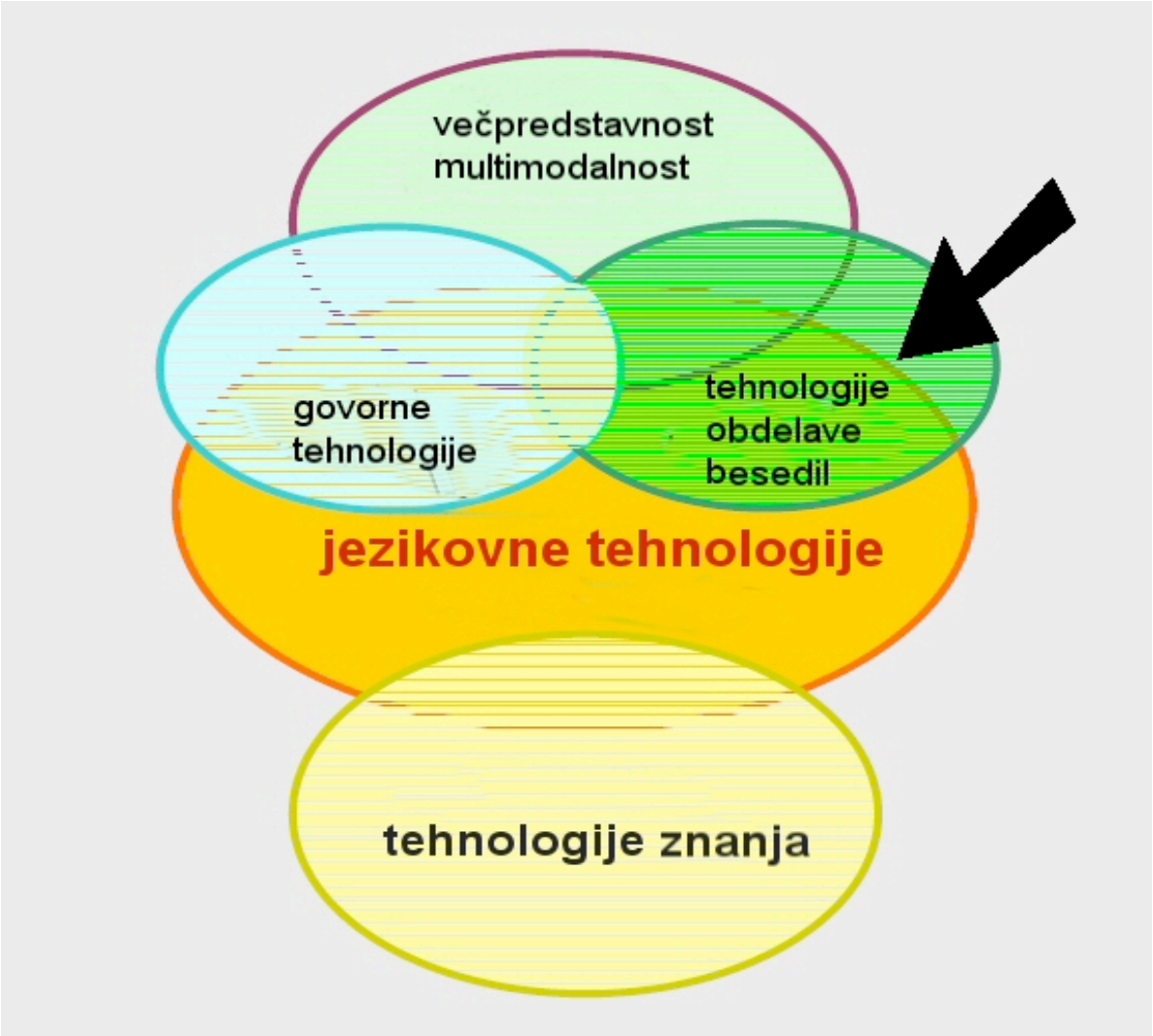
Osnovna opremljenost jezika s prostodostopnimi (pisnimi) viri in njihova standardizacija

Simon Krek,
Odsek za tehnologije znanja,
Institut Jožef Stefan

Maribor, 26. 10. 2007

O čem govorimo?



Nekaj pomembnejših področij...

- avtomatski prevajalni sistemi
- zbirke besedil (besedilni korpusi)
  - leksikografija, korpusno jezikoslovje...
- rudarjenje podatkov
  - klasifikacija besedil, ekstrakcija terminologije, sumarizacija...
- komunikacija med človekom in računalnikom

Mednarodno – lokalno

- mednarodni standardi zapisa jezikovnih podatkov
  - XML (Extensible Markup Language)
  - TEI (Text Encoding Initiative)
  - OSCAR (Open Standards for Container/ Content Allowing Re-use)
    - TMX (Translation Memory eXchange)
    - TBX (Term-Base eXchange)
    - SRX (Segmentation Rules eXchange)

Jezikovno specifično

- EAGLES (Expert Advisory Group on Language Engineering Standards)
- ISLE (International Standard for Language Engineering)
- osnova:
  - nabor oznak (oblikovnih, skladenjskih…)
  - leksikon (besednih oblik)
  - parser (oblikovni in skladenjski, semantični?)

Splošna priporočila – standardizacija...

- predpostavlja poznavanje področja, ki temelji na izstopajočih primerkih najboljših obstoječih praks
- mora temeljiti na kritični masi strokovnjakov s tega področja
- dokazati mora svojo relevantnost tako pri praktični uporabi kot pri razvoju tehnologije
- predlogi standardov morajo biti dostopni na način, da jih lahko uporabniki preizkusijo pri svojem delu in ocenijo njihovo relevantnost

Nabor oznak - ME

- Multext – Multext-EAST
  - Institut “Jožef Stefan”
    - http://nl.ijs.si/ME/
  - V.3, 7. julij 2004
  - prost dostop
- spremenjeni ME: FERI Maribor
- individualizirani ME: Amebis d.o.o.
- revidirani ME: JOS, IJS

Nabor oznak – ZRC SAZU

- ZRC SAZU
  - P. Jakopin and A. Bizjak. 1997. O strojno podprtem oblikoslovnem označevanju slovenskega besedila. Slavistična revija, 3–4:513–532
  - strojno berljiv?
  - dokumentiran?

Leksikon

- Multext-EAST (Amebis)
  - 15.000 enot
  - dostopen na spletu za raziskovalne namene ob podpisu licence
- Amebis
  - obsežen
  - komercialni dostop
- ZRC SAZU
  - ni javno dostopnih podatkov?

Učni korpus

- Multext-EAST
  - korpus ME = dostopen na spletu za raziskovalne namene – ob podpisu licenčne pogodbe
  - FidaPLUS = dostopen na podatkovnem nosilcu za raziskovalne namene – ob podpisu pogodbe
- spremenjeni ME: FERI Maribor
- individualizirani ME: Amebis d.o.o.
- revidirani ME: JOS, IJS
- nabor ZRC SAZU

Parser (POS)

- Amebis d.o.o.
  - FidaPLUS
  - komercialen
- IJS:
  - korpus Acquis
  - nedokumentiran
  - ni “prosto dostopen”

Iztočnice za debato?

- standardizacija
  - da ali ne?
  - stopenjskost?
  - združljivost
- prosta dostopnost
  - odprta koda
  - programski paket (OS?)
- dokumentiranost