# Korpusi v jezikoslovju in korpusi slovenskega jezika

Izr. prof. dr. Vojko Gorjanc
Oddelek za prevajalstvo FF UL
vojko.gorjanc@guest.arnes.si

# Lingvistika 20. stoletja

- Evropa in Severna Amerika
  - Ženevski strukturalizem
    - Ferdinand de Saussure (1857-1913), *Cours de linguistique générale*, 1916
  - Amerika: sinhroni pristop – opazovanje, zapisovanje in analiza govorjenega jezika (antropologija)
    - Franz Boas (1858-1942), *Handbook of American Indian Languages*, 1911
    - Edward Sapir (1884-1939), *Language*, 1921

# Lingvistika 20. stoletja

- Evropa in Severna Amerika
  - V Evropi močan vpliv ženevskega strukturalizma
    - Praški linvistični krožek, 1926 (magičnost navadnega jezika)
      - Vilém Mathesius, Roman Jakobson, Nikolai Sergeevič Trubetzkoy
    - Anglija – ameriški vpliv
      - John Rupret Firth
  - Severna Amerika
    - Tradicija ameriških antropologov (magičnost govorjenega jezika)
      - Leonard Bloomfield, *Language*, 1933

# Lingvistika 20. stoletja

- Avram Noam Chomsky (1928- ); *Syntactic Structures*, 1957
  - generativna oz. tvorbno-pretvorbna slovnica
  - velika jezikoslovna šola

  - delitev jezikoslovja izrazito na dva pola
    - chomskijansko in njegovo opozicija
      - opozicija je vrsta šol, ki temeljijo na realnosti komunikacije

# Gradivno usmerjeni jezikoslovni opisi

- raziskave mrtvih jezikov – korpus razpoložljivih besedil
- analiza jezikov brez pisne tradicije – govorni korpusi (Franz Boas)
  - načrtno zbiranje jezikovnih podatkov
    - od končnega nabora jezikovnih podatkov (mrtvi jeziki) do zbiranja in izbiranja besedil kot osnove jezikovne analize
    - od opazovanja pisnega k opazovanju govorjenega jezika

# N. Chomsky in “njegova opozicija”

N. Chomsky zavrača obsežne načrtno zbrane jezikovne podatke : jezikoslovje, utemeljeno na jezikovnih podatkih kot “organizirana opozicija”

Dileme

# Dileme

- absolutna zadostnost korpusa : absolutna zadostnost intuicije
- intuicija rojenega govorca
  - idealni govorec
- normativistični pristop
  - jezikovna raba : odklon od norme : napaka

  “KJER SE LOMIJO KOPJA”
- jezikovna realnost in uporabno jezikoslovje
- jezikovna realnost in procesiranje naravnih jezikov

# Korpusi prve generacije

- Dva osnovna principa
  - splošni korpus
  - specializirani korpus (korpus za posebne namene)

- Prvi korpus še ne pravi korpus
  - SEU (Survey of English Usage), 1955, 1959
    - še klasični neračunalniški, britanska angleščina
    - polovica besedil je transkribcija govora
    - kasneje prenesen v elektronsko obliko

    V TEM ČASU GRADIVNA ZBIRKA ZA SSKJ

    klasična listkovna kartoteka

# Korpusi prve generacije

- Korpus Brown, 1961-1964
  - ameriška angleščina
  - prvi elektronski korpus
  - standard je milijon besed

  V TEM ČASU GRADIVNA ZBIRKA ZA SSKJ

  klasična listkovna kartoteka

  Na lingvističnem inštitutu FF Univerze v Zagrebu

  prevod polovice korpusa Brown

  desetletje kasneje Mogušev korpus

# Korpusi prve generacije

- Korpus LOB (Lancaster, Oslo, Bergen) 1970-1978
  - angleška varianta korpusa Brown

# Korpusi druge generacije

- Povezani z razvojem tehnologije
- Povezava raziskovalnih, univerzitetnih in komercialnih (predvsem založniških) okolij
  - Birmingham University in Collins Cobuild
- Vprašanje zastarevanja podatkov in njihova neprenosljivost in netrajnost
  - standardizacija

