



A nyelvtechnológia alapjai

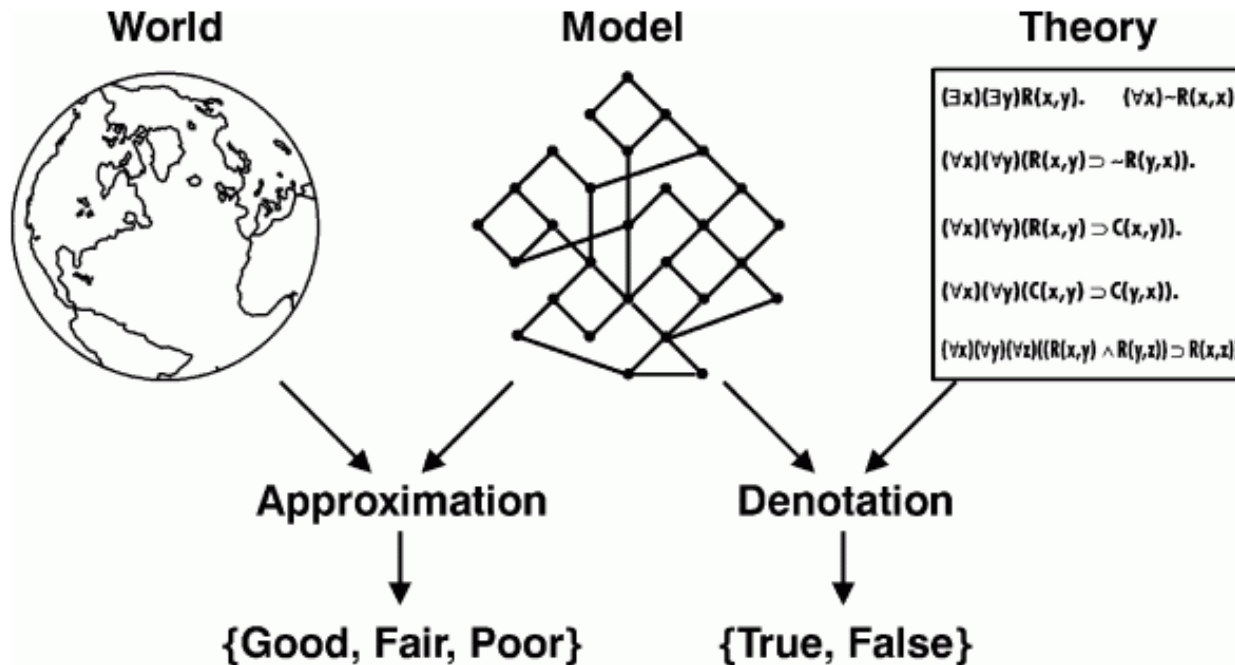
6.

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

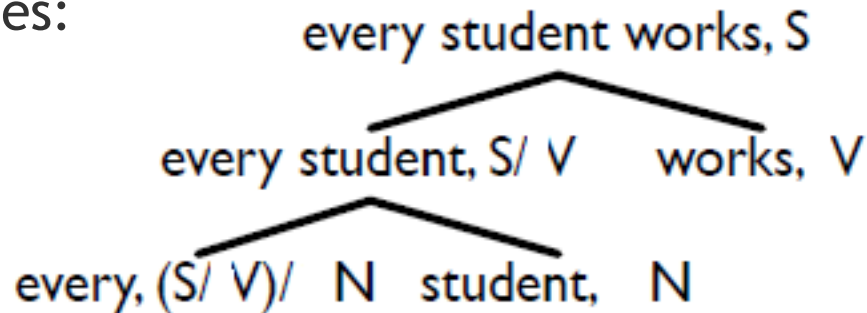
❑ Levezetett kategóriák:

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 levezetésekről 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

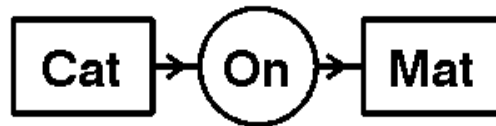
Sowa fogalmi gráfjai

- ❑ 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:

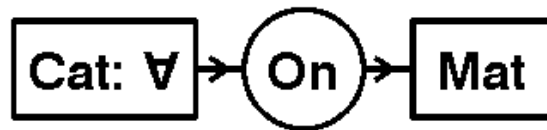
$[Cat: *x] [Mat: *y] (On ?x ?y)$

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:

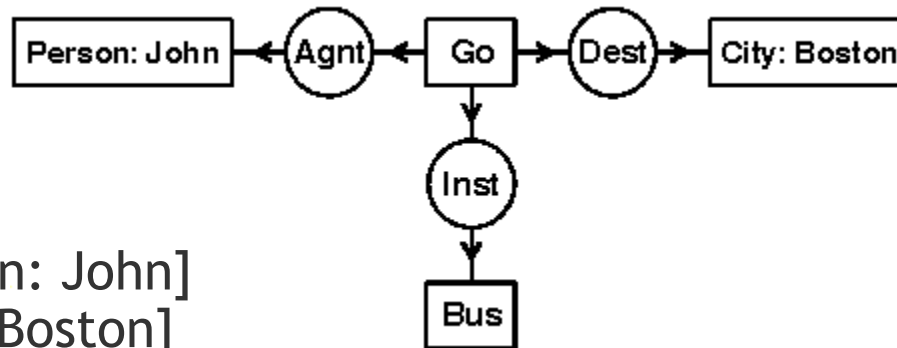
$[Cat: @every*x] [Mat: *y] (On ?x ?y)$

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)

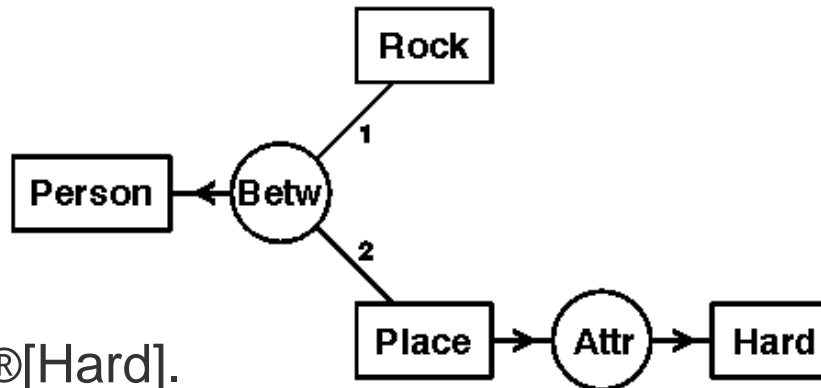
(Dest ?x ?z) (Inst ?x ?w)))

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.

