Avtomatsko indeksiranje 2

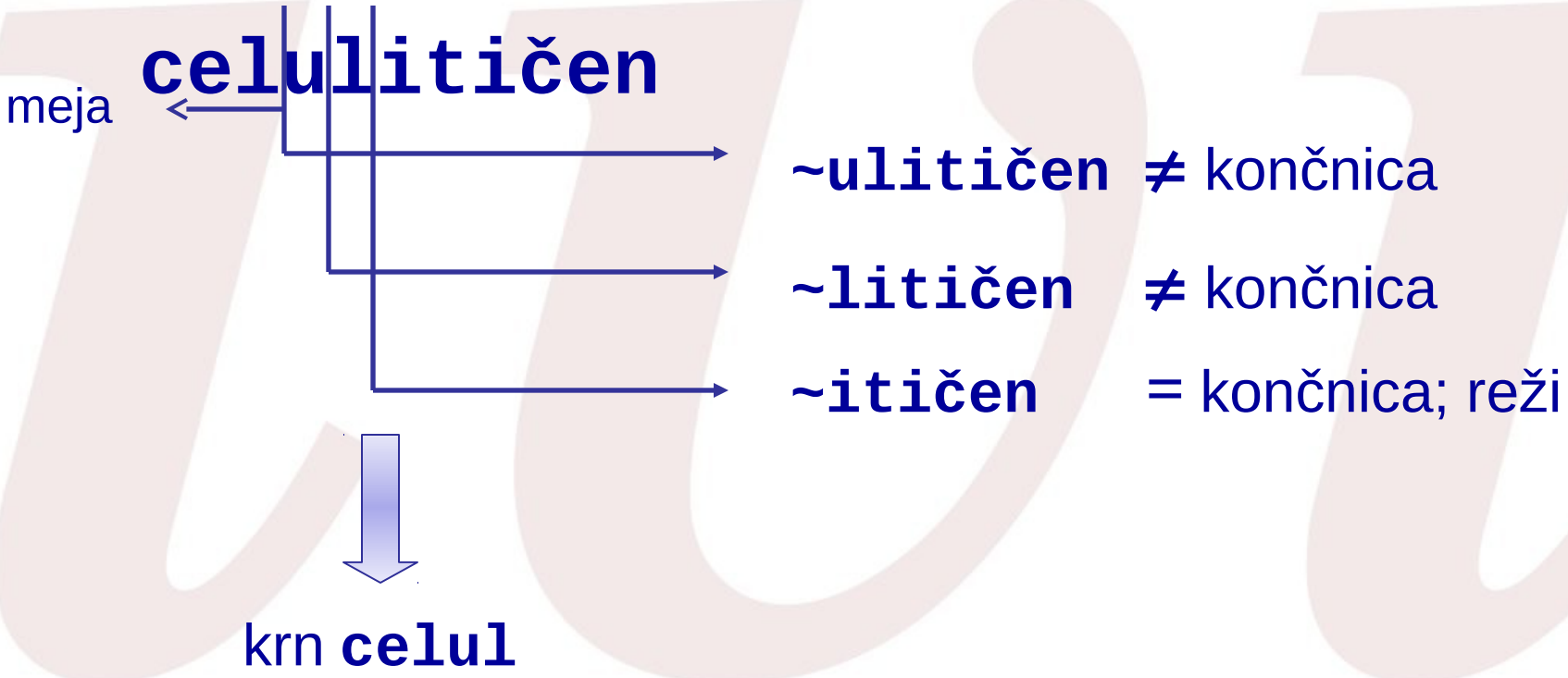
Krnjenje - nadaljevanje,
algoritmi s seznami končnic,
krnjenje slovenščine…

Algoritmi s seznamom končnic

- Algoritmi uporabljajo spisek možnih končnic v nekem jeziku.
- Za vsako besedo v dokumentu določijo s pomočjo spiska najustreznejšo končnico, ki jo potem odrežejo.
- Najenostavnejši so algoritmi z najdaljšim ujemanjem: najustreznejša končnica je najdaljša končnica.
- Spiske končnic dobimo v dovolj velikem korpusu s sortiranjem po obrnjenih besedah.

