Számítógépes lexikográfia

Computational Lexicography for Natural Language Processing

Bran Boguraev & Ted Briscoe

1989

Longman Publishing Group White Plains, NY, USA

- szavak ismerete mindenféle NLP task alapja
- lexikon kellene, gond: mi legyen benne? Túl sok a szó.
- *Machine Readable Dictionary* (MRD) előnyei és hátrányai:
  - hagyomány - megfelelő kiindulási pont a lexikon tartalmának meghatározásához
  - nagyok - szerkesztési munka nagy része kész
  - a humán felhasználónak készült - olyan előfeltevésekkel él, amik a gépi feldolgozásra nem igazak; pl. hogy érteni fogja a szavak angol nyelven leírt magyarázatát
- két cél:
  - automatikus módszerek, hogy a gép hozzáférjen az *MRD*-kben lévő infókhoz
  - ezeknek az infóknak a használata, különböző NLP rendszerek és mögöttük lévő nyelvészeti elméletek kiértékelésére és fejlesztésére

- Longman (LDOCE)
- számítógépes lexikográfia
  - a lexikonok jelentősen eltérnek majd a hagyományos, nyomtatott szótáraktól, az információ tárolásában és reprezentálásában is
  - ráadásul egyre nagyobb az igény, hogy mindenféle számítógépes technikát alkalmazzanak a humán szótárak létrehozásánál is

1.1 NLP - A tudás

5 fajta tudás releváns az NLP rendszereknek

1. fonológiai tudás
2. morfológiai tudás
3. szintaktikai tudás
4. szemantikai tudás: szavak jelentése, ezek hogyan kombinálódnak a mondatok jelentésévé
5. pragmatikai vagy enciklopédikus tudás: pl. anaforafeloldás, ellipszis megfejtése. Ezekhez fontos az intenció, a nyelven túli kontextus stb.

A pragmatikai kevéssé a lexikon dolga. (De nem éles a határ.)

1.1 NLP - A tudás

Általános szabályok halmazával jellemezhetők a tudástípusok, pl:

- angolban a melléknevek megjelenhetnek a létige után, ld. *That fish is beautiful* és főnév előtt, ld. *That beautiful fish.*
- *man-eating*? $\rightarrow$ lexikon
  - megjelöljük az összes elemet (*not-predicative*)
  - de a szabály eleve csak azokra vonatkozik, amik lehetnek melléknevek - ezt hol jelöljük?
  - van-e általános szabály?
  - szerintük a POS az a lexikonba megy

1.1 NLP - A tudás

A szintaktikai szabályok frázisstruktúra szabályokkal reprezentálhatók:

Mother $\rightarrow$ Daughter$_i$ Daughter$_j$ ... Daughter$_n$

1.	S	$\rightarrow$	NP VP
2.	VP	$\rightarrow$	V AP[prd +]
3.	NP	$\rightarrow$	Det N
4.	NP	$\rightarrow$	Det AP[prd χ] N
5.	AP[prd χ]	$\rightarrow$	A[prd χ]

Kell a lexikon is hozzá:

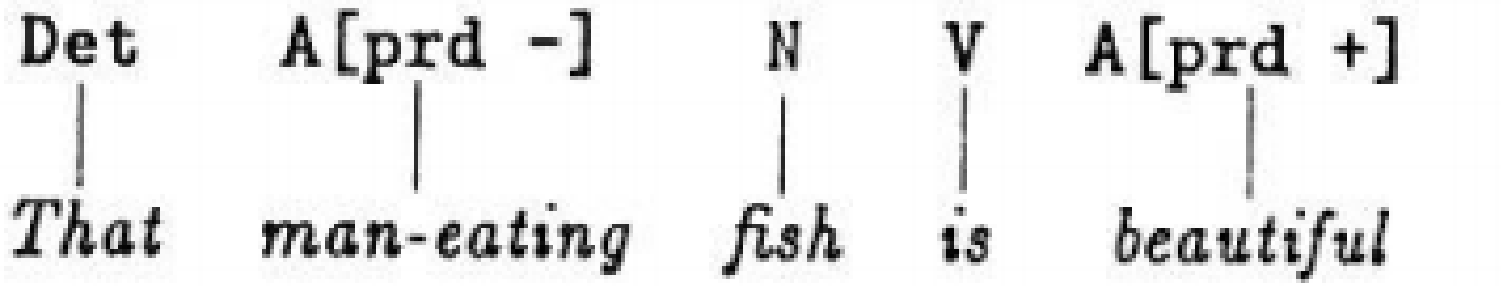
beautiful	:	A[prd +], A[prd -].
fish	:	N.
is	:	V.
man-eating	:	A[prd -].
that	:	Det.

