Nova generacija računalne obrade jezika

Damir Ćavar

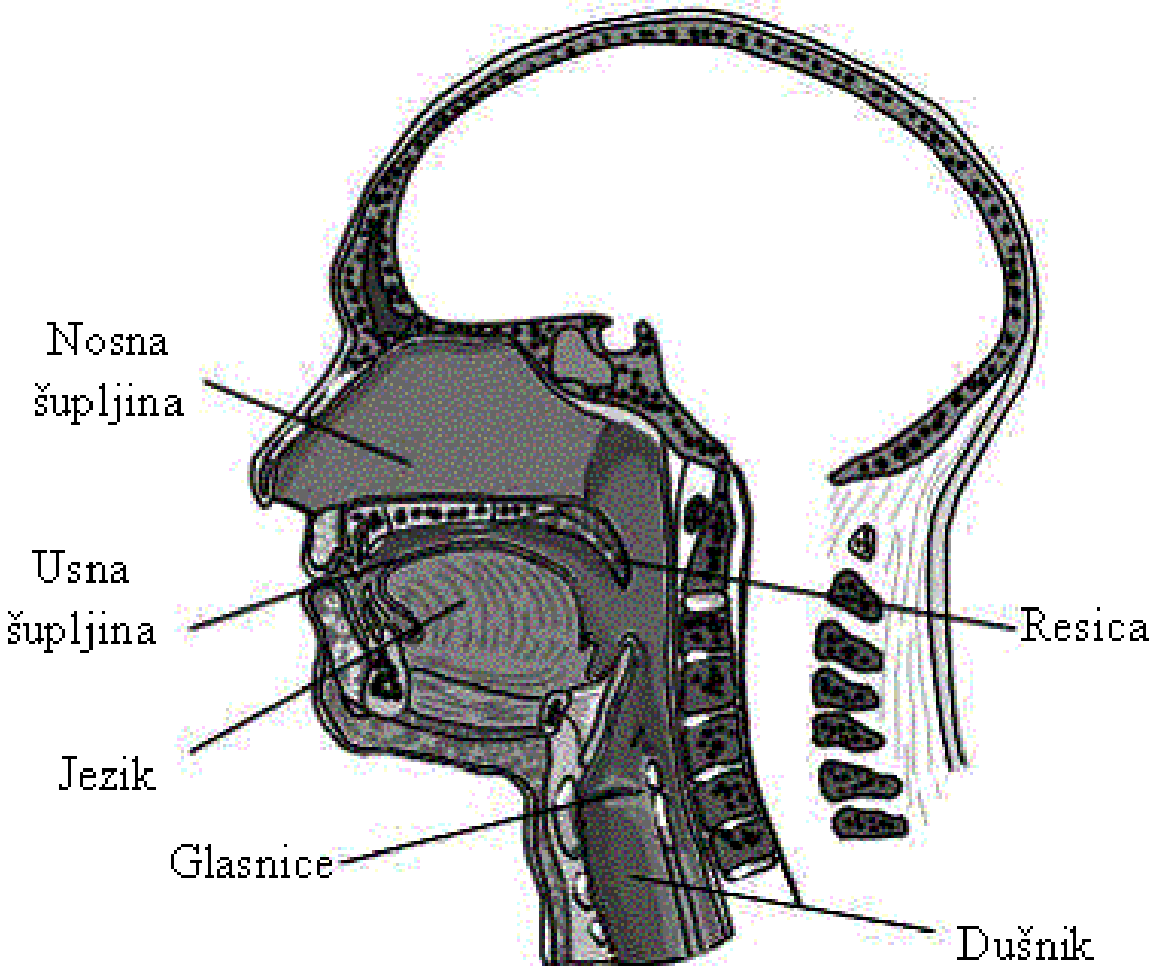
Odjel za lingvistiku u.o., Sveučilište u Zadru

34. skup IT profesionalaca u Splitu 2009

Što je jezik?

