

I. Vispārīgas lietas

1. Iepazīties ar www.korpuss.lv mājas lapu (Līdzsvarots miljons vārdlietojumu liels mūsdienu latviešu valodas korpuss)

1) **Sadalījums** (latviešu valodas korpusa koncepcija (LU MII, 2005); korpusa uzbūve – iepazīties, kādi funkcionālie stili un kādās proporcijās ir pārstāvēti korpusā. Izvēloties sīkāku statistiku, iegūstam informāciju par katru no stiliem, piem., kādās proporcijās ir pārstāvēta proza, dzeja un dramaturģija; kādi reģionālie laikraksti ir iekļauti korpusā u. tml.

2) **Izmantošana** – korpusu var brīvi izmantot pētniecībā un mācību mērķiem. Šobrīd pieejama korpusa beta versija, tas nozīmē, ka nedaudz korpusa statistiskie rādītāji tuvākā laikā var mainīties (tas jāņem vērā, ja Jūsu studenti jau tagad izmantos korpusu, ik pa brīdim jāpārbauda, vai korpusa apjoms nav mainījies). Korpusu izmanto ar pārlūkprogrammu Bonito. Šajā sadaļā var lejupielādēt arī lietošanas instrukciju ar ekrānskatiem, kā strādāt ar Bonito.

3) **Pielikumi** – bez „miljons” korpusa varam izmantot arī morfoloģiski marķētu paraugkorpusu „ledus”, kuru marķējusi Kristīne Levāne-Petrova (LU MII), un korpusu „timeklis”, kas ir veidots no Latvijas meklētāja savāktām tīmekļa lapām un kas ir (eksperimentāli) morfoloģiski marķēts, bet marķējumu cilvēks nav pārbaudījis.

4) **Citi elektroniskie resursi** – te var iepazīties ar FEB, SENIE, Periodika, kā arī ar internetā pieejamām latviešu valodas vārdnīcām (ME, LLVV u. c.).

2. Lejupielādēt un instalēt pārlūkprogrammu Bonito – programmatūra ir izstrādāta Čehijā, bet to lokalizējis Jānis Džeriņš (LU MII). Lejupielādējiet .zip failu un to atarhivējiet. Kad tas ir izdarīts, nospiediet uz ikonas.



3. Pieslēgties korpusam (lietotājevārds un parole atrodami mājas lapā).

NB Ja programmas lodziņa augšējā malā nav redzami latviešu burti, vajadzēs mainīt Jūsu datora iestatījumus (Start/Settings/Control Panel/Regional and language options/Advanced/Select language.../Latvian/Apply/OK). Vajadzēs pārstartēt datoru. Tas būs jāizdara tikai pirmo reizi.

4. Iepazīties ar informāciju par korpusu – Korpuss/Kopsavilkums vai Ctrl+I. Jautājums: vai korpusā ir miljons vārdlietojumu?

5. Turēt pa rokai un iepazīties ar „Lietošanas instrukciju ar piemēriem un ekrānskatiem”. Instrukciju sagatavojis Normunds Grūzītis (LU MII).

6. Vai ir aptuvena nojausma, ko nozīmē virsraksti ekrāna augšējā malā?

II. Tehniskas lietas

1. Ievadīt jebkādu konkrētu vārdformu (piem., *roka*) un iepazīties ar informāciju, ko iegūstam:

- atrasto vārdlietojumu skaits
- vaicājums
- cik daudz rindiņu redzam vienā lappusē?
- kā pāriet uz nākamo lapu?
- kā mainīt, cik daudz rindiņu skatīt vienā ekrānā? (Skatījums/Apgabals/Cik rindiņas rādīt)
 Tas ir svarīgi, ja atrasto vaicājumu skaits tikai nedaudz pārsniedz definēto rindiņu skaitu (piem., 25/29), lai skatītu visus atrastos rezultātus uzreiz.
- kā aplūkot izvērstu kontekstu ar meklējamo vārdu? (Dubultklišķis uz attiecīgās rindiņas)

2. Iegūto rezultātu rediģēšana

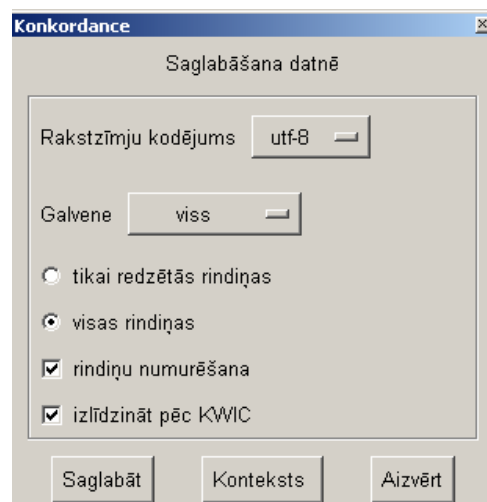
- kā iezīmēt atsevišķas rindiņas? (Peles klikšķis uz rindiņas)
- kā iezīmētās rindiņas izdzēst? (Rediģēšana/ Izmest iezīmētās rindiņas)
- kā, gluži otrādi, paturēt tikai iezīmētās rindiņas? (Rediģēšana/Inverss iezīmējums / Izmest iezīmētās rindiņas)

3. Saglabāto rindiņu saglabāšana

Saglabāt varam: Konkordance/ Saglabāt datnē F2. Rakstzīmju kodējums jāizvēlas utf-8 (ja būs palicis ascii, tad vēlāk latviešu diakritiskās zīmes nebūs redzamas, to vietā būs palikušas jautājuma zīmes). Nosaucot failu, jāpieņem paplašinājums .txt (citādi nevērs automātiski vaļā).

