



Szemantika, lexikográfia: jelentés-egyértelműsítés, wordnetek



Bevezetés

- **Lexikális szemantika**
 - Jelentés-egyértelműsítés
 - Szójelentés megadása
- **Lexikológia: szótártan**
 - Szavak jelentésének megadása
 - Mi a jelentés? Hogyan adható meg?

Jelentés-egyértelműsítés

- ***Word sense disambiguation:*** Egy szóalak jelentésének kiválasztása, egy előre megadott (lehetséges jelentés-) halmazból
 - A lehetséges jelentések ált. egy szótárból
 - Osztályozási modellek alkalmazhatók (előre egyértelműsített példák szükségesek)
- ***Word sense discrimination:*** Az adott szóalak különböző használati eseteit (~jelentés) elkülöníteni, anélkül, hogy a lehetséges jelentéseket kívülről megadnánk /és így címkézett példáink sem lehetnek/
 - Felügyelet nélküli statisztikai modellek

WSD

- **Típusai**

- All words

A cél az összes szóalak egyértelműsítése, folyó szövegben

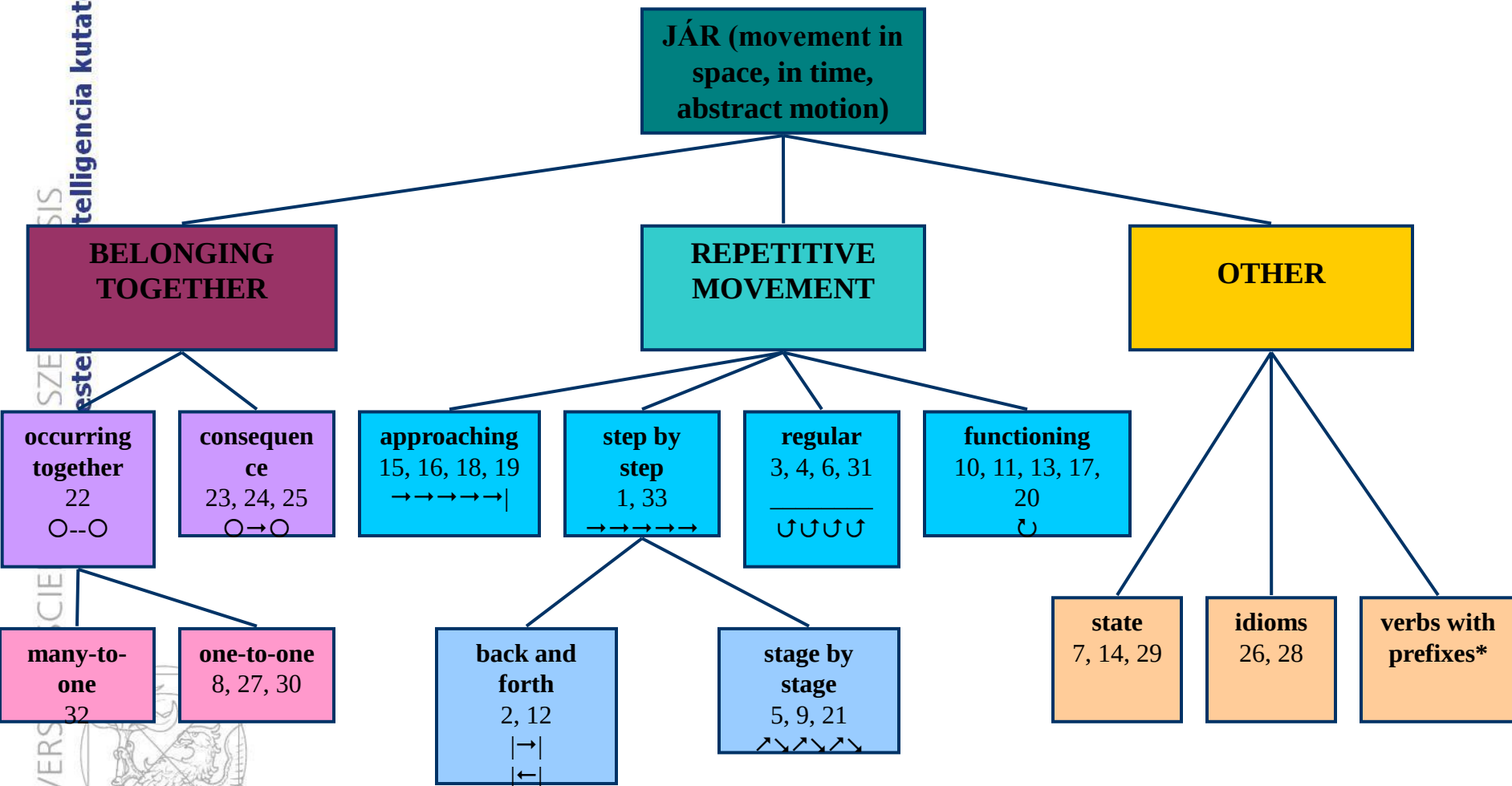
- Lexical sample

A cél bizonyos szóalakok egyértelműsítése, minden szóalakra önálló modell

- Finom jelentésmegkülönböztetés (min. 6-8 jelentés)

- Durva jelentésmegkülönböztetés (max. 3-4 jelentés)







WSD magyarra

- elkészült Szegeden az első (Lexical Sample) tanítókörpusz magyarra (finom jelentésmegkülönböztetés)
- melléknév:
anyagi, élő, erős, képes, pontos, szociális
- főnév:
civil, család, élet, ház, helyzet, intézmény, iskola, kép, képviselő, kormány, nap, oldal, ország, perc, pont, program, század, személy, szervezet, tanár, világ, víz
- ige:
függ, hat, jár, kap, kerül, marad, rendelkezik, szerepel, tart, tartozik, tud, válik

