségek kezelése

One minute has sixty seconds.

There is only minute difference between these pictures.

Part of the job of a secretary is to minute meetings.



PP-kapcsolások kezelése

The man saw the girl with the telescope.



Strukturális többértelműségek kezelése

Time flies like an arrow.



Többszavas kifejezések kezelése

Flying planes can be dangerous.



Idiómák kezelése

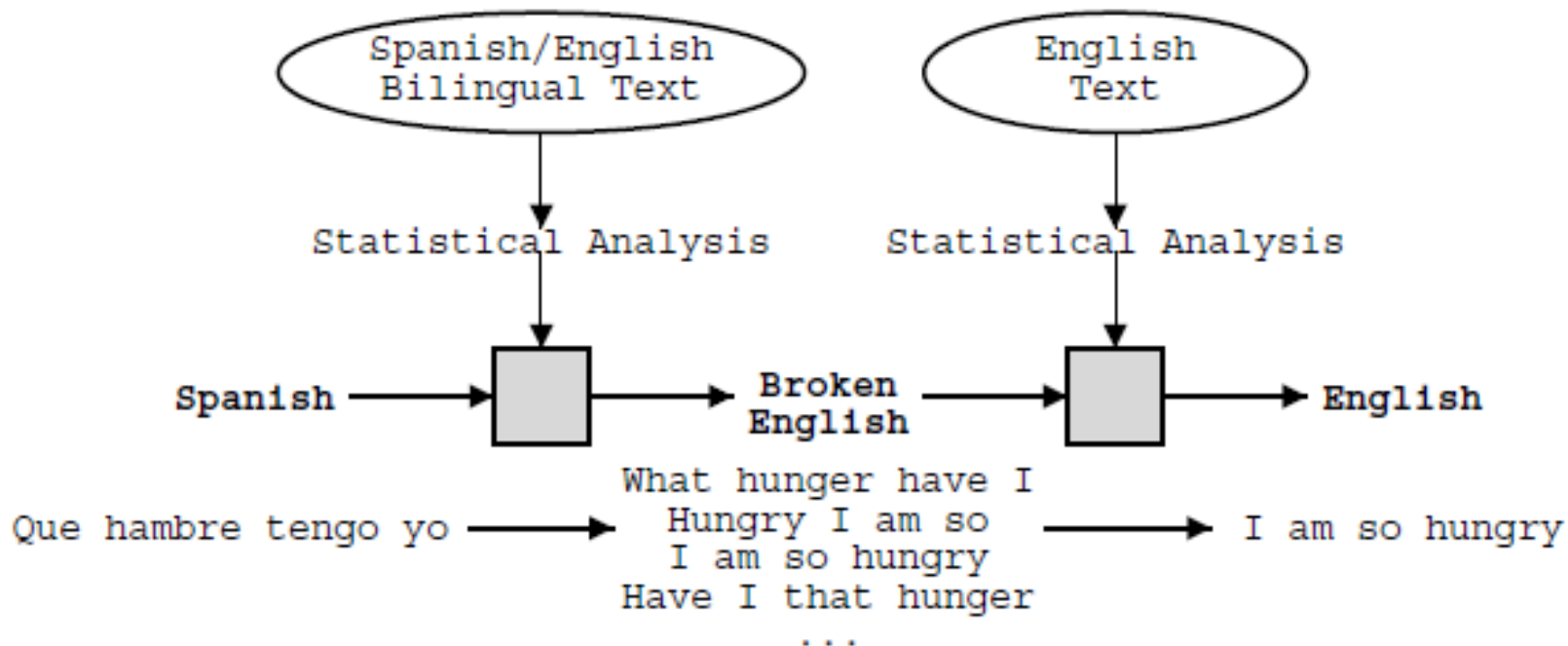
The country has gone to the dogs.



Statisztikai fordítás

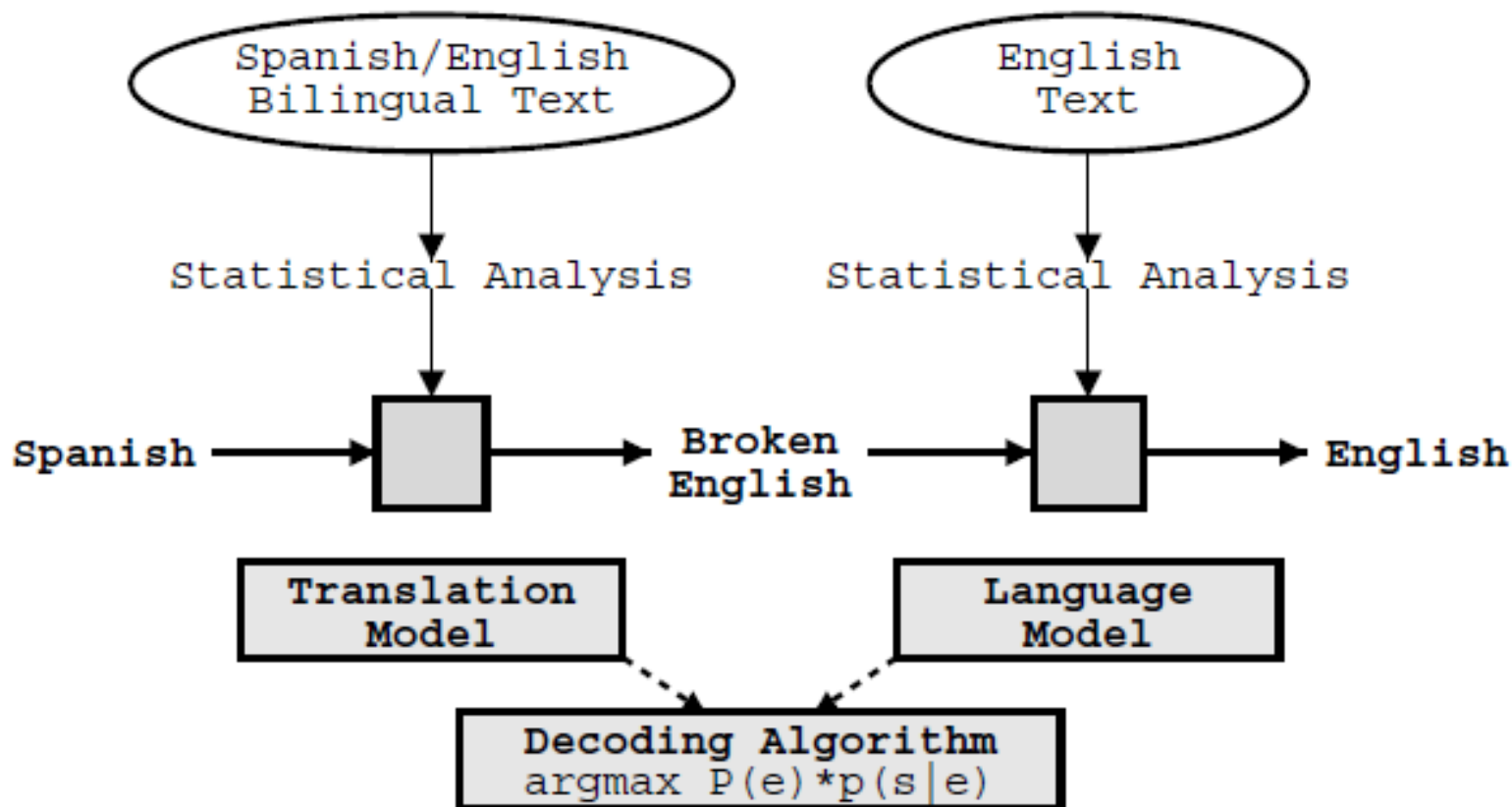
A statisztikai fordítás alapjai

(pl. spanyolról angolra fordítás)



A statisztikai fordítás alapjai (2)

(pl. spanyolról angolra fordítás)



