Language Resources and Technology for the Humanities in Latvia 2004–2010

Inguna Skadiņa, Ilze Auziņa, Normunds Grūzītis, Kristīne Levāne-Petrova,

Gunta Nešpore, Raivis Skadiņš, Andrejs Vasiļjevs

Institute of Mathematics and Computer Science
University of Latvia

TILDE

Background

- Language technologies in Latvia have a rather long history starting at the end of the 50s
- Overview of HLT in Latvia from 1988 till 2004 has been presented at two previous Baltic language technology events:
  - “Language and Technology in Europe 2000” in 1994
  - First Baltic conference on Human Language Technologies in 2004

State Language Policy

- The State language policy is defined in two major documents: *“Guidelines of the State Language Policy for 2005-2014”* and *“The State Language Policy Programme for 2006-2010”*
- Tasks related to language technology:
  - provide financial and administrative **support** to **research in computational linguistics** for the Latvian language;
  - organize and **create** a modern computer-aided Latvian language database and ensure its wide usage; the result of this task should be **corpora of the Latvian written and spoken language, tools for corpora management and lexicography**, standards and schemas for lexical and other data;
  - ensure **development of Latvian terminology**, creation of terminological databases and dictionaries, terminology harmonization and international cooperation in terminology development;
  - ensure **education in computational linguistics** in Latvian universities

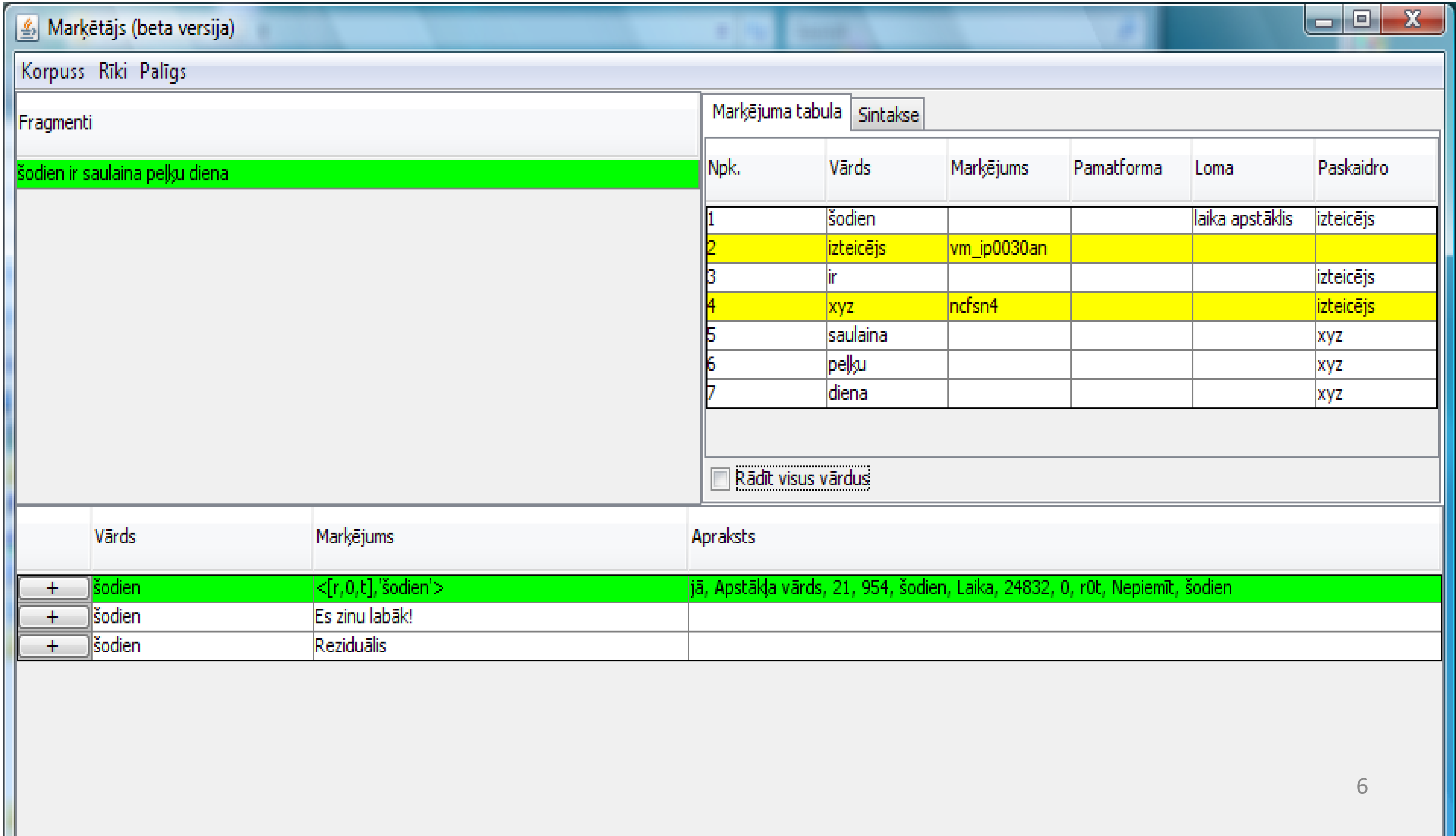
Latvian Council of Science and State Research Programmes