- Govor: percepcija
  - Kontinuirani nediskretni akustički događaji u vrijemenu
    - Spektrum varijacije energije na frekvencijama od 100–11000 Hz
    - Formanti: koncentracija energije na određenim frekvencijama
    - Prijelazi između šuma i tišine
- Govor: artikulacija
  - Kontinuirani nediskretne promjene u vokalnome traktu
    - Put zraka; položaj jezika, usana; stanje glasnice itd.
  - kao niz kompleksnih motornih instrukcija

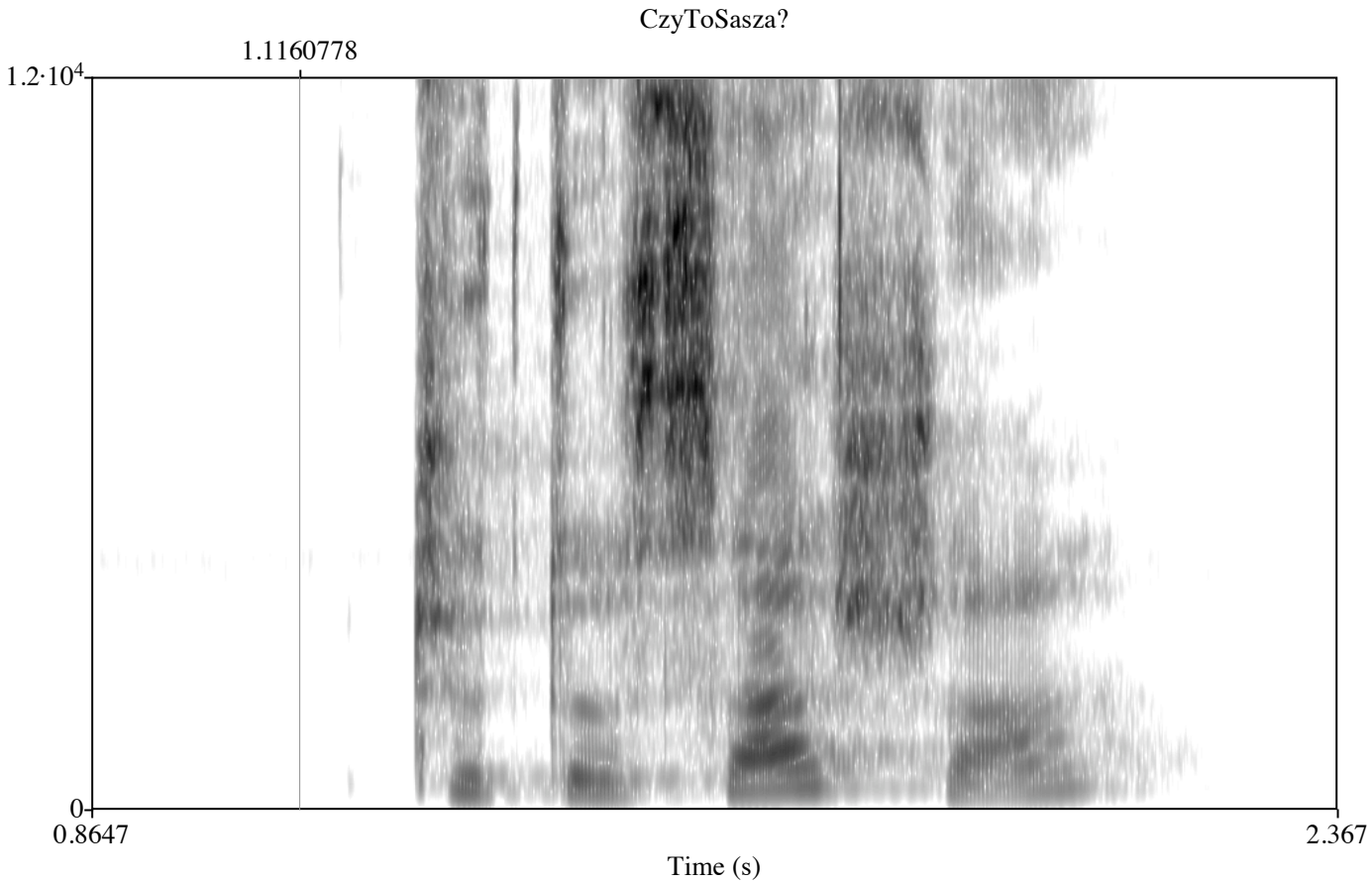
Što je jezik?