A három fő modul

☐ Nyelvmodell:

adott egy e angol sztring, hozzá kell rendelni egy $P(e)$ -t

jó angol sztring: magas $P(e)$

rossz angol sztring: alacsony $P(e)$

☐ Fordítási modell:

adott egy $\langle f, e \rangle$ sztringpár, a hozzárendelt $P(f|e)$ valószínűséggel

ha $\langle f, e \rangle$ egymás fordítása: magas $P(f|e)$

ha $\langle f, e \rangle$ nem egymás fordítása: alacsony $P(f|e)$

☐ Dekódoló algoritmus:

nyelvmodell + fordítási modell + bemenő f mondat, amikhez meg kell találni az e fordítást, a $P(e) \cdot P(f|e)$ maximalizálásával

A statisztikai GF lényege

- ❑ A naiv Bayes-tétel:

$$\tilde{e} = \arg \max_{e \in E^*} p(e|f) = \arg \max_{e \in E^*} p(f|e)p(e)$$

- ❑ Az ötlet a beszédfelismerésből jön
- ❑ A legtöbb „minőségi” probléma a korpusz méretéből adódik
- ❑ Csak egy fordítási modell
(feltételes valószínűségi rész) és
egy nyelvmmodell
(teljes valószínűségi rész) kell,
meg a dekóder, ami számol

A fordítási modell

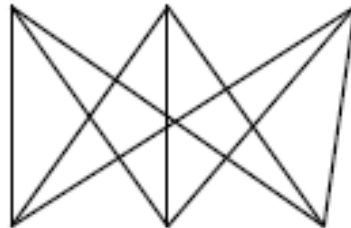
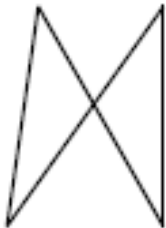
- ☐ A fordítási modell feladata megtalálni a legjobban illeszkedő bemenetet az e (angol) kimenethez
- ☐ $P(f|e)$ tanulását párhuzamos korpuszból lehet csak megoldani
- ☐ Tipikus helyzet: sokszor nincs elég adat a $P(f|e)$ közvetlen becslésére

Hogyan hozunk létre fordítási modellt?

- ☐ EM-algoritmus (expectation maximization)
- ☐ Inicializálás: minden kapcsolat azonos súlyú
- ☐ Valószínűség hozzárendelése a hiányzó adatokhoz
- ☐ Paraméterbecslés a teljes adatokból
- ☐ Iterálás

Inicializálás az EM algoritmusban

... la maison ... la maison blue ... la fleur ...

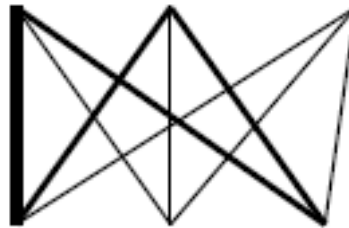


... the house ... the blue house ... the flower ...

Kezdőlépés: minden kapcsolat egyenlő valószínűségű

Iteráció az EM algoritmusban

... la maison ... la maison blue ... la fleur ...

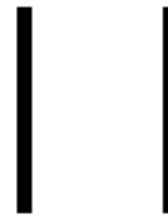


... the house ... the blue house ... the flower ...

Egy iterációs lépés után a modell megtanulja, hogy bizonyos kapcsolatok gyakoribbak, tehát valószínűbbek (pl. a *la* és a *the* közötti)

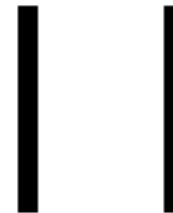
További iterálás az EM algoritmusban

... la maison ... la maison bleu ... la fleur ...



... the house ... the blue house ... the flower ...

... la maison ... la maison bleu ... la fleur ...



... the house ...| the blue house ... the flower ...

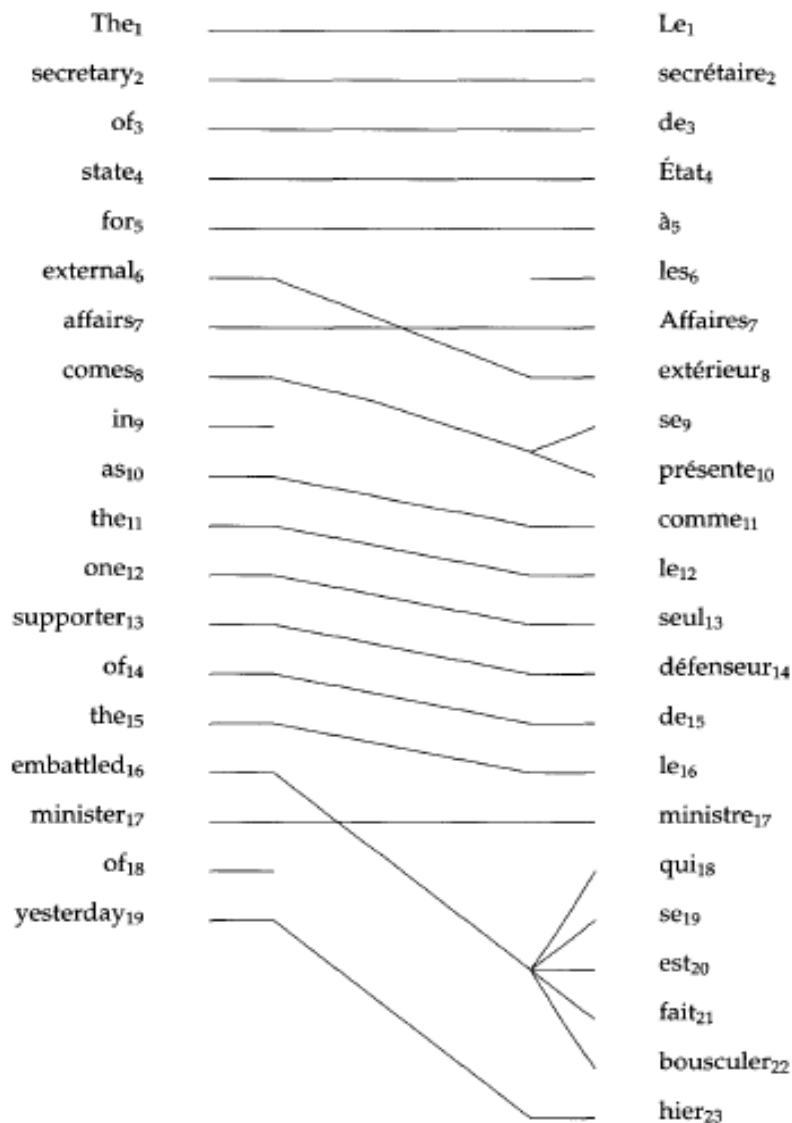
Paraméterbecslés az EM algoritmusban

... la maison ... la maison bleu ... la fleur ...
/ | | X | |
... the house ... the blue house ... the flower ...



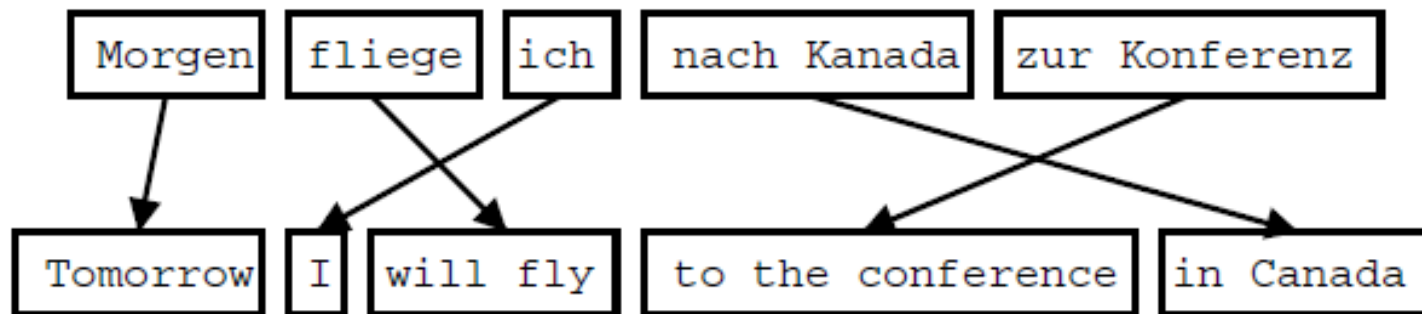
$p(\text{la}|\text{the}) = 0.453$
 $p(\text{le}|\text{the}) = 0.334$
 $p(\text{maison}|\text{house}) = 0.876$
 $p(\text{bleu}|\text{blue}) = 0.563$
...