# Korpusi druge generacije

- Angleški in ameriški
  - BNC, The Bank of English, ANC
- Nemški
  - Cosmas, Korpus nemščine 20. stol
  - Ideja nemškega nacionalnega korpusa
- Srednjeevropski
  - HNC, HNK, ČNK, SNC
- Drugi slovanski

# Delo s korpusi: jezikoslovje in jezikovne tehnologije

- gradnja korpusov
  - načela gradnje, elektronski zapis, označevanje
- orodja za delo s korpusom
- jezikovnotehnološke aplikacije: črkovalniki, elektronski slovarji, tezavri ...
- jezikoslovne raziskave
  - opisno jezikoslovje
  - uporabno jezikoslovje

# Tipi korpusov

- Referenčni
- Govorni
- Specializirani
- Vzorčni
- Spremljevalni ali dinamični
- Primerljivi
- Vzporedni

# Karakteristike korpuosv

- Količina
  - velikost
- Kakovost
  - avtentičnost besedil
  - avtentičnost zapisa
- Dokumentiranost (besedil)
- Enostavnost (zapisa)

# Projekt FIDA

- od 1997 do 2000

- Filozofska fakulteta Univerze v Ljubljani
- Institut Jožef Stefan
- DZS, d. d.
- Amebis, d. o. o.

# Karakteristike korpusa FIDA

- enojezikovni
- sinhroni
- referenčni
- (izhodiščno) pisni

**Reprezentativni korpus slovenskega jezika**
**velikost = 100 milijonov besed**

# Osnovna razmerja - zvrst

**Ft.Z zvrst**

- **Ft.Z.U umetnostna**
  - Ft.Z.U.P pesniška
  - Ft.Z.U.R prozna
  - Ft.Z.U.D dramska
- **Ft.Z.N neumetnostna**
  - **Ft.Z.N.S strokovna**
    - Ft.Z.N.S.H humanistično-družboslovna
    - Ft.Z.N.S.N naravoslovno-tehnična
  - **Ft.Z.N.N nestrokovna**

# Osnovna razmerja - lektura

- Avtentičnost besedil
  - lektura lahko poseg v avtentičnost
    - specifika slovenskega prostora

# Osnovna razmerja - prenosnik

| | | |
|---|---|---|
| Ft.P.P pisni | | |
| Ft.P.P.O objavljeno | | |
| **Ft.P.P.O.K knjižno** | **23 %** | |
| **Ft.P.P.O.P periodično** | **72 %** | |
| **Ft.P.P.O.P.C časopisno** | | **2/3** |
| Ft.P.P.O.P.C.D dnevno | | |
| Ft.P.P.O.P.C.V večkrat tedensko | | |
| **Ft.P.P.O.P.R revialno** | | **1/3** |
| Ft.P.P.O.P.R.T tedensko | | |
| Ft.P.P.O.P.R.S na štirinajst dni | | |
| Ft.P.P.O.P.R.M mesečno | | |
| Ft.P.P.O.P.R.D redkeje kot na mesec | | |
| Ft.P.P.N neobjavljeno | | |
| Ft.P.P.N.J javno | | |
| Ft.P.P.N.I interno | | |
| Ft.P.P.N.Z zasebno | | |

# Količinsko zajemanje govornih in pisnih besedil (Eagles)

RA in TV

lokalni RA in TV

zborovanje

predavanje

razred

diskusija

intervju

konverzacija

časopis

revija in knjiga

obvestilo

lokalna publikacija

delovni dokument

okrožnica

zapis del. skupine

privatna koresp.

# Regionalna uravnoteženost

# Dodatne informacije

- lematizacija (vsaki besedi je pripisana njena osnovna oblika)
- oblikoskladenjske oznake (pripisana je oblikoslovna kategorija – spol, sklon itd. – ali skladenjska vloga)

- izvor besedila
- čas nastanka
- druge informacije: spol, regija itd.