© Davor Petrinović

http://dog.zesoi.fer.hr/predavanja/HTML/Osnoveprocesanastajanjagovora.htm

Što je jezik?



Što je jezik?
Lingvističke osnove

- Razine lingvističke analize (teorije i modeli):
  - Fonetika i fonologija: zvukovi i fonemi
  - Morfologija: morfemi i riječi
  - Sintaksa: rečenica (i možda kontekst)
  - Semantika: značenje rečenice (možda u kontekstu)
  - Pragmatika: govorni čina, itd.
  - itd.

- Iluzija zato što:
  - lingvističke jedinice ne koreliraju nužno s fizičkim aspektima jezika,
  - nego su kognitivne interpretacije akustičkog događaja.

Što je jezik?
Lingvističke osnove

- Ekstralingvistička dimenzija:
  - Kognitivni sustav i njegove osobine (npr. *Lazy evaluation*, *Least Effort*, *Last Resort*, pamćenje)
  - Govorna situacija (npr. šum, događaji, biološki uvjeti)

- Lingvistička kognitivna jezgra:
  - Neovisne formalne osobine jezika

- Fonološka razina
  - Različiti (nediskretni) zvukovi klasificirani kao jedna linkvistička jedinica $\rightarrow$ fonem
  - Osnova: teorija ovisnosti i interdependencije zvukova i njihova kombinatorika
  - Primjer: hrvatski i španjolski "*r*" (*torero* – *onaj koji se bori s bikom*; *torrero* – npr. *stražar u svjetioniku*)

- Fonotaktička razina:
  - Hrvatski prihvaća "*dla*" a ne "*lda*" kao slog ili početak riječi

Što je jezik?
Lingvističke osnove

- Morfološka razina
  - Fonemi se slažu u morfeme, tj. najmanje jedinice koje imaju značaj ili neku funkciju, i koje se spajaju u riječi
  - Osnova: teorija značenja i funkcija, ovisnosti i interdependencije morfema i njihova kombinatorika
  - Primjer: hrvatski glagol "*čitati*" se može razdvojiti u dva minimalna dijela "*čita-*" i "*-ti*" s posebnim značenjem i funkcijama

- Morfotaktička razina:
  - Hrvatski glagoli tipa "*čita*" se mogu kombinirati sa sufiksima kao "*-m*" i "*-š*", ali ne s "*-om*", iako je "*-om*" legitiman sufiks hrvatskog jezika (npr. u riječi *ruk-om*)