Működhet-e a szószintű szinkronizálás?



Megjelennek a frázis-alapú fordítási modellek

- ❑ Szószinten nehéz a törlés és a beszúrás, ezért megéri a bemenetet frázisokra szegmentálni
- ❑ Itt a frázis nem nyelvtani fogalom, hanem csak a gyakran előforduló összefüggő szósorozatok
- ❑ Először minden frázist lefordítunk a célnyelvre
- ❑ Ezután a frázisokat átrendezzük:



- ❑ Így nem kompozicionális frázisok is fordíthatók
- ❑ Minél több az adat, annál több hosszú frázis tanulható meg

A nyelvmodell kialakítása

- ☐ Ez azoknak a mondatoknak a halmazából indul ki, amit az adott nyelven mondani szoktak
- ☐ Milyen a „jó” nyelvmodell? Milyen a „jó” nyelv?
- ☐ Trigram-valószínűségekkel szokás közelíteni
- ☐ Pl. $p(\text{witch} | \text{the green}) > p(\text{green} | \text{the witch})$
- ☐ A nyelvmodell kiindulhat a web n-gramjaiból is!

Ha nem elég nagy a korpusz

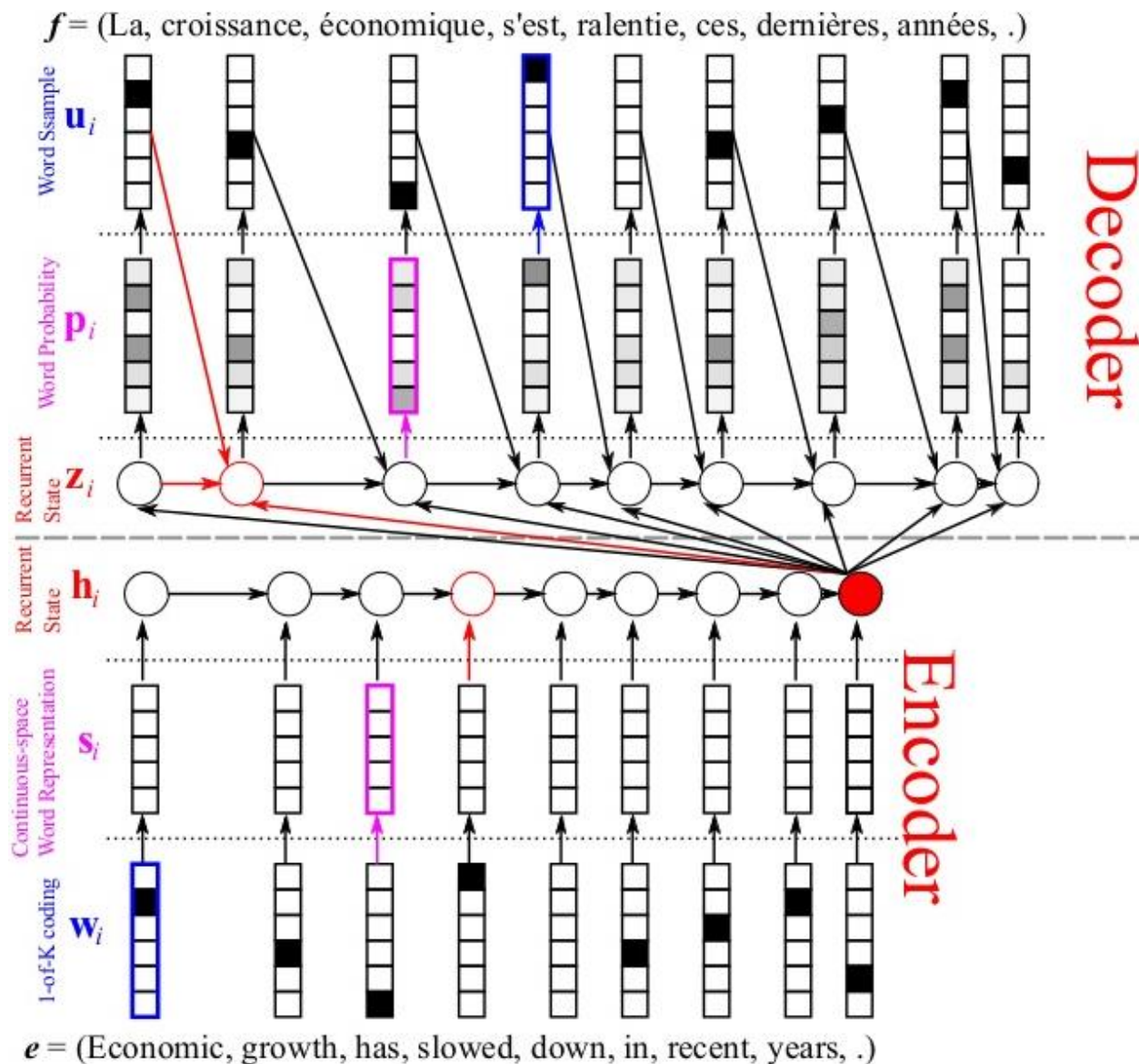
- ☐ Sokszor szükséges ún. simítás a nyelvmodellhez
 - ☐ Ha z sose követte xy -t a szövegeinkben, akkor még mindig megkérdezhetjük, hogy z legalább y -t követte-e?
 - ☐ Ha igen, akkor xyz talán nem olyan rossz
 - ☐ Ha nem követte, még mindig megkérdezhetjük, hogy z elfogadható hétköznapi szó-e?
- Ha nem, akkor xyz tényleg nagyon kis valószínűségű kell legyen!

- ☐ $a = \text{előfordulás}("xyz") / \text{előfordulás}("xy")$ simítva:
 $b(z|x y) =$
 $0.95 * \text{előfordulás}("xyz") / \text{előfordulás}("xy") +$
 $0.04 * \text{előfordulás}("yz") / \text{előfordulás}("z") +$
 $0.008 * \text{előfordulás}("z") / \text{összes-látott-szó} +$
 0.002

A statisztikai GF problémái és a továbblépés

- ☐ Tulajdonnevek: másolás, transliterálás, szótárból újraalkotás
- ☐ Számmal alkotott kifejezések, dátumok, mennyiségek: saját fordítási táblák kellenek
- ☐ Főnévi csoportok általában: ezeket érdemes nyelvileg „elő-elemezni”
- ☐ A hibrid rendszerek előtérbe kerülnek (pl. EuroMatrix)
- ☐ Alternatív elképzelések: létező rendszerek/ szolgáltatások kombinációja
- ☐ webforditas.hu → iTranslate4.eu
- ☐ Előtérbe kerül a géppel támogatott utószerkesztés
- ☐ 2010-es évek: elindul a világ a neurális hálós fordítás felé...

Neurális gépi fordítás





Bevezetés a nyelvtechnológiába

10.

Korpuszok

(és még néhány dolog, ami eddig kimaradt...)

