O indukciji gramatike: računalni i statistički modeli usvajanja jezika

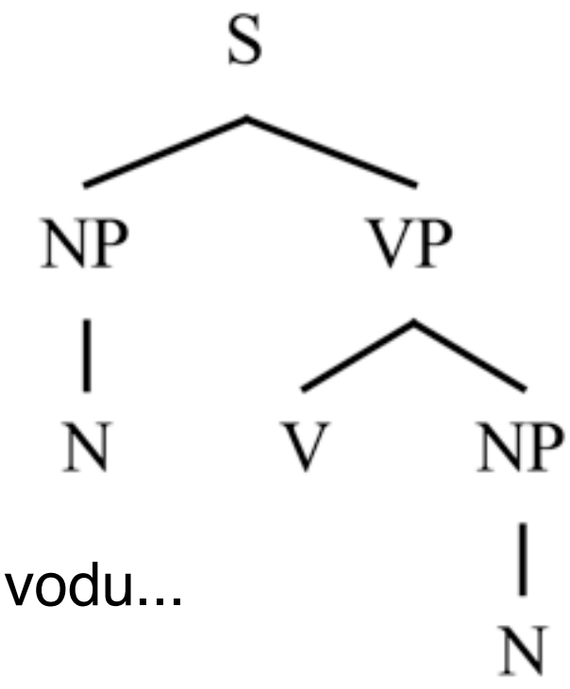
Damir Ćavar
Sveučilište u Zadru
Odjel za lingvistiku

Prolog

- Terminologija i okvir
  - Simbolistički, konekcionistički i empiristički modeli
  - Sintaktičke strukture
  - Statističke osnove

Modeli

- Urođeno:
  - Simbolistički: formalna pravila opisuju ograničenja kombinatorike različitih lingvističkih razina
    - S → NP VP
    - NP → N
    - VP → V NP
    - V → pije, čita, voli... N → Ivan, Marija, knjigu, vodu...



Modeli

- Empiristički, radikalni stav:
  - Opća kognitivna pravila ili funkcije opisuju asociativne osobine i dojam ili prepoznavanje obrasca ili mustre u uzorku
    - nema lingvističkog predznanja
    - nema a priori specifičnog lingvističkog predznanja

Stavovi

- Detaljno znanje o jeziku (svim jezicima) je urođeno, ili
- Moćne kognitivne funkcije učenja su urođene, koje omogućavaju učenje iz uzorka

Argumentacija

- Urođenost specifičnih jezičnih kompetencija:
  - Jezik ima kompleksne osobine, npr. sintaksa je formalno najmanje bezkontekstna, a bezkontekstne gramatike se ne mogu naučiti bez negativne evidencije (Gold, 1965, 1967)

Minimalistička statistika

- Vjerojatnost pojave dvaju događaja X i Y istovremeno
  - P(X) = vjerojatnost pojave X
  - P(Y) = vjerojatnost pojave Y
  - P(XY) = vjerojatnost pojave X i Y zajedno
  - Ako X i Y nisu povezani: $P(XY) = P(X)P(Y)$
  - Ako X i Y jesu povezani: $P(XY) \neq P(X)P(Y)$

N-gram modeli

- Unigrami: frekvencije pojavnica

i	2967
u	1585
se	1316
da	895
je	838
na	707
od	528
ne	445
s	416
...	

N-gram modeli

- Bigrami: frekvencije dvaju susjednih riječi
  - Pomičemo okvir dužine od dvije riječi kroz korpus, riječ po riječ, i brojimo bigram

N-gram modeli

- Relativiramo frekvencije u ngram modelima: unigrami i bigrami
  - Dijelimo frekvenciju pojedinačnih grama kroz broj grama u konačnom modelu
  - model vjerojatnosti: *maximum likelihood*

Bigram primjer

da je	20	0.004724
je u	16	0.003779
je da	8	0.001889
rekao je	8	0.001889
da će	8	0.001889
rekao da	7	0.001653
je rekao	6	0.001417
to je	6	0.001417
će se	5	0.001181
tijekom sezone	5	0.001181
ali je	5	0.001181
koji je	5	0.001181
kao i	5	0.001181

...

