

1. Definiáld az alábbi fogalmakat!

pragmatika: a szemiotika egyik ága, mely a jel és jelfelhasználó viszonyát, a jelek kommunikációs folyamatban való használatának kérdéseit vizsgálja

párhuzamos korpusz: az egyirányú párhuzamos korpuszok a forrásnyelvi szövegeket és azok egy vagy több célnyelven meglévő fordításait tartalmazzák. A kétirányú párhuzamos korpuszok egyaránt tartalmazzák az A nyelvű forrásnyelvi szövegeket és azok B nyelvű célnyelvi fordításait, illetve a B nyelvű forrásnyelvi szövegeket és azok A nyelvű célnyelvi fordításait.

poliszémia: többértelműség (ugyanannak a szónak több jelentése van)

RTN: Recursive Transition Network. 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 elemezhető. Az RTN valójában egy veremautomata, azaz környezetfüggetlen nyelvek elemzésére is alkalmas. Az RTN-nel tehát $O(n^3)$ elemzési idő garantálható.

linear precedence: Linear precedence is the order relationship of sister nodes. LP constraints specify in what order sister nodes under the same mother can appear. Nodes that surface earlier in strings precede their sisters.[2] LP can be shown in phrase structure rules in the form $\{A \rightarrow B \ C \ D\}$ to mean B precedes C precedes D, as shown in the tree below.

morfológia: Morfológia, avagy alaktan, a hagyományos elnevezése a szóalkotási eszközökkel foglalkozó nyelvészeti tudománynak.

jelentés-egyértelműsítés:

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

fordítómemória: (fordítási memória, translation memory, TM) szegmens párokat (általában mondatpárokat) tárol. A TM forrásnyelvi szegmensek és célnyelvi fordítások párait tárolja, némi adalék információval (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.

synset: Synonym set: A set of one or more synonyms that are interchangeable in some context without changing the truth value of the proposition in which they are embedded

fonetika: hangtan, a beszédhangok képzése, fizikai tulajdonságai

szemantika: jelentéstan, a jelentés

homonímia: „hason(ló)nevűség”, „azonosalakúság”) két **szó** – esetleg egyéb nyelvi elem, például szókapcsolat – közötti véletlen egybeesés, amikor íráskéjük (és többnyire hangalakjuk is) azonos, de a jelentésük különbözik

balsarok: algoritmusok melyek balról-jobbra működnek, és elkerülik ezen lehetetlen szerkezetek felépítését, balsarok-elemzőknek (left-corner parsers) nevezzük, mivel egy olyan elemzési fát építenek fel, amely a nyelvtan kezdő szimbólumával indul, és a mondat legszélén balra (a bal sarokban) álló szó felé terjeszkednek. Egy élt csak akkor adnak a diagramhoz, ha ezen elemzési fa kiterjesztését szolgálhatja

immediate dominance: the case where the two nodes in question are connected by a single branch without any intervening nodes. More formally, immediate dominance is defined:

A immediately dominates B iff (= if and only if)

- a. A dominates B, and
- b. there is no node C, distinct from A and B, such that A dominates C and C dominates B.

fonológia: vagy fonematika a nyelvtudománynak az az ága, amely a beszédhangoknak a nyelvi rendszerben elfoglalt helyét és szerepét vizsgálja, szemben a fonetikával, amely egy adott nyelv beszédhangjainak a leírásával foglalkozik.

kompozícionális elve: egy komplex kifejezés jelentése a részei jelentéseinek és az őket leíró szintaktikai szerkezetnek a függvénye

szintaxis: vagy mondattan a nyelvészetben a szavak szószervezetekké és mondatokká kapcsolódásának szabályait írja le.

2. Válaszolj az alábbi kérdésekre!

Mi a közvetlen összetevős nyelvtan és a függőségi szerkezet ? Mik az előnyei és hátrányai ? (6p)

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

Hasonlítsd össze a top-down (TD) és a bottom-up (BU) elemzést! (6p)

TD: nem fordulhat elő, hogy ne mondat-szimbólummal 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”

Mit jelent a szótár kifordíthatósága ? Mire jó? (5p)

Többféle jelentés kezelését teszi lehetővé, a bal oldali címszavak segítségével, új találati ablak behozásával. Elektronikusan érdemes végrehajtani.