A korpusznyelvészet alapgondolata

- ☐ Korpusz: „meghatározott szempontok alapján kiválasztott szövegmennyiség, amelyen a nyelvész vizsgálatát végzi”
- ☐ John R. Firth (1890-1960): „*You shall know a word by the company it keeps*”
- ☐ John M. Sinclair (1933-2007): „*Language cannot be invented; it can only be captured*”
- ☐ Korpusz-alapú módszer: a szövegtörzs segédeszköz, ami empirikus adataival támogatja az intuíciót, mérhetővé teszi a nyelvi jelenségeket, meglévő elméleteket bizonyít/cáfol
- ☐ Korpuszvezérelt módszer: a korpusz maga szolgáltatja az „elméletet”, a nyelvész minden előzetes feltevés és elvárás nélkül fordul az adatokhoz, és minden következtetést kizárólag a korpusz megfigyeléseiből von le
- ☐ Serendipity Principle: jelentős dolog véletlen felfedezése, tipikusan akkor, amikor valami egész másra figyelünk

Régi dilemma: nyelvtankönyv vs. szöveg

Riedl Frigyes: *Simonyi kis nyelvtana* (1882)

„Simonyi új grammatikai módszert akar behozni, könyve inductive halad, azaz a példákból kiindulva tanítja a szabályt, nem pedig dogmatica.

A grammaticát tehát valami olvasmány alapján akarja előadni, úgy hogy a szabályokat a tanár tanítványai közreműködésével vonhatja le ésszerű következtetések útján. Ilyenképp tehát ezen módszer véget vet a lelketlen magolásnak, és azt észfejlesztő inductióval pótolja. Eszerint a szabályok is mélyebben vésődnek be a gyermek emlékezetébe, mert amit magunk találunk, azt jobban tudjuk, mint amit más mond vagy más tanultat velünk.”

A korpusz méretének mérőszámai

- ☐ Hány „token” van benne? = Mekkora a korpusz?
- ☐ Mi egy token?
 - ☐ \$22.50
 - ☐ George W. Bush / George Bush / Bush
- ☐ Korpusznormalizálás
 - ☐ The / the / THE
 - ☐ Calif. / California
 - ☐ MTA / Magyar Tudományos Akadémia
- ☐ Hány „type” van benne?
= Hány különböző szó van benne?
- ☐ A „type”-ok tényleges gyakorisága

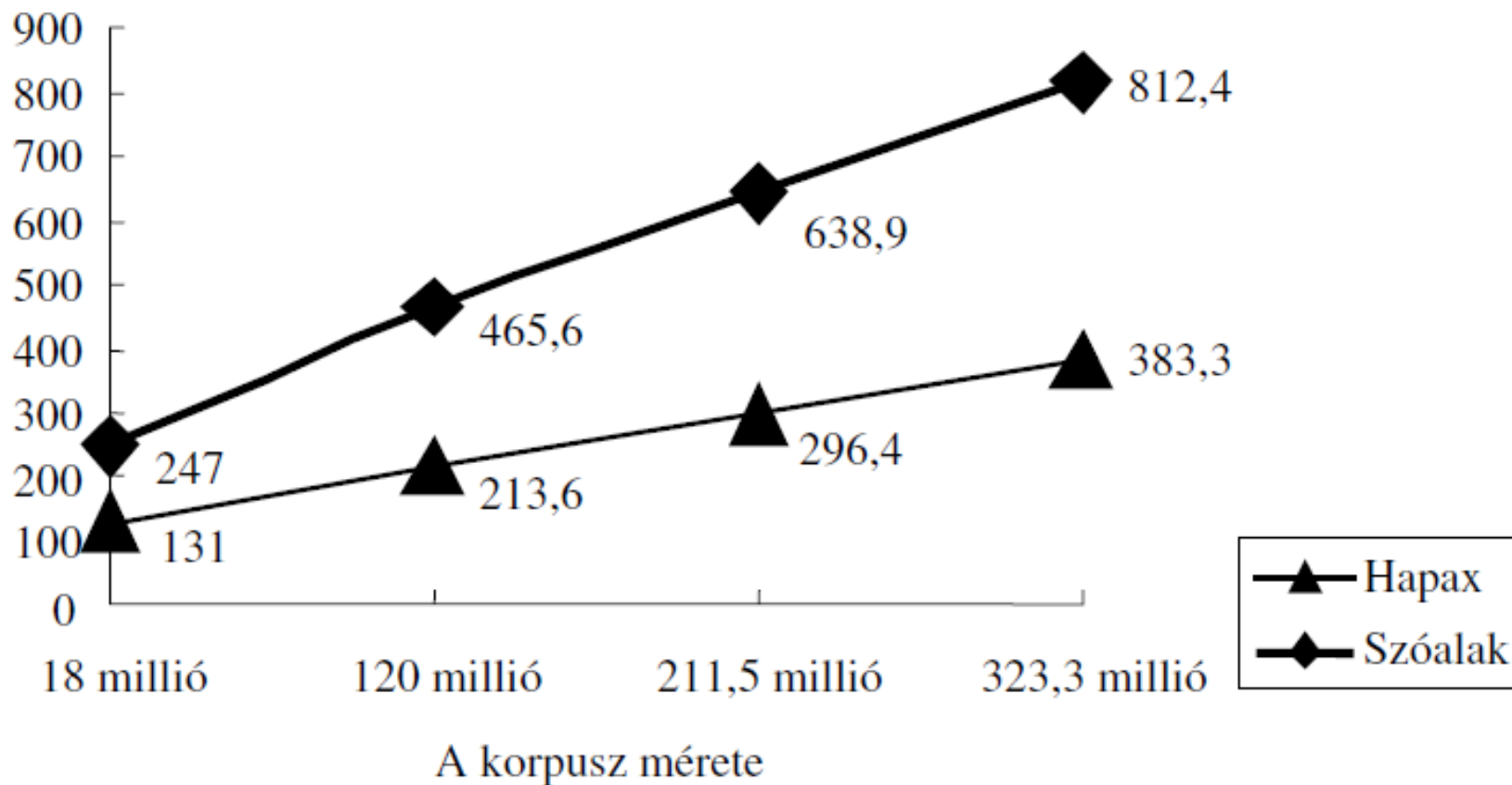
A korpuszok méretéről

(Szirmai M. alapján)

☐ A4 oldal (kettes sorközzel)	250 szó
☐ Gárdonyi: Egri csillagok	135 000 szó
☐ Brown Corpus (1961)	1 000 000 szó
☐ COBUILD Corpus (1987)	18 000 000 szó
☐ British National Corpus (2002)	100 000 000 szó
☐ Magyar Nemzeti Szövegtár (2005)	187 600 000 szó
☐ COBUILD Corpus (1996)	323 000 000 szó
☐ Corpus of Cont. Am. Engl. (2011)	425 000 000 szó
☐ MNSZ 2 (2014)	500 000 000 szó
☐ COBUILD Corpus (2012)	650 000 000 szó
☐ Pázmány Korpusz (2016)	1 200 000 000 szó

A korpuszok méretéről (2)

A szóalakok és hapaxok növekedése a Cobuild Korpuszban



Korpuszannotációk

- ☐ Egyszerű szöveg
- ☐ Különböző annotációk mindenféle hipotézisek teszteléséhez
- ☐ Szófaji egyértelműsítés
- ☐ Név kifejezések kezelése
- ☐ Szintaktikai szerkezetek bejelölése
- ☐ Jelentés-egyértelműsítés
- ☐ Koreferencia-kezelés
- ☐ Dialógus-jegyek



Néhány híres címkézett korpusz

Korpusz	# Tokenek	Megjegyzés
Brown	1 000 000	címkézett, kiegyensúlyozott
Susanne	120 000	a Brown szintaktikailag elemzett részhalmaza
Lancaster (-Oslo-Bergen)	1 000 000	„a UK válasza a Brown-ra”
Penn Treebank	2 000 000	szintaktikailag elemzett
British National Corpus (BNC)	100 000 000	szófajilag egyértelműsített
Szeged Korpusz	1 200 000	kiegyensúlyozott, morfológiailag és (2. verzió) szintaktikailag elemzett
Magyar Nemzeti Szövegtár	187 600 000	szófajilag egyértelműsített
Magyar Nemzeti Szövegtár 2	500 000 000	szófajilag egyértelműsített
Pázmány Korpusz	1 200 000 000	szófajilag egyértelműsített, NP-k

A Brown Corpus

- ☐ Standard American English
- ☐ Az első modern korpusz
(Francis and Kucera, 1961)
- ☐ 500 szöveg, darabja 2000 szó hosszú
- ☐ Amerikai könyvek, újságok, folyóiratok
- ☐ 15 témakör: tudományos-fantasztikus, regény, sajtó, tudományos művek
- ☐ POS-taggelt: 87 osztály



A CLAWS címkekészlet

(Constituent Likelihood Automatic Word-tagging System)

ABL pre-qualifier (quite, rather, such...)

