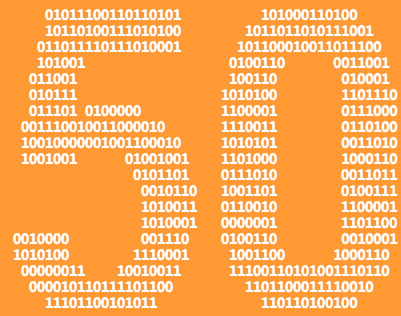
Latvijas Universitātes aģentūra
Latvijas Universitātes Matemātikas un informātikas institūts

Inguna Skadiņa un Andrejs Spektors

Latviešu valoda datorā — pētījumi, resursi, tehnoloģijas



Zinātniskā konference ”Matemātika un informātika pēc 50...”

2009. gada 9. novembris

Pirmsākumi

- Vēlme lietot datoru valodas apstrādei ir tikpat veca kā pirmais dators
- 60. gadu sākumā A.Gobzemis, V.Gorobecs, V.Juriks, T.Jakubaite izstrādā tiešās mašīntulkošanas sistēmu zinātnisko rakstu tulkošanai no krievu valodas latviešu valodā
- Valodas analīze, izmantojot kvantitatīvās metodes, uzsākta 60.gados (T.Jakubaite, S.Kļaviņa, A.Lorencs, Z.Nesaule)

Pirmsākumi

- Plašs kvantitatīvo metožu lietojums pētniecībā vērojams 70.gados (V.Drīzule, T.Jakubaite, S.Kļaviņa, V.Kuzina, N.Mecs, M.Oša, A.Rubīna, N.Sika, E.Soida, R.Zarovska)
- Veikta valodas kvantitatīvā analīze (vārdu krājuma sastāvs dažādu funkcionālo stilu tekstiem, vārdšķiru statistiskais raksturojums u.c.)
- Izstrādāta “Latviešu valodas biežuma vārdnīca” un “Latviešu valodas inversā biežuma vārdnīca”

Mākslīgā intelekta laboratorijas rašanās

- Tautasdziesmu tekstu ievade datorā
- 1988. gadā MII Lietišķās informātikas daļa uzsāk datorizētu latviešu valodas pētniecību:
  - Pētītas latviešu valodas kodēšanas iespējas, 1990. gadā izveidots kodu tabulas standarts (G. Lučkins un L. Vasermans)
  - Seno tekstu pētniecība (M.Baltiņa, E.Milčonoka, A.Ozoliņa)
  - Mūsdienu latviešu valodas pētniecība (A.Spektors, I.Āboliņa, R.Čevere, K.Dancis, I.Greitāne, B.Krauze-Krūze, U.Sarkans)
- 1992. gadā izveidota Mākslīgā intelekta laboratorija, kuru no tās pirmsākumiem līdz mūsdienām vada Dr. fiz. Andrejs Spektors

Galvenie darba virzieni

- **Korpuslingvistika** (M.Baltiņa, A.Spektors, E.Andronova, K.Levāne-Petrova, G.Nešpore, V. Krugļevskis, S. Reinsone)
- **Mašīnlasāmu vārdnīcu izveide** (A. Spektors, E.Andronova, G.Nešpore, N.Grūzīts, I. Ilziņa, V. Rostoks)
- **Programmrīki valodas automatizētai apstrādei** (N. Grūzītis, I. Skadiņa, P. Paikens, M. Virza, V. Krugļevskis, I. Poikāns)
- **Mašīntulkošana** (I.Skadiņa, E. Andronova, E. Brālīts, M. Virza)
- **Semantiskā tīmekļa tehnoloģijas** (G.Bārzdiņš, A. Spektors, G.Nešpore, N.Grūzīts, B. Valkovska, J. Džeriņš)
- **Runas tehnoloģijas** (I.Auziņa, M. Pinnis, R. Berzinskis, G. Rābante)
- **Mācīblīdzekļi** (A.Spektors, E.Andronova, K.Levāne-Petrova, G. Nešpore, I.Auziņa, N. Grūzītis)