Mi a hossz alapú szövegszinkronizáció lényege ? (3p)

- “length-based alignment”
- csak a szövegszövegek hosszait vizsgálja
- valószínűségi modell
- teljes szinkronizáció
- dinamikus programozás, globálisan optimális megoldás
- robosztus

Mik azok az X-vonás nyelvtanok? Mitől jók?

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 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):

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

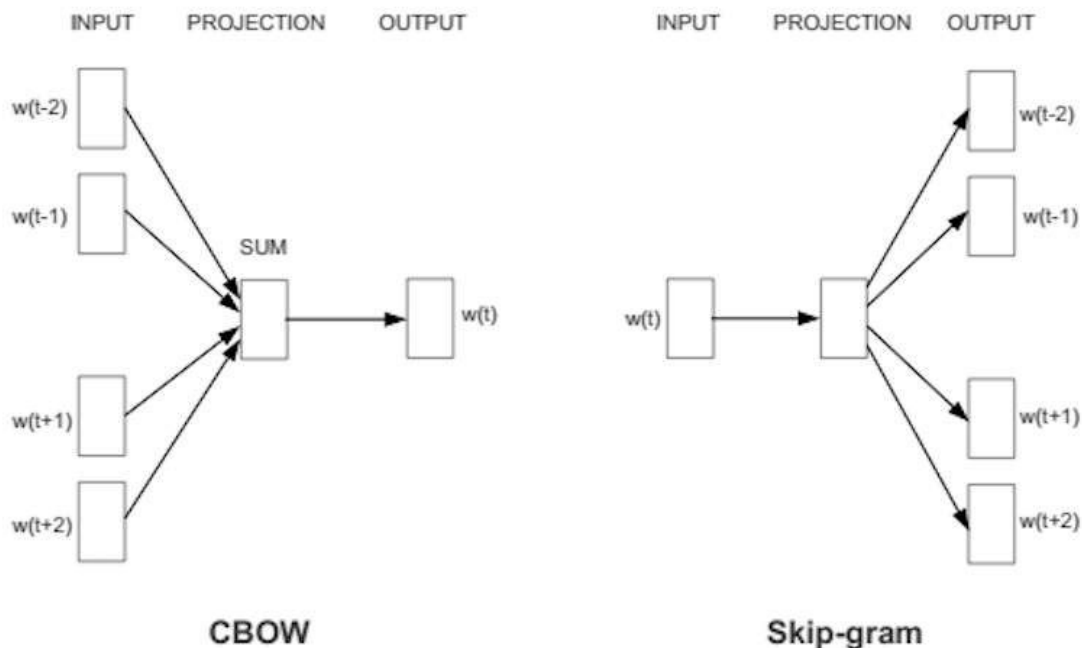
☐ Egy X-kétvonás (más néven XP)-szerkezet egy opcionális SPECIFIER elemből és egy X-vonásból áll:

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

Mi az a CBOW és a Skip-gram? Mire használják őket ?

(1) környezet→szó: folytonos szózsák (Continuous Bag Of Words, CBOW)

(2) szó→környezet: skip-gramm



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 szövegrepresentációs célokra, ahol a bemenet a szövegtörzs és a kimenet a törzs szavait reprezentáló jegyvektorok halmaza. A word2vec maga nem mélyneurális háló, de a kimenetét mélyneurális hálók is tudják kezelni.

Mi a horgony alapú szövegszinkronizáció lényege ?

A horgonyok a szövegegységek egymáshoz 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

Hogyan működik a statisztikai gépi fordítás és a szabályalapú gépi fordítás ? Hasonlítsd össze őket!