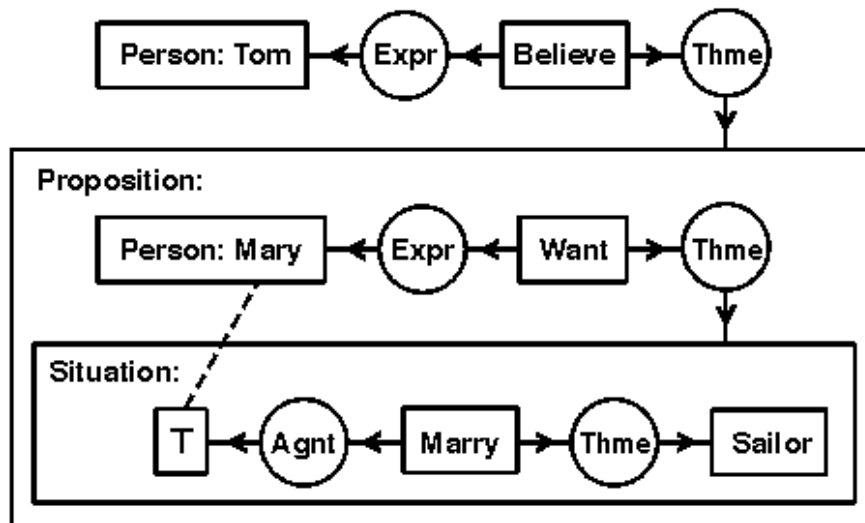
Tom believes that Mary wants to marry a sailor.

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.



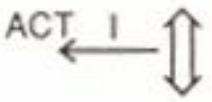
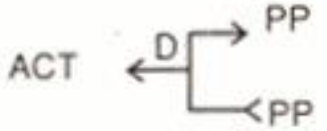
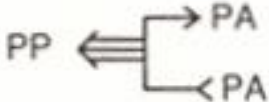

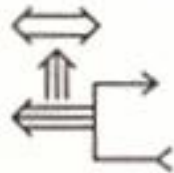

A fogalmi függőség állapotosztályai

<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 $\overset{P}{\longleftrightarrow} PTRANS$	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.	$ACT \xleftarrow{\theta} PP$	John $\overset{P}{\longleftrightarrow} PROPEL \xleftarrow{\theta}$ cart	John pushed the cart.
7.	ACT $\xleftarrow{R} \begin{cases} \rightarrow PP \\ \leftarrow PP \end{cases}$	John $\overset{P}{\longleftrightarrow} ATRANS$ $\xleftarrow{R} \begin{cases} \rightarrow \text{John} \\ \leftarrow \text{Mary} \end{cases}$ ↑ book	John took the book from Mary.

Schank eseményábrázolása (2)

8.  John ate ice cream.
9.  John fertilized the field.
10.  The plants grew
11. (a)  (b)  Bill shot Bob.
12.  John ran yesterday.

Forgatókönyvek

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

QUESTION:

Did Jane eat anything?



Az „étterem” forgatókönyve (a tipikus eseménysor)

entering

seating

ordering

serving

eating

paying

leaving

Az „étterem” forgatókönyve (alapismeretek)

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

COOK	...
MANAGER	...