Korpuslingvistika

- Tekstu uzkrāšana sākta 80. gadu beigās un 90. gados, kad tekstus ievadīja datorā. Kopš 90. gadu vidus teksti tiek skenēti un pēc tam pārbaudīti, skenēšanas precizitāte ir 98.5%
- Nozīmīgākās tekstu kolekcijas:
  - Latviešu literatūras klasika (3, 5 milj. vārdlietojumu)
  - Latviešu teiku un pasaku, ticējumu un sakāmvārdu datorfonds
  - “Rīgas Balss” latviešu un krievu valodās
- Vairāki paralēlie tekstu korpusi:
  - Sastatīts Orvela ”1984”
  - Platona “Valsts” tulkojums
  - Sadarbībā ar TTC veidots juridisko tekstu paralēlais korpuss

Seno tekstu korpuss

- Seno tekstu ievadīšana uzsākta 80. gadu un turpinās līdz pat mūsdienām:
  - **1989.g.** — M. Baltiņa ierodas LU MII, lai ievadītu Ījaba grāmatu datorā
  - **1992.–1994. g.** — M. Baltiņas vadībā datorā sāka ievadīt visu E.Glika vadībā tulkoto Bībeli

    .....
  - Kopš **2002. g.** Seno tekstu korpuss (http://www.korpuss.lv/senie/)

to layke/ kad Kyrenius Semmeßoge exkan Syrien by. Vnd Jckwens nogaya / ka the ßów mheſlotes lickte / yckwens exkan ßouwe Pille.

Tad nogay arridtczan Joſeph no Galilea / aran tás Pilles Nazareth / exkan to Judde ſemme/ vs to Pille Dauid / kattra tur dh̃eweey Bethlehem / Tapectcz ka tas no to Namme vnd Radde Dauid by/ vs to/ ka tás ßów mheſlote lickte / ar Maria ßouwe ßaloulata Gaſpaſſche / kattra tur apgrutenata by. Vnd kad the tur path by / nátce tas laix ka thay peczimpt by. Vnd tha peczimme ßouwe pirmoye Dhele/ vnd tinne to exkan autims / vnd licke to exkan wene Szille. Aeſto theins nhe by czittur neewena weta exkan to Maiawete.

Vnd tur by Ganne exkan to patte wete / tur prettybe wuerßon to louke/ te gannye py to Nackte ßouwes lopes. Vnd Rouge / tas Engels tha kunge apſpydey tós / vnd the iſſabyas ßów lote / Vnd tas Engels ßatey vs teins / nhe yßabyates yums/

Seno tekstu korpuss

- Pašlaik korpuss ietver 16. gs., 17. gs., kā arī vairākus 18. gs. sākuma tekstus latviešu valodā, kopā ap 1 miljonu vārdlietojumu
- Izstrādāti korpusam nepieciešamie meklēšanas, statistikas un konkordances rīki
- **2002. g**. izveidota darba grupa Seno tekstu korpusa un "Latviešu valodas vēsturiskās vārdnīcas (16.-18.gs.)" izveidei (http://www.tezaurs.lv/lvvv/)

SENIE

latviešu valodas seno tekstu korpuss

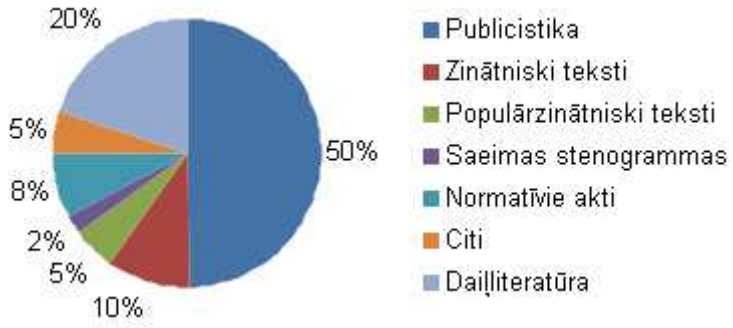
Konkordances rezultāts

Vārdformas šablons: **naud%**
Avots: SENIE
Statistika: vārdformas - 26, vārdlietojumi - 166