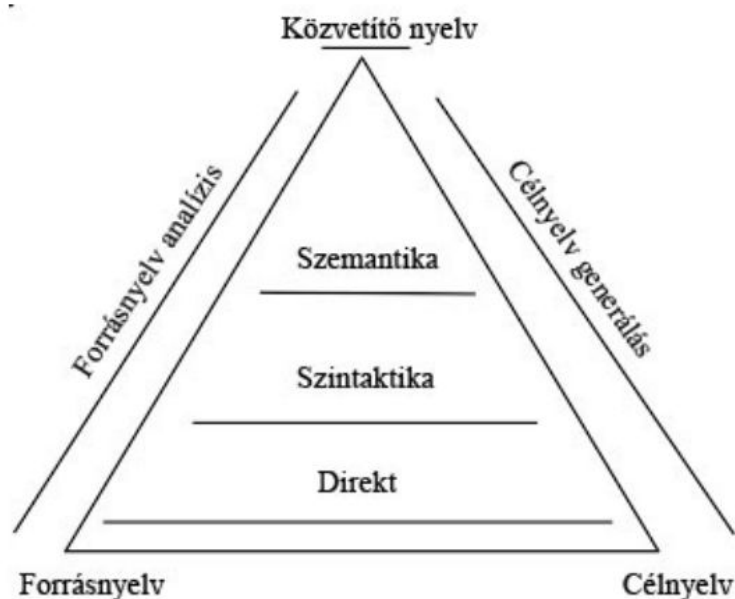
statisztikai gépi fordítás:

- a gyerek: példák megfigyelése
- párhuzamos korpuszból tanul
- emberek által előre lefordított szövegek felhasználásán alapszik
- beszédtechnológiában használt Shannon-féle zajoscsatorna-modelljét vették alapul

szabályalapú gépi fordítás:

- a felnőtt: szabályok megtanulása és szótár

- direkt fordítás
- transzfer
- interlingua



Szótárak esetén mi a perjel probléma ? Írj rá példát is!

a szótárak nem a gépnek készülnek, tudhatja-e egy program összefogni az egybetartozó kifejezést: arcára fagy/ráfagy az arcára/lefagy az arcáról a mosoly.

Hogyan működik a referenciafordítás nélküli kiértékelés ? Hasonlítsd össze a referenciafordítással történő módszerrel!

Hogyan működik a referenciafordítással történő kiértékelés ? Írj néhány példát, hasonlítsd össze a minőségbecslés módszerével!

Alap példa:

Forrás: Ez egy útmutató, amely biztosítja, hogy a hadsereg mindig betartsa a párt utasításait.

Gépi fordítás: It is a guide to action which ensures that the military always obeys the commands of the party.

1. referencia: It is a guide to action that ensures that the military will forever heed Party commands.
2. referencia: It is the guiding principle which guarantees the military forces always being under the command of the Party.
3. referencia: It is the practical guide for the army always to heed the directions of the party.

Def:

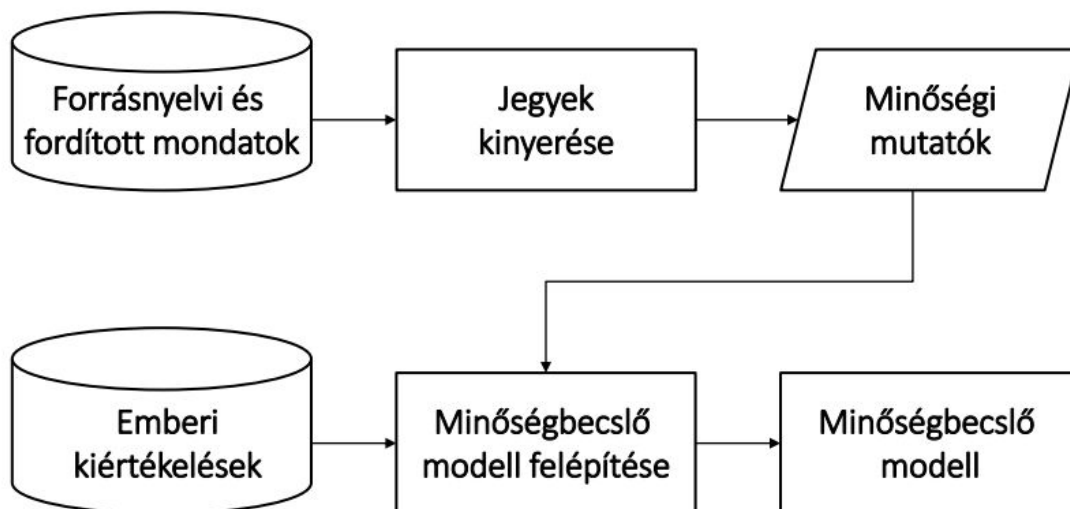
- Automatikus kiértékelő módszerek.
- Nincsen szükség emberi beavatkozásra.
- Nem kell ismerni a két nyelvet.
- Hasonlóságot és eltérést mér a fordítás és a referenciafordítás között.