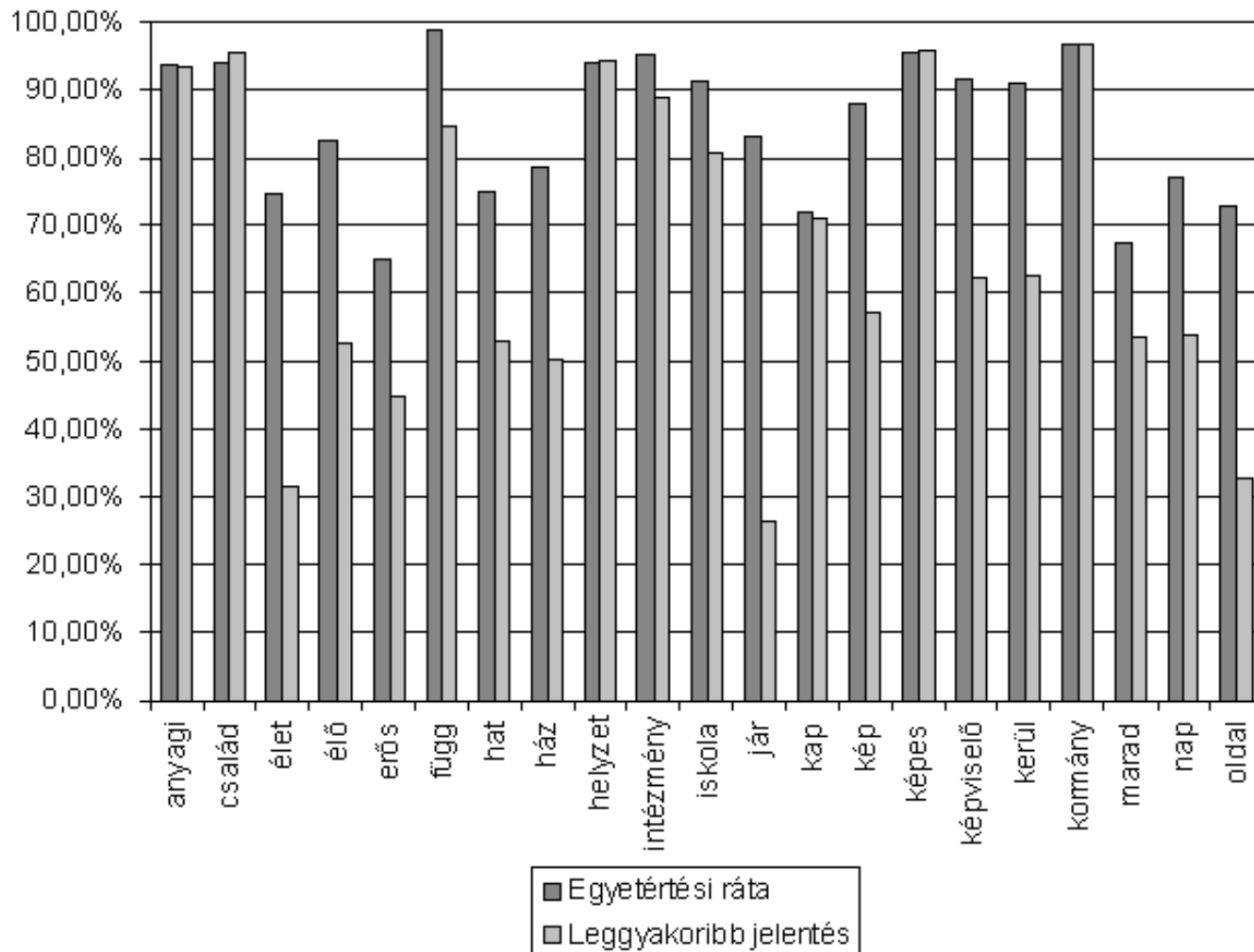
Magyar WSD-statisztikák

	összes példa	címkézett	összes jelentés	korpuszban megjelenő jelentés
anyagi (a)	372	371	3	3
család (n)	629	400	6	4
élet (n)	1401	398	7	6
elő (a)	609	386	9	9
erős (a)	583	398	10	9
függ (v)	278	278	4	3
hat (v)	162	157	6	6
ház (n)	749	400	9	8
helyzet (n)	1554	399	6	5
intézmény (n)	693	399	3	2
iskola (n)	343	343	7	6
jár (v)	1447	400	16	14
kap (v)	2349	400	8	7



Magyar WSD-statisztikák

Annotátor egyetértési ráta és leggyakoribb jelentés baseline



Példa: nap

- Hőt, fényt, stb. sugárzó égitest, amely körül a Föld is kering
- Napfény, napsütés
- Körülbelül 24 óra
- Reggeltől estig, pl. munkanap
- A hét napjai; dátumhoz köthető napok
- Egyéb: pl. más csillagok, kifejezések, stb.



Szöveg

Az éjféli **nap** a Föld sarkai körül észlelhető jelenség. Az év bizonyos részében a sarkköröknél ($66,33^\circ$) magasabb földrajzi szélességeken a **Nap** még éjfélkor is, azaz 24 órán át a horizont felett marad. Az említett $66,33^\circ$ szélességi fokon a jelenség minden évben egy-egy **nap** fordul elő. A jelenség megfigyelhető Alaszkában, Észak-Kanadában, Grönlandon, Izlandon, Észak-Norvégiában, Svédországban és Finnországban, valamint Oroszország északi részén.

Minél északabbra haladunk, a jelenség annál tovább tart, például Finnország legészakabbra fekvő pontján a nyár folyamán a **Nap** 73 napig nem bukik a horizont alá. A Föld két sarkán gyakorlatilag fél évig tart a **nappal**, és fél évig az éjszaka. Az északi sarkon március 21-e táján kel fel és szeptember 23. környékén nyugszik le a **Nap**. A délin fordítva. Tehát a sarkokon a **Nap** évente csak egyszer kel fel, és csak egyszer nyugszik le.

A fehér éjszaka a 60° szélességi fok felett az északi és déli féltekén egyaránt megfigyelhető, gyakorlatilag éjszakai alkonyinak is nevezhető. Az északi vidékeken a nyári **napforduló** körüli hetekben a **Nap** sehol sem süllyed 6° -nál mélyebben a látóhatár alá, ezért a **nappali** világosság nem oszlik el egészen. A déli féltekén, az Antarktisz táján december 22-e körül fehérlenek az éjszakák.

