Uvod u obradu prirodnog jezika

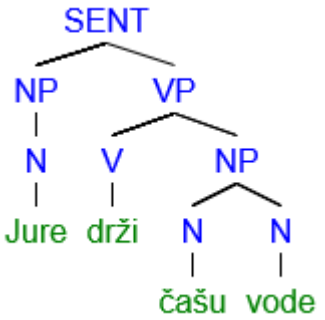
13.1. Sintaktičke strukture: sastavno vs. ovisnosno (Syntactic structures: consistuency vs. dependency)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

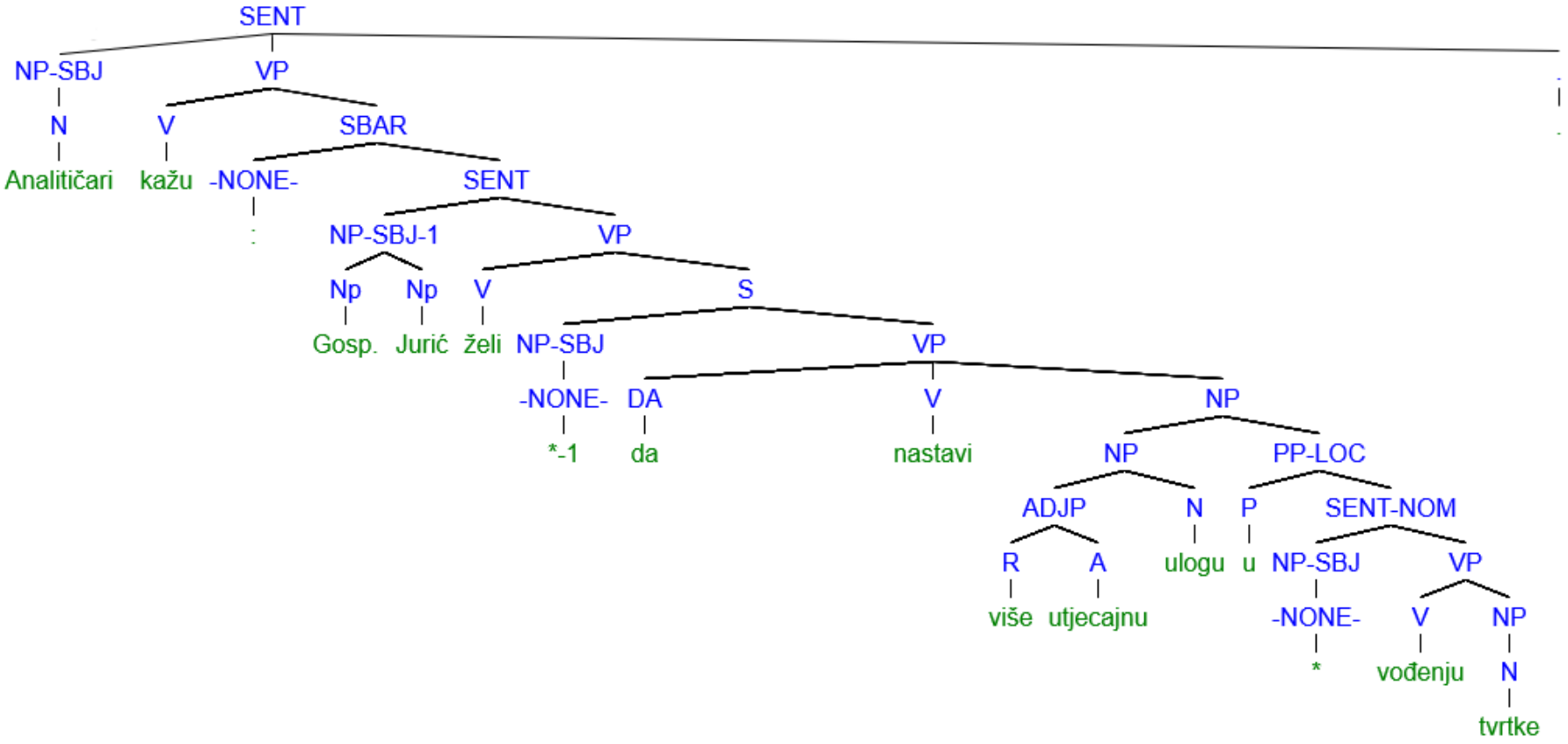
Dva pogleda na lingvističke strukture

1. Sastavni (constituency) - struktura fraze
   - Struktura fraze organizira riječi u ugniježđene sastavnice
   - Kako znamo što je sastavnica?
     - Distribucija: sastavnica se ponaša kao jedinica koja se može pojaviti na različitim mjestima:
       - Jure je govorio [svojoj djeci] o [štetnosti droga].
       - Jure je govorio o [štetnosti droga] [svojoj djeci].
       - *Jure je govorio droga svojoj štetnosti o djeci.