# Dostopnost

- http://www.fida.net
- glede na status uporabnika:
  - projektni partnerji – prosto
  - drugi – plačilo

- http://www.fidaplus.net
- prost dostop:
  - registracija

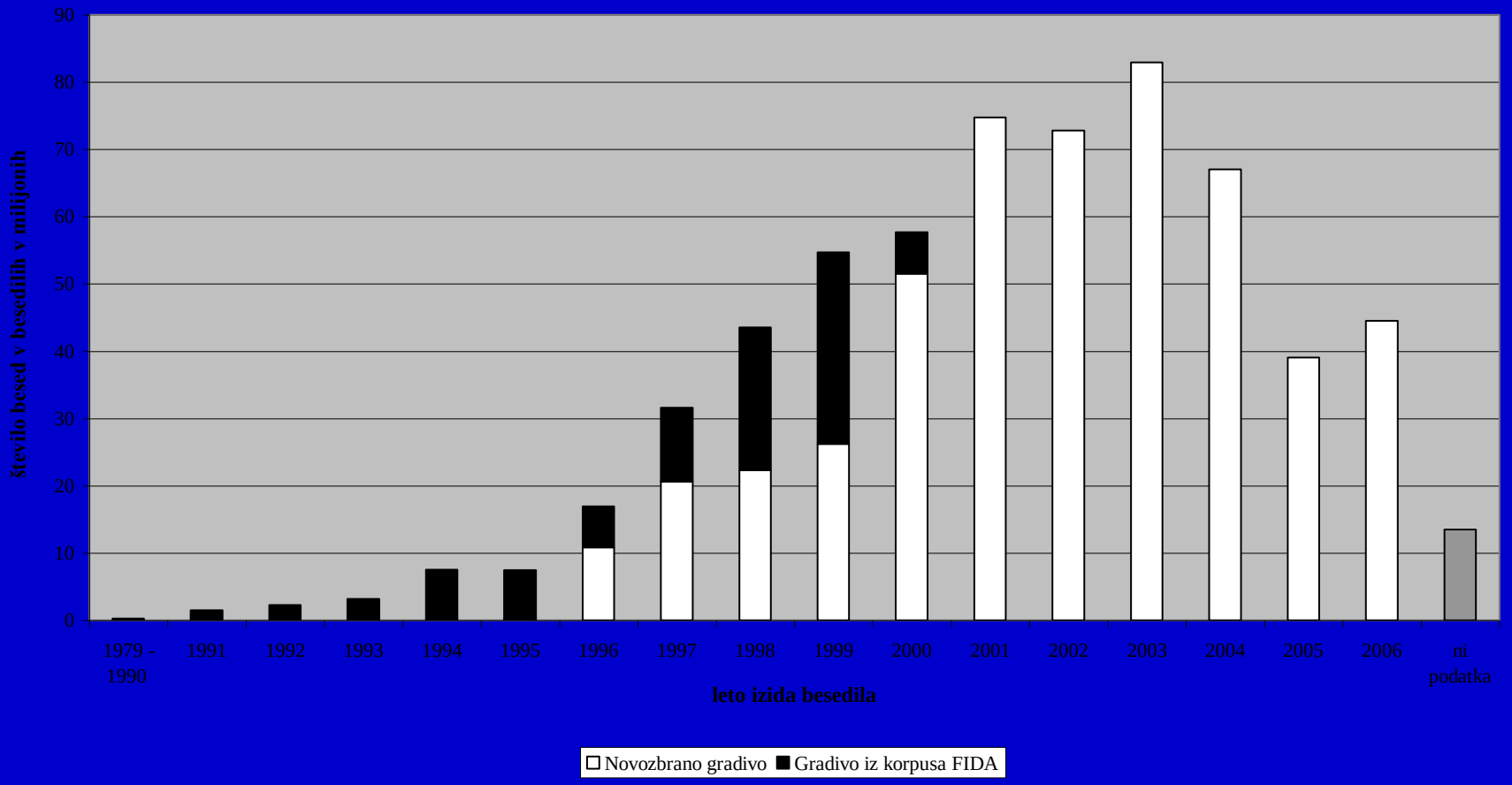
# FidaPLUS

- Projekt nadgradnje korpusa FIDA
  - Količinsko: nad 620.000.000 besed
  - Kvalitativno
    - Nova besedila: uravnoteženost, časovni razpon …
    - Novi postopki označevanja: lematizacija, MSD
    - Posodobljeno orodje za analizo: konkordančnik ASP32