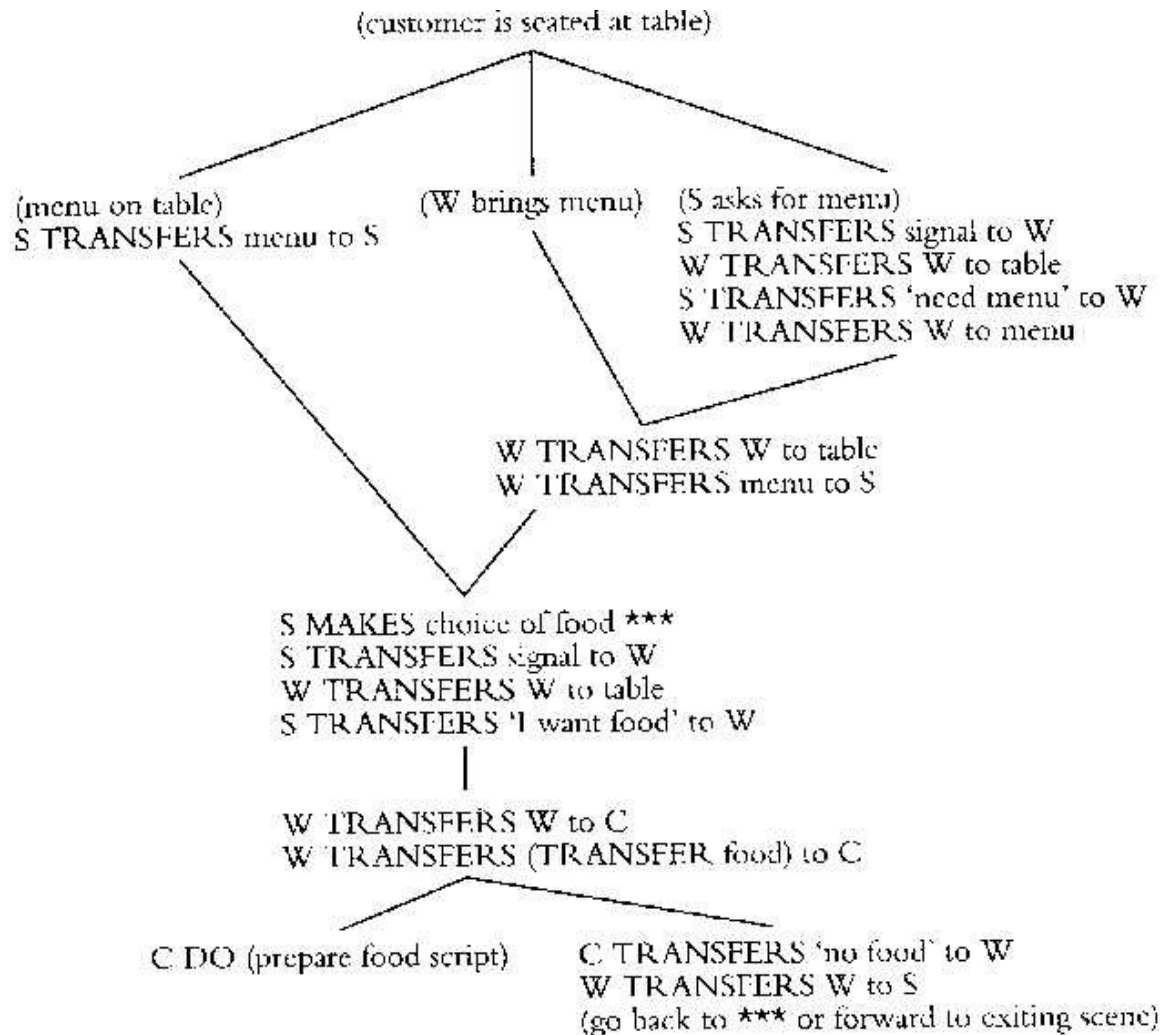
HEADER

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

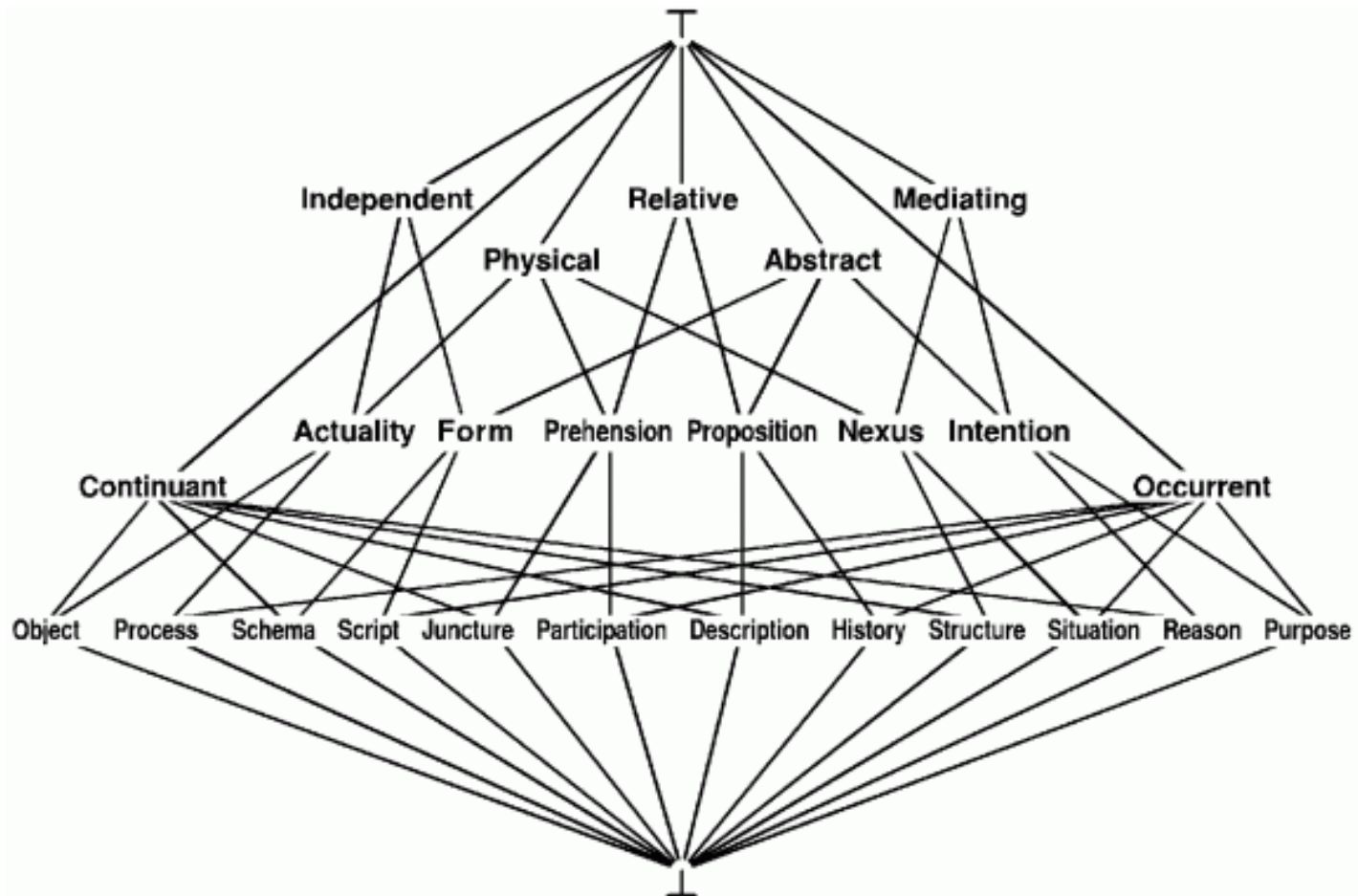
Lexikális szemantikus hálók

- ❑ 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, ...