- Latvian Council of Science (LCS) is responsible for the advancement, evaluation, financing, and coordination of research in Latvia
- Significant funding from the LCS has been received between 2005-2009
- Two HLT related projects were authorized as components of the State Research Programmes:
  - *“Scientific Foundations of Information Technology”*
  - *“Latvian Studies (Letonica): Culture, Language and History”*
- Each year 2-3 smaller projects related to HLT have funded by the Latvian Council of Science

SemTi-Kamols project

- Semti-kamols project (www.semti-kamols.lv) aimed at development and adaptation of the semantic web technologies for semantic analysis of the Latvian language
- Concept and methodology of „Semantic Latvia” is implemented in domain of medicine statistics: graphical conceptual ontologies for medicine domain serves as maps allowing doctors to formulate queries for ontological data bases
- Novel technique for „text-to-scene” conversion which in future will allow to convert text into schematic 3D animation
- Semi-automatic tool for morpho-syntacitc annotation

Semi-automatic tool for morpho-syntactic annotation



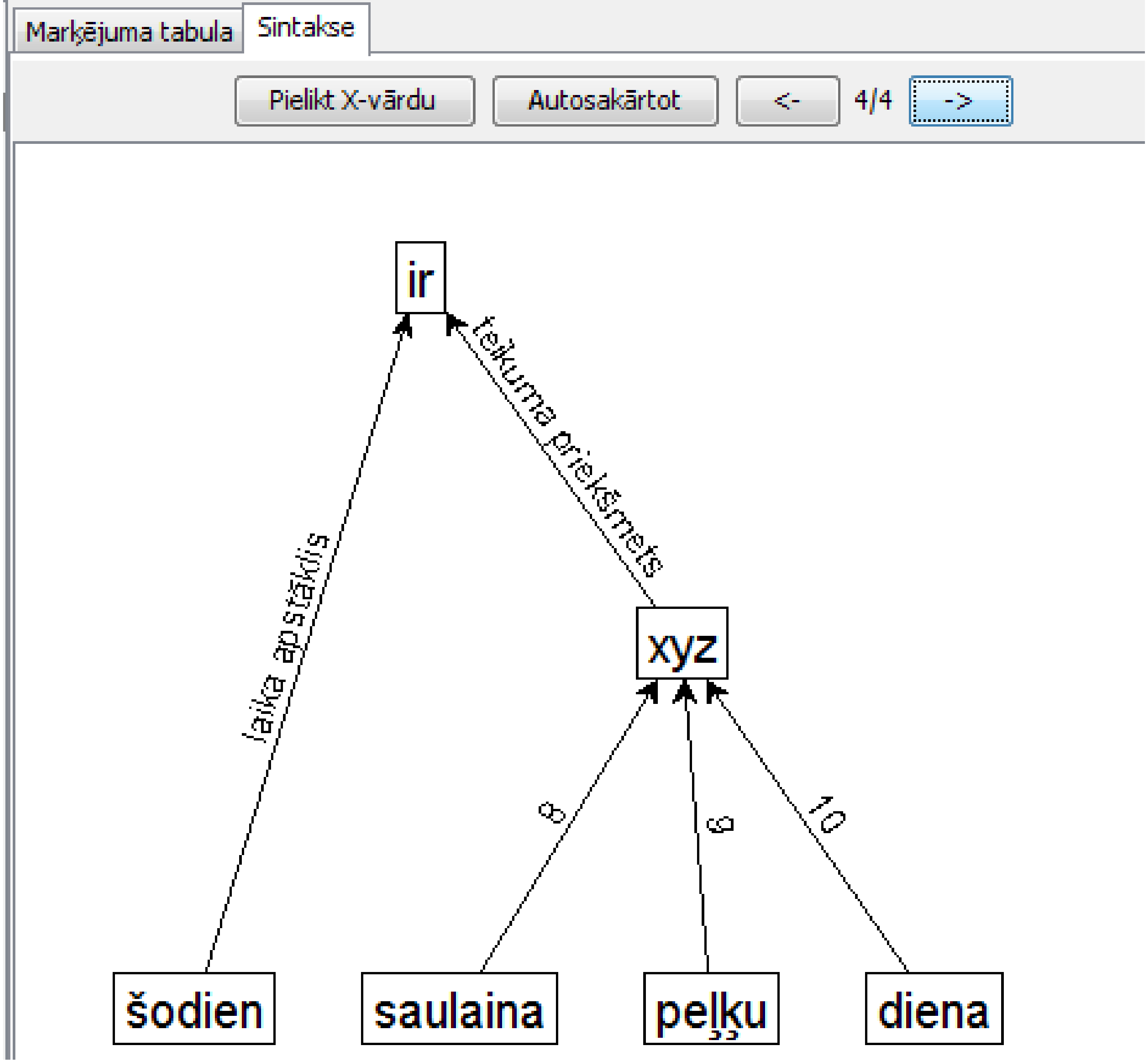
Fragmenti

šodien ir saulaina peļķu diena

Npk.	Vārds	Marķējums	Pamatforma	Loma	Paskaidro
1	šodien			laika apstāklis	izteicējs
2	izteicējs	vm_ip0030an			
3	ir				izteicējs
4	xyz	ncfsn4			izteicējs
5	saulaina				xyz
6	peļķu				xyz
7	diena				xyz

□ Rādīt visus vārdus

	Vārds	Marķējums	Apraksts
+	šodien	<[r,0,t],'šodien'>	jā, Apstākļa vārds, 21, 954, šodien, Laika, 24832, 0, r0t, Nepiemīt, šodien
+	šodien	Es zinu labāk!	
+	šodien	Reziduālis	


Marķējuma tabula
Sintakse
Pielikt X-vārdu
Autosakārtot
<-
4/4
->
ir
laika apstāklis
teikuma priekšmets
xyz
8
9
10
šodien
saulaina
pelķu
diena

Database of Latvian Explanatory Dictionaries and Recent Loanwords

The project "*Database of Latvian Explanatory Dictionaries and Recent Loanwords*" was mainly dealing with

- digitalization of dictionaries
- semi-automatic transformation of the dictionaries into a machine-readable format