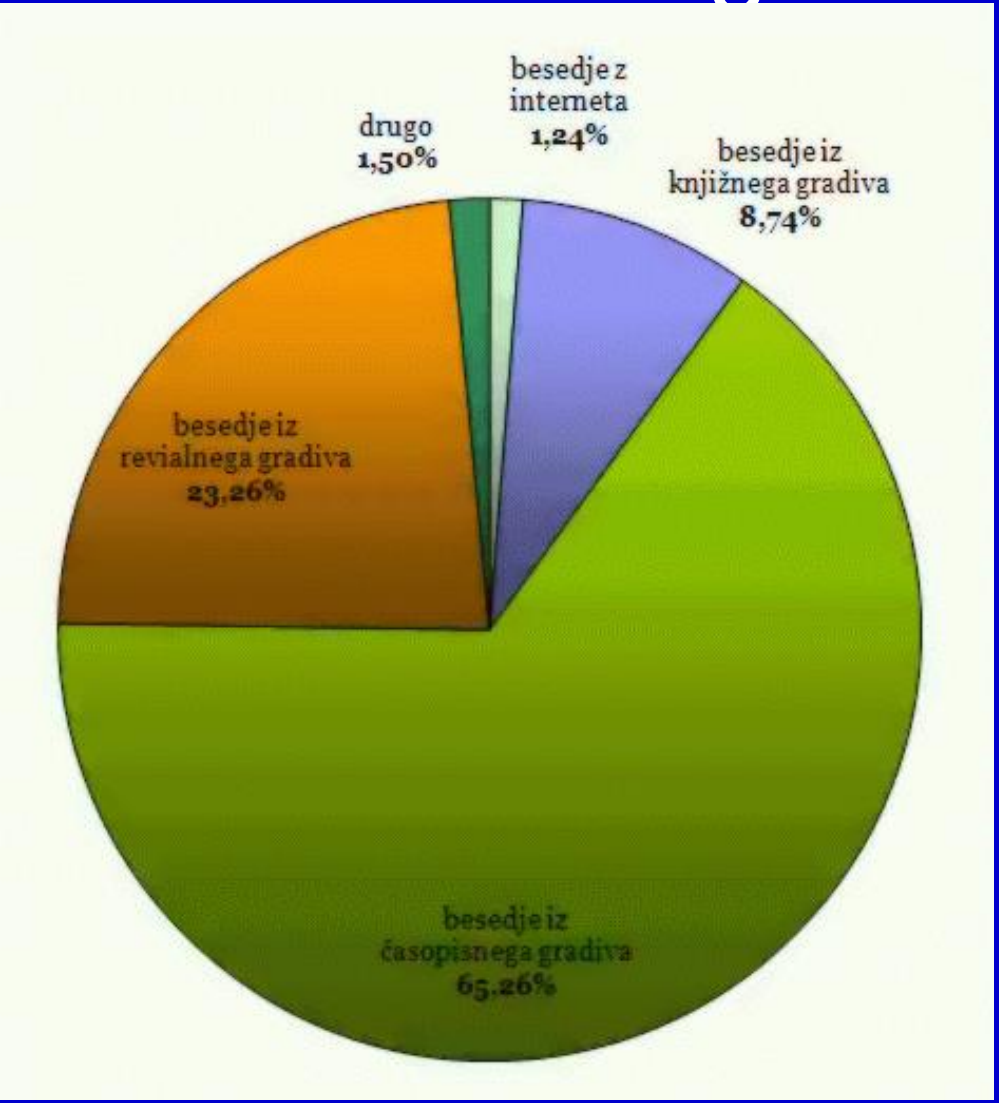
Vse relevantne informacije o zgradbi korpusa so na naslovu korpusa: http://www.fidaplus.net

in v članku Arhar & Gorjanc, JiS 2/2007

# FidaPLUS: besedila glede na čas izida



# FidaPLUS: besedila glede na



# Orodja za analizo korpusov – konkordančniki

- WordSmith, MonoConc
- CUE, Xkwic ...

- SARA - BNC
- ASP32 - FIDA
- Spletni iskalniki - IJS-Elan, Beseda ...