Módszerek:

- BLEU, NIST, METEOR, LEPOR, ROGUE, RIBES
- WER, TER, HTER

Minőségbecslés:

- Referenciafordítás nélküli módszer
- Valós időben tud kiértékelni
- Gépi tanuláson alapuló minőségbecslő modell
- Regressziós / osztályzási feladat



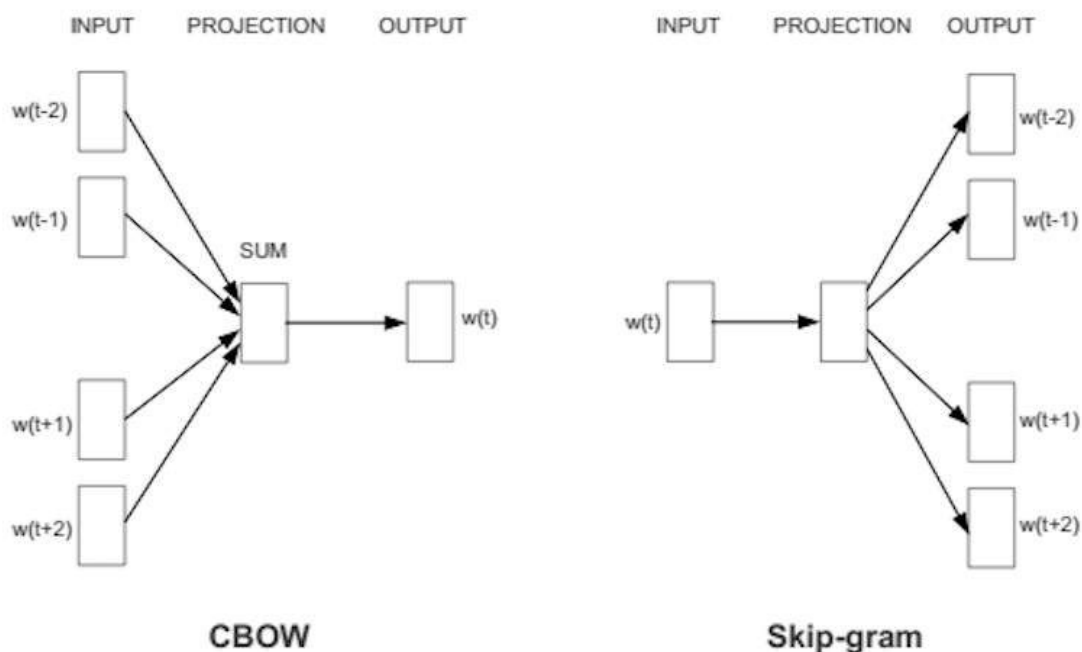
Mit jelent a szóbeágyazás (word embedding)? Milyen modellei vannak ? Mi a lényegük?

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 szövegrepresentációs 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-gramm



Gyakorlat

Levenshtein Distance

<https://planetcalc.com/1721/>

3. FELADATOK

1. Az alábbiak közül melyik tömorféma (bal oldal) unifikálható melyik toldalékmorfémával (jobb oldal)? Milyen szóalakok jönnek létre? Ezek közül melyek valódi magyar szavak? (6p)

hó	[-elől +DAT +INS:v]	nak	[+DAT -elől]
hav	[-elől -DAT -INS]	nek	[+DAT +elől]
füge	[+elől +INS:v]	val	[+INS:v -elől]
fügé	[+INS:v]	vel	[+INS:v +elől]

Dat f az elől
 elől képzett +
 INS: f k anylyan képzett
 képzett

hónak figének } értelmesek
 hóval figével }

figének } létrehozhatók
 figével }
 figéznek }

2. Mennyi a következő szavak Levenshtein-távolsága? Az egyikhez írd le a levezetést. (6p)

levél - lével : 2 lével / leved
 wood - would : 2 wood / would
 asztal - azatl : 3 asztal / azatal / azatl