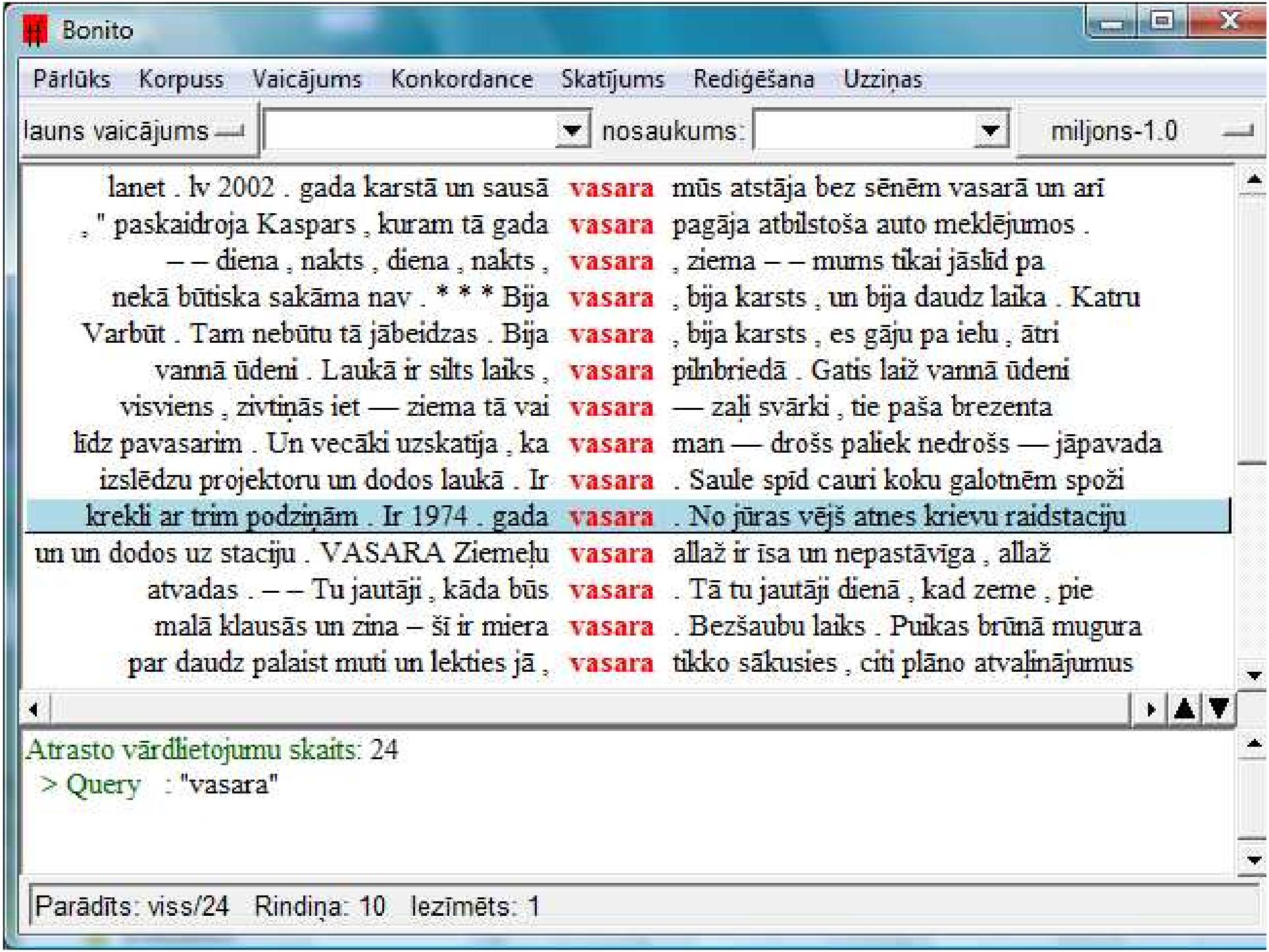
pirrzeeta beß	Naudas	und wellte tick lieds Peenu ka Wienu Nahzeeta
nhe ka dauds	Naudas	par Manntu kraht
tō wairs	Naudas	nhe gir Macka` us Addoſchanu jajämm
ſeeta nahzeeta ſchurr py Vhdeni und juhß kattreem	Naudas	nhe gir nahzeeta ſchurr pirrzeeteeß und ehdeeta
nahzeeta ſchurr pirrzeeteeß beß	Naudas	und wellte
wairahk	Naudas	dabbuit warr Winjam buhß py Chri§ti Eenaidneekeem
Sirrdi us	Naudas	d§iedammeeß Buhß tah Mahzekļam
Mehteli kaß jauns buhdams dauds	Naudas	maxajis bett nu
m Zilwehkam 80 Dalderus tas irr 40 Dalderus labbas	Naudas	tur klaht tam Wihriſchkam ſe§chi rei§ zaur Rihk§t
Zilwehkam 80 Dahlderus tas irr 40 Dahlderus labbas	Naudas	tur klaht tam Wihriſchkam ſe§chi rei§ zaur Rihk§t
pirrzetees Ghuddribu kad juhs beß	Naudas	to warraht dabbuit
mannas	naudas	sinnaiſchi kahd zits
Tam tahs	naudas	suit
ohtra` bahr§ta Bett nu tam Judas wi§ſi	Naudas=Ghabbali	par Wirrwehm ißdohdahß par ko wings pirrmahk pree
ja taß	Naudas=ghabbalings	winjai pa§uhd red§i ka winja
		Schähl gir nhe ka kahdam Baggatam kad