Modeli

- Iz modela za unigrame vadimo
  - vjerojatnost pojedinačnih pojavnica: P(X)
- Iz modela za bigrame vadimo
  - vjerojatnost za pojavu dviju susjednih pojavnica: P(XY)

Korpusi

- Iz korpusa vadimo statistiku: Brown-korpus (Francis i Kucera, 1964)

```
The/at Fulton/np-tl County/nn-tl Grand/jj-
tl Jury/nn-tl said/vbd Friday/nr an/at
investigation/nn of/in Atlanta's/np$
recent/jj primary/nn election/nn produced/
vbd ``/`` no/at evidence/nn ''/'' that/cs
any/dti irregularities/nns took/vbd place/
nn ./.
```

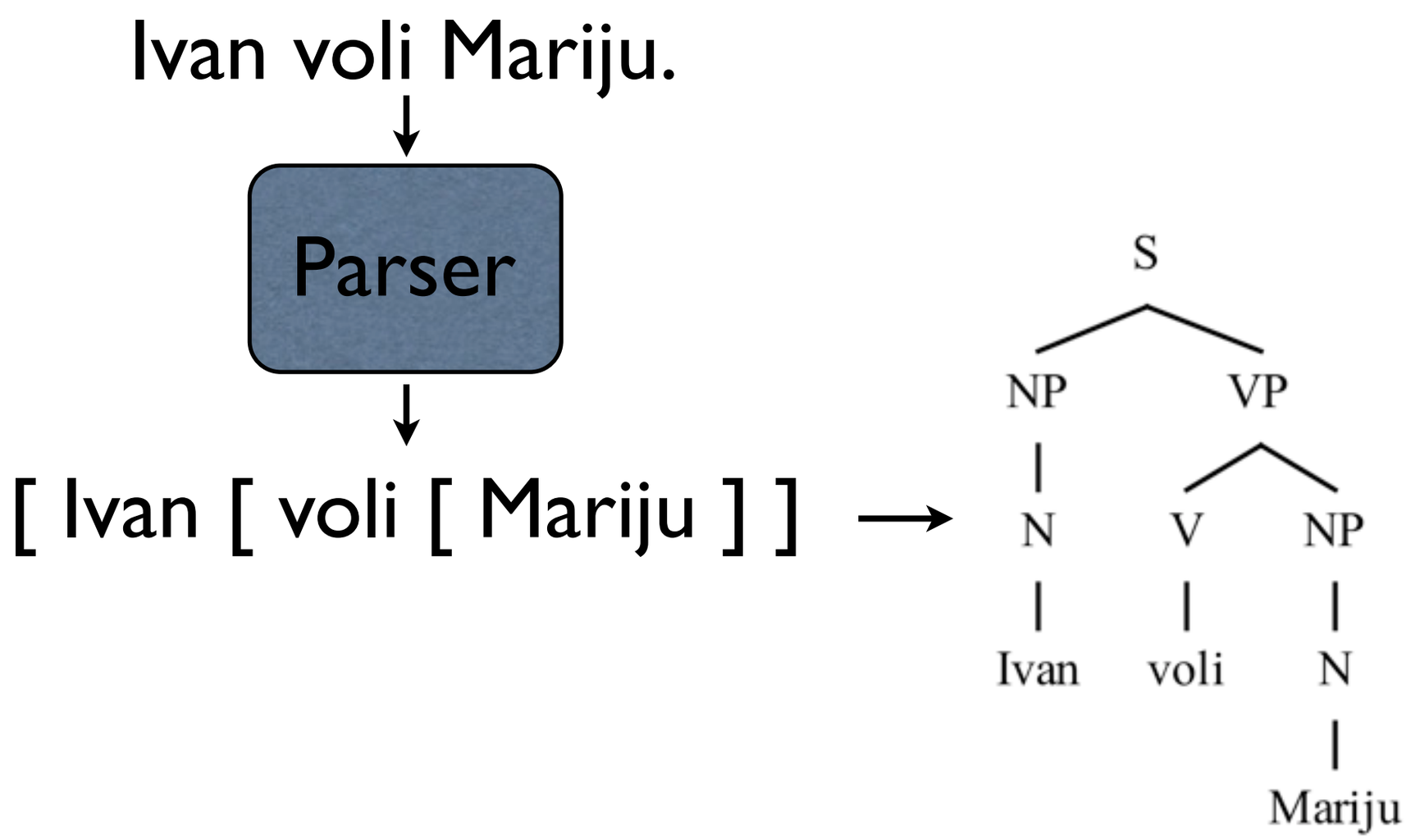
Korpusi

- Penn Treebank, (C) LDC 1995., za sintaktička stabla

```
( (S
    (NP-SBJ (NNP Mr.) (NNP Vinken) )
  (VP (VBZ is)
    (NP-PRD
      (NP (NN chairman) )
      (PP (IN of)
        (NP
          (NP (NNP Elsevier) (NNP N.V.) )
          (, ,)
          (NP (DT the) (NNP Dutch) (VBG publishing) (NN group) )))))
  (. .) ))
```