Načelo najdaljšega ujemanja končnic

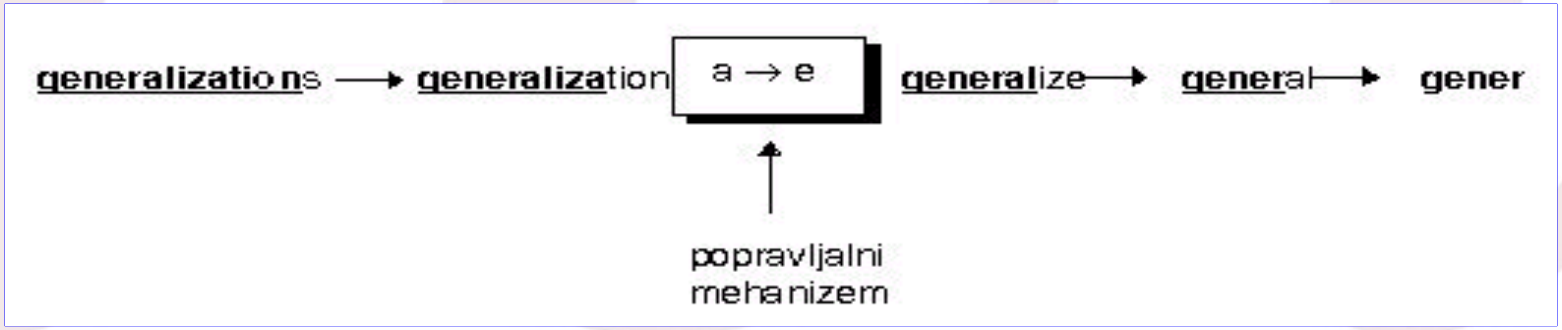
Obstaja najkrajši dovoljen krn, npr. 3 znaki;
krnimo besedo **`celuliti̇čen`**,
iščemo končnico od leve proti desni:



Algoritmi s seznamom končnic

Metoda s cikličnim rezanjem končnic:

- Sestavljene končnice nastanejo v procesu oblikovanja besede z nalaganjem krajših končnic.
- Algoritmi zaporedoma luščijo posamezne končnice in pri tem uporabljajo popravljalne mehanizme.
- Računalniško potratne metode, zahtevajo dobro modeliranje jezikovnih značilnosti, vendar dajo najboljše rezultate.



Napake pri krnjenju

- Prešibko krnjenje.
  Odrezane končnice so prekratke, za besedne oblike iste besede algoritem izdela različne krne:
  - **knjižnica → knjižn**
    knjižnični → knjižn
    knjižničnega → knjižničn
- Premočno krnjenje:
  Odrezane končnice so predolge, za različne besede algoritem izdela isti krn:

    knjižnica → knjiž
    književnost → knjiž

Krnjenje slovenskih besedil

- Krnjenje je še posebej pomembno za slovenske dokumente.
- Pri slovenščini ne smemo pričakovati, da bodo krni vedno del originalnih besed:
  - **`jetra, jeter, jetrom`** - edini skupni krn bi bil **`jet`**, vendar je prekratek;
  - dober krn je **`jetr`**, vendar ni del vseh besednih oblik;
  - potrebna so popravljalna pravila, predvsem e→0 (**`jeter → jetr`**).

Krnjenje slovenskih besedil

- Večina slovenskih iskalnikov ne uporablja pravega krnjenja pri gradnji zbirk dokumentov.
- Nekateri uporabljajo angleške algoritme za krnjenje - ☹☹☹.
- Nekateri shranjujejo v zbirko nekrnjene besede in se zanašajo na iskalce, ki morajo "ročno" krniti pri iskanju - ☹.
- Objavljena le dva prava algoritma za krnjenje slovenskih besedil: Popovič, Dimec.

Krnjenje slovenskih besedil: Popovič, 1991

- Algoritem z najdaljšim ujemanjem končnic.
- 5.276 različnih končnic.
- 8 pravil za rezanje končnic, na primer:
  - odreži končnico, če je v besedi znak pred njo soglasnik:
    ~**alna** nacionalna → nacion; socialna → social
  - odreži končnico, vendar ne, če sta v besedi pred njo dva zaporedna soglasnika:
    ~**ata** kandidata → kandid; kolovrata → kolovrat
  - odreži končnico, vendar ne, če sta v besedi pred njo "bl" ali "st":
    ~**em** hitrem → hitr; problem → problem

Krnjenje slovenskih besedil: Popovič, 1991

Rezanju končnic sledi:

- obravnava izjem,
- splošna pravila za popravljanje (20 pravil), naprimer:

~sež, ~seč → ~seg presež, preseč → preseg

~niš, ~nič → ~nik tehniš, tehnič → tehnik

- pravila za popravljanje alteracij e → 0, naprimer:

~**soglasnik+~r** → ~soglasnik+~er kadr → kader

Krnjenje slovenskih besedil: Dimec, 1988, 1995, 1999

- t.i. “Optimalni algoritem”. Razlogi za konstrukcijo:
  - vsi obstoječi algoritmi za slovenščino premočno krnijo (splošna značilnost algoritmov z najdaljšim ujemanjem končnic),
  - vsi obstoječi algoritmi za slovenščino modelirajo splošni jezik in niso prilagojeni izrazju v strokovnih podjezikih,
  - v strokovnih podjezikih so zelo pogoste besede z izvorom v grščini, latinščini in angleščini.

