Korpusno jezikoslovje
Uvod

Korpusno jezikoslovje

UNG
2008/2009

Nekaj besed o predavatelju

- Tomaž Erjavec
  Odsek za tehnologije znanja
  Institut "Jožef Stefan"
  Ljubljana
- http://nl.ijs.si/et/
- tomaz.erjavec@ijs.si
- jezikovne tehnologije
  - izdelava korpusov in drugih jezikovnih virov, predvsem za slovenski jezik
- konferenca IS-LTC 2008
- spletna stran za predmet:
  http://nl.ijs.si/et/teach/ung08-kj/

Študentje

- kaj že veste o korpusih?
- kaj pričakujete od predmeta?

Kaj je korpus?

- obsežna zbirka besedil
- jezik v resnični in sodobni podobi
- v elektronski obliki
- reprezentativnost za jezik, ki naj bi ga predstavljali -> **vzorec**
- služi za **opisovanje** jezika (deskriptivno/empirično jezikoslovje)

Definicija korpusa po EAGLES

Guidelines of the Expert Advisory Group on Language Engineering Standards:

Corpus *: A collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language.*

Computer corpus *: a corpus which is encoded in a standardised and homogeneous way for open-ended retrieval tasks. Its constituent pieces of language are documented as to their origins and provenance.*

Vrste korpusov

- pisni oz. govorni korpusi
- referenčni oz. korpusi podjezikov
- celoviti oz. vzorčni korpusi
- statični oz. spremljevalni korpusi
- enojezični oz. večjezični
- označeni oz. neoznačeni

Pisni oz. govorni korpusi

- pisni korpusi
  - teh je velika večina
  - cena, enostavnost obdelave
- podvrste “govornih” korpusov:
  - pisni, a namenjeni za govor: drame, predloge govorov
  - govorni korpus: transkripcija govora,
    - npr. predavanj, parlamentarnih razprav, intervjujev, klepeta ob kavi, ..
    - problemi transkripcije (spontanega) govora
  - govorjeni korpusi: posnetki govora
    - kvaliteta zvoka proti naravnosti govora
    - problemi varovanja pravice do zasebnosti
  - govorjeni korpusi za namene govornih tehnologij
    - omejeno besedišče - ogromno govorcev
    - rezervacija kino vstopnic, naročanje pice, …

Referenčni korpusi oz. korpusi podjezikov

- korpus podjezika oz. specializiran korpus
  - obravnava posameznega tipa besedil
  - terminološke študije
  - korpus posameznega avtorja, obdobja, besedilnega tipa,…
- referenčni korpus
  - vzorec “celotnega” jezika
  - veliki, dragi, skrbno sestavljeni
  - tipično sinhroni
  - dokumentirani, pravno čisti, označeni
- kriteriji pri izbiri besedil:
  - reprezentativnost:
    korpus zajema “vse” besedilne zvrsti
  - uravnoteženost:
    velikosti vzorcev besedilnih zvrsti so v sorazmerju z njihovo “pomembnostjo” za govorce jezika
- metodologija in statistika proti dejanski praksi…

Celoviti oz. vzorčni korpusi

- celoviti korpusi vsebujejo celotna besedila, korpusi vzorcev pa samo iztržke iz besedil
- v splošnem je bolje, da korpus vsebujejo celotna besedila, vendar:
  - zgodovinsko, problemi z velikostjo korpusov
  - pravni problemi
  - problem uravnoteženosti

Statični oz. spremljevalni korpusi

- statični korpusi:
  večina korpusov se, ko so narejeni, ne spreminja več
- spremljevalni korpusi (monitor corpora) se sproti dopolnjujejo:
  omogočajo opazovanje jezika v spreminjanju
  - spremljevalni korpusi so še vedno redki, saj je izdelavo potrebno dodatno avtomatizirati in vzdrževati

Enojezični oz. večjezični korpusi

večjezični korpusi koristni za prevajanje:

- vzporedni korpusi:
  besedilo skupaj s prevodom oz. prevodi
- primerljivi korpusi:
  različna, vendar primerljiva besedila v večih jezikih

Vzporedni korpusi

- isto besedilo v večih jezikih
- korpus se najprej poravna po stavkih
  - ali pa zajame iz pomnilnika prevodov
- izredno uporabni za:
  - prevodoslovne študije
  - “poceni” dvojezični slovar
  - (pol)avtomatsko luščenje prevodnih ustreznic
  - učne množice za strojne prevajalnike

Označeni oz. neoznačeni korpusi

- neoznačeni korpusi vsebujejo samo besedila in dokumentacijo o njih
- označeni korpusi dodatno vsebujejo v besedilih jezikoslovne oznake:
  - oblikoslovne oznake
  - leme
  - skladenjske povezave
  - imena
  - …

Značilnosti dobrih korpusov

- *avtentičnost*:
  korpus ustreza kriterijem, glede na katere je bil narejen
- *količina*:
  čim večji, tem boljši
- *kakovost*:
  zapis in oznake korpusa so pravilne
- *enostavnost*:
  računalniški zapis korpusa je razumljiv
- *dokumentiranost*:
  korpus je opremljen z bibliografskimi in drugimi podatki

Nekaj zgodovine

- 1964: korpus Brown
  - Kucera in Francis, 1964
  - ameriška angleščina
  - 1 milijon besed, 500 vzorcev po 2.000 besed
  - vzorci enakomerno razdeljeni na različne zvrsti besedil
  - vse besede ročno označene z oblikoskladenjskimi oznakami (part-of-speech tags)
- 1978: LOB
  - enak kot Brown, vendar za britansko angleščino