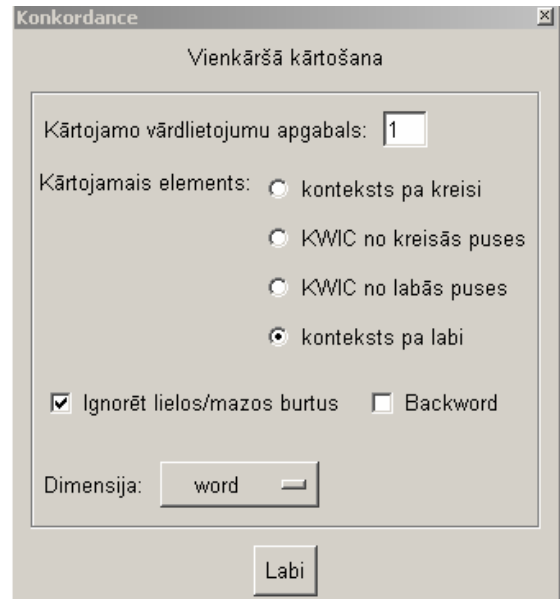
Otrs veids: Iezīmēt rindiņas /(Rediģēšana / Ielikt starpliktuvē) un iekopēt Word vai NotePad un saglabāt kā .txt ar utf-8 kodējumu.

Saglabājot var izvēlēties numurēt visas rindiņas, var saglabāt tikai redzētās rindiņas vai visas.

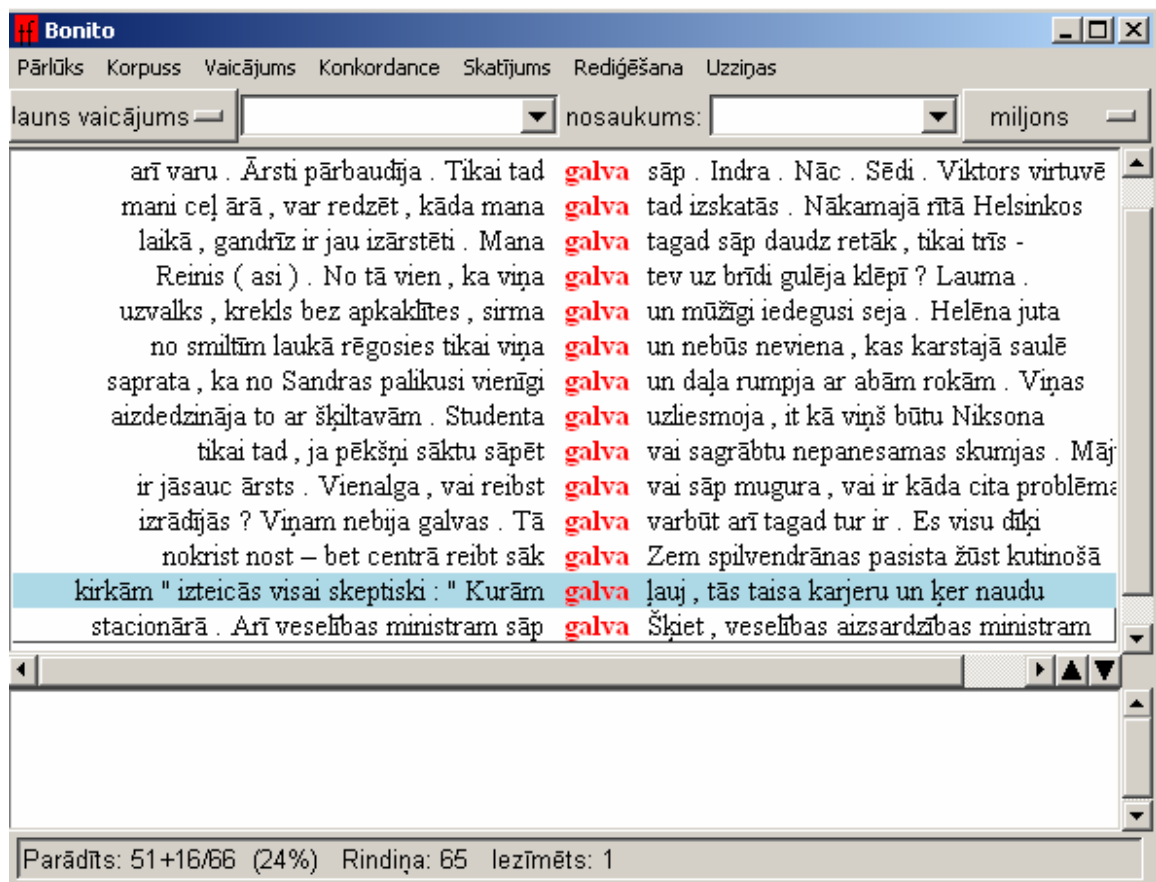


4. Kārtošanas iespējas

Lai kārtotu: Konkordance/Vienkāršā kārtošana. Varam izvēlēties vienkāršoto kārtošanu:



P. S. latviešu valodas alfabētu Bonito vēl nav līdz galam iemācījies, tāpēc, pirmkārt, visas pieturzīmes būs sākumā, otrkārt, vārdi ar diakritiskām zīmēm tiks kārtoti pēc „z”, skatīt pēdējās rindiņas ekrānā, kur kārtoti vaicājuma *galva* rezultāti.