Szöveg

- És Isten szólt: »Legyen világosság!« És lett világosság. Látta Isten, hogy a világosság jó. Elválasztotta a világosságot a sötétségtől, és elnevezte a világosságot **nappalnak**, a sötétséget pedig éjszakának. Akkor este és reggel lett: egy **nap**.
- Azt mondta ezután Isten: »Legyenek világítók az ég boltozatán! Válasszák el a **nappalt** az éjszakától, jelezzék az időket, a **napokat** és az esztendőket, ragyogjanak az ég boltozatán, és világítsanak a földre!« Úgy is lett. Megalkotta tehát Isten a két nagy világítót – a nagyobbik világítót, hogy uralkodjék a **nappalon**, meg a kisebbik világítót, hogy uralkodjék az éjszakán – és a csillagokat. Az ég boltozatára helyezte őket, hogy világítsanak a földre, s uralkodjanak a **nappalon** és az éjszakán, és válasszák el a világosságot a sötétségtől. És látta Isten, hogy jó. És lett este és reggel: a negyedik **nap**.



Lexikográfia

- **Lexikográfia:** szókincs (szavak és kifejezések) kutatása, gyakorlati szempontú leírása
- **Szótár:** szavak (és jellemzőinek) gyűjteménye
- **Ontológia:** Lételemélet? Tudás megosztása? Számítógépes szótár? Fogalmi háló?



Elektronikus szótárak

- Hagyományos (papír)szótár
- Elektronikus szótár
 - Könnyebb és gyorsabb a keresés (szófaji információ is)
 - Toldalékolt alakok is megtalálhatók
 - Nincsenek terjedelmi korlátok



Többszavas kifejezések

- Hol találhatók meg?
 - Alaptagnál, de nem egységes a lexikográfiai gyakorlat
 - Melyik az alaptag? (*kutya* vagy *szalonna*)
- Elektronikus szótárnál nem releváns kérdés



Ontológiák

- Fogalmi hálók
- Viszonyok alapján rendezett fogalmak (alá- és fölérendeltség, hierarchiák...)
- Taxonómiák (pl. biológiai rendszertan)
- Nyelvi ontológiák: nyelvi információt is tartalmaznak a hálón kívül



WordNet

- **Lexikális adatbázis**
- **Fogalmak hálóba rendezve
különféle relációk alapján**
- **Angol: Princeton WordNet**
- **Más nyelvekre is: EuroWordNet
stb.**
- **Synset: szinonimahalmaz**