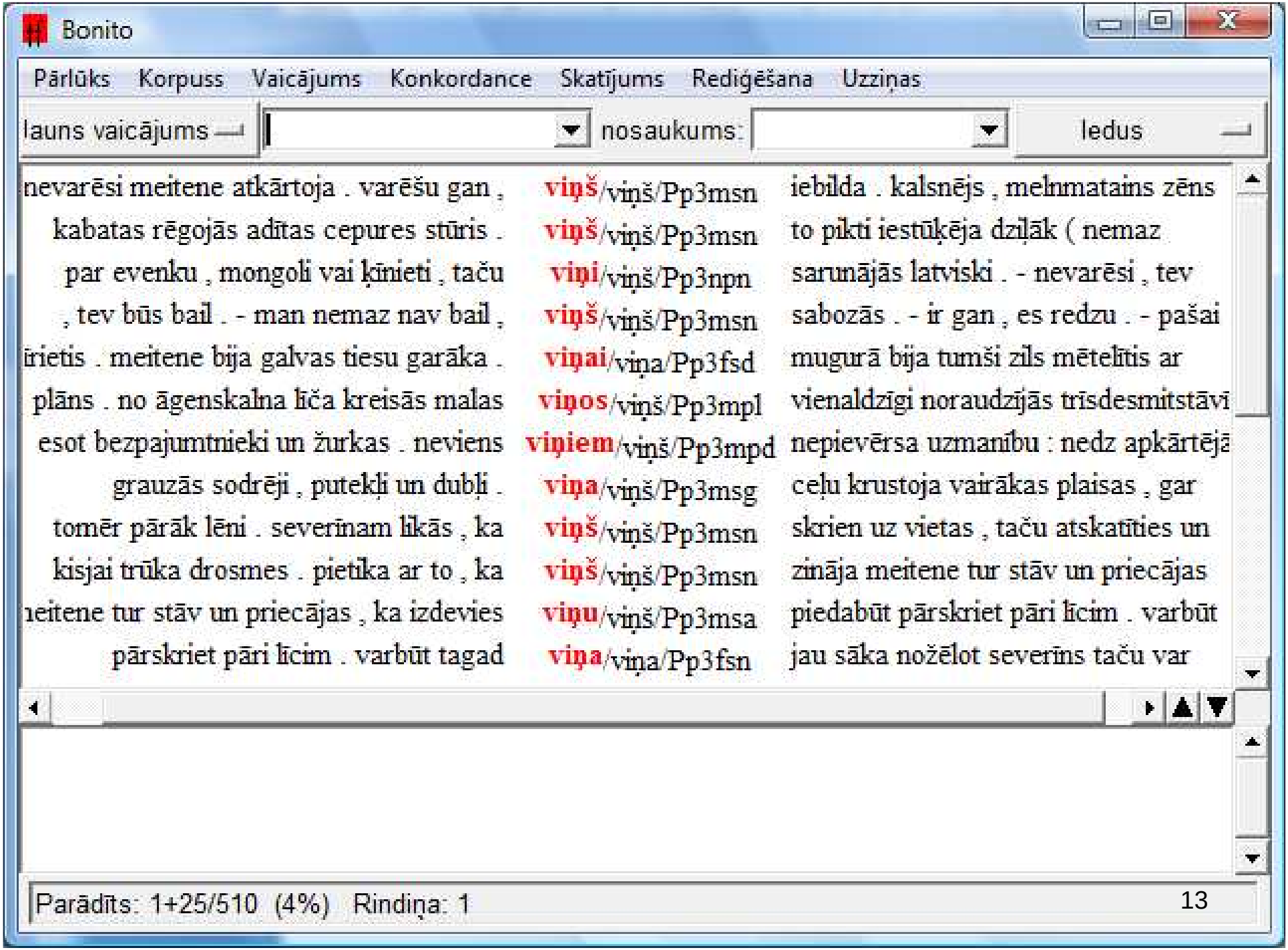
Mūsdienu latviešu valodas korpuss

- Pašlaik uzkrāti teksti ar vairāk nekā 100 milj. vārdlietojumu
- Izveidots līdzsvarots 3, 5 miljonus vārdlietojumu liels mūsdienu latviešu valodas tekstu korpuss (www.korpuss.lv)



- Morfoloģiski marķēts korpuss (~30 tūkst. vārdlietojumu)
- Saeimas stenogrammu korpuss
- Latvijas tīmekļa korpuss


Bonito
Pārlūks Korpuss Vaicājums Konkordance Skatījums Rediģēšana Uzziņas
launs vaicājums
nosaukums:
miljons-1.0
lanet . lv 2002 . gada karstā un sausā vasara mūs atstāja bez sēnēm vasarā un arī
, " paskaidroja Kaspars , kuram tā gada vasara pagāja atbilstoša auto meklējumos .
– – diena , nakts , diena , nakts , vasara , ziema – – mums tikai jāslīd pa
nekā būtiska sakāma nav . * * * Bija vasara , bija karsts , un bija daudz laika . Katru
Varbūt . Tam nebūtu tā jābeidzas . Bija vasara , bija karsts , es gāju pa ielu , ātri
vannā ūdeni . Laukā ir silts laiks , vasara pilnbriedā . Gatis laiž vannā ūdeni
visviens , zivtiņās iet — ziema tā vai vasara — zaļi svārki , tie paša brezenta
līdz pavasarim . Un vecāki uzskatīja , ka vasara man — drošs paliek nedrošs — jāpavada
izslēdzu projektoru un dodos laukā . Ir vasara . Saule spīd cauri koku galotnēm spoži
krekli ar trim podziņām . Ir 1974 . gada vasara . No jūras vējš atnes krievu raidstaciju
un un dodos uz staciju . VASARA Ziemeļu vasara allaž ir īsa un nepastāvīga , allaž
atvadas . – – Tu jautāji , kāda būs vasara . Tā tu jautāji dienā , kad zeme , pie
malā klausās un zina – šī ir miera vasara . Bezšaubu laiks . Puikas brūnā mugura
par daudz palaist muti un lekties jā , vasara tikko sākusies , citi plāno atvaļinājumus
Atrasto vārdlietojumu skaits: 24
> Query : "vasara"
Parādīts: viss/24 Rindiņa: 10 Iezīmēts: 1