Jautājumi: kādu kārtošanu izvēlēties

- ja skatīsities verba saistījumu?
- ja meklēsiet prievārdu pārvaldījumu?
- ja skatīsities, kādi apzīmētāji var būt formai „galva”?

III. Meklēšana, izmantojot regulārās izteiksmes (skat. izdali angļiski un *Bonito* lietošanas instrukciju)

A. Iespējams, ka korpusā atrodam ne tikai informāciju par latviešu valodu, bet varbūt arī par latviešiem vispār? :)

1. uzdevums. *Laimīgi* un *nelaimīgi*.

Ievadām formu *laimīgi*. Cik daudz gadījumu atrodam?

Jautājums: Kādi ir tipiskie darbības vārdi, kas sastopami savienojumā ar *laimīgi*? (Kārtojam pēc labā konteksta).

Jautājums: Vai *laimīgi* vairāk sastapts ar darbības vārdiem tagadnes, nākotnes, pagātnes laikā? Vai varbūt atstāstījuma vai vēlējumā izteiksmē?

Jautājums: Vai korpusā ir sastopama arī forma *nelaimīgi*? Cik bieži?

Jautājums: Vai varam domāt, ka latvieši ir ...? :)

2. uzdevums. *Rudens* un *pavasaris / pavasara*.

Paskatīsimies, cik bieži korpusā sastapta forma *rudens*. Vai ir kādi tipiski konteksti, kur *rudens* sastopams?

Tālāk mēs gribētu noskaidrot, cik daudz korpusā ir gadījumu ar *pavasaris* un *pavasara*. (Lai būtu godīgi aptvertas gan nominatīva, gan ģenitīva formas, jo *rudens* ir gan Nom., gan Ģen.).

Lai to izdarītu, vajag izmantot regulārās izteiksmes (skat. izdales lapu)

Pats minimums, ar ko sākt:

. jebkurš viens simbols
.+ – no 1 ...līdz bezgalībai
.* – no 0 ..līdz bezgalībai
? – 0 vai 1
} – liekam apgabalu iekšā
[] – izvēlamies elementu
– vai (šīs vērtības liekam apaļajā iekavās)

Jautājums: Kā uzrakstīsim formas *pavasaris* un *pavasara*?

Risinājums: pavasar((is)|a)

Pārbaudām, vai patiešām esam atraduši tikai formas *pavasaris* un *pavasara*.

Jautājums: cik reizes lietots rudens un pavasaris/pavasara?

P.S. Lai arī tas varētu šķist vieglprātīgs un nenopietns uzdevums, ir vairāki pētījumi, piem., par Britu nacionālo korpusu, tieši no ekstralingvistiska skatu punkta.

Padomājiet, kas vēl Jūs varētu interesēt latviešu valodas korpusā?

B. Korpusa veidotāji tic, ka korpuss ataino patiesos valodas datus. Paskatīsimies, kā korpusā fiksētais atbilst latviešu valodas vārdnīcu datiem.

1. uzdevums. *Islams* un *islāms*.

Noskaidrojiet, vai korpusā ir sastopams rakstījums ar īso –a- vai garo –ā-. Aplūkojiet ne tikai nominatīva formas, bet visus locījumus (un arī salikteņus).

Risinājums: isl((a)|ā)m.* (N.B. zvaigznīte beigās rāda, ka vārda beigās var būt bezgalīga burtu (simbolu) virkne).

Jautājums: kādas formas ir skaitliski nozīmīgākas?

Uzdevums: salīdzināt latviešu valodas korpusa rezultātus ar „Latviešu valodas pareizrakstības un pareizrunas vārdnīcu” un citām latviešu valodas vārdnīcām.

Skat., piem., LLVV šķirkli:



The screenshot shows the website 'Latviešu literārās valodas vārdnīca'. The search bar contains 'islams' and the search button is 'Meklēt'. Below the search bar are checkboxes for 'meklēt vārdus, kas sākas ar...' and 'meklēt vārdus, kas ir līdzīgi pēc izrunas'. The search results for 'islams' are displayed below, including the word 'islams' with its grammatical information and a definition: 'Musulmaņu reliģija, Muhameda mācība, kas radās Rietumārābijā 7. gadsimtā. ... iekarotāji arābi ieviesa tadžiku tautā Muhameda ticību - islamu. Zv 57, 13, 16. Sektantisms pavada ikvienu lielāku reliģiju - gan kristiānismu, gan islamu, budismu u. c. Pad J 60, 24, 2.'