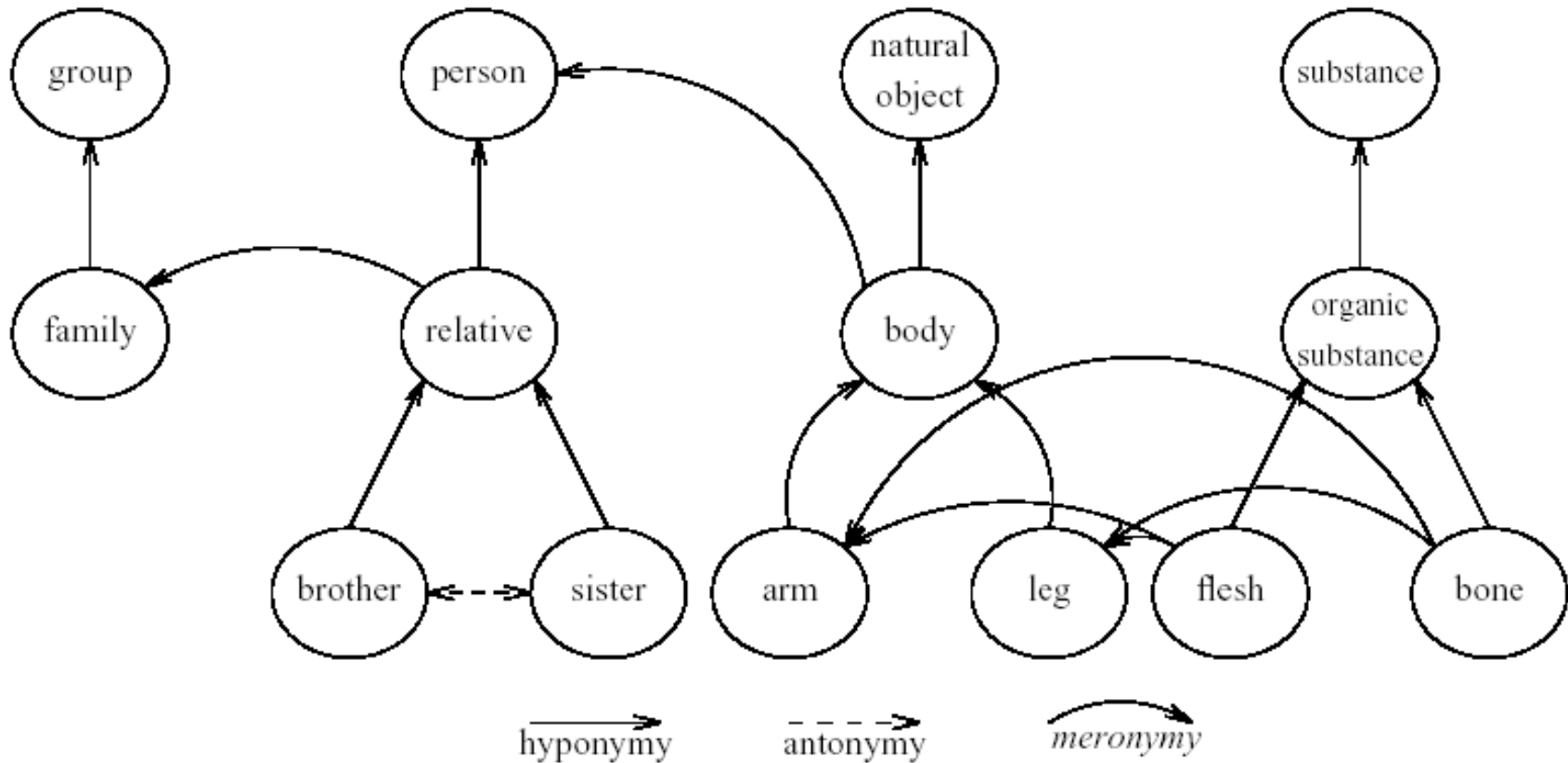


Princeton Wordnet

- **Eredetileg az emberi agy nyelvi tudásreprezentációjának modellje**
 - Ennek szánták
(inkább elméleti nyelvészeti / filozófiai indíttatás)
 - Más, mint egy szótár – szavak szintjén redundancia
- **A legnagyobb, egységes lexikai adatbázis**
 - Ingyenes
- **Gépileg feldolgozható információ**
 - Számítógépes nyelvészet egyik alapvető eszköze lett



Szemléltető ábra





A HuWN bemutatása

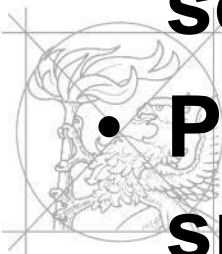
- 40000 synset (általános ontológia) + 2000 üzleti nyelvi synset (szakontológia) + 650 jogi nyelvi synset (szakontológia)
- Főnevek
- Igék
- Melléknevek
- Határozószók
- Ahol csak lehet, a PWN-nek megfeleltetni a synseteket
- Szerkesztéshez használt program: VisDic

Szófaji eloszlás

Szófaj	Synsetek száma
Főnév	33.778
Ige	3.310
Melléknév	4.083
Határozószó	1.038
Összesen	42.209

Synsetek

- Megközelítőleg azonos jelentésű szavak (literálok) alkotnak egy synsetet
- Synseten belül a fő szervező reláció a szinonímia
- Jelentésmegkülönböztetés miatt sorszámozott literálok
- Példa: {zsineg1, zsinór1, madzag1, spárga1}





Synsetek - 2

- Szófaj
- Szinonimák (literálok)
- Definíció
- Tipikus használat – példamondat
- ÉKSz.-beli megfelelő
- Domén
- Relációk más synsetekkel

POS: n ID: ENG20-04397101-n BCS: 2
Synonyms: izé:1, micsoda:1, hogyishívják:1, valami:2

Definition: Amit nem tudunk vagy nem akarunk megnevezni.