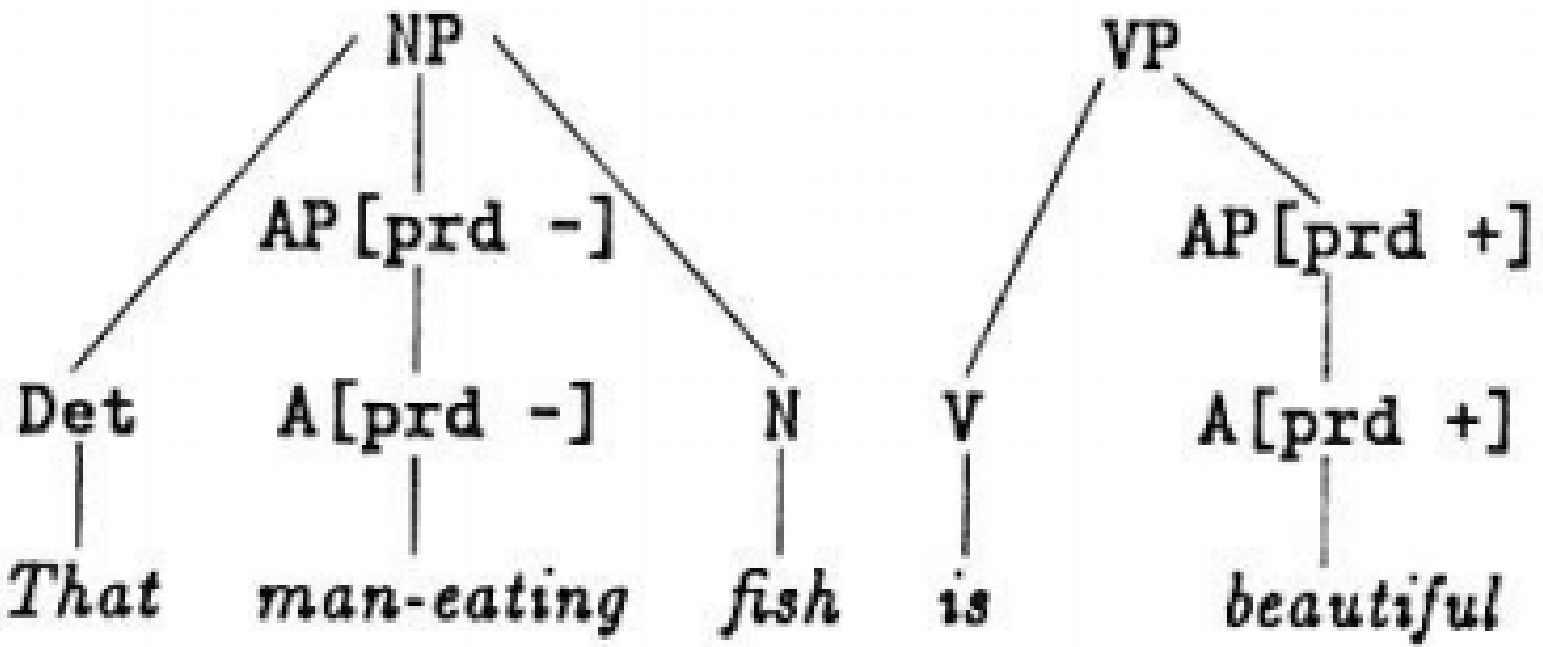
1.1 NLP - A tudás

Az elemzés:

1. Minden szót keress meg a lexikonban, és adj neki egy szintaktikai kategóriát.
2. Balról jobbra haladva próbáld meg illeszteni a szabályokat
3. Próbálj meg egy teljes fát építeni S csomópont alá

That man-eating fish is beautiful.





1.1 NLP - A tudás

Lényeg: tök jó ez a lexikon, de kb. minden új szabállyal változtatni kell rajta (pl. létige-nem létige)
A lexikonba az kell, ami megjósolhatatlan vagy idioszinkratikus. Ami szabályalapú, az nem. (*re*-)
Tehát addig a lexikont nem tudom megcsinálni, míg nincsenek kész a szabályok.
Szó? Nem, morféma. *re*

1.2 Számítógépes lexikográfia

NLP-rendszerek baja, hogy limitált az alkalmazhatóságuk - mert kevés lexikai *entry*-re íródtak.
Ráadásul az adott feladatra koncentrálva, specifikusan, nem általánosan: minden új feladathoz lényegében újra kell építeni a lexikont.
Hasonló infót ráadásul máshogy reprezentálnak.

1.2 Számítógépes lexikográfia

```
[ACKNOWLEDGE
  Category: V
  Base:     acknowledge
  Features: (TRANSITIVE (REALNP) (PASSIVIZES)) ;; 1. 2. 3. 4. 5.
            (CLAUSE (REALNP) (THATCOMP)        ;; 1.
             (INDICATIVE :TENSE) (WH-))
            (NP-VP :AGR :AGRX (REALNP) :AGRX   ;; 1.
             (PASSIVIZES) (INF) (WH-))]
```

```
[ACKNOWLEDGE
  FEATURES  (TRANS                    ;; 1. 2. 3. 4. 5.
             PASSIVE                  ;; 1. 2. 3. 4. 5.
             THATCOMP                 ;; 1.
             THATREQUIRED             ;; 1.
             NPTOCOMP)                ;; 1.
  V S-D]
```

```
(acknowledge                                          ;; 1, 2, 3, 4, 5
   ((v +) (n -) (subcat np)) acknowledge nil)         ;;
(acknowledge                                          ;; 1
   ((v +) (n -) (subcat sfin)) acknowledge nil)       ;;
   ; acknowledged that they were defeated
(acknowledge                                          ;; 1
   ((v +) (n -) (subcat se3)) acknowledge nil)        ;;
   ; acknowledge having been defeated
(acknowledge                                          ;; 1
   ((v +) (n -) (subcat or)) acknowledge nil)
   ; acknowledge him to be the best
```

1.2 Számítógépes lexikográfia

Megoldás: egy olyan szótár-adatbázis, amiből mindenféle speciális célra építhetők lexikonok.
Egy szóhoz minden olyan infó, amely bármely NLP feladat közben szükséges lehet.
Ideológiailag a lehető legkötetlenebb formában.
A legteljesebb szószedet kell.
Nem újat, kézzel, hanem a meglévőket módosítva és megfelelően konvertálva.
2 probléma lesz ekkor:

1. humán felhasználó nyelvi tudása, világismerete az alap
2. nem elég szisztematikus az információ, ráadásul nem megfelelő nyelvi modelleken alapul