Parsiranje

- Određuje jednoj rečenici odgovarajuću sintaktičku strukturu (ili više struktura)



Sintaktička stabla

- Reprezentacija hijerarhijskih relacija između sintagmi
- Semantičke relacije (*scopus*)
- Osnova za opisivanje sintaktičkih pravila, ograničenja, derivacija

Kraj prologa

Statističko parsiranje

- Može li se parsirati koristeći isključivo informaciju i znanje koje dobijemo iz lingvističkih uzoraka, znači empiristički?
- Mogu li se sintaktička stabla generirati statistički?
- Koliko informacije o strukturalnim osobinama jezika nalazimo u samome jeziku?

Intuicija

- Očekivanja kada čujemo:
  - *Ivan kaže*...
  - *Marija živi*...
  - *Petar voli*...
  - IMENICA GLAGOL PRIJEDLOG...

Lingvističke intuicije

- Selekcija i restrikcija tipa i sintaktičke kategorije
  - prijedlozi: imeničke fraze
  - glagoli: ovisi o konkretnom glagolu - valencija, c- i s-selekcija
- Selekcija u sintaksi je direkcionalna!

Lingvističke intuicije

- Selekcija mijenja distribucijske statističke osobine:
  - lijevo od prijedloga: nema značajnog ograničenja
  - desno od prijedloga: samo elementi sintagmi unutar imeničkih fraza

Statistički model intuicije

- Varijacija na jednoj strani pojavnice razlikuje se od varijacije na drugoj, ili ne

Ivan Marija mi svi oni često ...	kazati	da kako ...

Statistički model intuicije

- Funkcionalne riječi bilježe početak ili kraj sintagme
- Jako su frekventne: statistički dominantne
- Imaju veliku distribuciju: ca. 50% teksta/govora

Radovi do sada

- Relevantni:
  - Rens Bod (2009) u Cognitive Science
  - Magerman i Marcus (1990)
  - Church (1988)
  - itd.

Magerman i Marcus

- Koriste unigram i bigram modele iz morfosintaktičkih oznaka, tj. frekvencije pojedinačnih oznaka P, N, itd., i istih u kombinaciji P N, V N, itd.
- Traže najnižu vrijednost uzajamne informacije (*Mutual Information*)

$$MI(x,y) = log \frac{P(x,y)}{P(x)P(y)}$$

Magerman i Marcus

- Općenito za ngrame bilo koje dužine:

$$GMI_{i+j}(x_1 \ldots x_i, y_1 \ldots y_j) = \sum_{\substack{X \text{ završava s } x_i \\ Y \text{ počinje s } y_1}} \frac{1}{\delta XY} MI(X, Y)$$

Magerman i Marcus

- Gdje god se nalazi najmanja vrijednost međusobne informacije pretpostavljaju sintaktičku granicu (kraj i početak sintagme), tako nastavljaju rekurzivno za sve ostale dijelove rečenice:

  He directed the cortege of autos to the dunes near Santa Monica.

  pro 3.28 verb 3.13 det 11.18 noun 11.14 prep 1.20 noun 7.41 prep 16.89 det 16.43 noun 12.73 prep 7.36 noun

  [he] [directed [[the cortege] [of autos]]] [[to [the dunes]] [near Santa Monica]]

Magerman i Marcus

- Stipulacije:
  - Moraju ugraditi nadzor u obliku rječnika distituenata (obratno od konstituenata), znači listu bigrama gdje 100% mora biti sintaktička granica
  - Implicitni nadzor u smislu da statistički parsiraju sekvencije oznaka, a ne riječi i pojavnice, tj. lingvistički nadzor i podizanje statističkih osobina modela
  - Statistički manipuliraju model tako da svaku rečenicu popune lijevo i desno do maksimalne dužine (najduže rečenice) s praznim pojavnicama (bolji ngram modeli)

