# Učni korpus govorjene slovenščine

## Napovednik

- Govorni korpusi
- Učni korpus govorjene slovenščine
    - sestava
    - namembnost in uporabnost

## Govorni korpusi

Govorne zbirke, pravimo jim tudi zbirke govorjenega jezika, so nepogrešljive pri raziskovalnem delu na področju tehnologij govorjenega jezika. Vsebujejo predvsem računalniško berljive posnetke govora.
Posnetkom so vedno priloženi še podatki, ki na različne načine opisujejo posneti govor. Priloženi podatki so opisi govornih dejavnikov (npr. spol, starost, poklic, narodno-narečno območje), dejavnikov govorcev (npr. govorčevo trenutno razpoloženje, zvočne lastnosti okolja ob snemanju) ter zapisi in označitve posnetega govora. Pogosto so priložena še navodila za uporabo zbirke ter rezultati izbranih analiz govornih posnetkov.

- so računalniške zbirke transkribiranih posnetkov spontanega govora
- govorni korpusi proti korpusom govora
- za preverjanje hipotez o jeziku
- za opis jezika, predvsem v leksikografiji in slovnici
- za učenje jezika
- za sintezo in razpoznavanje govora

- BNC (10 000 000), BOE (10 000 000), nizozemski (cca. 9 500 000)
- švedski (cca. Malo nad 100 000)
- pod 100 000: ICE, London-Lund, COLT, Santa Barbara, C-ORAL-ROM, IBM-Lancaster, ČNK

## Gradnja govornega korpusa

Postopek pridobivanja govorne zbirke lahko logično razstavimo v tri zaporedne korake:

1. izbira besedila, potrebnega za snemanje govorne zbirke, oz. izbira govorne situacije v primeru, ko gre za snemanje spontanega govora
2. snemanje govornega gradiva
3. segmentacija, označevanje in analiza govornega gradiva.

Zajem besedil:

- demografska metoda zbiranja gradiva
- metoda taksonomije besedil

- **načrtovanje vsebine zbirk**: naj bodo čim bolj obsežne; neomejeno veliko količin podatkov v zbirki si ne moremo privoščiti, pazljivo načrtovanje vsebine zbirke, da ta čimbolj predstavlja celotno izbrano območje govorjenega jezika
- **pridobivanje govornih posnetkov**: dva načina, izbira besedil je ključnega pomena in je odvisna od namena zbirke
    1.) govorci izgovorijo vnaprej pripravljeno besedilo v snemalno napravo
    2.) snemanje spontano govorjenega besedila, ki je lahko bodisi monolog ali v pogovor
- **segmentacija in označevanje govorne zbirke**:
    - **segmentacija**: posneti govorni signal predstavlja le en del govorne zbirke, ta je brez ustreznih oznak govornih odsekov večinoma neuporabna za nadaljnje raziskave; govorni signal je v postopku segmentacije potrebno razmejiti oz. segmentirati na posamezne segmente ali govorne odseke in jim v postopku označevanja ali anotacije pripisati oznake na različnih anotacijskih nivojih: grafskem, fonetičnem, prozodijskem (prozodijske značilnosti

govora-nanašajo na trenutno razpoloženje govorca ter njihov pristop k tvorjenju govornih posnetkov) (odvisno od namena uporabe zbirke)

- **označevanje** (taksonomija?):

  - samodejno grobo označevanje govorne zbirke: avtomatsko, potem še ročno, dobra lastnost postopka je to, da se je sposoben učiti (predhodne ročne korekcije se upoštevajo pri naslednjem avtomatskem označevanju)

  - fino ročno označevanje govorne zbirke: za ročno pregledovanje in označevanje govorne zbirke ter popravljanje oznak govornih segmentov se uporabljajo raznovrstna programska orodja, namenjena delu z govornimi signali (primer: Sigmark)

**Sestava UKGS**

- **govorci**: najbolj preproste govorne zbirke so zbirke z nekaj deset govorci, število govorcev znotraj tega razpona že omogoča statistično ustreznost zbirke, to zagotovimo z upoštevanjem standardnih statističnih postopkov pri izbiri vzorca populacije vseh govorcev
  - govorci UKGS glede na : spol, starost, izobrazbo,regijo
- Pri izbiri govorcev moramo upoštevati nekatere značilnosti, ki so povezane z njihovim govorom, te značilnosti delimo na prehodne in trajne.
  - *prehodne* so morebitne psihološke ali fiziološke motnje (npr. počutje, bolezen, psihično stanje)
  - *trajne* so fiziološke in anatomske značilnosti (npr. spol, starost, teža, okvare na govorilih, kadilske in pivske navade)
- velikost: 15.000 pojavnic ??
- besedila glede na:
  - strukturo (D : M)
  - javnost/zasebnost
  - osebni stik/medij
  - formalnost/neformalnost
  - posneto z vednostjo/na skrivaj

**Transkribiranje**

- **Osnovna načela:**

- priporočila mednarodnih organizacij za standardizacijo korpusov (TEI, EAGLES)
- osnovna enota je izjava, ki jo omejuje premor ali menjava govorcev
- razširjena ortografska transkripcija
- brez ločil
- velika začetnica samo v lastnih imenih

**Kritični pogled na UKGS**

- oportunistične metode zbiranja
- pomanjkljiva demografska sestava govorcev
- nepopoln zajem besedil glede na taksonomijo
- nujno brisanje osebnih podatkov iz posnetkov
- ni lematiziran in morfosintaktično označen
- potrebna je korekcija transkripcijskih načel