3. Az alábbi nyelvtanban mi lehet az NP balsarka? (Terminálisok: Det, N, V) (4p)

S -> NP VP
 NP -> Det Nom
 NP -> Nom
 Nom -> N Nom
 Nom -> N
 VP -> V
 VP -> V NP

bal/sarok Det N
 NP -> Det N
 NP -> Det Nom
 ↓
 N

Feladatok:

1) Unifikáció:

$hó [-elől +DAT +INS : v]$
 $hav [-elől -DAT -INS]$
 $füge [+elől +INS : v]$
 $füge' [+INS : v]$

$hónak \neq fügenek$
 $hóval fügével$ } értelmezés

$fügenek$
 $fügéval$
 $fügenek$ } létrehozható

	a	az	igen (e,é,i,í,ö,ó,u,ü)	nem (a,á,ú,ü)
DAT	+	-	+	-
elől	előképzett			
INS: ?	+ amilyen hanggal kezdődik			

$nak [+DAT -elől]$
 $nek [+DAT +elől]$
 $val [+INS : v -elől]$
 $vel [+INS : v +elől]$

$hó [+DAT +INS : v]$
 $hav [-elől -DAT -INS]$
 $füge [+elől +INS : v]$
 $füge' [+elől +INS : v]$

$nak [+DAT -elől]$
 $nek [+DAT +elől]$
 $val [+INS : v -elől]$
 $vel [+INS : v +elől]$

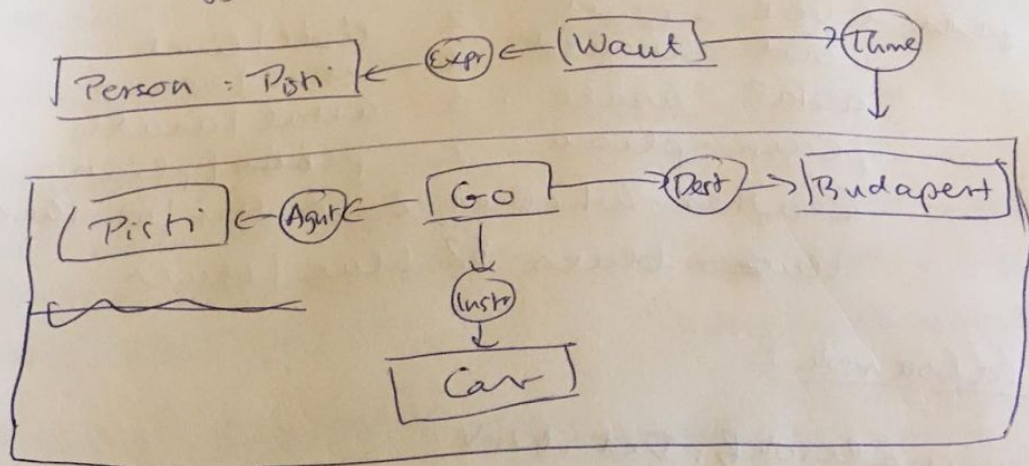
$hó \rightarrow \underline{hónak}, \underline{hóval}, \underline{hóval}$

$hav \rightarrow \emptyset$

$füge \rightarrow \underline{fügenek}, \underline{fügéval}$

$füge' \rightarrow \underline{fügenek}, \underline{fügéval}$

Pisti egy autóval szeretne eljutni Bp-re.



5 WordNet

fürdőszoba, mosdó, körte, sárgabarack, kajszibarack,
gyümölcs, vilmoskörte, lelegzik el
~~szaporodik~~ szaporodik,