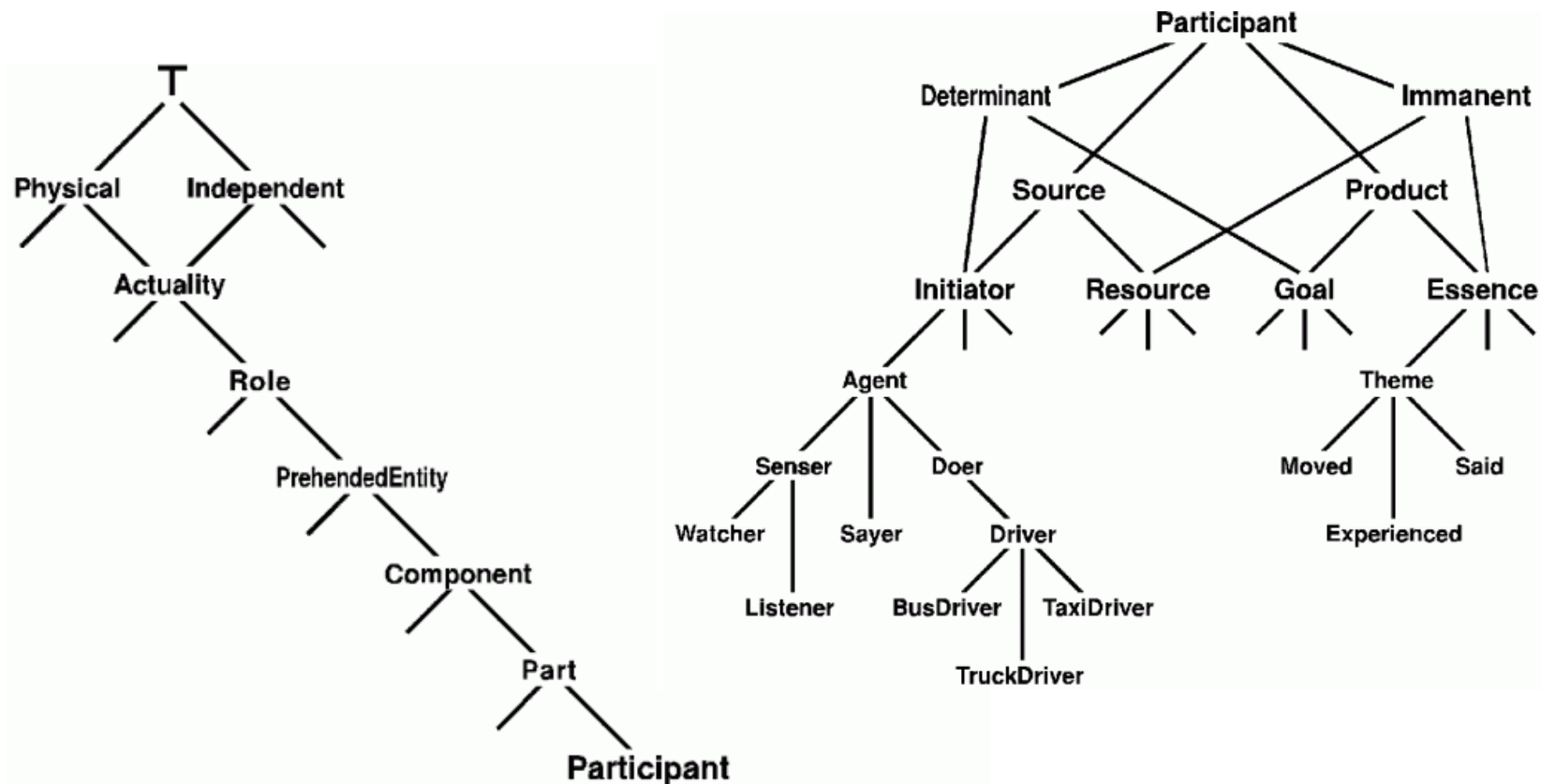
Ontológia

- ❑ 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)