Magerman i Marcus

- Rezultati parsera:
  - stabla s *n* ogranaka po dominirajućoj sintagmi (kategorijalnim simbolu)
    - stabla nisu označena, nije vidljivo što je *glava*, što *komplement*
- Evaluacija:
  - Treniraju i evaluiraju na Brown korpusu

Magerman i Marcus

- Rezultati:
  - “Dobri rezultati” kod parsiranja kratkih rečenica
    - prosječni broj pogrešaka: 1 u rečenici
  - U rečenicama s veznicima
    - prosječni broj pogrešaka: 2 u rečenici

Magerman i Marcus

- Rezultati:
  - U rečenicama s 16 do 30 riječi:
    - prosječno između 5 i 6 pogrešaka u rečenici
- Nema detaljne evaluacije rezultata u članku, ni drugdje

Alternativa

- Ćavar, Rodrigues i Schrementi (2006)
  - Dodatno mjerilo selektivnosti pojavnica s Relativnom entropijom
  - Inkrementalno učenje i parsiranje: dinamički statistički modeli
  - Evaluacija s ngram modelima: pojavnica - pojavnica, pojavnica - oznaka, oznaka - oznaka, oznaka - pojavnica

Relativna entropija

- Direkcionalna je (za razliku od međusobne informacije)
- Pouzdanija kod niskih frekvencija pojavnica
- Uspoređuje distribuciju jedne pojavnice s distribucijom te pojavnice u kontekstu jedne druge
- Manji RE = veća vjerojatnost da su X i Y ovisne varijable

$$RE(xy) = P(y)lg\frac{P(y)}{P(y|x)}$$

Međusobna informacija

- Uzimamo samo u obzir statistiku bigrama u kojima se neka pojavnica X nalazi lijevo
- Uspoređujemo distribuciju pod pretpostavkom da su X i Y ovisni, i neovisni
- Veći MI znači veće vjerojatnost da su X i Y ovisne varijable

$$MI(xy) = P(< xy > |x) lg \frac{P(< xy >)}{P(x)P(y)}$$

Eksperiment 1

- Parsiranje od-vrha-prema-dnu (*top-down*):
  - sječemo rečenicu u dva dijela gdje je:
    - MI lokalno najmanji
    - RE lokalno najveći
  - i nastavljamo sjeći odlomke dokle god sadrže više od dvije pojavnice

Eksperiment 1

- Strategija od-vrha-prema-dnu ne razlikuje se od strategije od-dna-prema-vrhu u rezultatu:
  - Spajamo dvije pojavnice u konstituente gdje je:
    - MI lokalno maksimalan
    - RE lokalno minimalan
- I nastavljamo tako rekurzivno spajati u preostalim dijelovima