(körte, sárgabarack, kajszibarack, vilmoskörte) → gyümölcs

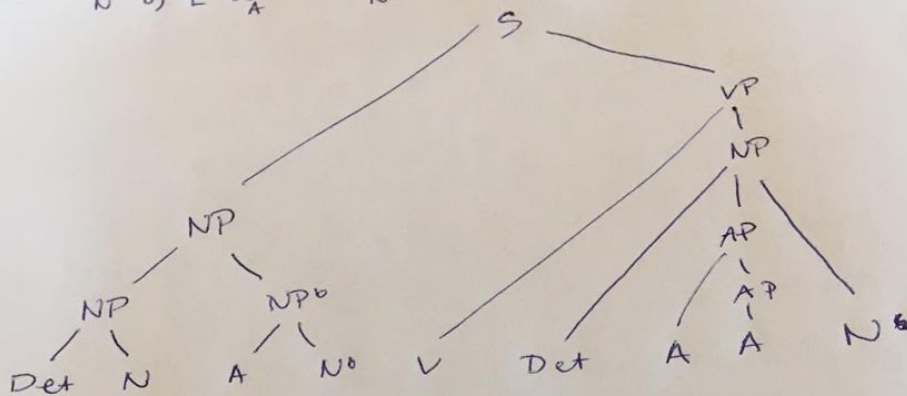
vilmoskörte → körte → gyümölcs

entailment: lelegzik → él

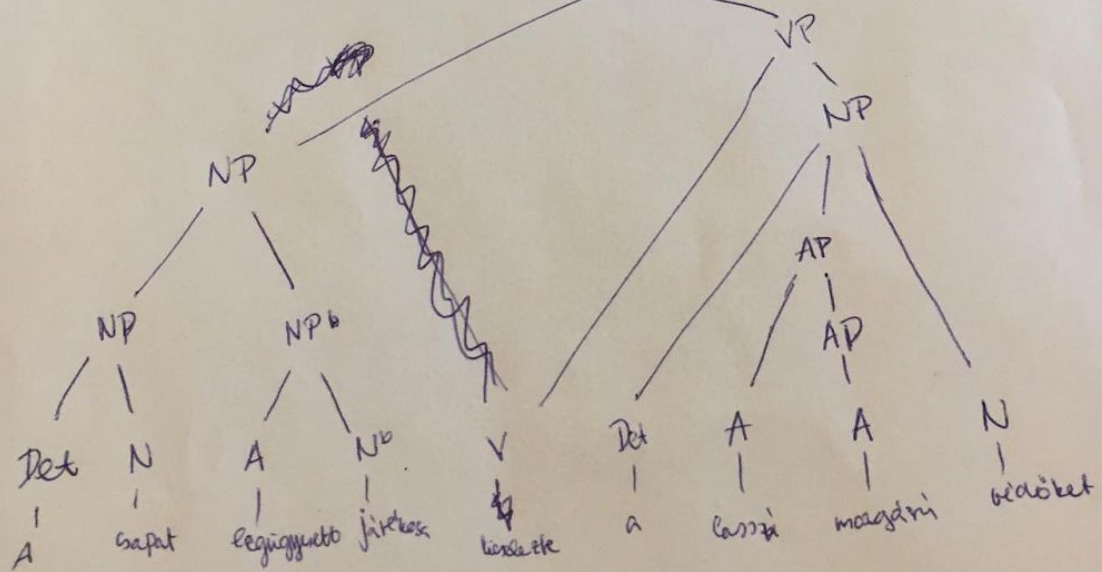
Synonym: fürdőszoba ↔ mosdó

6. Székely

[Az_{Det} ontológ_N] [legjobb_A tanulója_{Nb}] [becszi_V] [a_{Det} [két_A [pócs_A] barátával_{Ns}].



[A_{Det} csapat_N] [leggyengébb_A játékos_{Nb}] [kicsélté_V] [a_{Det} [lassú_A [mozgás_A] nézőket_N]



[2] Levenshtein távolság:

level - level : 2	level / level
wood - would : 2	woud / would
asztal - asatl : 3	asztal / asatl
plain - pian : 2	plein / pian
weather - whether : 2	whather / whether
blue - blues : 2	blue / blues

[3] Balsarok:

Terminalisok: Det, N, V

$S \rightarrow NP VP$

$NP \rightarrow Det Nom$

$NP \rightarrow Nom$

$Nom \rightarrow N Nom$

$Nom \rightarrow N$

$VP \rightarrow V$

$VP \rightarrow V NP$

balsarok Det N

$NP \rightarrow Det N$

$NP \rightarrow Det Nom$

↓
N

[4] koncepciógráf:

Feri be akarta tenni az ablakot egy kárpács