     - Supstitucija/ekspanzija/zamjenične-forme
       - Sjedio sam [na kutiji/baš na toj kutiji/tamo]
     - Koordinacija, regularna interna struktura, nema upada, fragmenti, semantike, …

Struktura fraze



Struktura fraze s glavom

- VP → … V* …
- NP → … N* …
- ADJP → … A* …
- ADVP → … Q* …

X-bar teorija

- SBAR(Q) → S|SINV|SQ→… NP VP …

- Još manjinski tipovi fraza
  - QP (kvantifikatorska fraza u NP)
  - CONJP(višeriječne konstrukcije kao "i tako dalje")
  - INTJ (uzvici ah, bravo, …)

Ovisnost

- Ovisnost pokazuje koje riječi ovise (mijenjaju ili su argumenti) o drugim riječima

Taj dječak stavlja malu kornjaču na mekani tepih.

Ovisnost

- Ovisnost pokazuje koje riječi ovise (mijenjaju ili su argumenti) o drugim riječima

Uvod u obradu prirodnog jezika

13.2. Parsiranje (Parsing)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

"Klasično" parsiranje

- Napravljena gramatika (CFG) i leksikon

S → NP VP	NN → *interest*
NP → (DT) NN	NNS → *rates*
NP → NN NNS	NNS → *raises*
NP → NNP	VBP → *interest*
VP → V NP	VBZ → *rates*

- Korištenje gramatika/dokaza za dokazivanje stabla parsiranja iz riječi
- Skaliranje se pokazuje vrlo lošim, i ne obuhvaća opće primjere:

Fed raises interest rates 0.5% in effort to control inflation

- Minimalna gramatika: 36 stabala
- Jednostavna gramatika s 10 pravila: 592 stabla
- Realna sveobuhvatna gramatika: milioni stabala

"Klasično" parsiranje: problem i rješenje

- Dodavanje kategoričkih ograničenja u gramatici radi ograničavanja malo vjerojatnih parsiranja rečenice
  - ali gramatika gubi na robusnosti
    - oko 30% rečenica neće biti parsirane
- Manje ograničena gramatika može parsirati više rečenica
  - ali jednostavne rečenice dobivaju više različitih stabala bez načina da se izabere jedno od njih
- Potreban je mehanizam koji pronalazi najvjerojatnije stablo parsiranja rečenice
  - Statističko parsiranje omogućava rad s malim gramatikama koje dozvoljavaju milione stabala parsiranja rečenica i vrlo brzo pronalazi najbolje stablo.

Banka stabala

- Izgradnja banke stabala se čini mnogo sporijim i manje korisnim od izgradnje gramatike

- Ali banka stabala ima brojne koristi:
  - ponovna upotrebljivost
    - mnogi parseri, POS označavaći, …
    - Vrijedni resursi za lingviste
  - Široka pokrivenost
  - Frekvencije i distribucije
  - Način evaluacije sustava

Penn TreeBank

```
( (S
    (NP-SBJ (DT The) (NN move))
    (VP (VBD followed)
      (NP
        (NP (DT a) (NN round))
        (PP (IN of)
          (NP
            (NP (JJ similar) (NNS increases))
            (PP (IN by)
              (NP (JJ other) (NNS lenders)))
            (PP (IN against)
              (NP (NNP Arizona) (JJ real) (NN estate) (NNS loans))))))
      (, ,)
      (S-ADV
        (NP-SBJ (-NONE- *))
        (VP (VBG reflecting)
          (NP
            (NP (DT a) (VBG continuing) (NN decline))
            (PP-LOC (IN in)
              (NP (DT that) (NN market)))))))
    (. .)))
```

[Marcus et al. 1993, *Computational Linguistics*]

Uvod u obradu prirodnog jezika

13.3. Eksponencijalni problem parsiranja

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

"Klasično" parsiranje

- Ključna odluka kod parsiranja: na koji način "povezati" različite sastavnice
  - prijedložne, priložne fraze, fraze čestica, infinitive, koordinacije, …