Main Resources and Tools

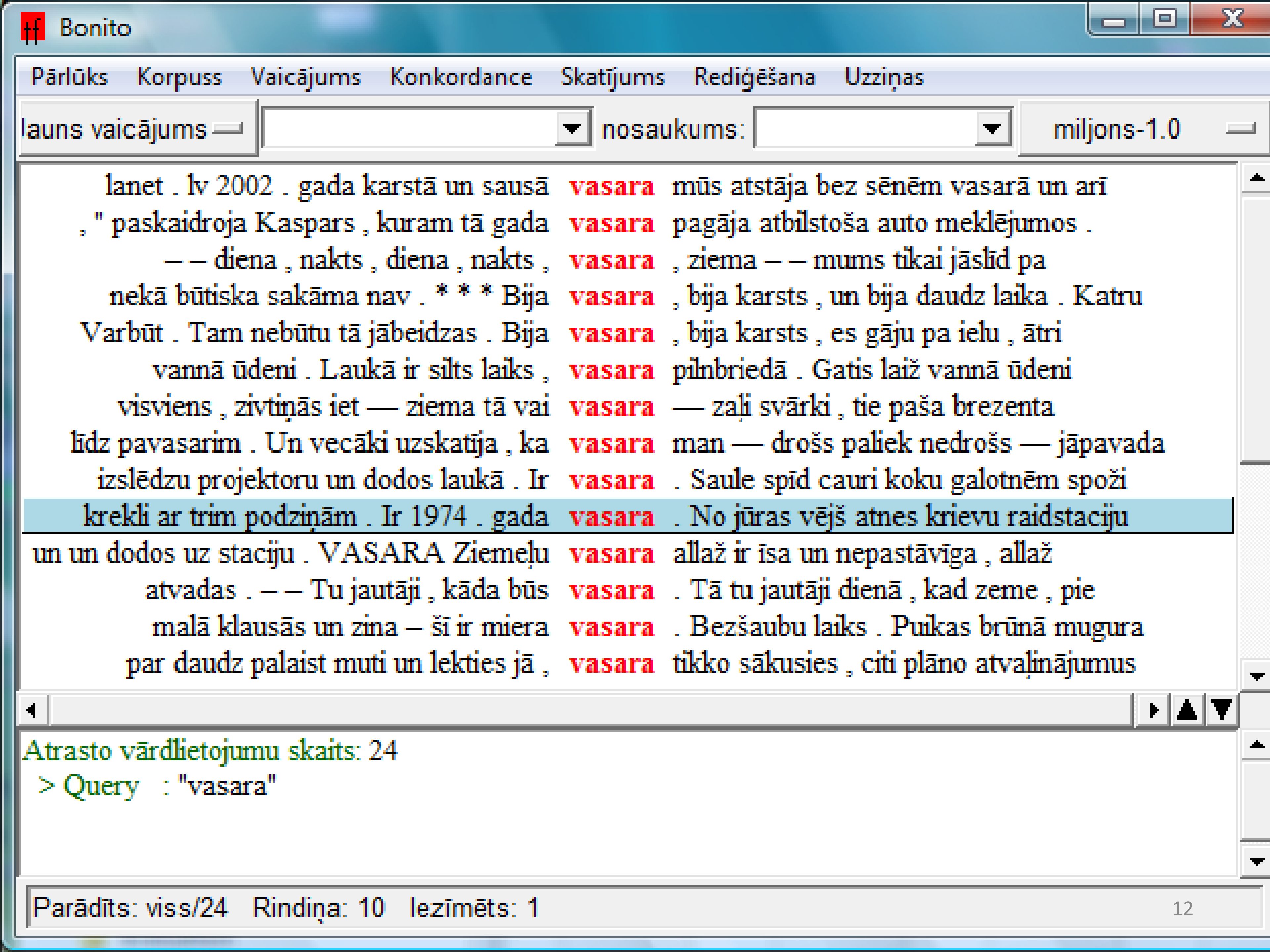
- Latvian Language Corpora Resources
- Electronic Dictionaries and Terminology Resources
- Machine Translation Tools and Prototypes
- Speech Technologies
- Tools for Natural Language Processing

Latvian National Corpus Initiative

- The development of the Latvian National Corpus was initiated by the State Language Commission in 2004
- Latvian National Corpus Initiative envisions establishing an umbrella for all the available corpora of the Latvian language
- The Agreement of Intention between the main language resource developers and holders in Latvia, both academic and industry, has been signed