<http://www.tezaurs.lv/llvv/>

Un „Mūsdienu latviešu valodas vārdnīcas” šķirklis:

Mūsdienu latviešu valodas vārdnīca
© LU Latviešu valodas institūts
Red. Dr. philol. I. Zuicena
Atbalsta Valsts valodas aģentūra
2003-2008

[Ievads](#), [saīsinājumi](#), [apzīmējumi](#)

Meklēt!

islāms v. *lietv.* Viena no trim visvairāk izplatītākajām pasaules reliģijām, ko nodibinājis pravietis Muhameds 7. gs. Rietumārābijā; musulmanisms; muhamedānisms. *Islāma priekšraksti. Islāma fundamentālisti. Cīlme no arābu islām 'padevība'.*

<http://www.tezaurs.lv/mlvv/>

2. uzdevums. *Balzams* vai *balzāms*?

Noskaidrojiet, vai korpusā sastopamas formas *balzāms* (ne tikai *balzams*). (Izmantojiet līdzīgu meklēšanas modeli kā iepriekš.)

Jautājums: cik liela ir iespēja, ka *Latvijas* un *balzams* ir skaitliski nozīmīgs vārdu savienojums?

Risinājums: Izmantojiet kolokāciju novērtēšanas iespēju: **Korpuss/ Kolokācijas**

- var izvēlēties apkaimes biežumu
- var rakstīt konkrētu vārdformu vai arī
- atzīmēt, ka izmantosit regulāras izteiksmes, un tad meklēt visas vārdformas

Jautājums:
vai ir gadījumi,
kad *Latvijas* un
balzāms
sastopami blakus?

Statistika

Kolokāciju novērtēšana

Dimensija 1: word Vērtība 1: Latvijas

Dimensija 2: word Vērtība 2: balz((a)|ā)m.*

Pieļaujamā distance - no: 1 līdz: 1

Lietotas regulārās izteiksmes

f(Latvijas) = 2773
f(balz((a)|ā)m.*) = 9
f(Latvijas, balz((a)|ā)m.*) = 4
MI = 7.671
T = 1.990

Novērtēt Aizvērt

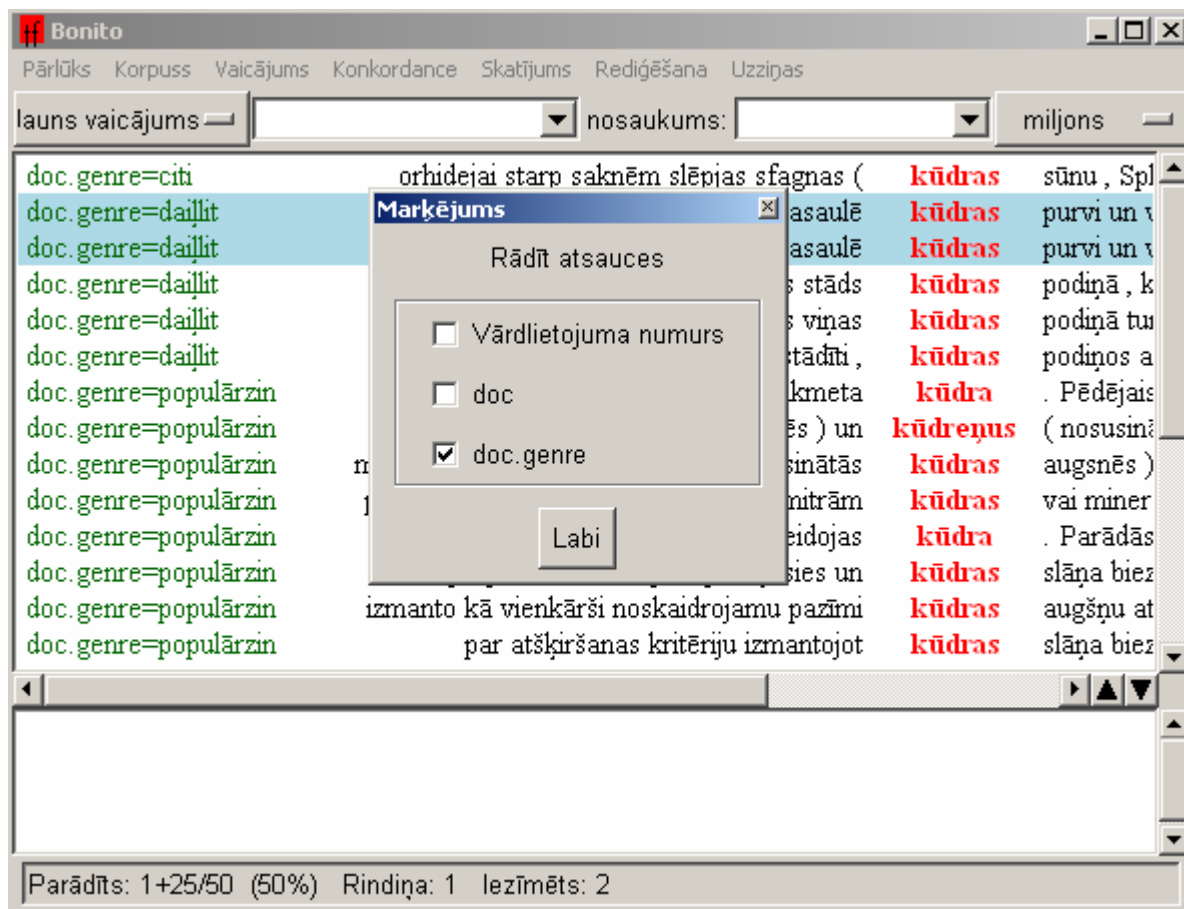
3. uzdevums. *Bij* un *bija*.