Bonito
Pārlūks Korpuss Vaicājums Konkordance Skatījums Rediģēšana Uzziņas
launs vaicājums
nosaukums:
ledus
nevarēsi meitene atkārtoja . varēšu gan , viņš/viņš/Pp3msn iebilda . kalsnējs , melnmatains zēns
kabatas rēgojās adītas cepures stūris . viņš/viņš/Pp3msn to pikti iestūķēja dziļāk (nemaz
par evenku , mongoli vai ķīnieti , taču viņi/viņš/Pp3npn sarunājās latviski . - nevarēsi , tev
, tev būs bail . - man nemaz nav bail , viņš/viņš/Pp3msn sabozās . - ir gan , es redzu . - pašai
īrietis . meitene bija galvas tiesu garāka . viņai/viņa/Pp3fsd mugurā bija tumši zils mētelītis ar
plāns . no āgenskalna līča kreisās malas viņos/viņš/Pp3mpl vienaldzīgi noraudzījās trīsdesmitstāvī
esot bezpajumtnieki un žurkas . neviens viņiem/viņš/Pp3mpd nepievērsa uzmanību : nedz apkārtējā
grauzās sodrēji , putekļi un dubļi . viņa/viņš/Pp3msg ceļu krustoja vairākas plaisas , gar
tomēr pārāk lēni . severīnam likās , ka viņš/viņš/Pp3msn skrien uz vietas , taču atskatīties un
kisjai trūka drosmes . pietika ar to , ka viņš/viņš/Pp3msn zināja meitene tur stāv un priecājas
meitene tur stāv un priecājas , ka izdevies viņu/viņš/Pp3msa piedabūt pārskriet pāri līcim . varbūt
pārskriet pāri līcim . varbūt tagad viņa/viņa/Pp3fsn jau sāka nožēlot severīns taču var
Parādīts: 1+25/510 (4%) Rindiņa: 1
13

Mašīnlasāmu vārdnīcu izveide

- Esošo latviešu valodas vārdnīcu digitalizēšana un konvertēšana mašīnlasāmā formā
- Mērķis – izveidot visaptverošu vārdnīcu datubāzi, kas ietvertu pēc iespējas vairāk latviešu valodas vārdu un to nozīmju aprakstu, ko var izmantot gan galalietotāji, gan automatizēti valodas apstrādes rīki

Internetā publicētās vārdnīcas

- Skaidrojošā vārdnīca (~140 000 šķirkļu, http://www.tezaurs.lv/sv/)
- Latviešu literārās valodas vārdnīca (~64 000 šķirkļu, http://www.tezaurs.lv/llvv/)
- Daudzvalodu terminoloģijas datu bāze TTC vajadzībām (~115 000 šķirkļu)
- Latviešu valodas vārdnīca (~30 000 šķirkļu, http://www.tezaurs.lv/lvv/)
- Mīlenbaha-Endzelīna vārdnīca (~77 000 + ~55 000 šķirkļu, http://www.ailab.lv/mev/)

www.tezaurs.lv/**llvv**, www.tezaurs.lv/**wap**

Programmrīki valodas apstrādei

- Morfoloģiskās analīzes un vārda formu ģenerēšanas programmas
- Automātiskas morfoloģiskās marķēšanas rīki

pirmā <Mosfsny> pirmā
nodaļa <Ncfsn4> nodaļa
Atmosfēra <Ncfsn4> atmosfēra
tu <Pp2nsn> tu
to <Pdnmsa> tas
nevarēsi <Vmnifti32say> nevarēt
meitene <Ncfsn5> meitene
atkārtoja <Vmnipti23san> atkārtot
.

Programmrīki valodas apstrādei

- Morfēmiskās analīzes rīki
- Rīki terminoloģijas izguvei
- Sintaktiskās analīzes modeļi un rīki