Latvian Language Corpora Resources

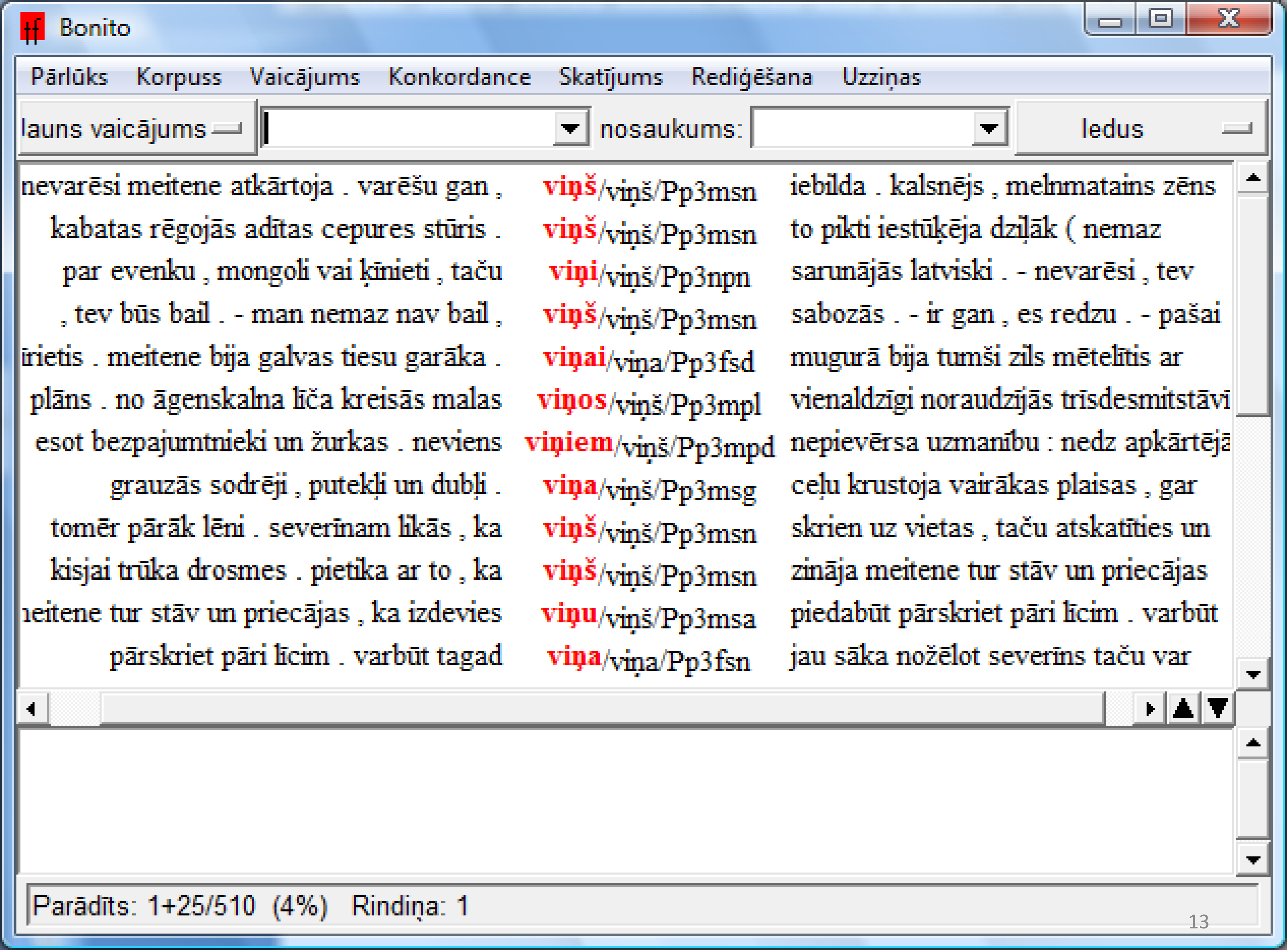
- Since 2006, The National Library of Latvia has been working on the creation of the *Latvian National Digital Library "Letonica":*
  - Digital Library holds collections of newspapers, pictures, maps, books, sheet music and audio recordings
  - Collection *Periodicals* (www.periodika.lv) offers 41 newspaper and magazines in Latvian, German, and Russian from 1895 to 1957
- Three corpora have been developed at Institute of Mathematics and Computer Science (IMCS) (www.korpuss.lv)
  - *Balanced Corpus of Modern Latvian* (~3.5 million running words)
  - *Web Corpus* (~100 million running words)
  - *Corpus of the Transcripts of the Saeima's (Parliament of Latvia) Sessions* (more than 20 million running words)
- Pilot morpho-syntactically annotated corpus has been developed at IMCS, it covers approximately 30 000 words of modern written Latvian manually annotated



```
          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
```