1.2.1 A szótári bejegyzés természetéről

Több szótárban megnézzük a bejegyzésekben található információkat, különös tekintettel a Longmanre.
Ezt összehasonlítjuk azzal, hogy milyen infó kell a különböző NLP-rendszereknek.
A legtöbb szótárban a bejegyzések egy adott szóalak távoli homográfjait sorolják fel: az adott alak jelentéseinek halmaza, ha éppen ige, főnév vagy akármi.
Minden bejegyzés felosztható a következőkre:

1. alak (*form*): címszó, kiejtés, kötőjelezés mindenféle variációkkal; esetleg a használatról (szleng), ragozással kapcs. infó, hangsúlyok...
2. funkció (*function*): a viselkedés; sima szófajcímke, vagy részletes grammatikai alkategorizáció
3. jelentés: definíciók, példák, keresztreferenciák
4. sok egyéb: ábrák, szinonimák, antonimák, etimológia stb.

sand-wich[1] /ˈsændwɪtʃ, ˈsænwɪtʃ / *n* **1 2** flat pieces of bread with some other usu. cold food between them, eaten with the hands **2** *BrE* a cake of 2 flat parts with JAM[3] (1) or cream between them

sandwich[2] *v* [X9, esp. IN, *between*] *not fml* **1** to put tighly in between 2 things of different kind: *a film of plastic sandwiched between 2 pieces of glass* **2** to find time for, among other events: *I am very busy today but I'll try to sandwich that job in after tea*

1.2.1 A szótári bejegyzés természetéről

A Longman MRD verziója egyéb infókat is tartalmaz, pl. *box codes* a *sandwich*-nek: - - L - X - - - - S

- X: absztrakt vagy emberi alany
- S: szilárd tárgy

Van még *subject code* is, ami a használati környezetről szól; a *sandwich* első jelentésének *subject code*-ja FO = food.
Mind szép és jó, de az NLP-rendszereknek mi kell?
Jelentések - eleve feltételez "szövegértést". Persze kisebb szótárral (Longman: 2200).
Szintaktikai és szemantikai elemzések és a kétféle tudás: hol a hangsúly?
Van rendszer, ami mindent a lexikonba rak - nem ugyanaz a szótári bejegyzés kell egy olyan rendszernek, ami mondjuk élesen elválasztja a szintaktikai, a szemantikai, és a pragmatikai tudást.
És akkor még mindig kérdés a világismeret, ami az alap.

1.2.2 Szótár-elrendezés és reprezentáció

"Szemetes" az anyag. Pl.

- beilleszti a nyelvi kódba a lexikográfus, hogy *usu. pass*; hol középen, hol a végén ⇒ nem egységes a szerkezet
- néha esztétikai okokból történik valami (pl. ilyen-olyan szóköz), néha meg azért, mert jelentése van (pl. betűtípus-váltás)
- sima elütések
- jelentésnél: keresztbe-kasul hivatkozunk
- morfológia hiányzik! (Kay & Kaplan véges állapotú automatája, Koskenniemi kétszintes morfológiája stb.)
- címszavak, ABC: ez az egy elérési útvonal (pl. beszédfelismerő inkább a fonetikát keresné)

1.3 Az MRD-vel való munka az NLP-ben

1. Szólista; pl. helyesírás-ellenőrzéshez, gépi fordítás támogatásához stb.; általában csak a szó, semmi egyéb infó, nem érdekes
2. Taxonómiák
3. Elérés,keresés; WordSmith, WORDNET; feladatok pl.: normalizálás
4. Beszédfeldolgozás; hangsúly, fonetikai átírás
5. Szintaktikai elemzés; CRITIQUE, IBM: nyelvi és stílusellenőrző üzleti szövegekhez. Itt különösen izgalmas kérdés a nyelvelmélet.

6. Szemantikai feldolgozás; (eddigiek: forma és funkció) egyértelműsítéshez már a definíció mezőjéhez kell hozzáférnünk. Lexikai bázis vs. tudásbázis. Cél: természetes nyelvi szavakat kapcsoljunk fogalmak taxonómiájába.
   - NLP rendszerekben a lexikális szemantikát és a speciális világismeretet el kell választani - az utóbbi nem igazán mehet MRD-k alapján, az előbbi inkább.
   - Nem lehet olyan automatikus az előállítás, mint egy szintaktikai komponensnél. Inkább okos elérés.
   - Milyen struktúra legyen? Általában hierarchikus (*frame*-ek, szerepek stb.)
   - *Oxford Dictionary of Medicine* $\rightarrow$ medical KR
   - lexikai egyértelműsítéshez: implicit kapcsolatok a jelentések között, keresgélés a szótári definíciók mögötti fogalmak hálózatában. Mindezt szintaktikai feldolgozás nélkül.
7. Generálás

1.4 Az MRD-k megbízhatósága és használhatósága

1. milyen információt találunk a szótárban?
    - mennyire reprezentatív? (méret)
    - mennyire megbízható? (tipográfia)
2. hogy van az információ rendszerezve és tárolva?, Hogy kezeljük a lekérdezését?
    - nyelvtan leírása?
    - belső inkonzisztenciák, p. *box*, *container* és *contain*, *hold back*, *hold*, *back*

Egy könnyen kezelhető gépi szótár mint a számítógépes szemantika erőforrása

Wilks, Fass, Guo, McDonald, Plate, Slator (New Mexico)