Usage: Mi ez az izé a szekrényen?

EKSZ: = izé_1_1

EKSZ: = micsoda_1_7

EKSZ: = hogyishívják_1_1

EKSZ: = valami_1_16

Domain: factotum

SUMO/MILO: + Artifact

Last Edit: nkfp 2005/09/08

--> [hypernym] [n] dolog:3, fizikai dolog:1, objektum:1

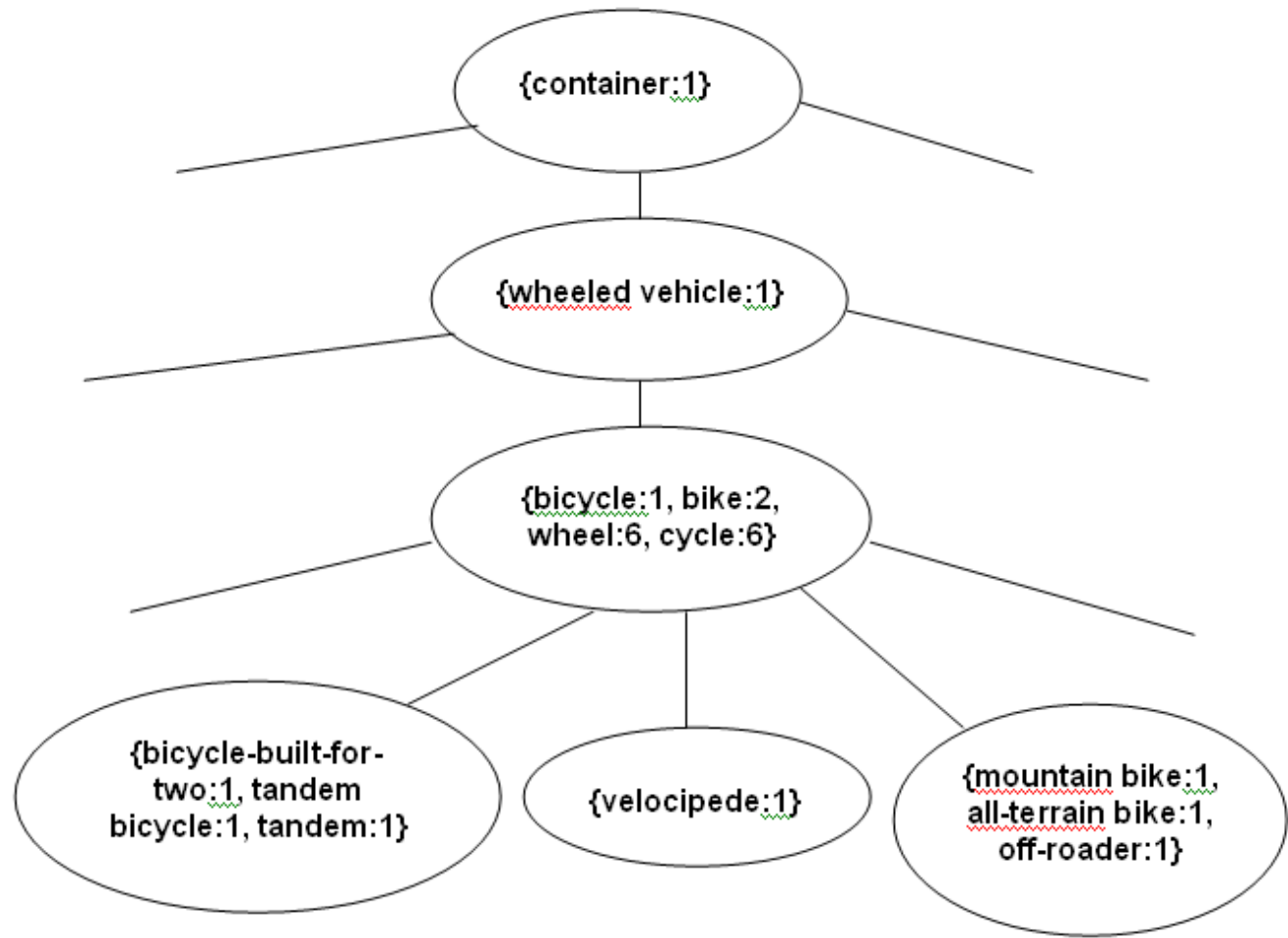
Relációk a HuWN-ben: főnevek

- Alapvető rendező reláció a hipernima-hiponima
- Alá- és fölérendeltségi viszonyokat mutató fák
- Példa: *bicikli*