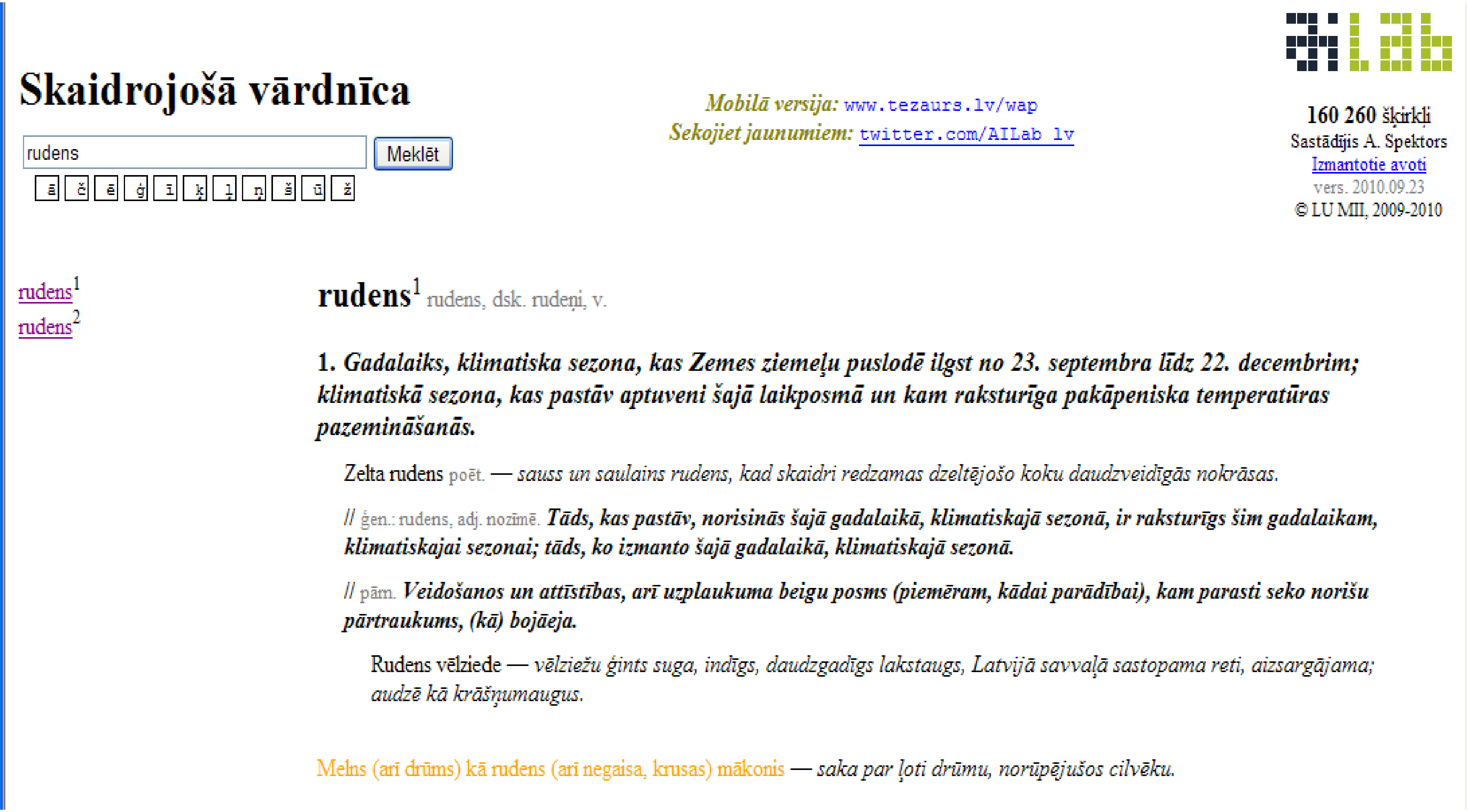
```
   nevarēsi meitene atkārtoja . varēšu gan ,   viņš/viņš/Pp3msn     iebilda . kalsnējs , melnmatains zēns
    kabatas rēgojās adītas cepures stūris .    viņš/viņš/Pp3msn     to pikti iestūķēja dziļāk ( nemaz
     par evenku , mongoli vai ķīnieti , taču   viņi/viņš/Pp3npn     sarunājās latviski . - 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 trīsdesmitstāvi
     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
ieitene 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

Electronic Dictionaries

Several machine-readable versions of monolingual dictionaries of modern Latvian have been created by IMCS in cooperation with other research institutions (www.tezaurs.lv):

- The *Dictionary of Standard Latvian Language* - largest Latvian monolingual dictionary of the second half of the 20th century (~64 000 entries in 8 volumes)
- *The Explanatory Dictionary* (more than 150 000 entries from about 120 Latvian dictionaries of different times and domains)
- New *Dictionary of the Modern Latvian* (~20 000 entries from A–Ļ)