Što je jezik?
Lingvističke osnove

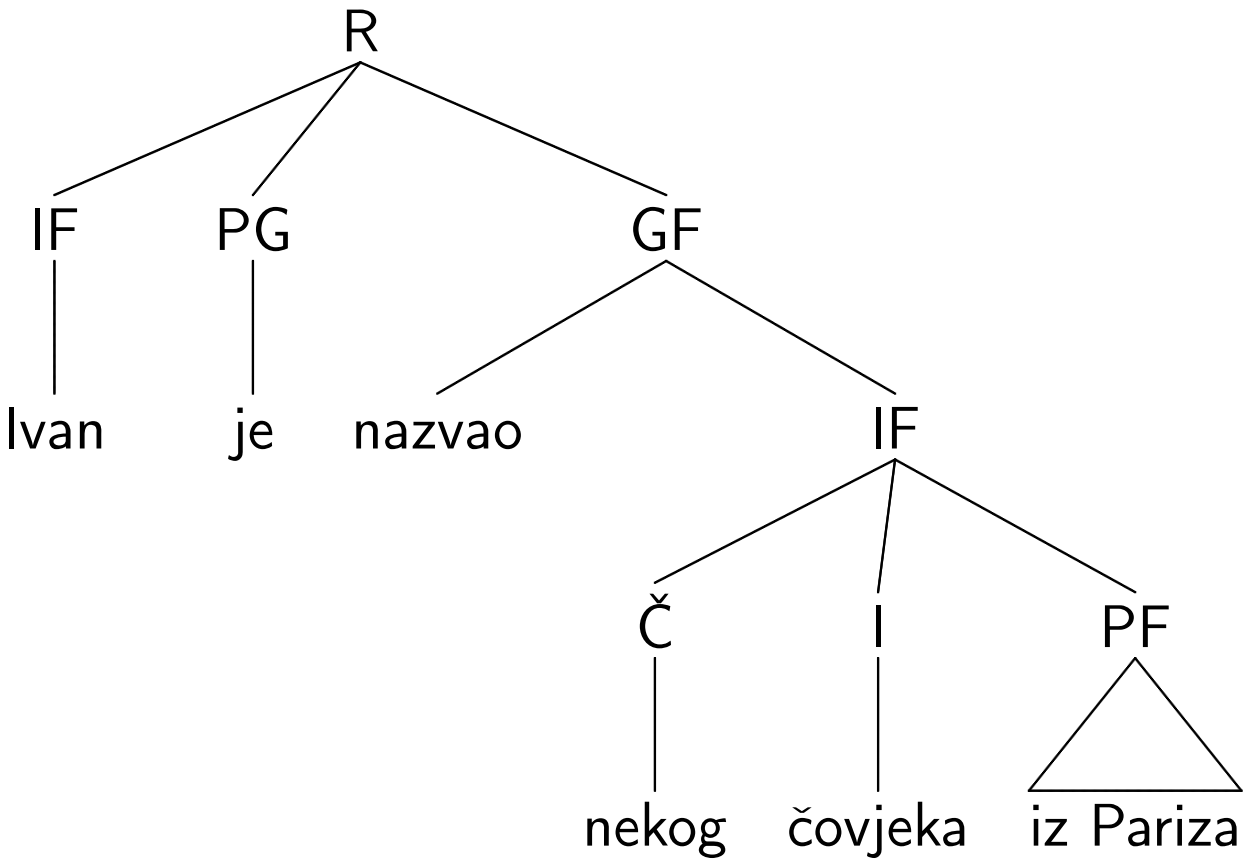
- Sintaktiča razina
  - Riječi se slažu u rečenice
  - Iako imamo dojam da su skoro sve kombinacije moguće, hrvatski je sintaktički jako ograničen

- Primjer:
  - Može biti: *Ivan se penje na krov.*
  - Ne može biti: *Krov Ivan se penje na.* ili *Ivan se krov penje na.* itd.

- Dodatni problemi:
  - Što znači: *Ivan je nazvao nekog čovjeka iz Pariza.*
  - Tko je *on* u: *Ivan* ***ga*** *je nazvao.* i *Ivan tvrdi da* ***ga*** *je Marija nazvala.*