Eksperiment 1

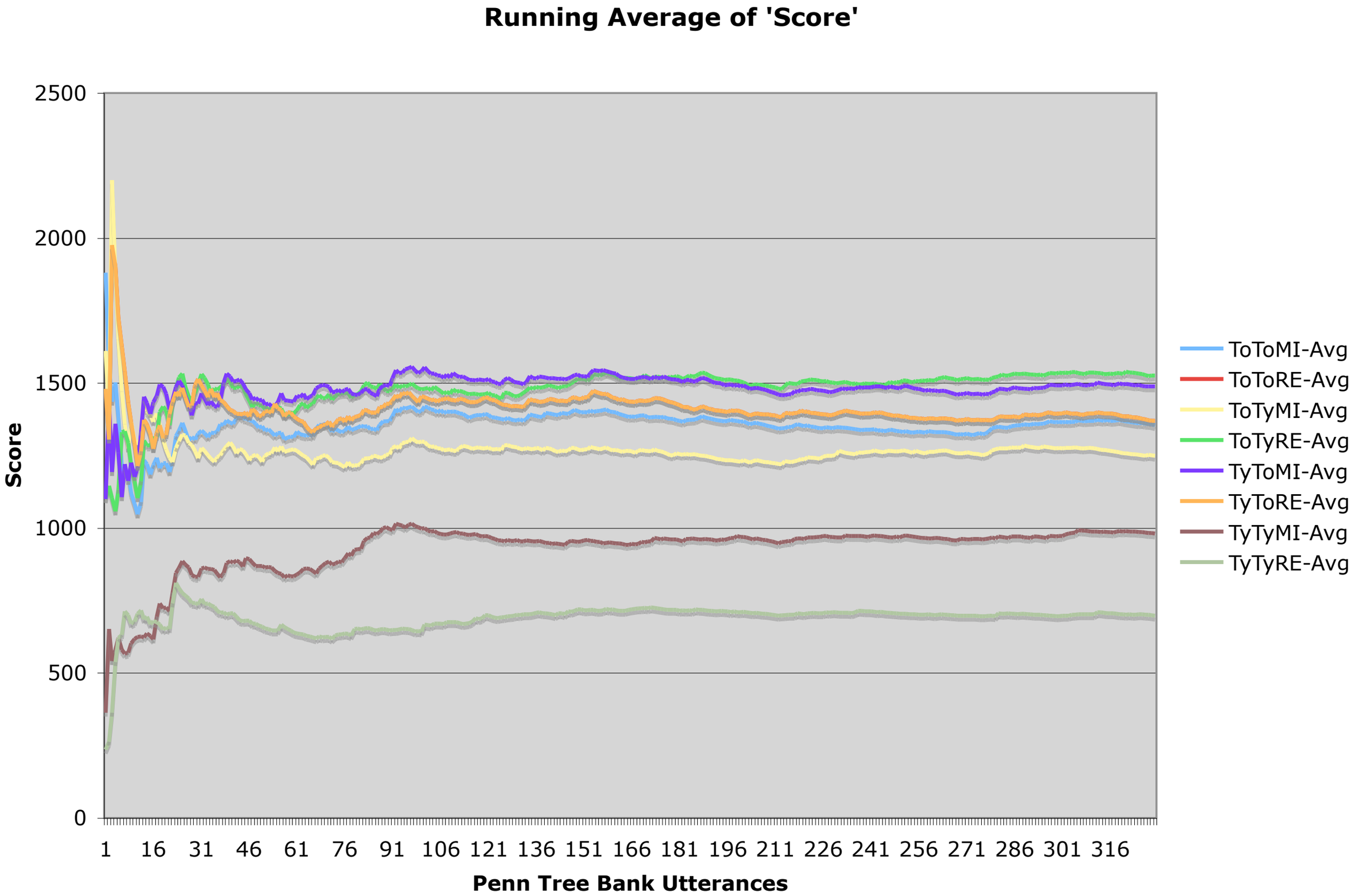
- Osobine:
  - Rezultirajuća stabla se granaju binarno (što je pozitivno za neke teorije sintakse)
  - Sa svakom ulaznom rečenicom automatski širimo statističke unigram i ngram modele (inkrementalno učenje)
  - Uspoređujemo za svako mjerilo i svaki tip kombinatorike u ngram modelima rezultat ili kvalitetu rezultirajućeg sintaktičkog stabla

Eksperiment 1

- Evaluacija:
  - Za svaku rečenicu iz Penn Treebank korpusa stabala:
    - parsiramo rečenicu bez strukturalnih informacija
    - uspoređujemo automatski generiranu strukturu s strukturom u korpusu

Eksperiment 1

- Problemi:
  - Penn Treebank nije ograničen na binarno granajuća stabla
  - isto tako nije konsistentan (za jednu rečenicu postoje različite sintaktičke strukture)
  - a i za jednu rečenicu: N sintaktičara nudi ca. N različitih analiza


Running Average of 'Score'
Score
2500
2000
1500
1000
500
0
1
16
31
46
61
76
91
106
121
136
151
166
181
196
211
226
241
256
271
286
301
316
Penn Tree Bank Utterances
ToToMI-Avg
ToToRE-Avg
ToTyMI-Avg
ToTyRE-Avg
TyToMI-Avg
TyToRE-Avg
TyTyMI-Avg
TyTyRE-Avg

Rezultat

- Očekivano:
  - Relativna entropija je značajno bolja (direkcionalnost u sintaktičkim relacijama)
  - Modeli preko morfosintaktičkih oznaka daju najmanje pogrešaka

Diskusija

- Tip jezika je složen: novinski članci, nizak koeficijent relacije između broja tipa riječi i pojavnica
    - Analiza korpusa s razgovorima između roditelja i djece (CHILDES corpus), očekuje se, dala bi puno bolje rezultate i na razini modela s pojavnicama
- Lingvistička analiza struktura je neophodna

Diskusija

- Distribucijska statistika ne samo da daje dobre rezultate za morfološke analize (Ćavar et al. 2004-2007), nego i za sintaksu
- Statistika kao način inicijalnog pristupa jeziku u prvom usvajanju: *Bootstrapping*