Skaidrojošā vārdnīca

Mobilā versija: www.tezaurs.lv/wap
Sekojiet jaunumiem: twitter.com/AILab_lv

160 260 šķirkļi
Sastādījis A. Spektors
vers. 2010.09.23
© LU MII, 2009-2010

rudens[1]
rudens[2]

rudens[1] rudens, dsk. rudeņi, v.

1. ***Gadalaiks, klimatiska sezona, kas Zemes ziemeļu puslodē ilgst no 23. septembra līdz 22. decembrim; klimatiskā sezona, kas pastāv aptuveni šajā laikposmā un kam raksturīga pakāpeniska temperatūras pazemināšanās.***

Zelta rudens poēt. — *sauss un saulains rudens, kad skaidri redzamas dzeltējošo koku daudzveidīgās nokrāsas.*

// ģen.: rudens, adj. nozīmē. ***Tāds, kas pastāv, norisinās šajā gadalaikā, klimatiskajā sezonā, ir raksturīgs šim gadalaikam, klimatiskajai sezonai; tāds, ko izmanto šajā gadalaikā, klimatiskajā sezonā.***

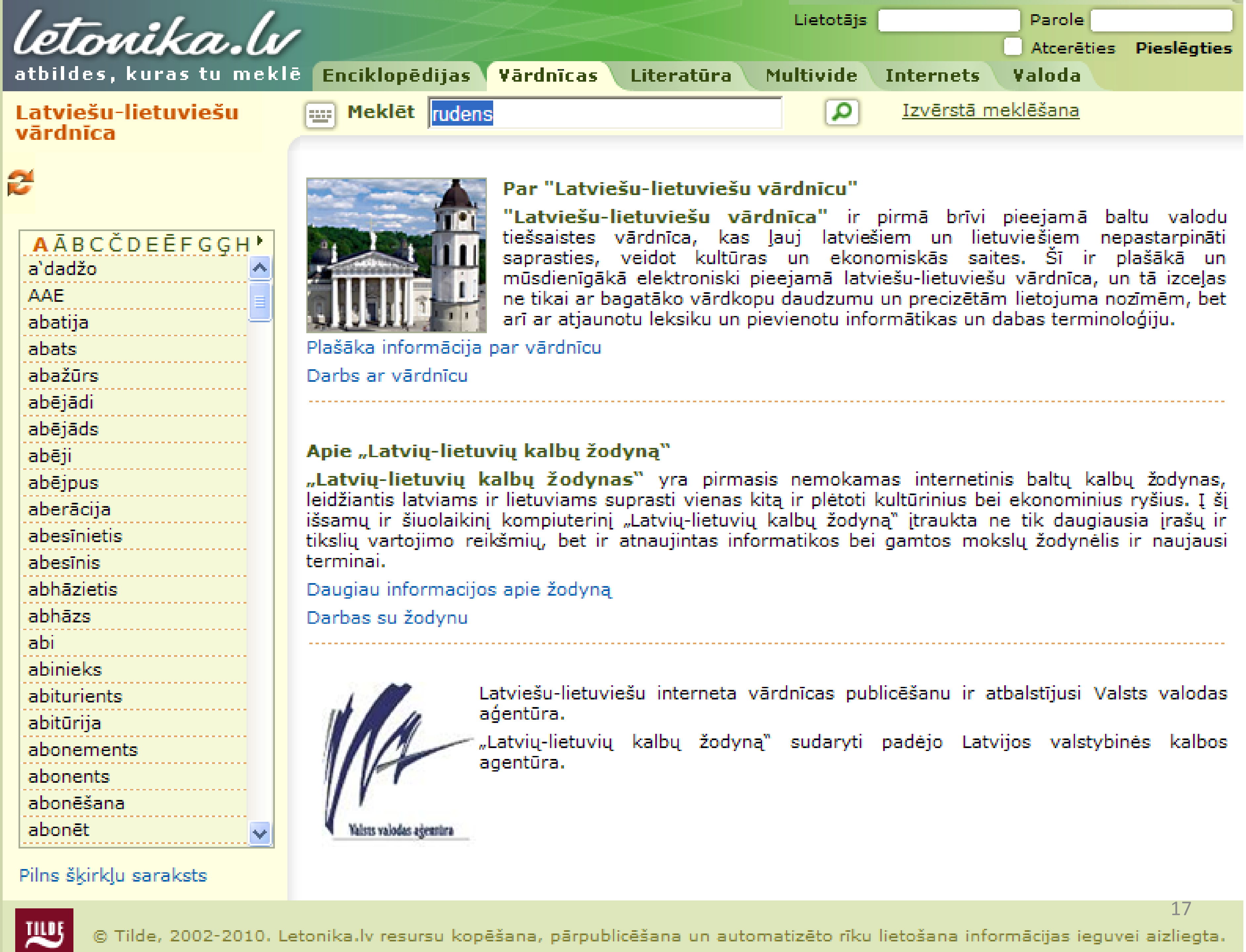
// pām. ***Veidošanos un attīstības, arī uzplaukuma beigu posms (piemēram, kādai parādībai), kam parasti seko norišu pārtraukums, (kā) bojāeja.***

Rudens vēlziede — *vēlziežu ģints suga, indīgs, daudzgadīgs lakstaugs, Latvijā savvaļā sastopama reti, aizsargājama; audzē kā krāšņumaugus.*