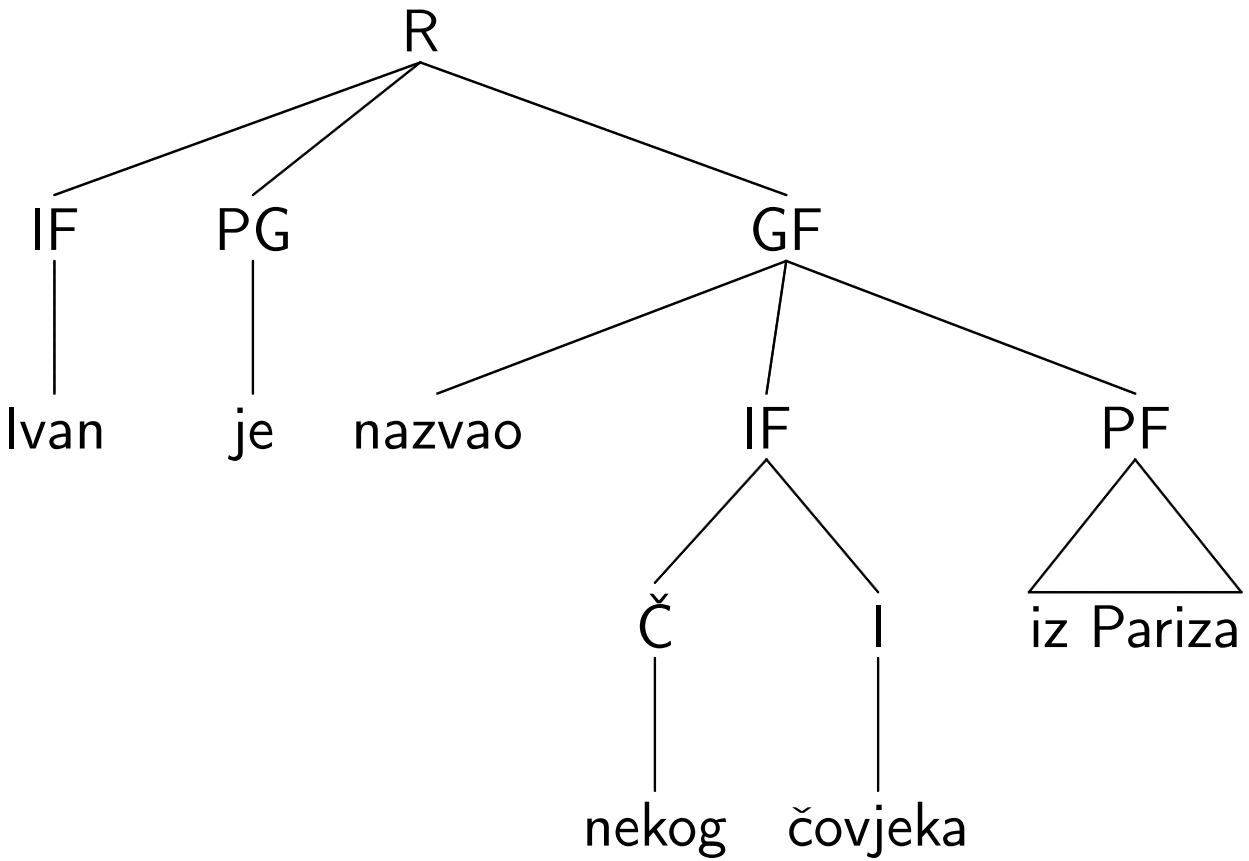
Što je jezik?

Sintaktiča stabla i hijerarhijska struktura



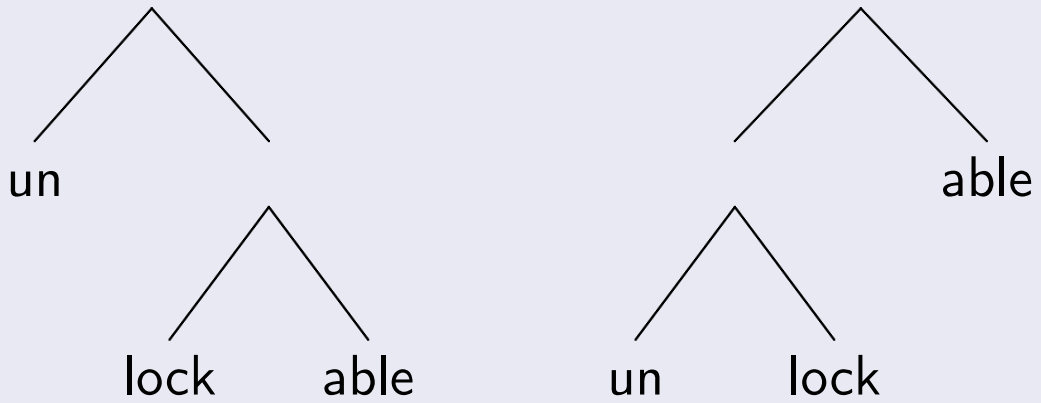
Što je jezik?

Sintaktiča stabla i hijerarhijska struktura



Što je jezik?
Sintaktiča stabla i hijerarhijska struktura

- Strukturalna višeznačnost
  - Za jednu rečenicu ili riječ postoji više struktura u skladu s gramatikom, npr.