Mašīntulkošana

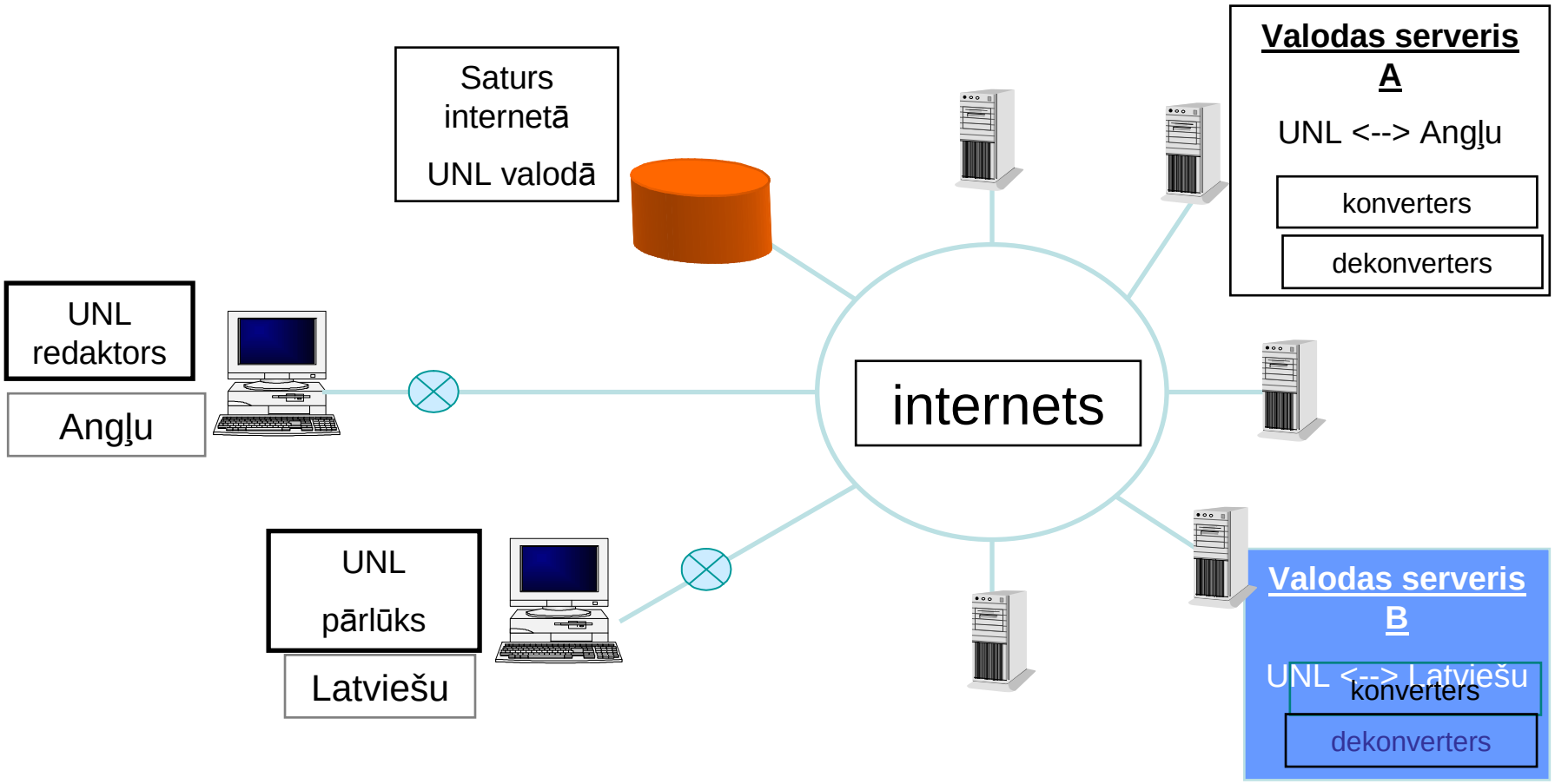
- 1994.g. LU MII Mākslīgā intelekta laboratorijā tiek uzsākta interlingvas tulkošanas sistēmas modeļa LATRA izstrāde

Gada sākumu iezīmēja arī struktūrkapitāla palielināšanas maratons, kuru ir spiesta uzsākt komercbanku lielākā daļa.

The marathon of the increase of statuary capital which the most of the commercials bank must begin characterised the beginning of the year also.