9.1 Bevezetés - review

- minden korai AI lényegében számítógépes szemantika; Schank, Wilks, Winograd, Charniak
- három dologgal lehet összehasonlítani: szintaktikai elemzéssel, logikai szemantikával és *expert systems*
- konnekcionizmus fontos (apró számítási egységek, sok-sok kapcsolattal, tapasztalati úton tanulással (=súlyozás)
  - hangsúly a nem-szintaktikai módszereken
  - folytonosság világismeretek más formáival
  - nehézségek a logikai-alapú megközelítésekkel szembeni pozícióerősítésben
- eltérés két konnekcionista irányzatban: szubszimbolikus megközelítés (Smolensky) és a lokalista;előbbiben nincs külön reprezentációja a szójelentéseknek, hanem egy nemszimbolikus reprezentáció különböző aspektusai lennének, válogatott, súlyozott élek gyűjteménye. Az utóbbinál a szójelentések is jelen vannak a szimbolikus reprezentációkban

9.1 Bevezetés - review

- diszkrét jelentések vs. folytonosság; és ha választunk a jelentések között, akkor mikor történik ez?
- fontos alapvetések:
  - tudás és nyelv szétválaszthatatlansága
  - a nyelvi struktúrák, szövegstruktúrák valójáűban a tudásstruktúrák paradigmái
  - ebből következik, hogy a tudást szöveg formájában kell tárolni (nem pl. elsőrendű logika nyelvén)
- nyelvi szöveg egy példája a szótár; olyan szöveg, amelynek tárgya a nyelv
- épp ezért várható, hogy a nyelv szemantikai struktúrája ezekben lesz a leginkább explicit; így elérhető és összehasonlítható a tudásreprezentácuiók szemantikai szerkezetével
- és valóban: hasonlították már össze szótárak szemantikai szerkezetét tudásreprezentációk szerveződésével, sok hasonlóság, pl. *genus term*-ök, amik hierarchiába rendezhetők

9.1 Bevezetés - review

Tehát: közös alapelvek a nyelvi szöveg szemantikai struktúrájában és a tudásreprezentációk szemantikai struktúrájában.
Milyen elvek?
Tudáselsajátításban fontos.
Szemantikai infó kinyerése MRD-kből. Automatikusan.
A számítógépes szemantika elméleti kérdései és a tudáselsajátítás és a számítógépes lexikográfia kérdései között van átfedés.
Közös kérdések:

9.1 Bevezetés - review

1. helyes-e feltételeznünk a szavaknak "értelmet", vizsgálat nélkül, egyenesen a hagyományos lexikográfiából eredeztetve, amely az MRD-ket létrehozta? Alapozható a számítógépes szemantika erre a fogalomra? (Igen)
   - sok NLP feladat, pl. MT azért bukott meg, mert a lexikális többértelműséggel nem tudtak mit kezdeni a programok.
   - ellenérv: nem is volt igazi többértelműség, míg el nem kezdtünk szótárakat írni
   - nincs is többértelműség, *play1, play2* és *play3* van
   - ez sem teljesen oké; nem lehet géppel szimulálni a fordítás folyamatát a lexikális többértelműség valamilyen reprezentációja nélkül.
   - de van még baj: az A szótár *play3* jelentésének köze se lesz a B szótár nyolc különböző *play* jelentésének egyikéhez sem
   - megoldás: tanulnia kell tudni a rendszereknek!
   - még egy baj: a magyarázat is tartalmazhat többértelmű szavakat
     1.1 utánozzuk az embert, fusson le egy egyértelműsítő program a definíción, mielőtt nekiállnánk megérteni
     1.2 ne legyen többértelműség a definíciókban
     1.3 ne foglalkozzunk a problémával, amíg nem lesz feltétlen szükséges

9.1 Bevezetés - review

2. elégséges-e a szótár? elég erős tudásbázis-e?
    - nem elég; van olyan szemantikai infó, ami nem elérhető a szótárból, ezért máshonnan kell megszerezni
    - elég, csak sok infó implicit, a szótár más részeiből kell összevadászni
3. kivonatolható-e? Lehet ezt géppel?
4. *bootstrapping*? Amivel indul egy algoritmus.
    - belső: magából a szótárból szedi az induláshoz az infót
    - külső: máshonnan, más eljárásból (*noun*-ról máshonnan tudás, hogy szintaktikai kategóriák stb.)

9.2 Szemantikai információ kinyerése az LDOCE-ből: az LDOCE

Az LDOCE

- angolul tanulóknak, 55 000 bejegyzés könyvként, 41 100 bejegyzés MRD-ként
- bejegyzés (*entry*): egy vagy több jelentés gyűjteménye, melyek a következő címszóig tartanak
- címszó (*head*): a szó, kifejezés vagy kötőjeles szó, melyet egy bejegyzés meghatároz
- jelentésmeghatározás (*sense definition*): definíciók, példák és más, a címszó egy jelentésével kapcsolatos szövegek halmaza. Ha egy bejegyzés egynél több jelentésmeghatározást tartalmaz, akkor minden jelentésmeghatározás kap egy sorszámot.
- primitívek és nem-primitívek: elvileg a bejegyzésekhez egy 2000 szavas szótárat használtak, ennek elemei a primitívek

9.2 Az LDOCE

Heads	Words	Entries	Sense Definitions
Primitives	2166	8413	24 115
Non-Primitives	25 592	32 687	49 998
Totals	27 758	41 100	74 113

2219 szó az alapszótárban; ebből 59 prefix és suffix ki és 35 szó ki, mert nincs is címszóként. Kb. 30 szó hozzáadva, mert gyakran szerepel definíciókban, hiába nem alapszó. = 2166

Mi a nagyon meglepő a táblázatban?

110 szintaktikai kategória (*noun/count/followed-by-infinitive-with-to*) *box codes*: ABSTRACT, CONCRETE, ANIMATE, típushierarchiába rendezve

pragmatic codes = subject codes: ENGINEERING, ELECTRICAL

Három használatot mutatnak be. Elégséges és kivonatolható az infó, de a *bootstrapping* kapcsán eltérnek.