- Leksička višeznačnost:
  - Jedna riječ ima više značenja: npr. *duga*, *pita*, *je*

- Gramatike (elementi i pravila) opisuju moguću kombinatoriku na svim lingvističkim razinama
- Deskriptivne gramatike
  - Opis zvučnih osobina jezika
  - Riječnici
  - Preskriptivne gramatike za standardni jezik
  - Dijalektološke gramatike

Lingvistička jezgra
Formalni aspekti jezika

- Formalne gramatike
  - koriste eksplicitnu teoriju i formalizam i omogućavaju falsifikaciju, teoretske predikcije itd.
  - Automati: generatori i prepoznavači jezičnih izraza (na svim lingvističkim razinama) (niski generativni kapacitet)
  - Parseri: analizatori jezičnih izraza (visoki generativni kapacitet)
- Palatalizacija (poljski): *krok* – *kroczek*; *mózg* – *móżdżek*; *duch* – *duszek*

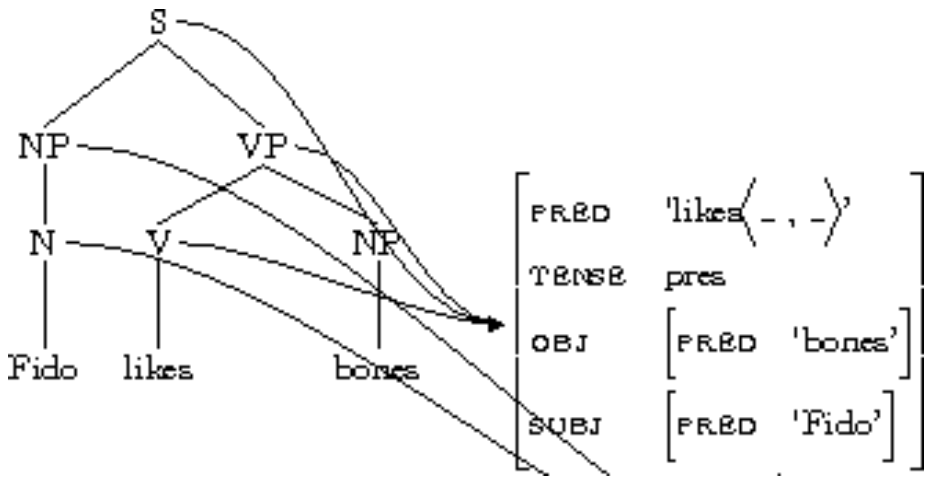
  `k,g,h → č,dž,š/___ i,e [ Deminutiv | Vokativ ]`
- Sintaksa:

  `S → NP VP`

  `NP → (Adj) N (PP)`

  `...`

- Dodatni formalizmi: Unifikacijske gramatike (LFG, HPSG itd.)



© Andrew Bredenkamp http://www.essex.ac.uk/linguistics/clmt/latex4ling/avms/

Lingvistički modeli
Razlike: formalni i prirodni jezici

- Prirodni jezici su višeznačni na svim višim lingvističkim razinama
- Sve razine su povezane i doprinose riješenju problema višeznačnosti $\rightarrow$ paralelizam u analizi, procesiranju itd.
- Gramatike prirodnih jezika su rekurzivne (tj. regularne, kontekstno neovisne i ovisne), što objašnjava neograničen broj izraza, rečenica itd.