# Iskanja z ASP32

- Več o tem na vajah: samo preletimo in ponovimo

- Enostavno iskanje:
  - nadomestni znaki (*, ?)
  - lema
  - MSD
  - kombinacije iskalnih možnosti, npr. izključevalnost: *#3čudovit~#1čudovito*

# Iskanja z ASP32

- Razširjeno iskanje
  - taksonomije: prenosnik, zvrst, lektorirano
  - cobiss
    - letnica izida

# Postopki korpusne analize

- Lista besed
- Konkordance
  - Korpusni vzorec
  - Variantnost konkordančnega niza
- Statistične vrednosti
- Besedni skupi
- Kombinacija različnih postopkov

# Lista besed

biti
v
in
na
za
da
ta
ki
pa
z
tudi
s
po
kot
še

biti
ta
pa
še
ves
tako
imeti
jaz
lahko
drug
nov
morati
slovenski
prvi
čas

# Različnica s korpusnimi podatki

**slovenski** 174941

slovenski (37626), slovenske (34241), slovenskih (24829), slovenska (21682), slovenskega (18814), slovensko, (18091), slovenskem (11793), slovenskim (3585), slovenskimi (2579), slovenskemu (1604);

v (12181), na (8276), biti (8802), za (4351), s (2993), prvi (1616), dober (1595), tudi (2362);

vojska (2728), ljudski (2233), jezik (2547), narod (2150), gorica (2035), razvojen (1668), vlada (2723), filharmonija (579)

# Lista besed: korpus FIDA in SSKJ

brk
brka
brkač
brkačka
brkast
brkat
brkati
brkec
brkež
brkica
brklja
brkljalnik
brkljanje
brkljarija
brkljariti
brkljarnica
brkljati
brkončica

bianko
biatlon
biatlonec
biatlonka
biatlonski

brokat
brokatast
brokaten
broker
brokerski
brokola
brokoli

# Konkordance

- Konkordančno jedro
  - Obseg
  - Variantnost
- Veriženje jedrne besede

Največ podatkov iz konkordančnih nizov

# Statistika v korpusni analizi

## Besedilno okolje

- Absolutne vrednosti
- Vzajemne vrednosti
  - MI
  - $MI^3$
  - druge: npr. Z in T

# Različnica “kava” in njeno besedilno okolje (okvir +/-3)

| **Absolutne vrednosti** | **MI** | **MI$^3$** |
|---|---|---|
| biti | =kavairska | skodelica |
| in | =javah | čaj |
| v | =fromaže | piti |
| na | =idotskih | in |
| za | =broodje | skuhati |
| skodelica | =poscaline | biti |
| s | =cikorijevo | popiti |
| ali | =rinocaffe | =instant |
| se | =korifanju | pitje |
| čaj | =znojmu | cigareta |
| pa | =okari | srebati |
| ob | =litrus | jutranja |
| z | =poopoldanski | na |

# Lema "kava"

**pomeni s kolokacijami  stalne besedne zveze**

1. pijača iz zrn kavovca
skodelica, termovka, požirek/
vroča, močna/
piti, skuhati, srebati, naročiti/

instant ~, jutranja ~, turška ~, črna ~, bela ~,
ledena ~, ~ s smetano, ~ z mlekom,
~ brez kofeina