Odbor je potvrdio [svoja preuzimanja] [s vanjskom tvrtkom]

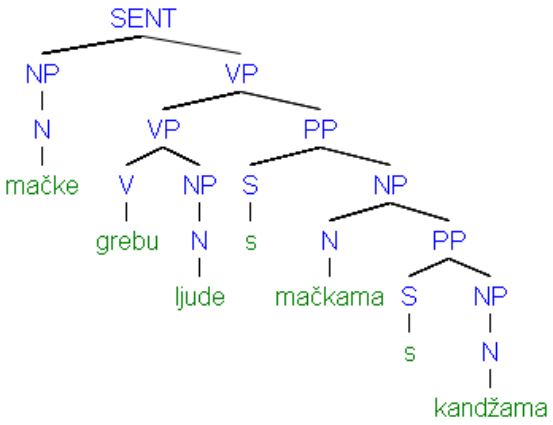
[iz Splita]

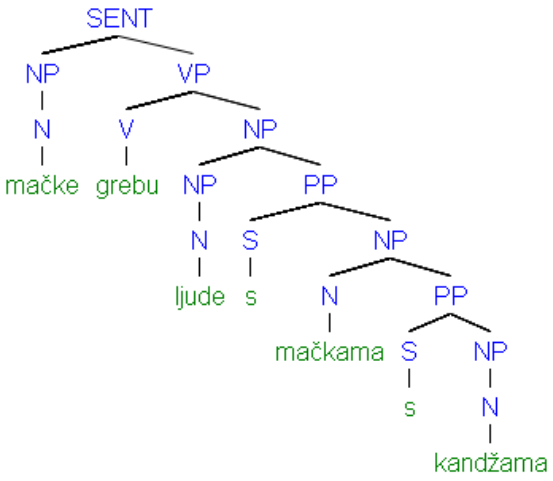
[za 25kn po dionici]

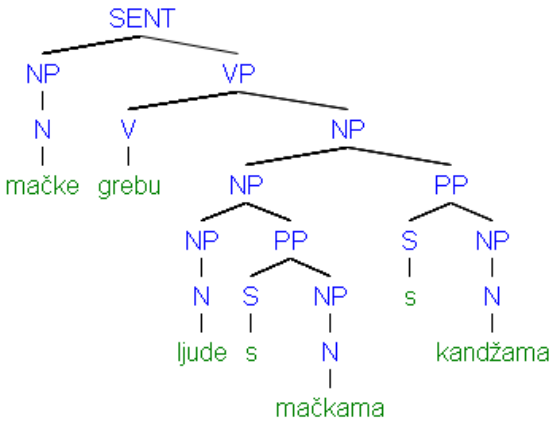
[na svom mjesečnom sastanku].

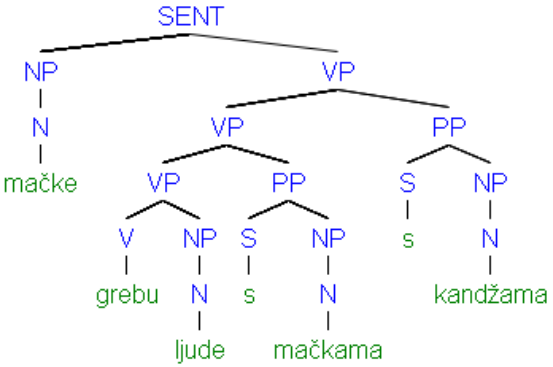
Catalanovi brojevi: $C_n = (2n)!/[(n+1)!n!]$ – eksponencijalni rast