ABN pre-quantifier (all, half)

...

AT singular article (a, an, every)

ATI article (the, ze, no)

BE (be)

BED (were)

BEDZ (was)

...

BEZ (is)

...

CD cardinal number (two, dozen, hundred...)

...

CD-CD hyphenated cardinal number (1985-1995...)

...

DO (do)

DOD (did)

...

HV (have)

...

IN preposition (after, by, of, for, since ...)

JJ general adjective (turquoise, happy...)

...

NN singular common noun (boy, pencil...)

NN\$ genitive singular common noun (boy's, parliament's ...)

...

RB adverb (else, about, hopefully ...)

...

TO infinitival TO

VB base form of lexical verb

VBD past tense of lexical verb

...

POS-taggelés a Brown-korpuszban

Television/NN has/HVZ yet/RB to/TO work/VB
out/RP a/AT living/RBG arrangement/NN with/IN
jazz/NN ,/, which/VDT comes/VBZ to/IN the/AT
medium/NN more/QL as/CS an/AT uneasy/JJ
guest/NN than/CS as/CS a/AT relaxed/VBN
member/NN of/IN the/AT family/NN ./.

<http://ilk.uvt.nl/~zavrel/tagtest.html>

Többértelműség a POS-taggelésben

The	AT		
man	NN	VB	
still	NN	VB	RB
saw	NN	VBD	
her	PPO	PP\$	

Egyértelműsítés rákövetkezés-gyakorisággal

The man still saw her.

The AT

man NN VB

still NN VB RB

saw NN VBD

her PPO PP\$

$$p(AT)=1.000$$

$$p(AT-NN)=0.186$$

$$p(AT-VB)=0.001$$

$$p(NN-NN)=0.040$$

$$p(NN-VB)=0.009$$

$$p(NN-RB)=0.040$$

$$p(AT-NN-NN)=p(AT-NN)*p(NN-NN)=0.07440$$

$$p(AT-NN-VB) = p(AT-NN)*p(NN-VB) = 0.01674$$

$$p(AT-NN-RB) = p(AT-NN)*p(NN-RB) = 0.07440$$

...

	NN	PPO	PP\$	RB	VB	VBD	.
AT	186	0	0	8	1	8	9
NN	40	1	3	40	9	66	186
PPO	7	3	16	164	109	16	313
PP\$	176	0	0	5	1	1	2
RB	5	3	16	164	109	16	313
VB	22	694	146	98	9	1	59
VBD	11	584	143	160	2	1	91

A Susanne-formátum

A01:0010b	AT	The	the	[O[S[Nns:s.
A01:0010c	NP1s	Fulton	Fulton	[Nns.
A01:0010d	NNL1cb	County	county	.Nns]
A01:0010e	JJ	Grand	grand	.
A01:0010f	NN1c	Jury	jury	.Nns:s]
A01:0010g	VVDv	said	say	[Vd.Vd]
A01:0010h	NPD1	Friday	Friday	[Nns:t.Nns:t]
A01:0010i	AT1	an	an	[Fn:o[Ns:s.
A01:0010j	NN1n	investigation	investigation	.
A01:0020a	IO	of	of	[Po.
A01:0020b	NP1t	Atlanta	Atlanta	[Ns[G[Nns.Nns]
A01:0020c	GG	+<apos>s	-	.G]
A01:0020d	JJ	recent	recent	.
A01:0020e	JJ	primary	primary	.
A01:0020f	NN1n	election	election	.Ns]Po]Ns:s]
A01:0020g	VVDv	produced	produce	[Vd.Vd]
A01:0020h	YIL	<ldquo>	-	.
A01:0020i	ATn	+no	no	[Ns:o.
A01:0020j	NN1u	evidence	evidence	.
A01:0020k	YIR	+<rdquo>	-	.
A01:0020m	CST	that	that	[Fn.
A01:0030a	DDy	any	any	[Np:s.
A01:0030b	NN2	irregularities	irregularity	.Np:s]

A Lancaster/IBM-formátum

B01000 1 v
[N The_AT world_NN1 N][V owes_VVZ [N a_AT1
considerable_JJ debt_NN1 [P of_IO [N gratitude_NN1
N]P]N][P[P&g to_II [N Mr._NNSB1 Byles_NP1 [N
the_AT Butcher_NN1 N]N]P&g] ,_. or_CC [P+ to_II [N
the_AT [less_RGR generic_JJ] alias_NN1 [Fr[P
under_II [N which_DDQ N]P][N this_DD1 indispensable_JJ
tradesman_NN1 N][V supplied_VVD [N the_AT needs_NN2
[P of_IO [N the_AT Stevenson_NP1 family_NN1 N]P]N]
[P in_II [N Bournemouth_NP1 N]P][P in_II [N 1885_MC
N]P]V]Fr]N]P+]P] ,_. [Fa since_ICS [N Stevenson_NP1 N]
[V explained_VVD [Fn that_CST [[N it_PPH1 N][V was_VBDZ
[N the_AT bills_NN2 [P of_IO [N Mr._NNSB1 Byles_NP1 N]
P]N]V]]][N which_DDQ N][V drove_VVD [N him_PPHO1 N]
[P to_II [N the_AT writing_NN1 [P of_IO [N The_AT
Strange_JJ Case_NN1 [P of_IO [N[N&g Dr._NNSB1 Jekyll_NP1
N&g] and_CC [N+ Mr._NNSB1 Hyde_NP1 ,_.
N+]N]P]N]P]N]P]V]]Fn]V]Fa]V]

A BNC-formátum

Eredeti:

Jordan lifts 22 years of martial law.

By Wafa Amr in Amman

JORDAN'S Prime Minister, Mr Mudar Badran, yesterday announced the freezing of martial law for the first time since 1967 as a prelude for formally revoking most of its provisions and abolishing it.

BNC-formátum:

```
<head type=BYLINE><s n=0002 p=Y><w II>By <w NP1>Wafa <w  
NP1>Amr <w II>in<w NP1>Amman </s></head>  
<p><s n=0003 p=Y><w NP1>JORDAN<w GE>'S <w JJ>Prime <w  
NN1>Minister<c YCOM>, <w NNB>Mr <w NP1>Mudar <w  
NP1>Badran<c YCOM>, <w RT>yesterday <w VVD>announced <w  
AT>the <w NN1>freezing <w IO>of <w JJ>martial <w  
NN1>law <w IF>for <w AT>the <w MD>first <w NNT1>time <w  
II>since <w MC>1967 <w II>as <w AT1>a <w N1>prelude <w  
IF>for <w RR>formally <w VVG>revoking <w DAT>most <w  
IO>of <w APPGE>its <w NN2>provisions <w CC>and <w  
VVG>abolishing <w PPH1>it<c YSTP>. </s></p>
```