Noskaidrojiet, kurā korpusa daļā sastopam formu „bij”. Un salīdziniet, cik daudz ir „bija” lietojumu.

Ieteikums: Pamēģiniet gan konkrētu formu vaicājumus, gan vaicājumu ar regulāro izteiksmi *bij[a]*? – tā tiks parādīti visi *bij* un *bija* gadījumi.

Ieteikums: Lai ātri sagrupētu visus lietojumus, izmantojiet Konkordance / Kārtošana / Vienkāršā kārtošana / KWIC no labās puses. Tādējādi visi *bij* gadījumi būs sākumā.

Risinājums: korpusa platforma ļauj uzzināt, no kurienes konkrētais piemērs ir (daiļliteratūras, zinātniskajiem tekstiem utt.). Lai to izdarītu, izvēlieties **Skatījums/ Atsauces F4 / Marķējums / Rādīt atsauces**, ieklikšķiniet pie **doc.genre**. Kreisajā stūrī redzēsiet informāciju, no kuras korpusa daļas šis piemērs ir., piem.,



Jautājums: kā noņemt informāciju par korpusa daļu? (Izklikšķinot.)

C. Papildu uzdevumi ar korpusu „miljons”

1. uzdevums. Atrodiet piemērus ar *māja* kā darbības vārda formu un saglabājiēt rezultātus failā. Atvērt to ar MSWord.

Atceramies, kā izdzēst liekās rindiņas (te noderēs inversā iezīmēšana), kā saglabāt failu (kāds paplašinājums; vai latviešu burti ir redzami).

2. uzdevums. Cik reižu vārdforma *biju* ir lietota ar vietniekvārdu, cik reižu bez?

3. uzdevums. Cik reižu vārdforma *biju* sastopama nedaiļliteratūras tekstos?

4. uzdevums. Kura stila tekstos nav sastopams vārds *es*?

5. uzdevums. Atlasiet piemērus ar saikļa vārdu *a* ‘bet’.

6. uzdevums. Noskaidrojiet, kādi vārdlietojumi ar sakni *rok-* tiek lietoti 5 un vairāk reižu.

Ieteikums:

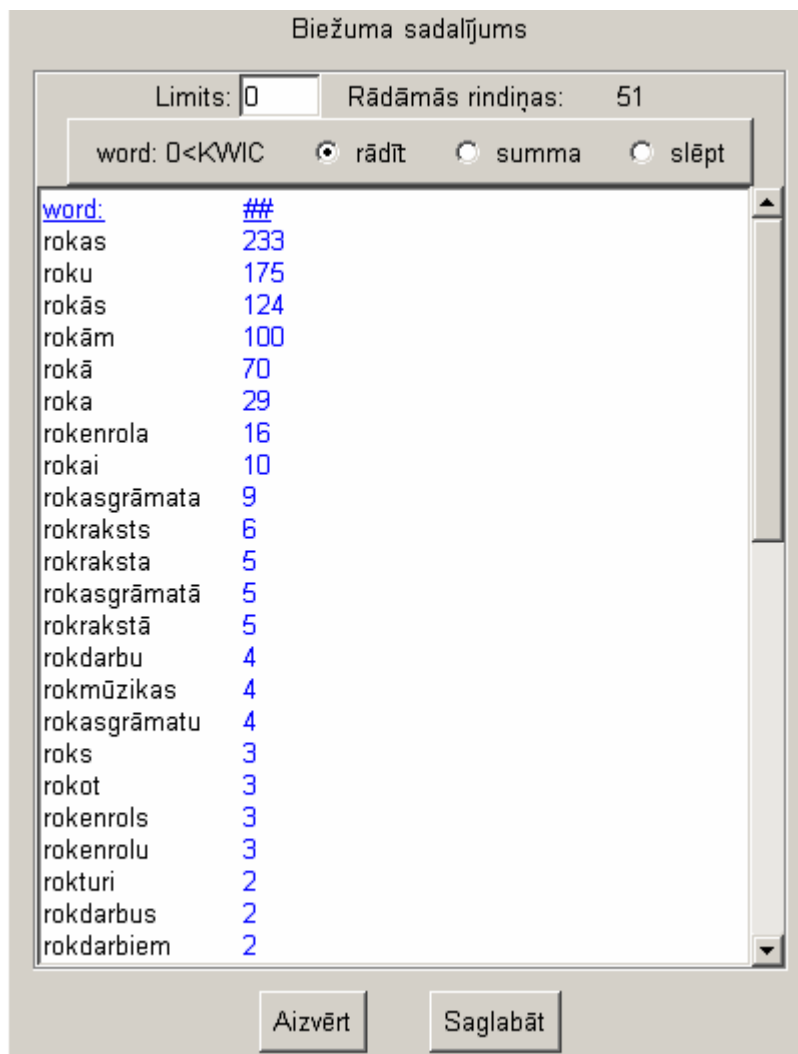
Izmantojiet iespēju

Konkordance/

Statistika/Biežuma

sadalījums (vai

Ctrl+F).