Különböző szintű hiponimák:

kutyaféle – emlős – állat – entitás

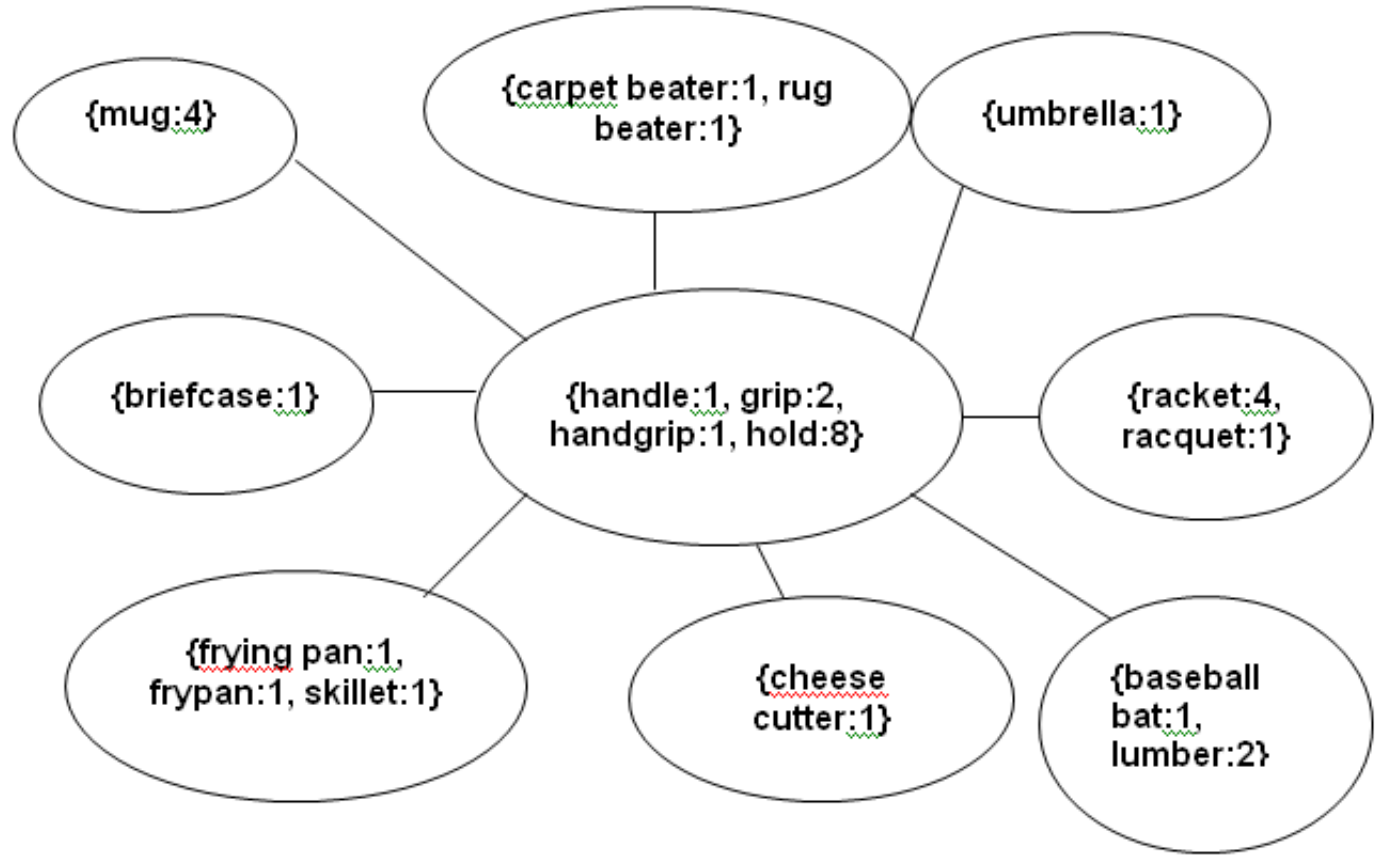




Relációk a HuWN-ben: főnevek - 2

- **Antonímia (near_antonym): ellentét**
 - Példa: *férj – feleség*
- **Holonímia – meronímia: rész-egész viszonyok**
 - Példa: *kalap – karima* (rész)
 - *újság – papír* (anyag)
 - *csatár – focicsapat* (tag)