Együtt-előfordulások a BNC-ban

turn British National Corpus freq = 44051

object	10205	2.3	subject	10255	4.6	modifier	7350	and/or	1947	0.3	part_trans	3430	15.4	
back	597	43.78	tide	71	30.29	0.6		twist	60	39.7	down	813	51.97	
attention	505	43.14	wheel	53	22.65	left	396	77.77	walk	108	36.3	over	500	51.14
head	752	39.77	Wycliffe	17	20.81	right	462	60.02	look	181	31.73	off	418	45.35
corner	232	36.86	Robyn	19	19.84	away	990	59.97	loss	29	31.07	round	169	41.3
page	178	31.08	Sheldukher	9	19.64	then	895	50.36	smile	36	26.82	out	469	36.74
handle	66	30.66	Rex	16	18.81	upside down	62	48.74	flee	23	26.35	up	489	35.98
somersault	19	29.8	Cardiff	21	18.11	now	517	41.73	stop	57	24.95	around	108	33.79
tap	56	29.33	Athelstan	14	17.42	inside out	26	41.02	run	74	24.3	on	210	31.56
knob	33	28.64	poacher	9	17.36	slowly	116	36.79	stare	25	22.46	back	191	30.26
key	97	27.02	Devore	7	16.64	half	52	35.93	shrug	15	21.69	in	44	15.39
eye	292	26.18	Luke	21	16.54	quickly	122	34.55	stride	11	21.1	about	10	8.83

A Penn Treebank

- ❑ Az első szintaktikailag elemzett korpusz
- ❑ Tartalma: kb. 3 millió szó
- ❑ Alapjai:
 - ❑ Brown-korpusz (Treebank I)
 - ❑ Wall Street Journal Corpus (Treebank II)
 - ❑ ATIS corpus
 - ❑ www.cis.upenn.edu/~treebank
 - ❑ tgrep-interfésszel:
www ldc.upenn.edu/ldc/online/treebank/

A Penn Treebank formátuma

[S[N Nemo_NP1 ,_, [N the_AT killer_NN1
whale_NN1 N] ,_, [Fr[N who_PNQS N][V
'd_VHD grown_VVN [J too_RG big_JJ [P
for_IF [N his_APP\$ pool_NN1 [P on_II [N
Clacton_NP1 Pier_NNL1
N]P]N]P]J]V]Fr]N] ,_, [V has_VHZ
arrived_VVN safely_RR [P at_II [N
his_APP\$ new_JJ home_NN1 [P in_II [N
Windsor_NP1 [I safari_NN1 park_NNL1
I]N]P]N]P]V] ._. S]

Néhány treebank-kereső

☐ Közvetlen összetevős szerkezetekre

VP << /[^]believe/ < (S < (/[^]NP/ !<< /[^{*}]/ !< (-NONE- < T)) < (VP|AUX << to))

(pl. tgrep, Penn Treebank)

(QP iDominates CONJ)

(pl. CorpusSearch, Penn, Helsinki)

☐ Függőségi szerkezetekre

[pos = "NN" & lemma = /Be.+ung/]

(pl. TigerSearch, Potsdam)

cat="NP" & #1:arity=4

(pl. ANNIS, Humboldt)

Grafikus keresők

(pl. PNL-TQ, Netgraph, Prága)



Magyar szövegtörzsek

Szeged Korpusz

- ☐ Az első magyar annotált korpusz
(SZTE + MTA NYTI + MorphoLogic közös projektje)
- ☐ szépirodalom (részletek Rejtő Jenő Pizkos Fred, a kapitány, Szerb Antal Utas és holdvilág, ill. George Orwell 1984 c. műveiből)
- ☐ 14-16 éves korú tanulók fogalmazásai
- ☐ újságcikkek (részletek a Népszabadság, a Népszava, a Magyar Hírlap, ill. a HVG egyes számaiból)
- ☐ számítógépes szövegek (részletek Kis Balázs Windows 2000 c. könyvéből, ill. a ComputerWorld/Számítástechnika magazin egyes számaiból)
- ☐ jogi szövegek (részletek a gazdasági társaságokról, ill. a szerzői jogokról szóló törvényekből)
- ☐ Morfológiailag egyértelműsített korpusz
 - ☐ Brill: 96,52%
 - ☐ TnT (HMM): 96,18%
 - ☐ RGLearn (szabály-alapú, saját): 94,54%
 - ☐ Kombinációjuk: 96,95%

Szeged Treebank

- ☐ Az első magyar treebank (kézi annotáció!)
 - ☐ határozói szerkezetek (ADVP)
 - ☐ jelzős szerkezetek (ADJP)
 - ☐ névutós szerkezetek (PP)
 - ☐ az igeneves szerkezetek (PA, INF)
 - ☐ tagadószók, igekötők, igék és kötőszavak
-
- ☐ A szabályok egy részét a konzorcium nyelvész szakértői készítették el és a kézzel definiált szakértői szabályokat az annotált treebankból számítógépes tanulási módszerekkel kinyert szabályokkal egészítették ki

A (nem címkézett) Magyar Webkorporusz

- ☐ Szószablya projekt (BME MOKK, 2003-2004)
- ☐ A .hu domén weboldalai
- ☐ 1,48 milliárd token (szűrés nélkül)
- ☐ Szűrések (az ismeretlen szavak százaléka):
 - 40%: gyakorlatilag nincs már benne nem magyar szöveg
 - 8%: gyakorlatilag nincs már benne ékezet nélküli szöveg
 - 4%: egy átlagszöveg elütéseinél kevesebb van benne
- ☐ A 4%-os letölthető: az 1220 millió magyar weboldal 589 millió szava tömörítve is 4 GB

	millió oldal	millió token	millió type
teljes	3,5	1486	19,1
40%	3,125	1310	15,4
8%	1,918	928	10,9
4%	1,221	589	7,2

A Magyar Nemzeti Szövegtár

- ❑ 1998 óta: <http://corpus.nytud.hu/mnsz/>
- ❑ A fele sajtó, benne van a DIA, és részben a MEK, továbbá index.hu-s fórumok szövegei és joganyagok
- ❑ 187,6 millió szó:

	magyaró.-i	szlovákiai	kárpátaljai	erdélyi	vajdasági	összesen
sajtó	71,0	5,7	0,7	5,5	1,5	84,5
szépirodalom	35,5	1,4	0,4	0,8	0,2	38,2
tudományos	20,5	2,3	0,7	1,6	0,3	25,5
hivatalos	19,9	0,2	0,3	0,6	0,1	20,9
személyes	17,8	—	0,4	0,4	0,1	18,6
összesen	164,7	9,5	2,5	8,9	2,0	187,6

- ❑ Magyar Nemzeti Szövegtár 2: 500 millió szó



Az MNSZ szerinti szógyakoriság

Főnevek

...
30. év
...
54. ember
...
58. idő
...
62. rész
...
65. szó
...

Melléknevek

...
36. nagy
...
44. jó
...
46. magyar
...
62. új
...

Igék

szótő	szófaj	db
1. a	<i>Det</i>	11128421
2. az	<i>Det</i>	3716414
3. és	<i>Con</i>	2544751
4. hogy	<i>Con</i>	2166004
5. A	<i>Det</i>	2103970
6. az	<i>Pro</i>	1803814
7. nem	<i>Adv</i>	1693748
8. is	<i>Con</i>	1677108
9. van	<i>V</i>	1418113
10. ez	<i>Pro</i>	1204269
11. egy	<i>Num</i>	899832
12. Az	<i>Det</i>	730287
13. meg	<i>Pre</i>	592986
14. kell	<i>V</i>	499659
15. csak	<i>Adv</i>	477956
16. lesz	<i>V</i>	469189
17. de	<i>Con</i>	462508
18. már	<i>Adv</i>	452814
19. Ez	<i>Pro</i>	447310
20. amely	<i>Pro</i>	417945

Prószéky Gábor

A Mazsola vonzatkereső


Körpusz:

Ikerd:

Nem: ☐ Esetnévutó: Nem: ☐ Lemma:

Nem: ☐ Esetnévutó: Nem: ☐ Lemma:

Nem: ☐ String:

Mehet

Ikerd-stat: ☐

Lemma-stat: ☐

Lemma-stat: ☐

Nincs stat: ☐

1010 találat. Stat: [bocsánat](#) [51] [segítség](#) [53] [elnézés](#) [32] [az](#) [136] [engedély](#) [32] [tájékoztatás](#) [21] [támogatás](#) [25] [péng](#) [20] [felmentés](#) [12] [válasz](#) [14] [tanács](#) [13] [íróint](#) [16] [magyarul](#) [9] [igazolás](#) [8] [átírástól](#) [8] [kutatás](#) [7] [normatorkod](#) [6] [információ](#) [6] [teljesítmény](#) [7] [készen](#) [6] [megállapítás](#) [7] [készen](#) [6] [gazdaság](#) [6] [belső](#) [6] [szó](#) [12] [vissza](#) [7] [vissza](#) [6] [az](#) [7] [az](#) [6]

adat

Az információk ellenőrzésére perazsa a kórháztól kér adatokat, Feladatai teljesítéséhez adatokat kérhet a bíróságtól, az ügyvédségtől, a nemzetbiztonsági szolgálatoktól, a társadalombiztosítási igazgatási szervezettől. ha maximum öt főről kér adatot a hivataltól. hogy a rendőrség az adóhatóságtól, telefonaszolgálatoktól, bankoktól ügyézi jóváhagyás nélkül kérjen adatokat, hogy kérje tőlük a művelődési tárca a bérfejlesztéshez szükséges adatokat.

adatgyűjtés

hanes Orbán Viktor és Deutsch Tamás ellen is adatgyűjtést kértek az ügyben érintett magánnyomozótól.

adóigazolás

amely szerint a gépkocsi átíratásakor adóigazolást kértek a polgároktól.

adókezelés

Korábban Budapest és a vidéki nagyvárosok külföldi adókezelésüket is kértek az eldől kabinettől,

adózás

Ha a bérbeadó magánszemély, akkor adószámlát kell kérnie az APEH-től,

ha a magánszemély adószámlát kér az APEH-től

aggregátor

A hadseregétől kértek aggregátort,

ajánlat

A helyreállítással megbízott Szabolcs-Szatmár-Bereg Megyei Községi Kht. tíz vállalkozótól másfél milliárd forintos felújítási munkára kért a közelmúltban kivitelezési ajánlat. Az eredménytelen pályázatok után a bizottság új ajánlatot kért a pályázóktól. és csatolva a pályázóktól - meghatározott műszaki paraméterekkel rendelkező - új kocsikra kért ajánlatot. hogy az MTK vezetői már ajánlatokat kértek néhány légitársaságtól a norvégiai repülőterrel kapcsolatban,

aláírás

A MÁV nyilatkozat aláírását kérte a szakaszvezetőktől, hogy az aláírásokat lehetőleg olyan személyektől kell kérni.

alkotmányértelmezés

Az MDF alkotmányértelmezést kér a legfőbb ügyésztől

Az Országgyűlés alkotmányügyi bizottsága abstract alkotmányértelmezést kér az Alkotmánybíróságtól az ellenzék által kezdeményezett népszavazás ügyében. így alkotmányértelmezést sem kérhet az Alkotmánybíróságtól.

Stílus-összehasonlítás a Mazsolával

PÉLDA

'ad -t' keret

a Magyar Nemzetben és az Index fórumban

8674 találat. hang [477] lehetőség [412] válasz [331] otthon [381] tájékoztatás [170] hír [309] ok [180] áldás [85]
esély [114] lendület [74] hely [150] pénz [129] magyarázat [87] szám [128] alkalom [101] tanács [85] koncert [71] munka [116]
utasítás [61] felvilágosítás [52] hangverseny [49] támogatás [93] kép [95] igaz [82] lökés [41] százalék [82] segítség [70]
engedély [56] megbízás [50] forint [85] információ [59] felhatalmazás [39] voks [38] ízelítő [29] alap [68] mód [59] jel [48] nyomaték [27] tér [46]

eltérések

7649 találat. igaz [30] válasz [275] tanács [172] hang [143] tipp [6] ami [212] lehetőség [155] pénz [151] magyarázat [88]
hála [67] ok [116] név [109] esély [93] ötlet [68] kegyelem [49] mi [118] hitel [61] vér [44] link [34] bizonyosság [31] hely [74] élet [66] cím [54]
felmentés [32] több [75] valami [59] fegyver [41] áldás [26] az [138] maga [62] alap [53] semmi [45] alkalom [40] fej [39] információ [37] jel [35] interjú [32]
parancs [28] megbízás [28] puszi [18] erő [40] hír [35] engedély [33] hit [30] kedvezmény [27] felvilágosítás [22] pofon [20] ez [74] sok [40] tér [34] kép [30]

Automatikus szemantikai osztályzás

nő

alany: szám arány ár

-bAn: mérték év

-rA: forint százalék dupla

-vAl: százalék forint

növekedik

alany: szám arány érdeklődés

-bAn: mérték év

-rA: év

-vAl: százalék forint

emelkedik

alany: szám ár árfolyam

-bAn: mérték év

-rA: forint jogerő százalék

-vAl: százalék forint

drágul

alany: részvény TVK kenyér

-bAn: mérték forgalom

-rA: forint

-vAl: százalék forint

Akár szótárgyártásra is használható

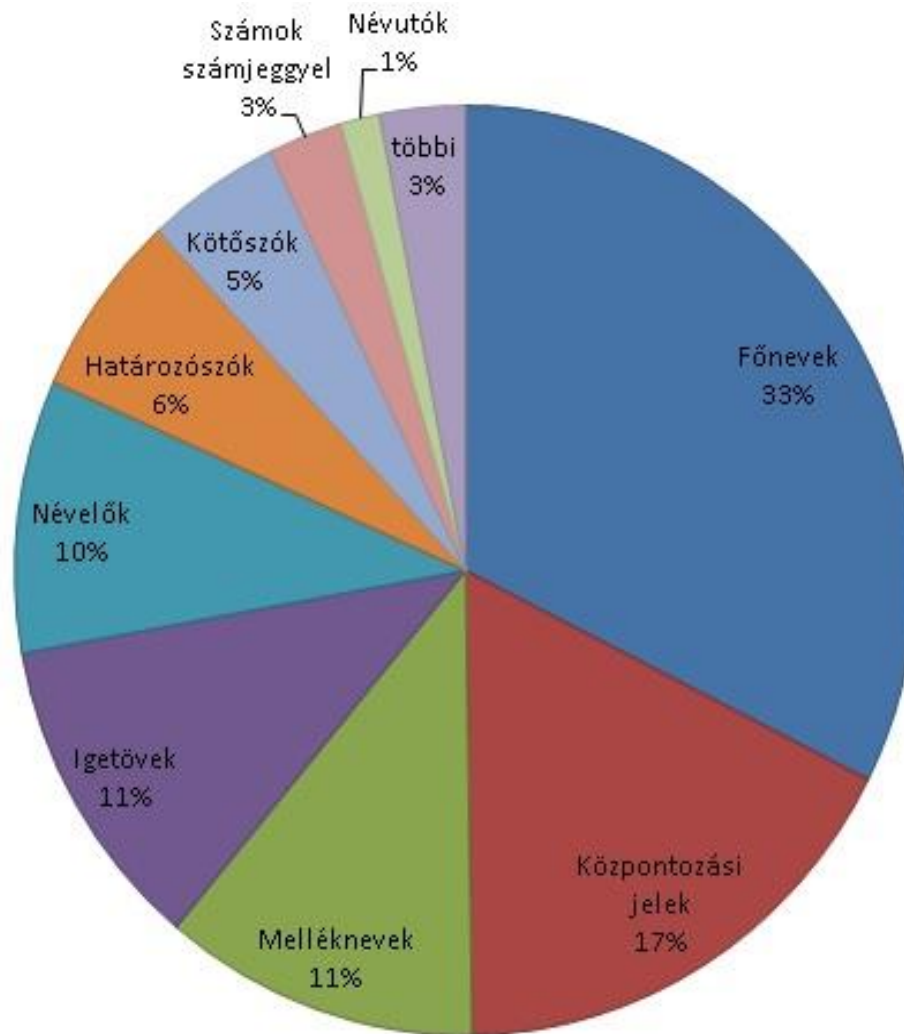
☐ -t hány	fitty
☐ -ra hány	szem
☐ alá kerül	víz, kalapács, fennhatóság
☐ alá rejt	véka
☐ alá hoz	tető
☐ alá helyez	vád
☐ alá vesz	górcső, kalap, tűz