9.2.1 Az LDOCE - együttes előfordulás statisztikája (Tony Plate)

- *bootstrap* nélkül
- alap: minden olyan mondat, ami tartalmaz egy adott szót, információforrás az adott szó használatáról, nem csak a definíció
- az alap alapja: erős a kapcsolat két szó együttes előfordulásának gyakorisága és a szemantikai kapcsolatuk között
- 6 téma:
  1. az együttes előfordulásos technika és a többi technika kapcsolata
  2. az együttes előfordulásos adatok gyűjtése
  3. vizsgálatok arról, hogy milyen kapcsolat rejtőzik az együttes előfordulásos adatokban
  4. pszichológiai vizsgálatok, amelyek alátámasztják az előfeltevést
  5. spekulációk arra vonatkozóan, hogy az NLP-rendszereknek hogyan kéne használniuk az együttes előfordulásos adatokat az egyértelműsítéses feladatokban
  6. kapcsolódás korábbi munkákhoz

9.2.1 Együttes előfordulásos rész

1. a fejezet más megközelítéseivel való viszony
   - elég a Longmanben lévő jelentésmeghatározás
   - elégséges
   - kivonatolható
   - teljesen belső *bootstrapping*: az együttes előfordulás statisztikájának első körös gyűjtése tekinthető annak. Két dolog teszi lehetővé a teljesen belső *bootstrapping*et:
     - az előfordulások feltételes valószínűségének a szemantikai kapcsolat mérőszámaként történő értelmezése
     - a szemantikai kapcsolat mértéknek használata bázisként, ahonnan a szóról és jelentéseiről további információt nyerünk ki

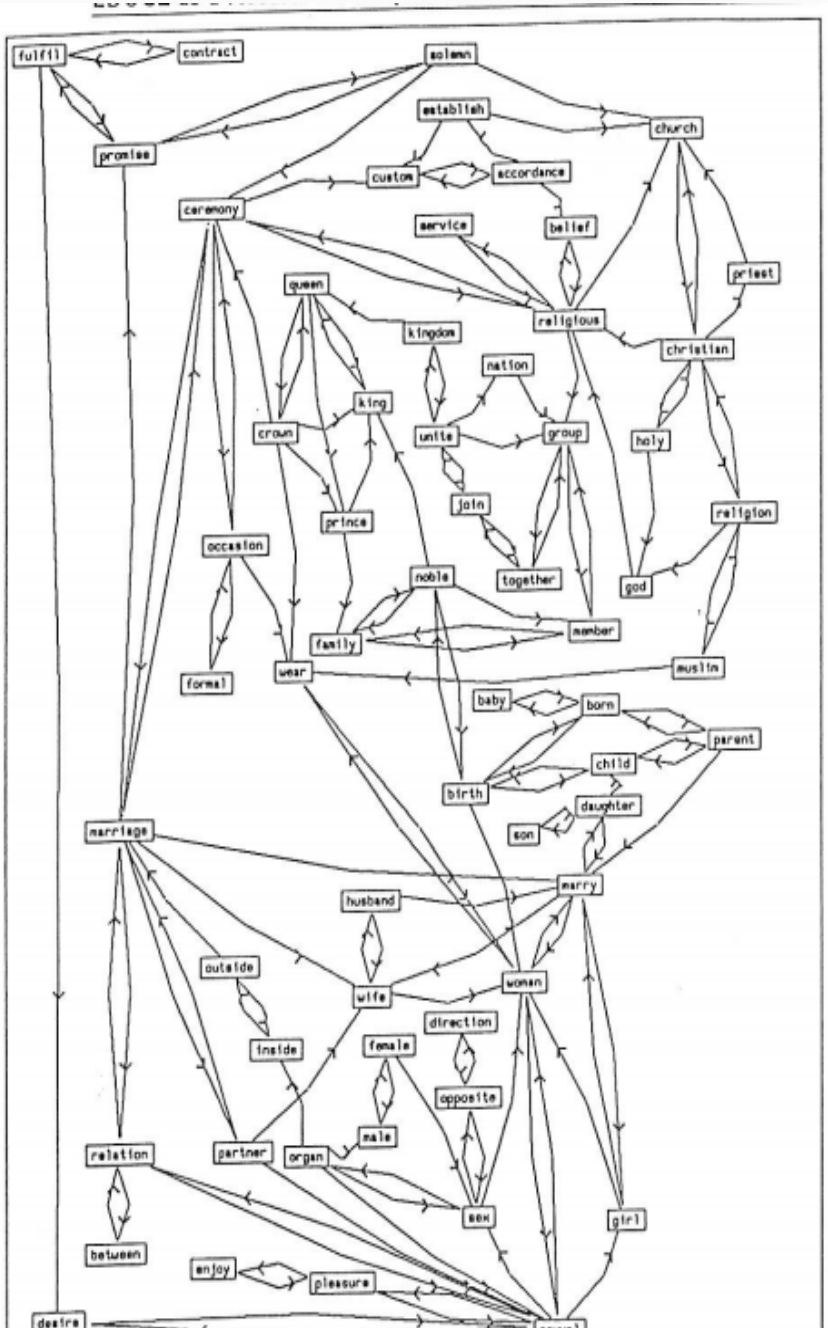
9.2.1 Együttes előfordulásos rész

2. az együtteselőfordulás-mátrixok létrehozása
Az együttes előfordulás adata szópárosok együttes előfordulásának gyakoriságát rögzíti egy adott szövegegységen belül (ha nincs külön megmondva, akkor a jelentésmeghatározás az).
A "sima" előfordulás is fontos.
LDOCE jó alapanyag: primitívek, meg a definíciók stb.
Gyűjtés: rendezetlen listaként tekintünk a definícióban lévő szavakra; szótövek; csak a primitívek.
Gond: kis méret. Van olyan primitív, ami 30-nál kevesebbszer fordul elő.