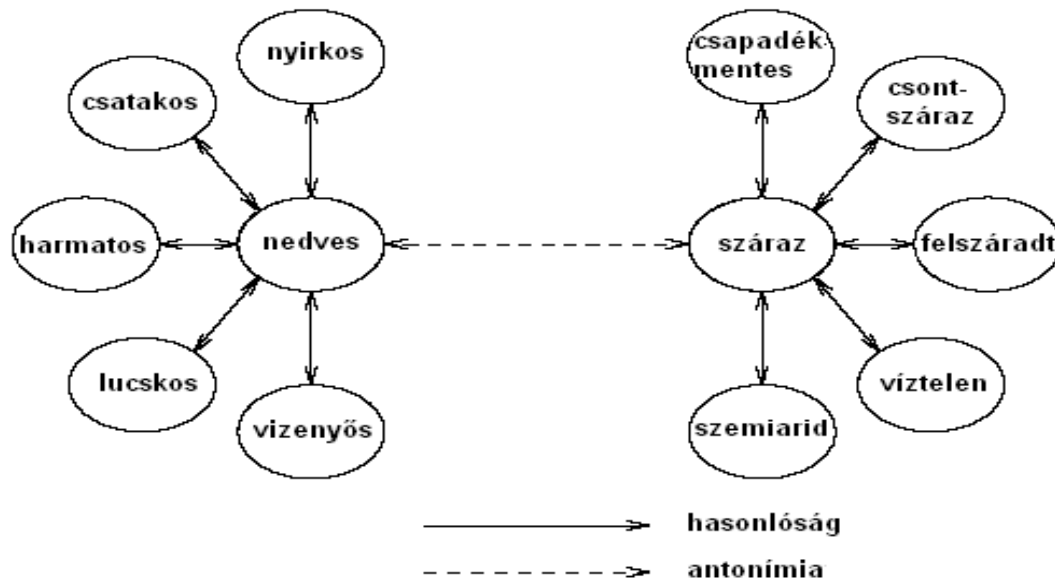






Melléknevek a HuWN-ben

- Alapvető szemantikai kapcsolat: antonímia
- *Hideg – meleg, szép – csúnya* stb.
- Központi literálok köré vannak összegyűjtve a hasonló jelentésű, specifikusabb melléknevek



Speciális igei relációk a HuWN-ben

- Példa: **FELMELEGSZIK**
- **is_preparatory_phase_of:**
előkészítő folyamat (*melegszik*)
- **is_telos_of:** **sikerpont**
(*felmelegszik*)
- **is_consequent_state_of:** **utóállapot**
(*meleg*)



Egyéb igei relációk

- **Antonímia: elad 1 – megvesz 1**
- **Hipernímia – hiponímia: {szerez 4, kap 5, vesz 7 ...} hiponimái bérel 1, megtalál 2, belebotlik 1, kölcsönvesz 1 stb.**



Nyelvek közti relációk

- **eq_xpos_synonym**: ha az angolban és a magyarban eltérő szófaj fejezi ki ugyanazt a fogalmat, ezért közvetlenül nem feleltethetők meg egymásnak

fél (ige) – *afraid* (melléknév)



Mire jó a WordNet?

- Szinonimalisták
- WSD: jelentések előre meg vannak adva
- Dokumentumosztályozás: szavak közti relációk
- Többnyelvű IE
- Gépi(leg segített) fordítás: nyelvek közti megfeleltetések





Synsetek építése

- **Aszparágusz**
- **Madzag**
- **Csirág**
- **Zsinór**
- **Nyúlárnyék**
- **Spárga**
- **Zsineg**