word:	##
rokas	233
roku	175
rokās	124
rokām	100
rokā	70
roka	29
rokenrola	16
rokai	10
rokasgrāmata	9
rokraksts	6
rokraksta	5
rokasgrāmatā	5
rokrakstā	5
rokdarbu	4
rokmūzikas	4
rokasgrāmatu	4
roks	3
rokot	3
rokenrols	3
rokenrolu	3
rokturi	2
rokdarbus	2
rokdarbiem	2

IV. Morfoloģiski marķēta korpusa analīze (izvēlamies korpusu *ledus*)

Morfoloģiski marķētam korpusam ir sava pievienotā vērtība – katrai vārdformai ir pievienota pazīmju kopa, kas līdzīga vienai garai virknei, kur katrai pozīcijai atbilst noteikts raksturojums. Lai iepazītos ar morfoloģisko pazīmju apzīmējumiem, izpētiet izdales materiālus vai lejupielādējiet to (http://www.korpuss.lv/uzzinas/plans_ledus.pdf).

Uzziniet, cik (ne)liels ir šis korpuss.

Tādējādi šis korpuss kalpo tikai kā paraugs tam, ko mēs varētu darīt ar morfoloģiski marķētu korpusu. Un parāda, cik svarīgi ir atrisināt teorētiskus latviešu valodas gramatikas jautājumus, jo mums ir jāpaļaujas uz korpusa marķētāja interpretāciju. Nenoliedzami, morfoloģiski marķēts korpuss ir izmantojams gramatikas pētījumos.

Šajā korpusā iespējams meklēt pēc [lemma="..."] (resp., vārda pamatformas), pēc [tag="V.*"] (resp., morfoloģiskā raksturojuma), pēc [word=".."] (resp., vārdformas vai vārdlietojuma). Varam meklēt pēc viena kritērija, vai tos kombinējot.

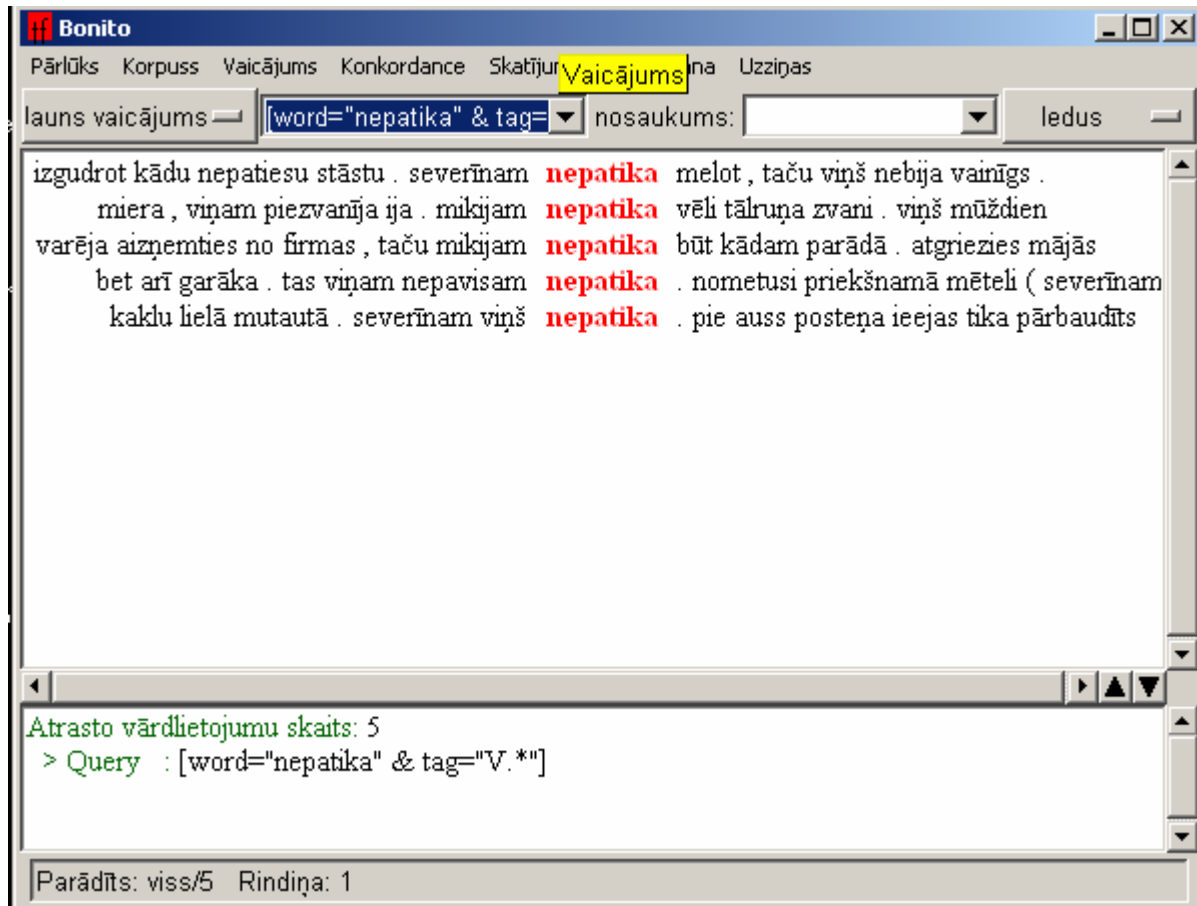
Ieteikums: Ieskatieties Bonito lietošanas instrukcijā un pamēģiniet veikt piemēru vaicājumus.