9.2.1 Együttes előfordulásos rész

3. Az együttes előfordulás adatainak kiaknázása
2.5 millió gyakoriság (2200x2200 mátrixháromszög)
Túl nagy. Töröljük a zajt.
PATH-FINDER program. Pszichológiai adatokban a hálózati szerkezet megtalálása. Élek megmagyarázása. (A $\leftrightarrow$ C és A $\leftrightarrow$ B és B $\leftrightarrow$ C, akkor ha lehet töröljük az elsőt.)
BROWSE. Kifejezetten együttes előfordulási mátrixokra. Részmátrixok, mindenféle küszöbértékek alapján.
Ennek eredménye odaadható a PATH-FINDER-nek, és a sűrű hálózatból ritkát csinál
Innen az ihlet, hogy az együttes előfordulás jó alapanyag szemantikai információkhoz.

9.2.1



X if $\max(\Pr(X \mid marriage) - \Pr(X),$
$\Pr(marriage \mid X) - \Pr(marriage)) \geq 0.0$
Y if $\max(\Pr(X \mid Y) - \Pr(X),$
$\Pr(Y \mid X) - \Pr(Y)) \geq 0.03(2),$
where X is one of the words satisfying 9.1

9.2.1 Együttes előfordulásos rész

4. Pszichológiai validálás Ezeket a mátrixokat összehasonlították kapcsolatok megítélési mátrixaival (humán döntések).
N szóra az alanyoknak értékelnie kellett az $N(N-1)/2$ szópár rokonságát (*relatedness).*
Több kísérlet. Egynek az eredményei röviden:

- 5 ember, 20 szavas halmaz
- alanyok közti korreláció $0.7 - 0.83$ (1: azonos, 0: nincs kapcsolat, -1: pont az ellentét)
- korreláció a humán értékelés átlaga és az együttes előfordulás feltételes valószínűsége között 0.66

9.2.1 Együttes előfordulásos rész

5. Lehetséges alkalmazások

- egy adott szóhoz szemantikailag kapcsolódó szavak halmazát meg lehet találni
- két szóhoz tartozó ilyen halmazok annyira fednek át, amennyire a két szó hasonló
- (ez most nem tudom, hogy jön ide:)miért jobb az LDOCE mint bármely szabad szöveg?
  - a primitívek (4.7 MB, 1 óra)
  - jelentésmeghatározások
  - nem csak a gyakran használt jelentések!

9.2.2 Gépesített szótár építése az LDOCE alapján (Cheng-Ming Guo)

A KDV (*key defining vocabulary*)

- 1200 szavas halmaz
- 4 meghatározási ciklussal ezekből előáll az egész szótár (fokozatosan adjuk hozzá a kontrollált szótárt, végül az egész szótárt a KDV-hez)
- fogunk egy szót, ez a jelölt, megnézzük az első három jelentésében a szótöveket; ha mindegyik benne van a KDV-ben, akkor a szó jöhet, amúgy nem. Így bővül körröl-körre. 3 ciklus végére bekerül az egész kontrollált szótár, és akkor elméletileg a negyedik kör végére a teljes szótár.

9.2.2 KDV

- mire jó ez a KDV?
  1. azért jó a KDV, mert kicsire csökkenti az előzetesen szükséges tudást (itt most a tudásstruktúrákat *integrated semantic units*-nak hívjuk (ISU))
  2. azonosíthatók az üres definíciók (*trip* 'journey' és *journey* 'trip')
  3. garantálja, hogy csak olyan szavakat használok egy új definíciónál, amiket már definiáltam korábban
  4. a KDV elemeinek jelentései lesznek a szemantikai primitívek
- még nem tiszta, mi legyen a kritériuma a KDV-kbe kerülő szavaknak; egyelőre gyakoriság és fogalmi egyszerűség
  - 4000 leggyakrabban használt szó az LDOCE definícióiban és a 850 Basic English szó metszete és egy másik vizsgálat 500 leggyakoribb szava
  - azok a szavak, amik egyetlen fő fogalmat fejeznek ki (*go* vs. *marry*)