Optimalni algoritem, 1999

- Algoritem deluje v treh korakih:
  - rezanje končnice (uporablja le končnice, ki se začenjajo na samoglasnik - 3650 končnic),
  - obdelava soglasniških parov na koncu krna (60 pravil),
  - pravila za popravljanje (večinoma e→0).

Optimalni algoritem, 1999

- Primer: krnjenje besed konec, končen, končnega
  - 1. korak (rezanje končnic):
    ec → ‘c’; en → ‘’; ega → ‘’:
    konec → konc; končen → konč; končnega → končn
  - 2. korak (obdelava soglasniških parov):
    čn → č; nč → nc:
    konč → konc; končn → konč → konc

konec → konc

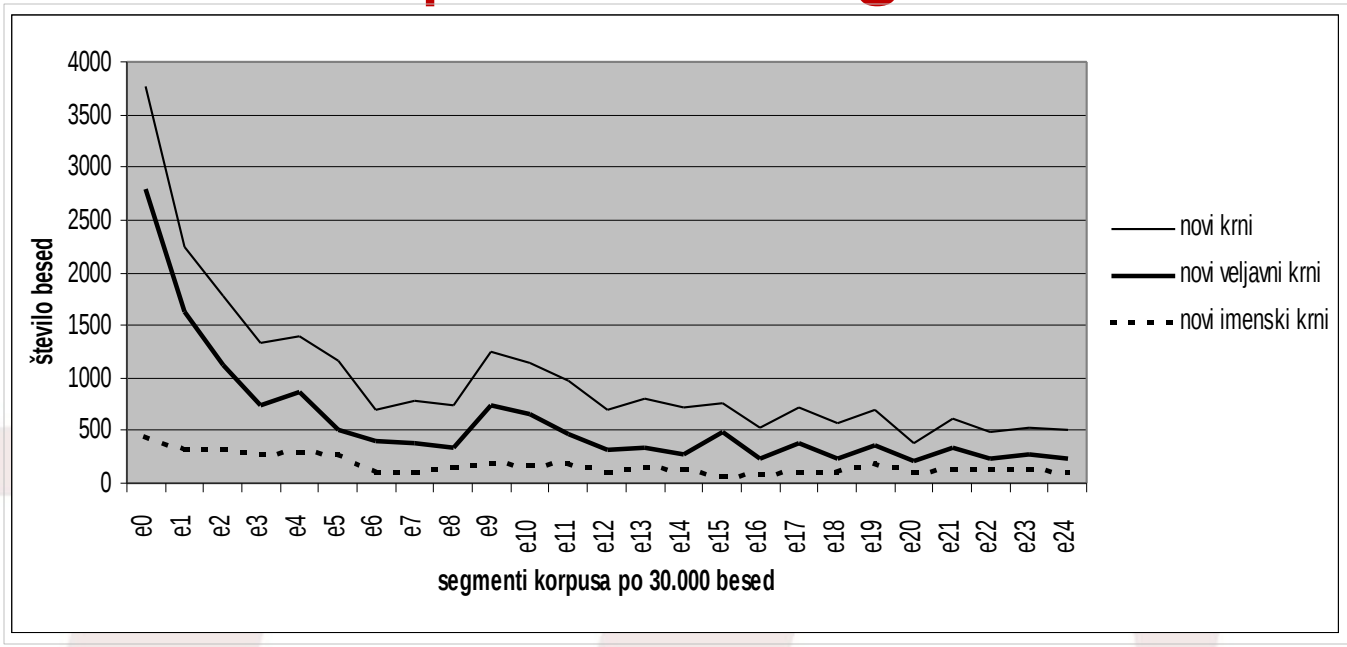
končen → konc

končnega → konc

Optimalni algoritem, 1999

- Optimalni algoritem ne uporablja načela najdaljšega ujemanja končnic, ampak je bližje cikličnemu algoritmu!
- Vedno poskuša najti najkrajšo končnico, ki da “optimalni” krn.
- Uporablja seznam “optimalnih” krnov za podjezik.
- Algoritem obdeluje besedo od desne proti levi.
- Algoritem je učljiv.

Optimalni algoritem, 1999



Krivulja učenja medicinskega podjezika v korpusu z 750.000 besedami.

Prvi segment: 12,54% novih krnov.

Zadnji segment: 0,81% novih krnov.

Vključevanje virov dokumentov:

a = ISIS, glasilo Slovenske zdravniške zbornice,

b = JAMA, Journal of American Medical Association, Slovenska izdaja,

c = Medicinski razgledi,

d = Zdravniški vestnik.

Optimalni algoritem, 1999

Primer:

- **`-itičen, -oza, -aren`** so običajne končnice v slovenskem medicinskem podjeziku.

besede	alg. z najdaljšim ujemanjem	optimalni algoritem
`celulitičen`	**`celul -itičen`**	**`celulit -ičen`**
`celuloza`	**`celul -oza`**	**`celuloz -a`**
`celularen`	**`celul -aren`** 	**`celular –en`**