2 zrna kavovca
mlinček za, pražarna/

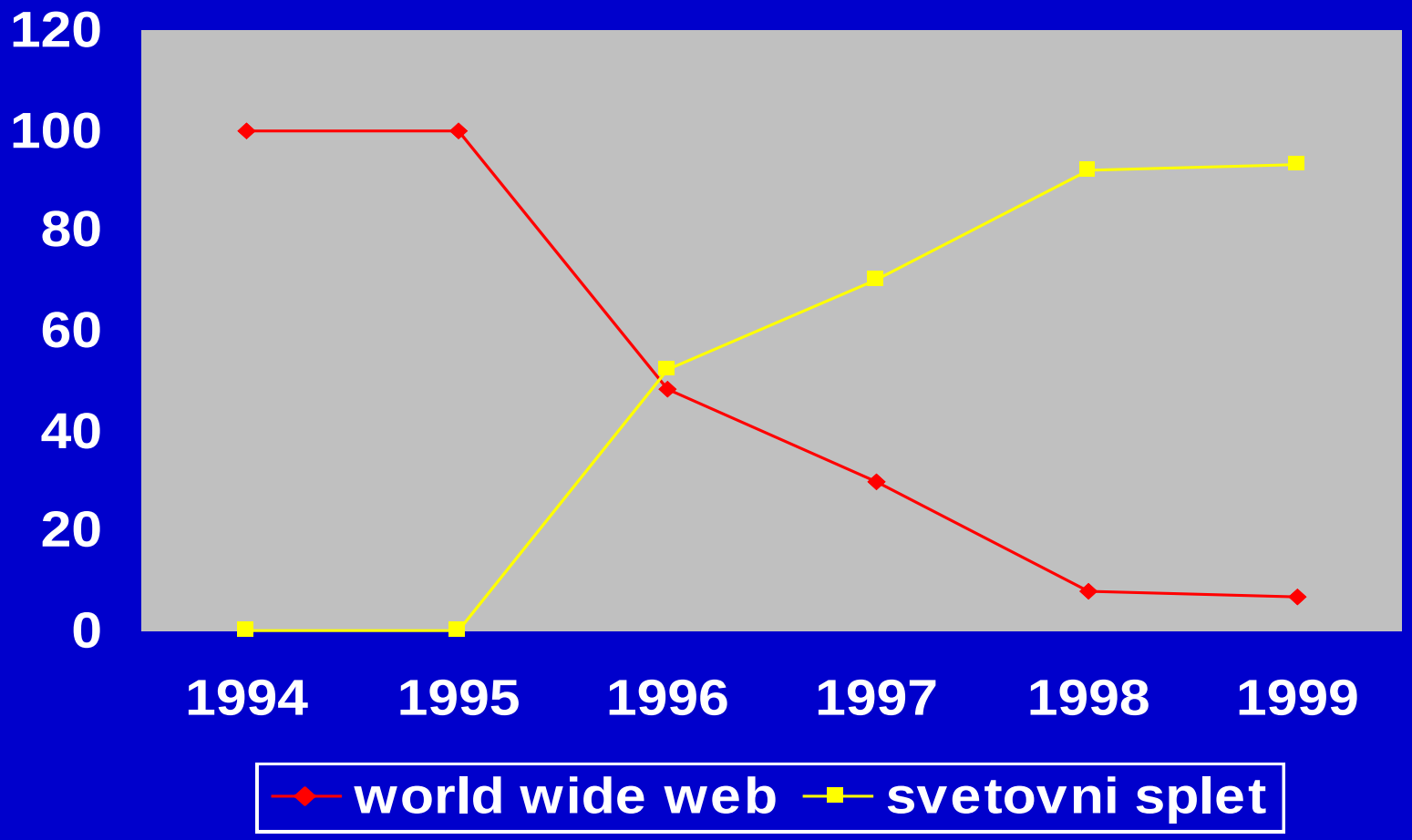
pražena ~, surova ~

3 pijača iz kavnega nadomestka

ječmenova ~, žitna ~

4 družabni dogodek
povabiti, priti/ na
sedeti, klepet/ ob

5 rastlina (kavovec)
plantaža/


120
100
80
60
40
20
0
1994
1995
1996
1997
1998
1999
world wide web
svetovni splet

# *Internet* in besedotovorje

- izpeljani vrstni pridevnik na *-ni* in *-ski*: internetni, internetski
- izpeljani pridevnik s pomenom vrstnosti na *-ov*: internetov
- izpeljani višjestopensjki vrstni pridevnik na *-ski*: internetovski
- izpeljani prislov iz vrstnega pridevnika: internetsko, internetno
- izpeljani samostalnik za poimenovanje nosilca povezave: internetar, internetovec
- zloženi samostalnik: internetdžanki

# Dodatna literatura

Vojko Gorjanc, Uvod v korpusno jezikoslovje. Domžale: Izolit, 2005.

# Navodilo za samostojno delo

- S pomočjo literatura (Gorjanc) preglejte spletne strani tam omenjenih tujejezičnih korpusov in korpusov slovenskega jezika
- Ugotovite, kako je z njihovo dostopnostjo in uporabnostjo za prevajalsko delo.

- Oglejte si spletno stran korpusa FidaPLUS http://www.fidaplus.net, možnost dostopa do korpusa (registracija, pridobitev gesla …) in ugotovite razlike med korpusoma FIDA in FidaPLUS.