1. uzdevums. *Nepatika*.

Noskaidrojiet, cik reizes vārdforma *nepatika* lietota kā lietvārds, bet cik – kā darbības vārds?

Ieteikums: Vaicājumu par darbības vārdu varat definēt divējādi:

1. [lemma="nepatikt"];
2. [word="nepatika" & tag="V.*"]



Līdzīgi veiciet vaicājumu par lietvārdu.

2. uzdevums. *Auga*.

Noskaidrojiet, vai vārdforma *auga* ir verba vai lietvārda forma?

3. uzdevums. *Debitīvs ar nominatīvu vai akuzatīvu?*

Noskaidrojiet, kāda lietvārda forma seko tieši pēc debitīva.

Ieteikums: Varat mēģināt definēt sarežģītāku vaicājumu, izmantojot gramatiskās pazīmes. Tātad:

1. `[tag="V..d.*"] [tag="N.{3}n.*"]` – atrodam gadījumus, kad seko nominatīvs;
2. `[tag="V..d.*"] [tag="N.{3}a.*"]` – definējam vaicājumu ar akuzatīvu.

Lai pārlicinātos, cik vispār un kādas debitīva konstrukcijas korpussā ir sastopamas, veiciet vienkāršu vaicājumu `jā.*` (neizmirstiet izdzēst *jā* kā partikulu).

Jautājums: Kāpēc šajā vaicājumā iegūstam vairāk rezultātu?

4. uzdevums. *Runāt...*

Noskaidrojiet, kādas prepozīcijas var sekot pēc verba *runāt*?

Risinājums: [lemma="runāt"][tag="Sp.*"]

5. uzdevums. *Darbības vārds un prievārds*

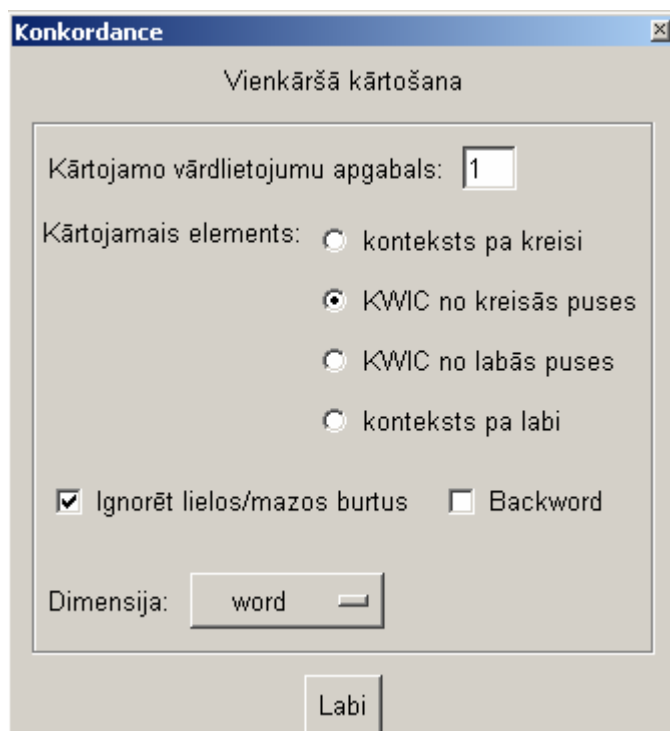
Noskaidrojiet, cik daudz darbības vārda un prepozīcijas savienojumi sastopami korpusā?

6. uzdevums. *Darbības vārdi ar prievārdu ‘ar’*

Noskaidrojiet, kādiem darbības vārdiem seko prievārds ‘ar’?

Kuri darbības vārdi lietoti ar prievārdu ‘ar’ vairāk nekā 2 reizes?