- Formalne osobine:
  - Regularna: fonologija i fonotaktika, morfologija
  - Kontekstno neovisna: sintaksa (možda djelomično kontekstno ovisna)
  - Semantika itd.: kontekstno ovisna

Lingvistički modeli
Formalne osobine

- Kontekstno neovisne i kompleksnije gramatike (i jezici) se formalno ne mogu usvojiti samo s pozitivnom evidencijom (Gold, 1967)
  iako sada postoje istraživanja koja to relativiraju
- Takve gramatike kompleksne su u procesiranju
- Ne pokrivaju nikada 100% podatke
- Ne predviđaju razlučivanje višeznačnosti

Lingvistički modeli
Statistička revolucija (ponovo)

- Generiranje gramatika, riječnika i lingvističkih modela iz obilježenih lingvističkih podataka (npr. korpusa)
- Kontekstno neovisne gramatike s vjerojatnosti pravila
  `S → NP VP` p:0.021
  `NP → (Adj) N (PP)` p:0.001
  ...
- Konačni automati s vjerojatnosti na prijelazima (i/ili prijelaznim akcijama kod transduktora)
- *n*-gram modeli
- Nesimbolički statistički modeli (npr. neuronske mreže)
- ...

Namjena lingvističkih modela

Osnovno procesiranje tekstualnih oblika jezika

- Transkripcija u fonetski i/ili fonemski oblik
  - IPA transkripcija
  - za npr. phonex, soundex pretraživanje, statističke fonetske modele za prepoznavanje govora i sintezu itd.
- Morfološka segmentacija i obilježje:
  - *izponapijali*: aspektualni prefiks – aspektualni prefiks – korijen i lema *napiti* – sufiks participa u množini
  - Dodatno obilježje:
    do neke mijere – malo – "opiti se" od korijenske leme *piti* – prošlost
  - za parsiranje i semantičku analizu

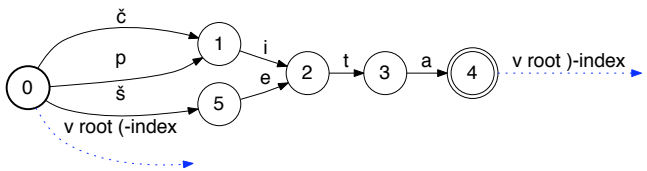
Namjena lingvističkih modela

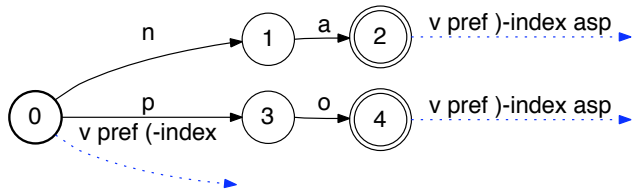
Osnovno procesiranje tekstualnih oblika jezika

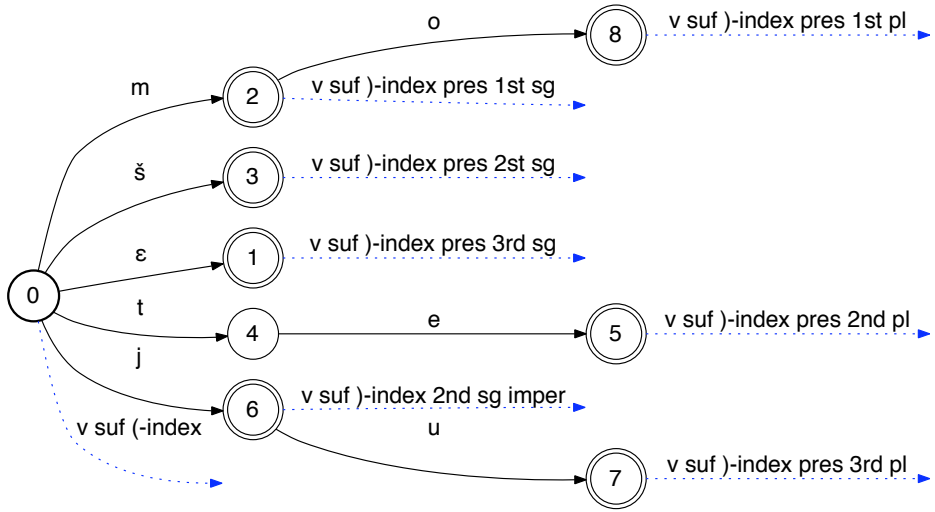
- Sintaktičko parsiranje
  - Stabla s kategorijama i hijerarhijskom strukturom skopusa sintaktičkih fraza i riječi
  - za npr. razlučivanje višeznačnosti, semantičku analizu
- Semantičko obilježje i analiza
  - Stabla i mreže relacija i povezivanje s reprezentacijom koncepata i funkcija
  - za npr. strojno prevođenje, prepoznavanje govornog čina, analizu sadržaja itd.