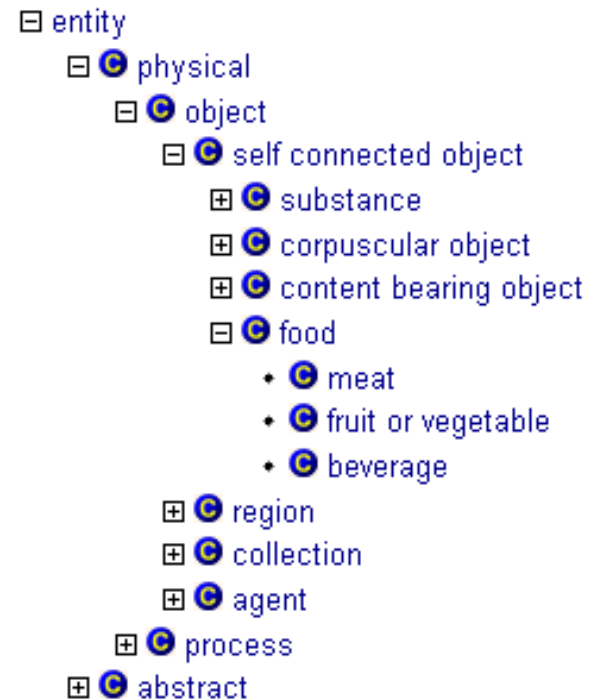


A felső és az alsó szintű ontológiák egy találkozása



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

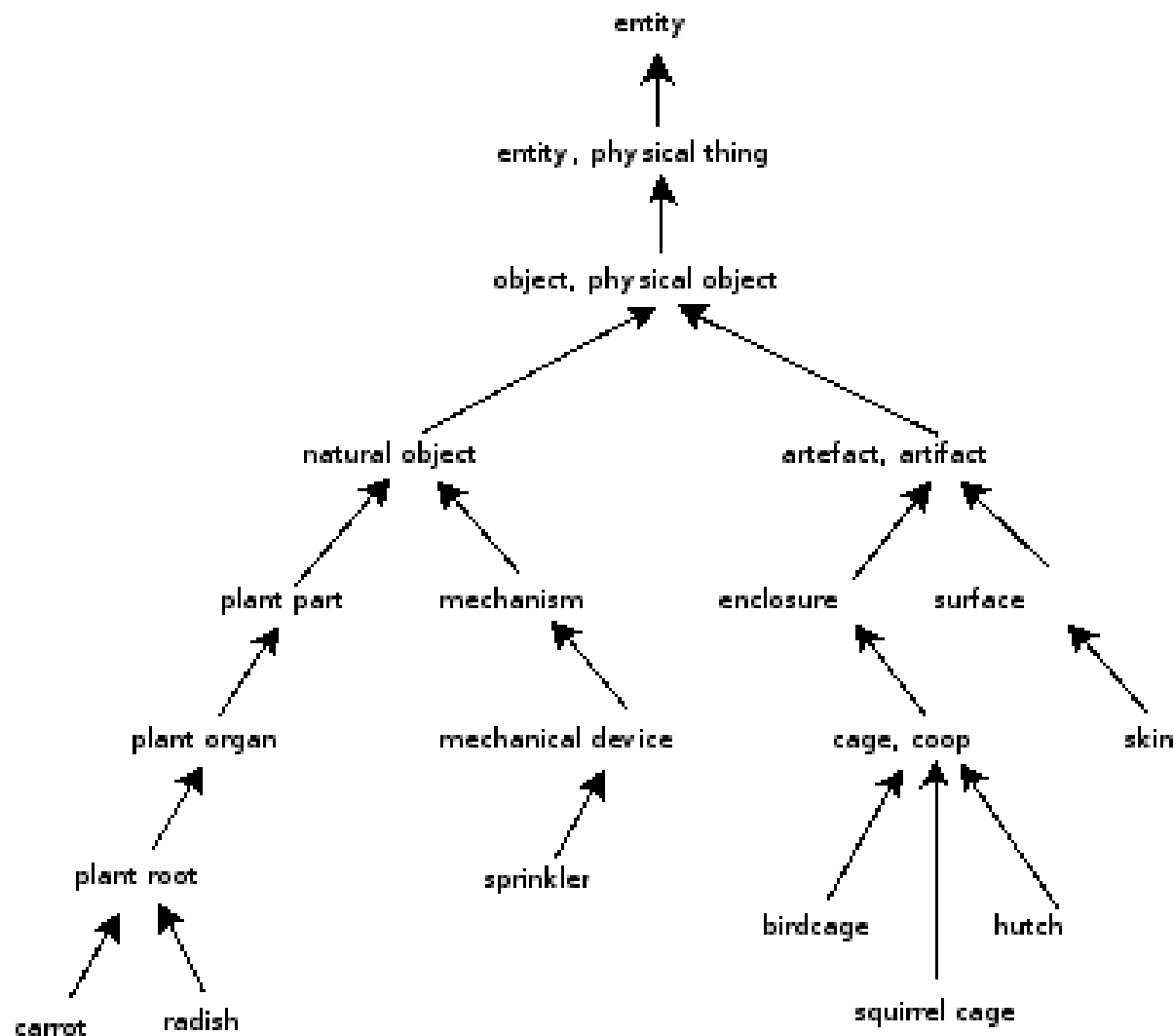
- *melléknévgyök*