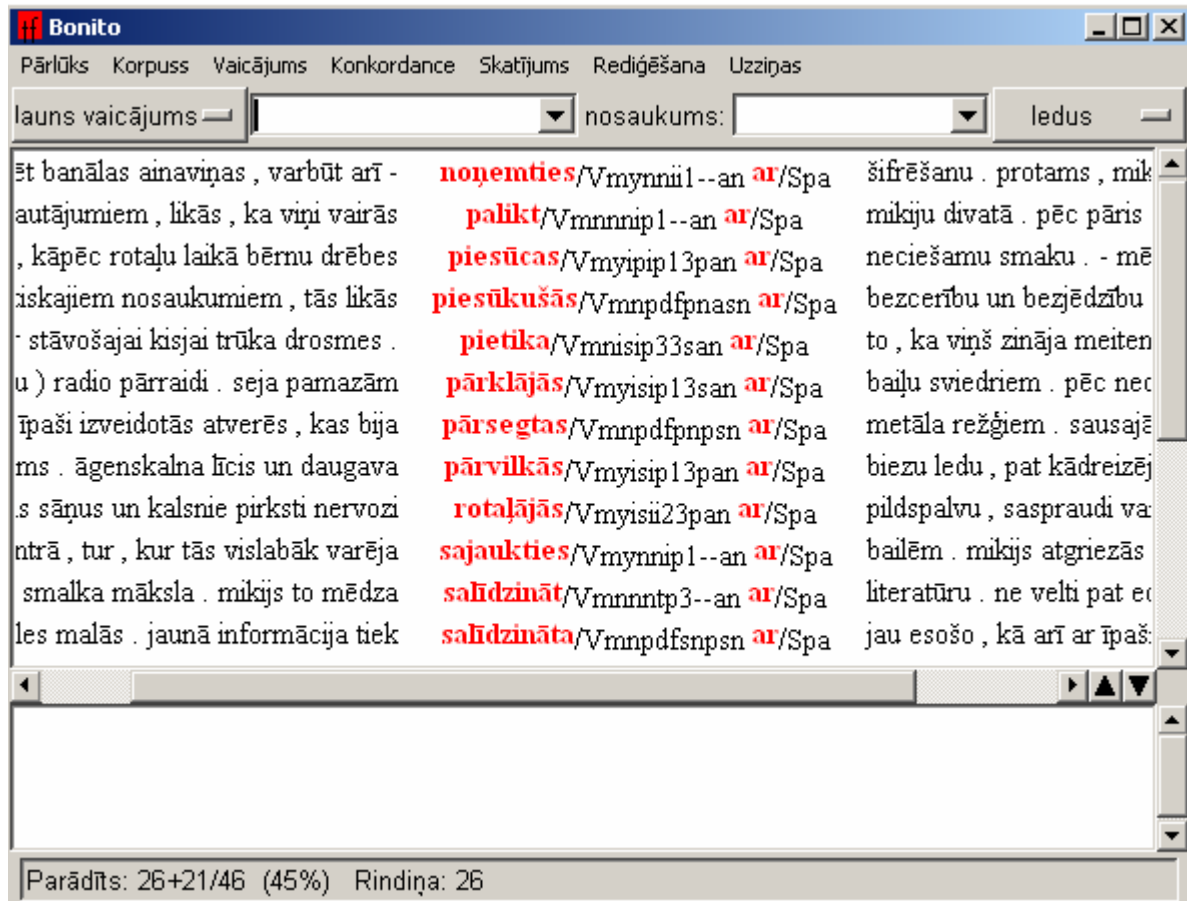
Ieteikums: Iegūtos rezultātos sakārtojiet, izmantojot Konkordance/
Vienkāršā kārtošana / KWIC no kreisās puses:



Papildu jautājums: Kuros gadījumos Jūs izmantotu KWIC kārtošanu no labās puses?

NB: Ja „miljons” korpusā mēs varējām uzzināt, no kuras korpusa daļas ir attiecīgais piemērs, tad „ledū” mēs varam uzzināt, kā īsti vārdforma ir nomarkēta. Lai to uzzinātu, izvēlamies **Skatījums /**

Rādīt dimensijas un ieklikšķinām pie tag (labāk tikai KWIC, bet nepieciešamības gadījumā Jūs varat pārlūkot visu marķējumu):



Ieteikums: Izpētīt Bonito lietošanas instrukciju un izmēģināt, kā varam veidot sintaktiskas konstrukcijas vaicājumu.

V. Eksperimenti ar tīmekļa korpusu – liels, nesakārtots, nepārbaudīts (automātiski marķēts, bet nepārbaudīts) – noder, lai redzētu tendences, bet pagaidām Neuzticams

Iepazīstamies ar korpusu „tīmeklis” – cik liels tas ir?

1. uzdevums. Jūs varat veikt jau iepriekšējos vaicājumus, lai salīdzinātu, kādas tendences var novērot valodā, piem., kas tiek lietots vairāk: *Islande* vai *Īslande*, kurš no rakstības variantiem ir dominējošais tīmeklī: sakne *slovak-* vai *slovāk-*?

2. uzdevums. Papētiet, kādos kontekstos (gramatiskās konstrukcijās) sastopami vārdi *grasās*, *gatavojas*, *domā*. Vai ir pamats uztraukties, ka palielinās *grasās* lietojumu skaist?

3. uzdevums. Papētiet vārdu savienojumus (izmantojiet iespēju Konkordance / Statistika / Biežāk sastopamās kolokācijas).

4. uzdevums. Paskatieties, kādas ir tendences ar vajadzības izteiksmi:
Ieteikums: izmantojiet šablonu, lai atrastu gan gan lietvārdus, gan vietniekvārdus un vietniekvārdus pēc debitīva:
[tag="v..d.*"] [tag="[na]...a.*" | tag="p.{4}a."]

Protams, ka tagad ir jāveltī laiks visa atlasei, lieko piemēru dzēšanai...

Ieteikums: Varbūt ir laiks vienkārši saglabāt failu kādai citai dienai?

Lai sokas!