Konačni automati

Morfemi kao deterministički konačni automati (DKA) (Mealy ili Moore automati):







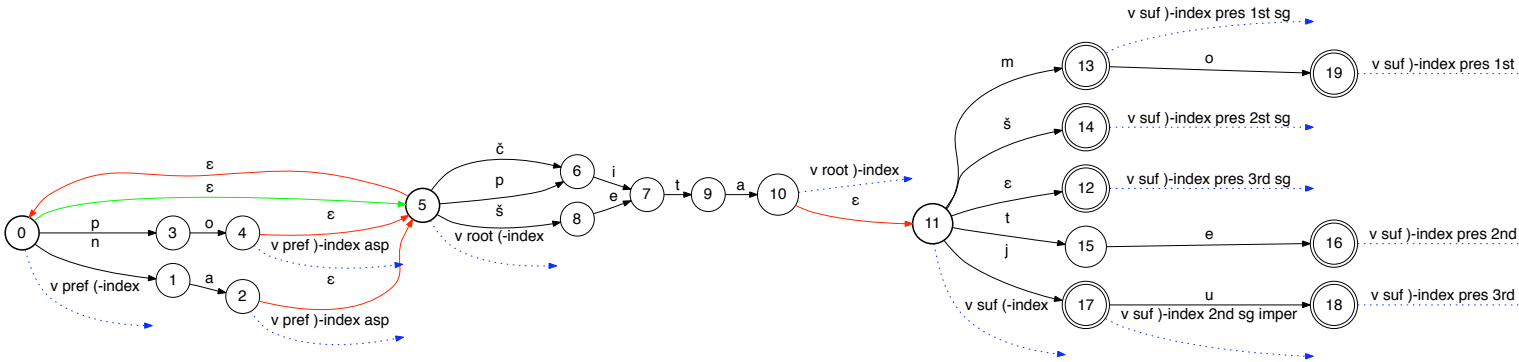
Spajanje u monolitičke automate uz regularne izraze

- Višeznačnost kao emisija više oznaka: lista emisija 1 do *n*
- Oznaka DKA s imenom
- Pravila koja koriste ta imena i modeliraju morfotaktička distribucijska pravila:

```
glagolAspektPref* . glagolAtiKorijeni . glagolFleksSuf
```

Generiranje monolitičkih automata

Monolitički automati, mogu biti ciklički DKA:



Sintaktičko parsiranje

Obično u računalnoj lingvistici Earley parser i varijacije tog algoritma (dinamično programiranje):

- s dodatnom vjerojatnosti za razlučivanje najvjerojatnije analize u slučaju višesnačnosti
- s unifikacijom obilježja za pravila kongruencije i perkolaciju oznaka
- s obilježjem semantičkih osobina i funkcija

Statistički modeli
za npr. obilježje i prepoznavanje

- n-gram modeli
  - lingvističko obilježje ovisno o (obično lokalnom) kontekstu
  - distribucijske osobine fonema, morfema, riječi u kontekstu odkriva osobine teksta, riječi itd.
  - za npr. klasifikaciju teksta u jezike, zadržajno; klasifikaciju nepoznatih riječi itd.

Glavni zadatci

- Stvaranje lingvističkih resursa
- Stvaranje alata za lingvističku analizu
  - lematizacija riječi u tekstu za pretraživače i daljnu analizu
  - gramatike i transfer pravila za strojno prevođenje
  - prepoznavanje jezičnih jedinica i klasifikacija u npr. ime osobe, ime tvrtke, ime produkta, datum i vrijeme, lokacija itd.
  - klasifikacija tekstova
  - analiza govora i procesiranje govornog dijaloga
  - prepoznavanje zadržaja za forenzičku analizu
  - ekstrakcija znanja i generiranje novih saznanja
  - itd.