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

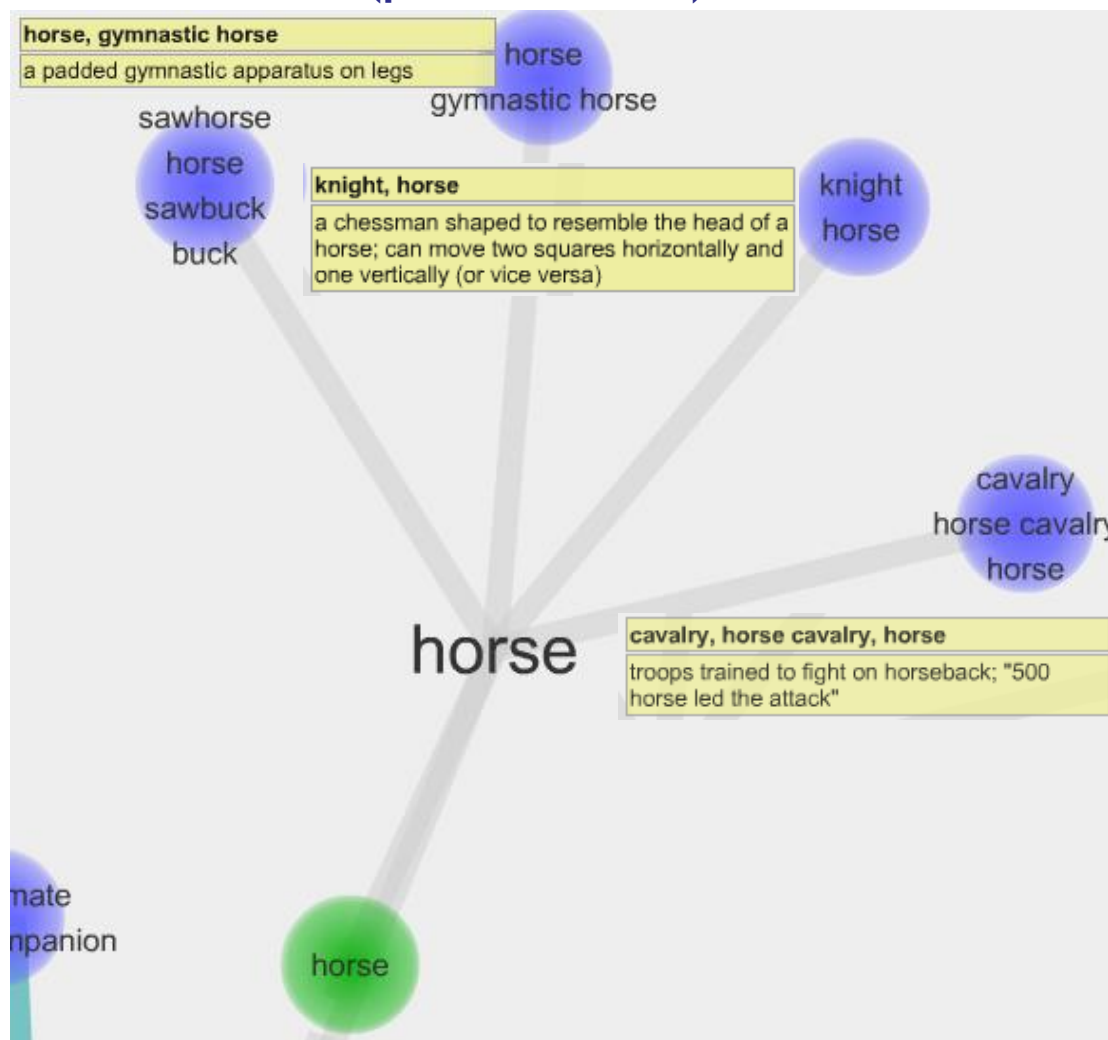
Noun

- [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

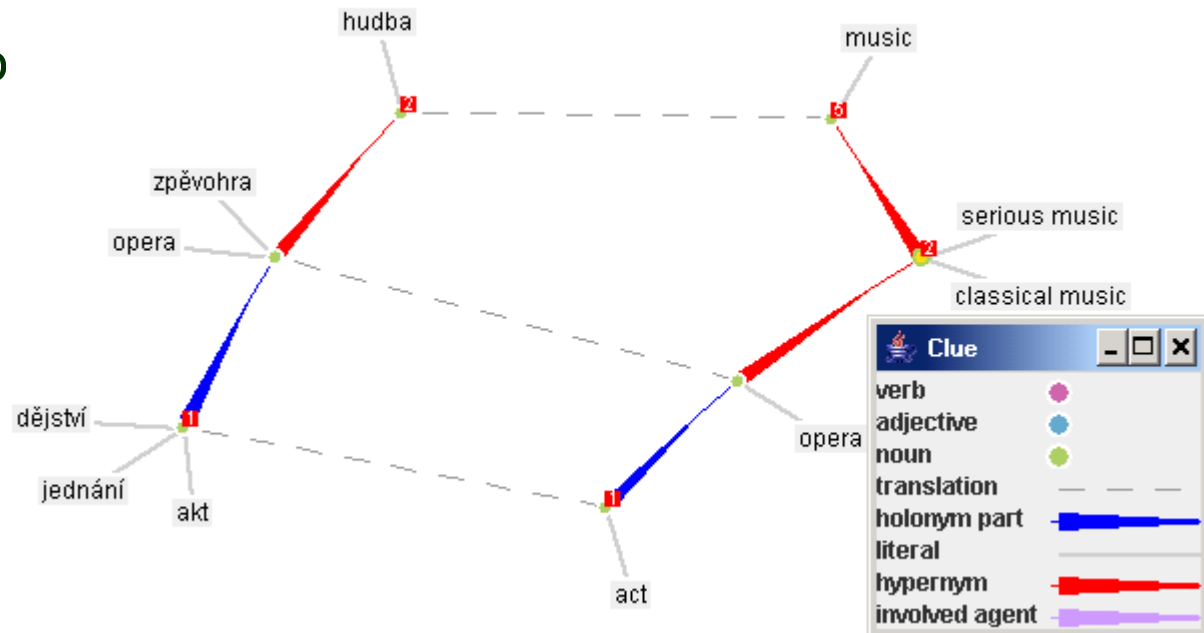
- [S:](#) [\(v\)](#) **horse** (provide with a horse or horses)