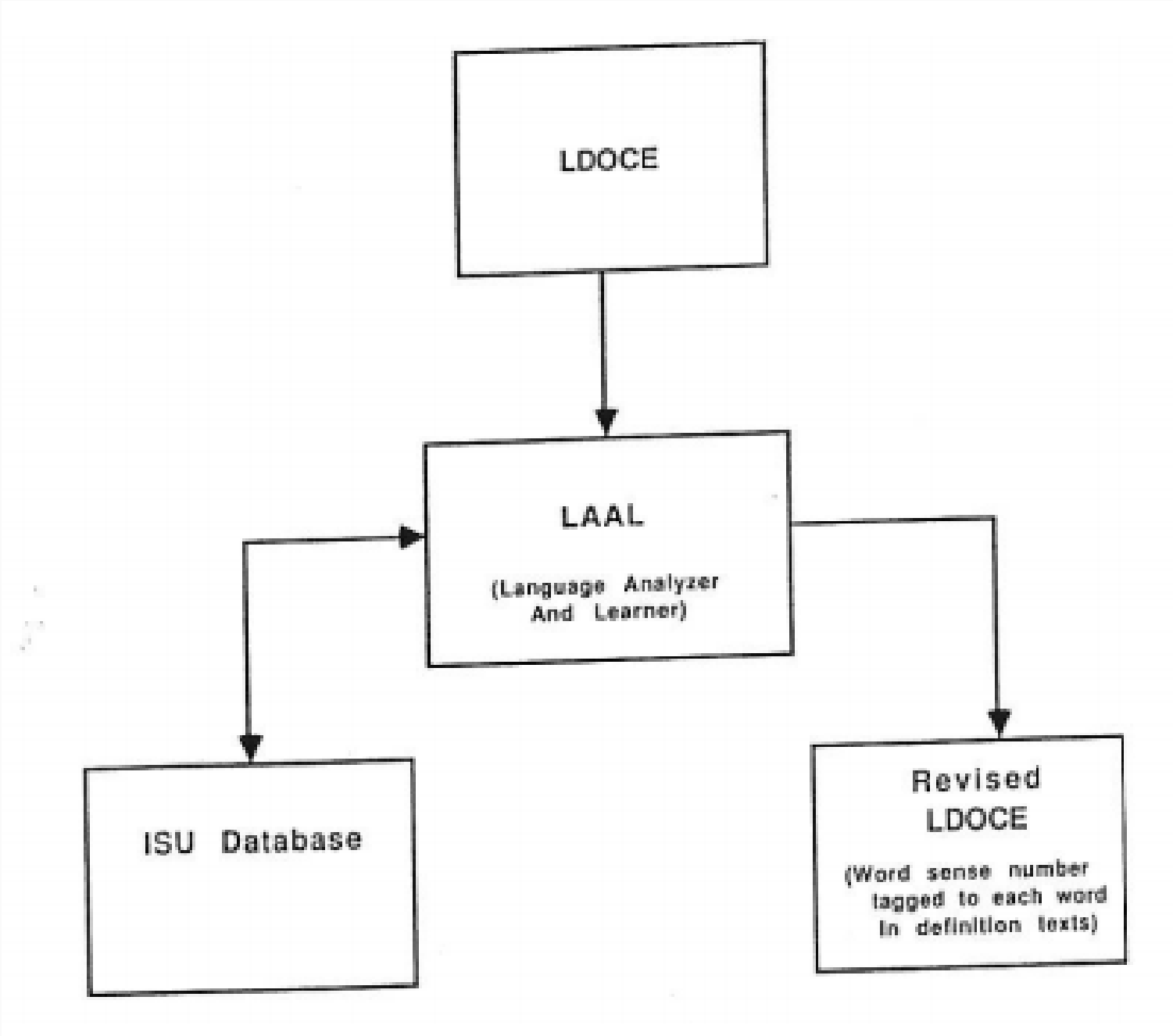
9.2.2 Az ISU-k kézi kódolása

ISU: gazdagabb reprezentációja a szójelentéseknek; nyelvi tudás, világismeret

```
isu(Wordsense,
    belong(Superordinate),
    ik(Integrated_knowledge)
   ).
```

3600 ISU-t kell kézzel megírni.

9.2.2 A *bootstrapping* folyamata



9.2.2 Jelentéshierarchiák és tanulás

- Baj a *genus term*-ökkel: több fölérendelt lehet: *bank* az *land* és *place*
- valójában *bank1* az *land1* és *bank2* az *place1*
- világismeretet a szótárból kell szedni - pl. ágens, páciens stb.
- egy példa: a *marry*
  - *marry1* 'to take a person in marriage'
  - *marry2* 'to perform the ceremony of marriage for (two people)'
  - *marry3* 'to cause to take in marriage'
  - mindegyiknél *H* a *box code* az ágensre és a páciensre
  - 85 helyen fordul elő még definíciókban, mindig *marry1* $\Rightarrow$ definiáló jelentése az ez
  - preferált ágens a *marry1*-re a *human1*
  - *marry2*-re zárójelben meg van adva, így *priest1* vagy *official1*
  - *marry3*-nál következtetek a példamondatból, ahol van egy *daughter* $\Rightarrow$ *parent1*

9.2.3 Egy lexikongyártó (Brian M. Slator)

LDOCE bejegyzések $\rightarrow$ **lexikális szemantikai struktúrák** $\Rightarrow$
tudásalapú elemzésre jók lesznek
Minden lexikális szemantikai struktúra (**keret**) része egy vagy több **hierarchiának**.
Hierarchiák lényege, hogy ne haljunk meg a *My car drinks gasoline* típusú példáknál. Felfelé mászunk a hierarchiában, lazítunk a megszorításainkon.

9.2.3 Egy lexikongyártó - a keretgyártás folyamata

1. lépés: a **pontos nyelvi kódok**at használjuk, de csak általános szemantikai és pragmatikai információkkal, amik könnyen kinyerhetők.
2.: Ha kell (egyértelműsítés stb.), akkor a kereteket módosítom az elemzési fák alapján.
Elemzési fákhoz: ***parser***. Szimpla sémájú mondatok. Csak a tartalmas szavakra, funkciószavakra nem! (Nem az angol nyelvre, hanem az LDOCE definícióira írt *parser*.
A lehető leglaposabbat, a lehető leghosszabb sztringre (a lehető legkevesebb összetevő).
Címszó és definíció viszonya: "IS-A". Keret = intenzió.
Lényeg: kimenet szójelentések kereteinek gyűjteménye, minden keret több hierarchiába illesztve.