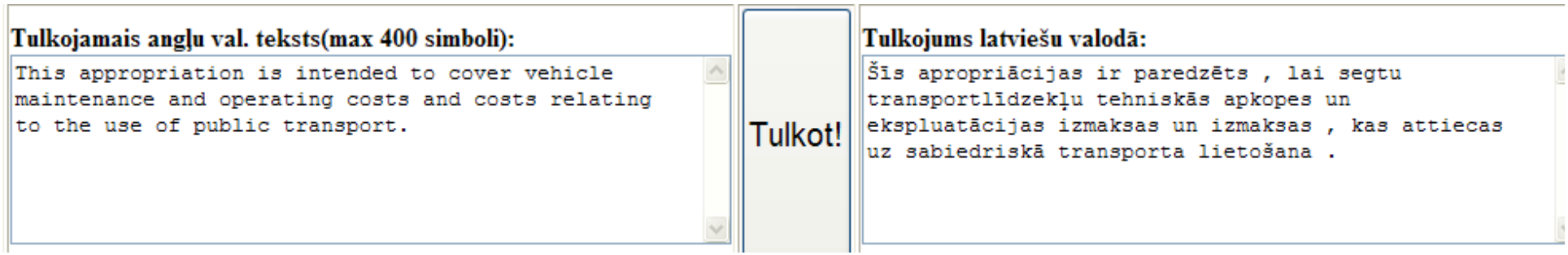
- 1997. g. LU MII Mākslīgā intelekta laboratorija iesaistās Apvienoto Nāciju universitātes projektā *Universal Networking Language*

Universālā tīkla valoda



Statistiskā mašīntulkošana

Pētījumi statistiskajā mašīntulkošanā uzsākti 2005. gadā LZP projekta „Statistisko metožu izvērtējums angļu-latviešu tulkošanas sistēmā” ietvaros un turpinās projektā “Faktorēto metožu lietojums angļu-latviešu statistiskajā mašīntulkošanas sistēmā”



eksperimenti.ailab.lv/smt

Runātās valodas resursi

- "ONOMASTICA-COPERNICUS" (1995-1997) projekta ietvaros transkribēti ~250 000 īpašvārdu
- 2001. gadā uzsākta runas korpusa izveide:
  - 15 runātāju ierunātas frāzes (1 300)
  - 8 stundu garš semināra ieraksts
  - 50 runātāju ierunāts teksts (1000 vārdlietojumu)
  - Teksti ir transkribēti (t.i., pierakstīti mašīnlasāmā formā) un segmentēti, izmantojot *Transcriber* un *WaveSurfer* programmatūru

Runas tehnoloģijas

- Izstrādāta grafēmu-fonēmu pārveides programmatūra un likumi. Automātiskās transkripcijas precizitāte ir 92%
- Izstrādāts eksperimentāls runas sintezators 
- Ir izstrādāts eksperimentāls runas analīzes modulis, kas spēj atpazīt aptuveni 80 vārdus

Latviešu valodas mācīblīdzekļi

http://valoda.ailab.lv/latval/

- Vispārējas ziņas par latviešu valodu
- Latviešu valodas multimediju mācīblīdzeklis iesācējiem
- Latviešu valoda sākumskolai
- Latviešu valoda pamatskolai
- Latviešu valoda vidusskolai
- Latviešu valoda bilingvālai apmācībai
- Latviešu valoda bērniem ar dzirdes traucējumiem
- Latviešu valodas vārdu analizators un sintezators


Vārdu saistāvs
Jaunu vārdu darināšana
Vārdu šķiras
Lietvārds
Īpašības vārds
Vietniekvārds
Apstākļa vārds
Skaitļa vārds
Darbības vārds
Divdabis
Saiklis
Prievārds
Izsauksmes vārds
Partikula
Uzdevumi

CLARIN - Vienota valodas resursu un tehnoloģiju infrastruktūra

- Izveidot integrētu, sadarbību veicinošu pētniecības infrastruktūru, kas ļautu viegli piekļūt un izmantot valodas resursus un tehnoloģijas
- Novērst pašreizējo sadrumstalotību un piedāvāt stabilu, pastāvīgu un paplašināmu infrastruktūru