(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

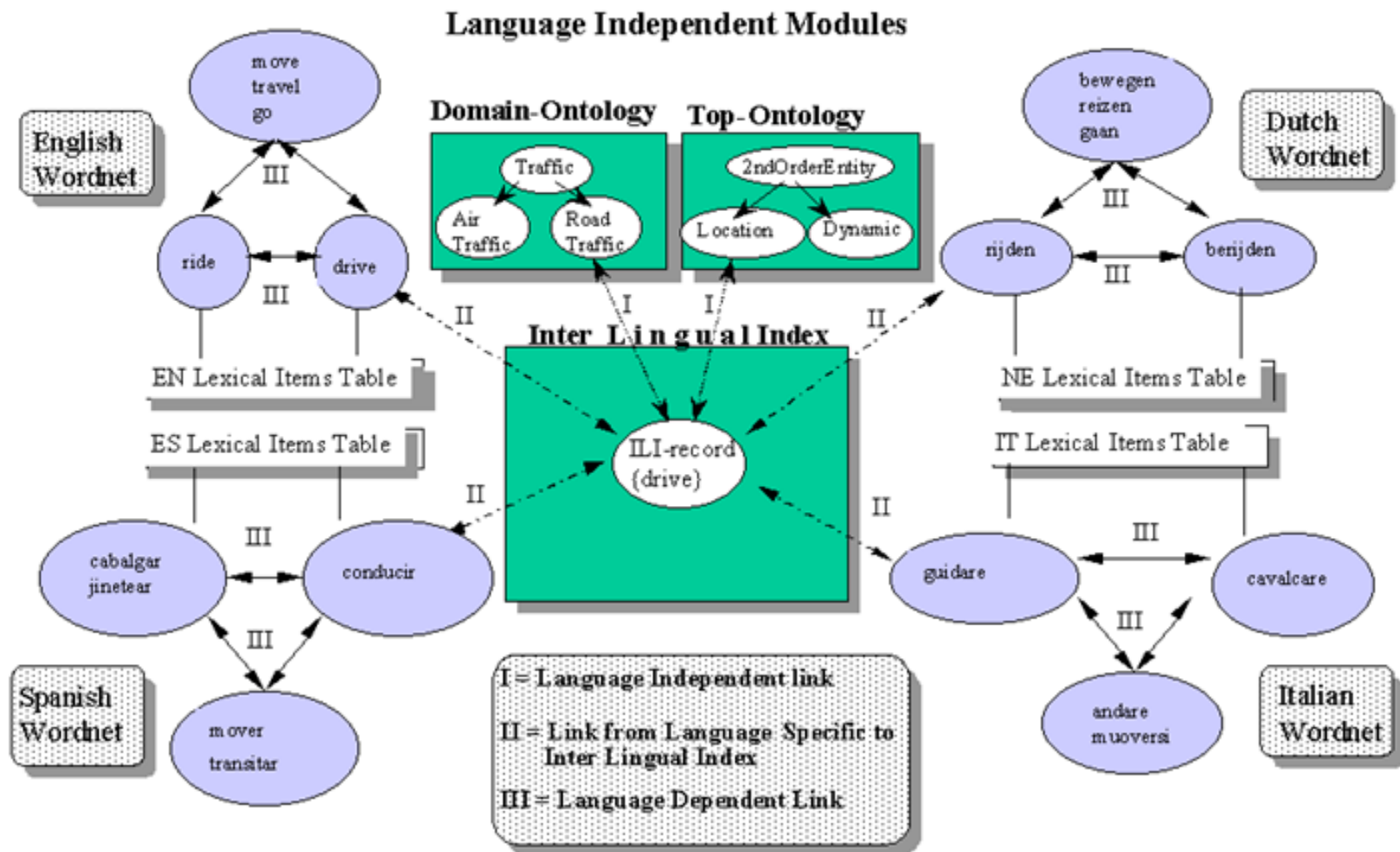




EWN Base Concepts & Top Ontology

- ❑ 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

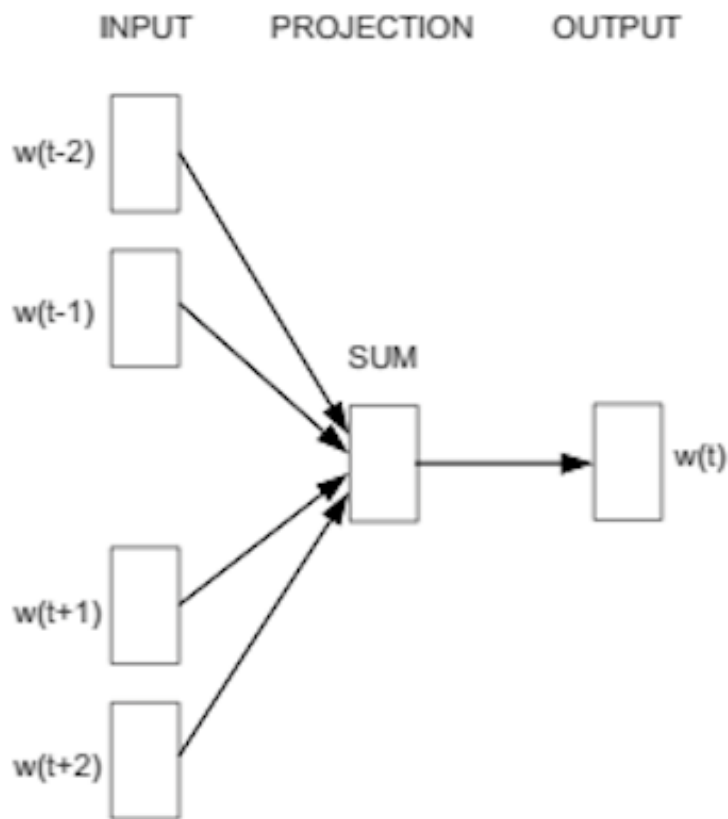
- ❑ MorphoLogic Kft, MTA NyTI, SzTE: 2007-ben fejeződött be a fejlesztése
- ❑ Kutatási célokra hozzáférhető: <https://github.com/dlt-rilmta/huwn.rdf>
- ❑ **Forrásai:**
 - idegennyelvűek (pl. más WordNetek)
 - kétnyelvűek (pl. elektronikus szótárak)
 - egynyelvűek (pl. ÉKSZ, szinonimaszótárak)
- ❑ 42.288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

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

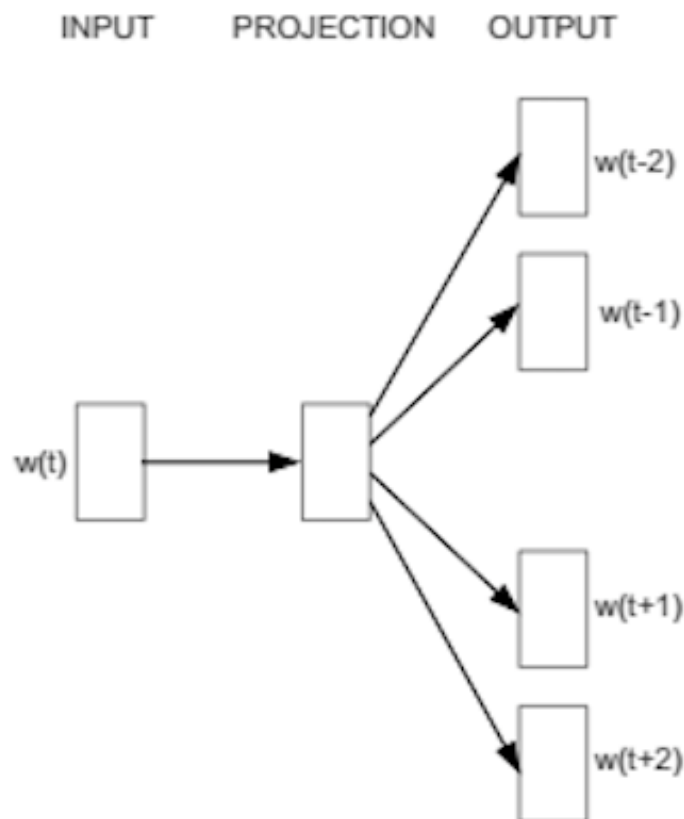
Mélytanulás és szójelentés

- 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 \rightarrow szó vagy szó \rightarrow környezet