Sinclairova revolucija

- 1980: začet projekt “Cobuild”
- sodelovanje Collins Publishers in Birmingham University
- projekt izdela spremljevalni korpus “Bank of English” (100..200..300M besed)
- namen: izdelati slovar, osnovan na računalniškem korpusu
- rezultat: Cobuild English Dictionary
- vodilni znanstvenik je bil John Sinclair, utemeljitelj korpusnega jezikoslovja
- dostopnost
  - sedaj proti plačilu, v sklopu ponudbe založnika Collins

BRITISH NATIONAL CORPUS

- 1994: BNC, prvi računalniški nacionalni referenčni korpus
- konzorcij pod vodstvom Oxford University Press
- 100 milijonov besed, vzorčen, sinhron
- uravnotežen in reprezentativen
- vsebuje tudi govorni del
- oblikoslovno označen
- dostopnost
  - enostavno spletno iskanje
  - naprednejše s programom SARA
  - dostopen tudi v celoti, za nekomercialno uporabo
  - zapisan v skladu z mednarodnimi standardi

Desetletje nacionalnih korpusov

- BNC: British National Corpus (100M, 1994)
- CNC: Czech National Corpus (100M, 1998)
- HNC: Hungarian National Corpus (100M, 1998)
- HNK: Croatian National Corpus (100M, 1999)
- SNK: Slovak National Corpus (100M, 2000)
- slovenski jezik:
  - Fida (100M, 1998) / FidaPLUS (600M, 2000)
  - Beseda (100M, 1998) / Nova Beseda (200M, 2000)

Korpusi za vsakogar (ki ima $)

- LDC: Linguistic Data Consortium (1992, ZDA)
- ELRA: European Language Resources Association (1995)
- korpusi za jezikovne tehnologije: npr. MULTEXT-East

Tretje tisočletje: splet kot korpus

- tradicionalno je bilo zbiranje besedil za korpus dolgotrajen in drag proces
- danes na spletu najdemo ogromno besedil iz raznovrstnih področij
- zakaj torej ne uporabiti spleta kot vira za izgradnjo korpusov?
- avtomatske metode selekcije, zajema in poenotenja formata medmrežnih strani
- korpusi dosegajo 1.000.000.000 besed
- ponovno omejitve računalniških zmogljivosti

Raziskovalne paradigme v (računalniškem) jezikoslovju

Performansa proti kompetenci:

- 1950 -- 1960: prvi "korpusi"
  - empirija, vendar šibki računalniki
- 1970 -- 1980: Chomsky
  - raziskovanje jezikovne kompetence
  - umetna inteligenca, pravila
  - globinske analize, temeljne raziskave
  - neuporabno v praksi
- 1990 -- 2000: renesansa empirije
  - korpusno jezikoslovje
  - strojno učenje, statistika
  - površinske analize, aplikativne raziskave
- 2010 -- : združevanje paradigem?

Kje se korpusi uporabljajo?

- teoretično jezikoslovje
  - korpusno podprte raziskave
  - na korpusih temelječe raziskave
- uporabno jezikoslovje
  - slovaropisje
  - poučevanje jezikov
  - prevodoslovje
- jezikovne tehnologije
  - učni podatki
  - testni podatki

Korpusi na spletu

- Angleščina:
  - British National Corpus [http://www.natcorp.ox.ac.uk/]
  - Bank of English [http://www.cobuild.collins.co.uk/form.html]
- Nemščina
  - COSMAS II Korpusauswahl [http://www.ids-mannheim.de/cosmas2/]
- Zbirke povezav na korpuse:
  - [http://devoted.to/corpora]
  - [http://www.clarin.eu/wp5-documents/wg-53-documents/survey-corpora]
  - [http://universal.elra.info/]
- Splet kot korpus
  - WebCorp [http://www.webcorp.org.uk/index.html]
  - KwicFinder

Slovenski korpusi na internetu

- Slovenščina:
  - FIDAplus [http://www.fidaplus.net]
  - Nova beseda [http://bos.zrc-sazu.si/s_beseda.html]
  - Specialni enojezični korpusi [http://nl2.ijs.si/index-mono.html]
- Slovensko-angleški vzporedni korpusi:
  - ELAN, TRANS, SVEZ [http://nl2.ijs.si/corpus/index-bi.html]
  - EVROKORPUS [http://www.gov.si/evrokorpus/]

FIDAPLUS
korpus slovenskega jezika

- referenčni korpus slovenskega jezika
  - uravnotežen in reprezentativen korpus slovenščine
- 600 milijonov besed
- sinhron korpus: 1990-2000
- avtomatsko oblikoslovno označen in lematiziran
- sodelavci projekta FIDA:
  Filozofska fakulteta
  Inštitut Jožefa Stefana
  DZS d.d.
  Amebis d.o.o.
- dostopnost
  - spletno iskanje
  - naprednejše skozi SketchEngine
  - ni dostopen kot podatkovna množica

Korpusa JOS

- Projekt jezikoslovno označevanje slovenskega jezika
- jezikovnotehnološki nameni: bogate oznake
- v delu
- jos100k
  - 100,000 besed
  - vzorčen iz FidaPLUS
  - ročno oblikoslovno označen in lematiziran
  - za učenje oblikoslovnih označevalnikov in lematizatorjev
  - v prihodnosti še skladenjsko in pomensko označen
- jos1M podoben, vendar 10x večji
  - in samo delno ročno popravljan
- dostopen
  - prototipno spletno iskanje
  - zastonj dostopen tudi v celoti, za nekomercialno uporabo
  - zapisan v skladu z mednarodnimi standardi