A Pázmány Korpusz

alkorpusz	tokenszám	NP-szám	mondatszám
szztenderd szövegek	954 424 307	223 347 534	48 536 849
egyéb tartalmak	228 920 436	52 865 889	15 802 499
kommentek	59 013 104	13 867 066	3 505 818
összesen	1 242 357 847	290 080 489	67 845 166

A Pázmány Korpusz szófaji megoszlása





A Pázmány Korpusz szerkezete

eredeti szóalak	szótő	Humor- szófajcímke	módosított Humor-szófajcímke	főnévicsoport- határok
<s>				
Kezdetkor	kezdet	N+TEM	FN+TEM	1-N_1
teremtette	teremt	V+TMe3	IGE+TMe3	O
Isten	isten	N+NOM	FN+NOM	1-N_1
az	az	DET	DET	B-N_1
eget	ég	N+ACC	FN+ACC	E-N_1
és	és	CON	KOT	O
a	a	DET	DET	B-N_1
földet	föld	N+ACC	FN+ACC	E-N_1
.	.	PUNCT PERIOD	PUNCT	O
</s>				

A Pázmány Korpusz főnévi csoportjai

	A főnévi csoport felépítése	Korpuszbeli gyakoriság	Példa
1	[N+Nom]	14 312 895	ház
2	[Det] [N+Nom]	11 632 547	a ház
3	[N Pron+Nom]	6 416 317	ez
4	[Adj+Nom]	6 279 294	okos
5	[Det] [N+Acc]	4 616 117	a házat
6	[N Pron+Acc]	4 206 866	ezt
7	[N+Acc]	3 639 354	házat
8	[Det] [Adj] [N+Nom]	3 389 714	a nagy ház
9	[N+Nom] [N+Nom]	2 920 072	műanyag ház
10	[Det] [N+Pl+Nom]	2 690 635	a házak



A Pázmány Korpusz lekérdezhetősége

Query vár, FN.* 103,303 (83.2 per million)

[First](#) | [Previous](#) Page of 4,133 [Go](#) [Next](#) | [Last](#)

doc#5 normann lovaggal ostromolta meg Galeria várát , bevette , és foglyul ejtette Benedeket
doc#5 , VII. a Matild birtokában lévő Canossa várába ment , hogy várja a fejleményeket . 1077.
doc#5 25 : Henrik vezeklő ruhában jelent meg a vár kapuja előtt és 3 napon át kérte a perc.
doc#5 Pozsony ostrománál nagykorúsították , amikor a várat Borisztól visszavívta , IX. 11 : a Boriszt
doc#5 ostromolták . A háromszög alaprajzú hegyi várból jelentős romok állnak . Fügedi 1977:55
doc#5 vízhez , földalatti folyosót vágtak Sion várában , mely egy 13 m mély , függőleges aknához
doc#5 merítsenek . Dávid ezen keresztül vette be a várat : " Azon a napon Dávid azt mondta : ' Aki
doc#5 fölkapaszkodott a függőleges aknában (fönn a belső várban találta magát , s kinyitotta az ostromlók
doc#5 másztak meg , amikor bevették a jebuzeusok várat . Az akna felső végében egy ferdén felfelé
doc#5 karitatív intézete , neki adományozta Zólyom várat , s megbízta a Szepesség és a bányavárosok
doc#5 Hunyadi Giskraval , kinek kezén maradtak várai és jövedelmei . V. László megerősítette
doc#5 Mátyás csapatai mindenfelől támadták Giskra várai

Query vár, IGE.* 599,506 (482.6 per million)

[First](#) | [Previous](#) Page of 23,981 [Go](#) [Next](#) | [Last](#)

doc#7 ember keze nyújt , veszendő : | Halhatatlant vár s keres a valódi , | A nemes virtus : maga
doc#7 uralkodónéja . " A női közönségtől játékos hatást vár a férfiúi nemre a hazai nyelv , az irodalom
doc#7 székvárosát , mert ebben az időben Napóleontól várta a monarchia és hazája feudális viszonyainak
doc#7 levegő árad ; olyasféle , amelyet már régen várt a kritika és a színház . A Rang és mód
doc#7 egyik grófi származású , rá nagy örökség vár . Sem a nagypapa , sem a két lány , se
doc#7 ellenezte az emigrációt is : tőle semmi jót nem várt , legfeljebb a Monarchia s vele együtt
doc#7 ugyanígy vélekedik : " Kisült , amit nem vártunk , | Mutatja a hatás : | Rossz drámaíró
doc#7 még hátra maradt , leküzdeni - ... mi vár tehát ránk ? Azt a jó isten tudja , ember
doc#7 fogantja lett . Nem volt más hátra , mint várni , és azt remélni , hogy az abszolutizmus
doc#7 kudarcra intő példa volt arra nézve , mi várna egy újabb magyar szabadságharcra . A kiegyezési
doc#7 az emberiség minden kérdésének megoldását várták . A harmincas , negyvenes évek nemzedékeinek
doc#7 mélységeket , melyeknél halál és kárhozat vár ránk ! " Martinuzzi alakjánál hangsúlyosabbnak



A nyelvtechnológia további területei

Néhány további kutatási terület

- ☐ Névelem-felismerés
- ☐ Anafora-felismerés
- ☐ Koreferencia-feloldás
- ☐ Témafelismerés
- ☐ Véleményelemzés
- ☐ Automatikus kivonatolás
- ☐ Diskurzuselemzés
- ☐ Természetesnyelv-generálás
- ☐ OCR-támogatás
- ☐ Kereséstámogatás
- ☐ Beszéd-alapú rendszerek támogatása
- ☐ Kérdés–válasz rendszerek
- ☐ ...

További információk a nyelvtechnológiáról

Association for Computational Linguistics

<http://www.aclweb.org>

ACL Video Archive

<http://www.fask.uni-mainz.de/lk/videoarchive>

ACL Anthology

<http://www.aclweb.org/anthology-new/>

DFKI

<http://registry.dfki.de>

PBML

<http://ufal.mff.cuni.cz/pbml.html>

META

<http://www.meta-net.eu/>

További információk, de már a vizsgáról

A végső osztályzat az alábbiakból alakul ki:

jelenlét az előadáson
(erre én emlékszem...)

+

aktív jelenlét a gyakorlaton
(ezt a gyakorlatvezetők tudják...)

+

házi feladatok megoldása
(ha volt ilyen, nyoma van...)

+

2 zárthelyi dolgozat eredménye
(a pontszámok adottak...)

+

vizsga az előadások anyagából
(amivel feljavítják a fentiek alapján kialakult osztályzatot...)

Vizsgatételek

1. Karakterek, kódolási szabványok, rendezések, n-gramok, karakter- és n-gram-alapú statisztikai megfigyelések
2. Véges állapotú automaták (és kiterjesztéseik) nyelvtechnológiai alkalmazásai
3. A számítógépes morfológia alapvető módszerei (kétszintes, unifikációs)
4. A számítógépes morfológia alkalmazásai (helyesírás, elválasztás, intelligens keresés, intelligens csere)
5. Formális nyelvtanok, szintaktikai reprezentációk (közvetlen összetevős, függőségi, X-vonás), mondatelemzési módszerek (TD, BU, kombinált)
6. Az unifikáció és alkalmazásai a nyelvtechnológiában (morfológiában, szintaxisban)
7. Jelentésreprezentációk, szemantikus hálók, fogalmi gráfok, ontológiák, WordNet
8. Számítógépes lexikográfia: szótárépítés, szótárátalakítás, szótárhasználat támogatása nyelvtechnológiával
9. Párhuzamos korpuszok, szövegszinkronizálás, fordítómemóriák
10. A gépi fordítás és módszerei