CBOW



Skip-gram

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

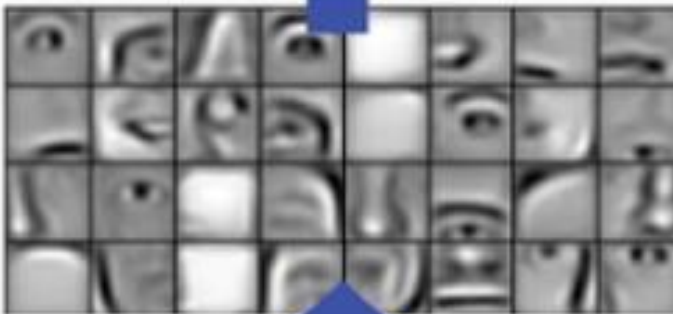
Mélytanulás

- 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

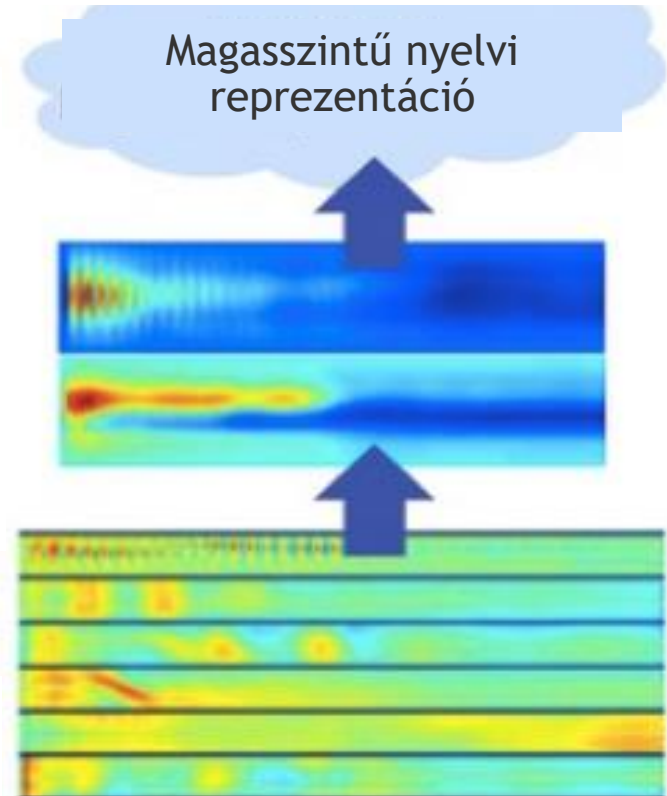


A részek összekapcsolása objektumokká

Második réteg

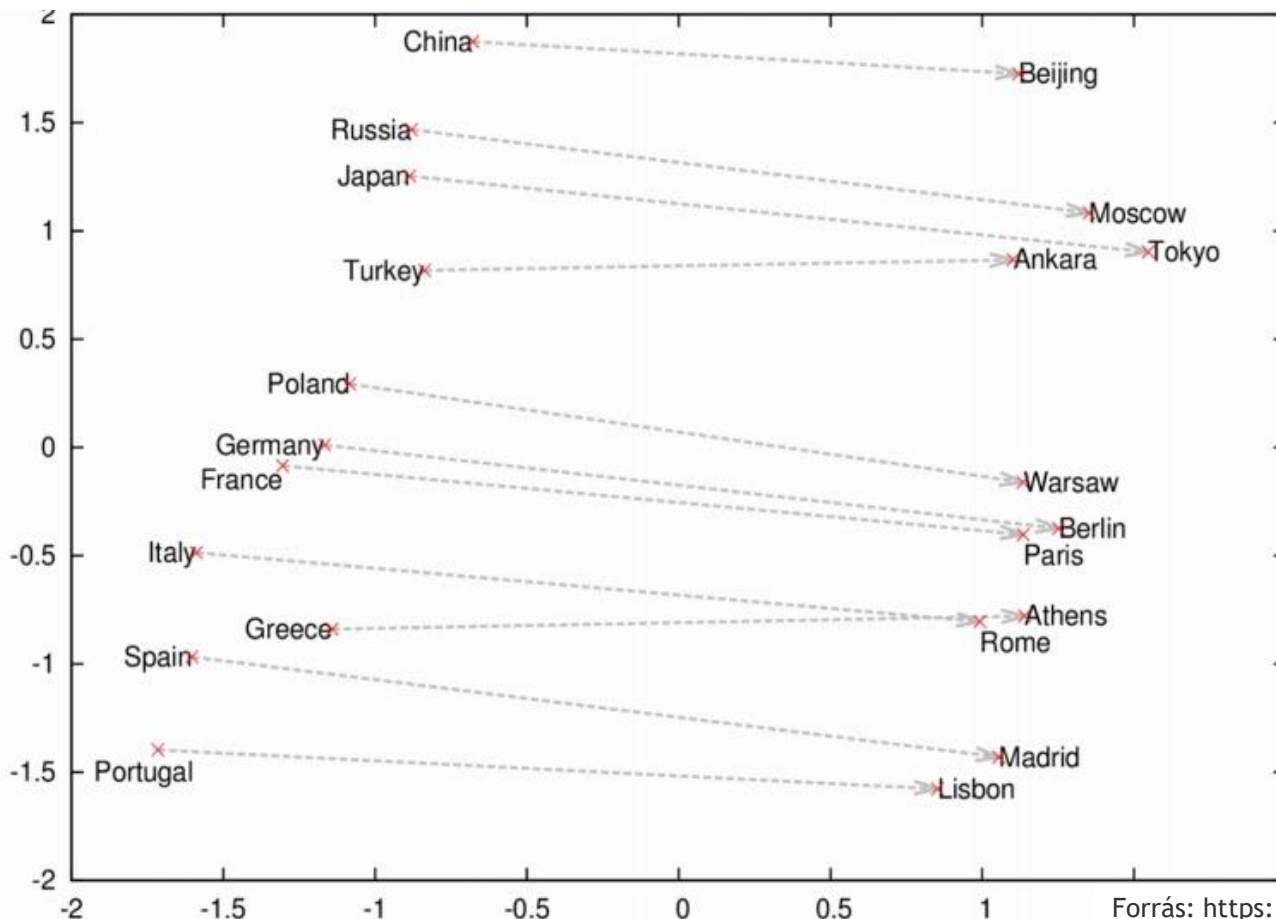


Első réteg



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	zsivány	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	nap szemcs	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