9.3 Az LDOCE-ből kinyert szemantikai információ felhasználása

Preference-semantics (preferencia-szemantika):

- egy szöveg jelentését egy komplex szemantikai struktúra reprezentálja, amely kisebb szemantikai összetevőkből áll elő
- a komponensek közötti **kapcsolatok koherencia** és **szemantikai preferencia** alapján épülnek
- az a szemantikai reprezentáció lesz a győztes, ami a **legsűrűbb**
- sűrű reprezentáció: erősek a belső preferenciák
- preferencia számítása: 1) preferencia-illesztő tulajdonságok jelenléte, 2) preferencia-törő tulajdonságok hiánya, és 3) a következtetési láncok hossza, amelyek a jelentésválasztások igazolásához és az összetevő-illesztési döntések igazolásához kellenek.

9.3 Az LDOCE-ből kinyert szemantikai információ felhasználása

Collative semantics (egyeztető szemantika):

- lexikai kétértelműség és szemantikai kapcsolatok kérdését vizsgálja
- 7 vizsgált szemantikai kapcsolat: literális, metonimikus, metaforikus, rendellenes (*anomalous*), újszerű, inkonzisztens és redundáns kapcsolat
- egy meta5 nevű NLP-programban van implementálva
- mondatokat elemzi, szavak jelentései között megkülönbözteti ezeket a kapcsolatokat, és feloldja a többértelműségeket

9.3.1 Egy lexikon-fogyasztó (Brian M. Slator)

A preferencia-szemantikai elemző feladata: válasszunk az előállított interpretációkból. A legsűrűbbet.
Kézhez kapja a szöveget és a jelentés-keretek lexikonát.
(Még csak elképzelés)
Balról-jobbra halad
Minden bemenetre mond valamit, mindegy, milyen kertiösvény.
Előfeltevés:

- LDOCE elég infót tartalmaz tudásforrásnak
- Az információ hatékonyan feldolgozható géppel

Külső információ: nyelvtan a jelentésmeghatározásokhoz és értelmes fa-mintázatok halmaza. Csak a szótár!
Csak az adott jelentésmeghatározás (vs. együttes előfordulások).

9.3.2 Egyeztető szemantika (Dan Fass)

4 összetevő:

1. jelentéskeretek: tudásreprezentációs sémák, egy szójelentés reprezentációja
2. szemantikai vektorok: a leképezések rendszerét reprezentálják (és ezáltal az azokban kódolt szemantikus kapcsolatokat, kivéve a metonimikust)
3. egyeztetés: két szójelentés jelentéskereteit egyezteti és elkülöníti a szójelentések közti szemantikus relációkat a jelentéskereteik közötti leképezések komplex rendszereként
4. *screening*: két szemantikai vektor közül választ, rangsorolva a szemantikai kapcsolatokat és mérve a fogalmi hasonlóságot - ezáltal feloldja a többértelműséget

9.3.2 Egyeztető szemantika

A jelentéskeretek:

- két fő rész: élek (*arcs*) és a csomópont (*node*)
- élek: a *genus term*-re mutató címkézett élt tartalmaz
- az összes jelentéskeret összes éle egy sűrű szemantikus hálót alkot, ez a jelentés-hálózat
- a csomópont: az infó, ami megkülönbözteti másoktól
- a csomópont olyan cellákból áll, amelyek egy természetes nyelvi nyelvtan szerint alakulnak
- három típusú csomópont:

```
sf(crook1,
   [[arcs,
     [[supertype, criminal1]]],
    [node0,
     [[it1, steal1,
        valuables1]]]]).
```

```
sf(crook2,
   [[arcs,
     [[supertype, stick1]]],
    [node0,
     [[shepherd1, use1, it1],
      [it1, shepherd1,
       sheep1]]]]).
```

```
sf(record_player1,
   [[arcs,
     [[supertype, playback1]]],
    [node0,
     [[it1, play10,
       record1]]]]).
```

```
sf(tape_recorder,
   [[arcs,
     [[supertype, playback1]]],
    [node0,
     [[it1, play10,
        audio_tape1]]]]).
```

```
sf(female1,
   [[arcs,
     [[superproperty, sex1],
      [property, female1]]],
    [node1,
     [[preference, organism1],
      [assertion,
       [[sex1, female1]]]]]]).
```

```
sf(male1,
   [[arcs,
     [[superproperty, sex1],
      [property, male1]]],
    [node1,
     [[preference, organism1],
      [assertion,
       [[sex1, male1]]]]]]).
```

```
sf(play6,
   [[arcs,
     [[supertype, strike1]]],
    [node2,
     [[agent,
       [preference,
        human_being1]],
      [object,
       [preference,
        ball1]]]]]).
```

```
sf(play7,
   [[arcs,
     [[supertype, use1]]],
    [node2,
     [[agent,
       [preference,
        human_being1]],
      [object,
       [preference,
        playing_card1]]]]]).
```

```
af(play10,
   [[arcs,
     [[supertype,
       [[create1, operate1]]]]],
    [node2,
     [[agent,
       [preference,
        human_being1]],
      [object,
       [preference, recording1]],
      [instrument,
       [preference,
        playback1]]]]]).
```

```
af(play11,
   [[arcs,
     [[supertype,
       [[perform1, use1]]]]],
    [node2,
     [[agent,
       [preference,
        human_being1]],
      [instrument,
       [preference,
        musical_instrument]]]]]).
```

9.3.2 Egyeztető szemantika

Mi kell ahhoz, hogy automatikusan lehessen ilyet építeni az LDOCE-ből?

- alapvetően tök jó, mert a *genus* és a *differentiae* bele van írva
- nem elég, hogy *genus term*, az kell, hogy a 12 fajta élcímke egyike legyen rajta
- igék elég nehéz ügy: három fajta differencia kell, nehéz megtalálni a definícióbam
- mellékneveknél, határozószóknál a legkönnyebb a *genus*