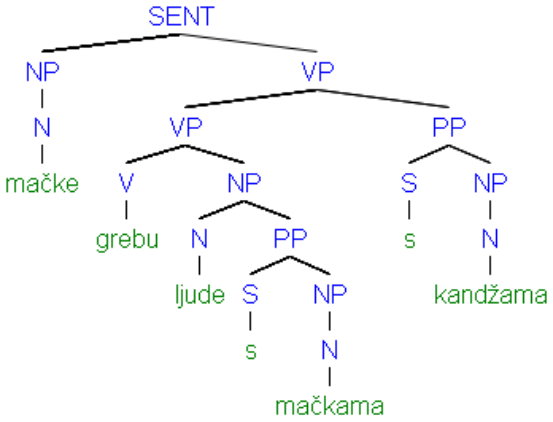
Dva problema: 1. ponavljanje posla



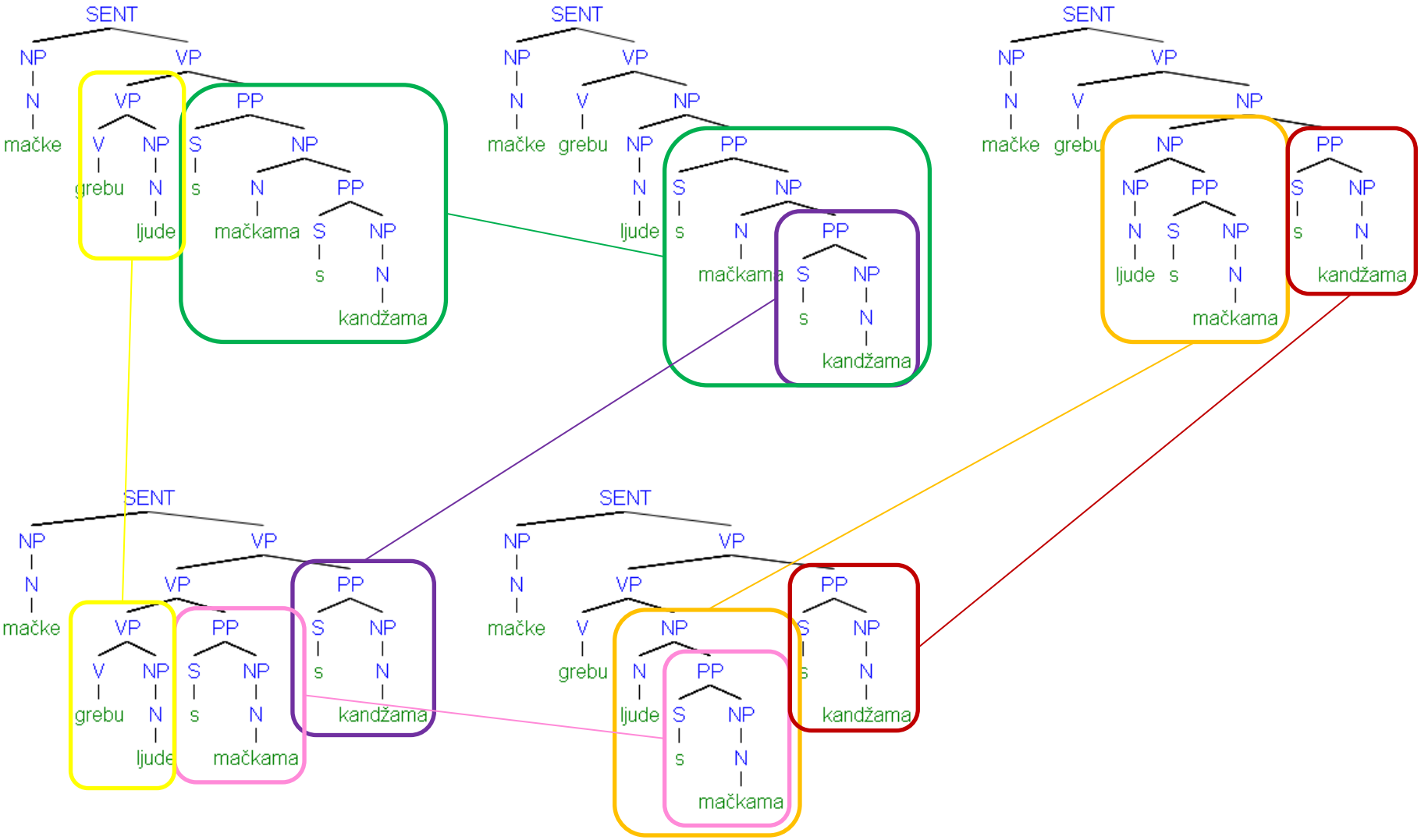








Dva problema: 1. ponavljanje posla



Dva problema: 2. izbor stabla parsiranja

- Kako dobro povezati

  Ona vidi čovjeka s teleskopom

- Riječi su dobri prediktori povezivanja
  - Čak i sa odsustvom potpunog razumijevanja

  Susjedna Saudijska Arabija je poslala više od 1000 vojnika u Bahrein …

  Izraelske vlasti su prekršile dogovor s washingtonskom administracijom …

- Statistički parseri će pokušati iskoristiti takvu statistiku