Melns (arī drūms) kā rudens (arī negaisa, krusas) mākonis — *saka par ļoti drūmu, norūpējušos cilvēku.*

Electronic Dictionaries

- Tilde’s electronic dictionaries include 20 translation routes: from English, French, German and Russian into Latvian and Lithuanian and vice versa as well as Latvian-Lithuanian, Lithuanian-Latvian and Estonian-Latvian
- Included in online internet resource in reference portal www.letonika.lv



Latviešu-lietuviešu vārdnīca

Par "Latviešu-lietuviešu vārdnīcu"

"Latviešu-lietuviešu vārdnīca" ir pirmā brīvi pieejamā baltu valodu tiešsaistes vārdnīca, kas ļauj latviešiem un lietuviešiem nepastarpināti saprasties, veidot kultūras un ekonomiskās saites. Šī ir plašākā un mūsdienīgākā elektroniski pieejamā latviešu-lietuviešu vārdnīca, un tā izceļas ne tikai ar bagatāko vārdkopu daudzumu un precizētām lietojuma nozīmēm, bet arī ar atjaunotu leksiku un pievienotu informātikas un dabas terminoloģiju.

Plašāka informācija par vārdnīcu

Darbs ar vārdnīcu

Apie „Latvių-lietuvių kalbų žodyną"

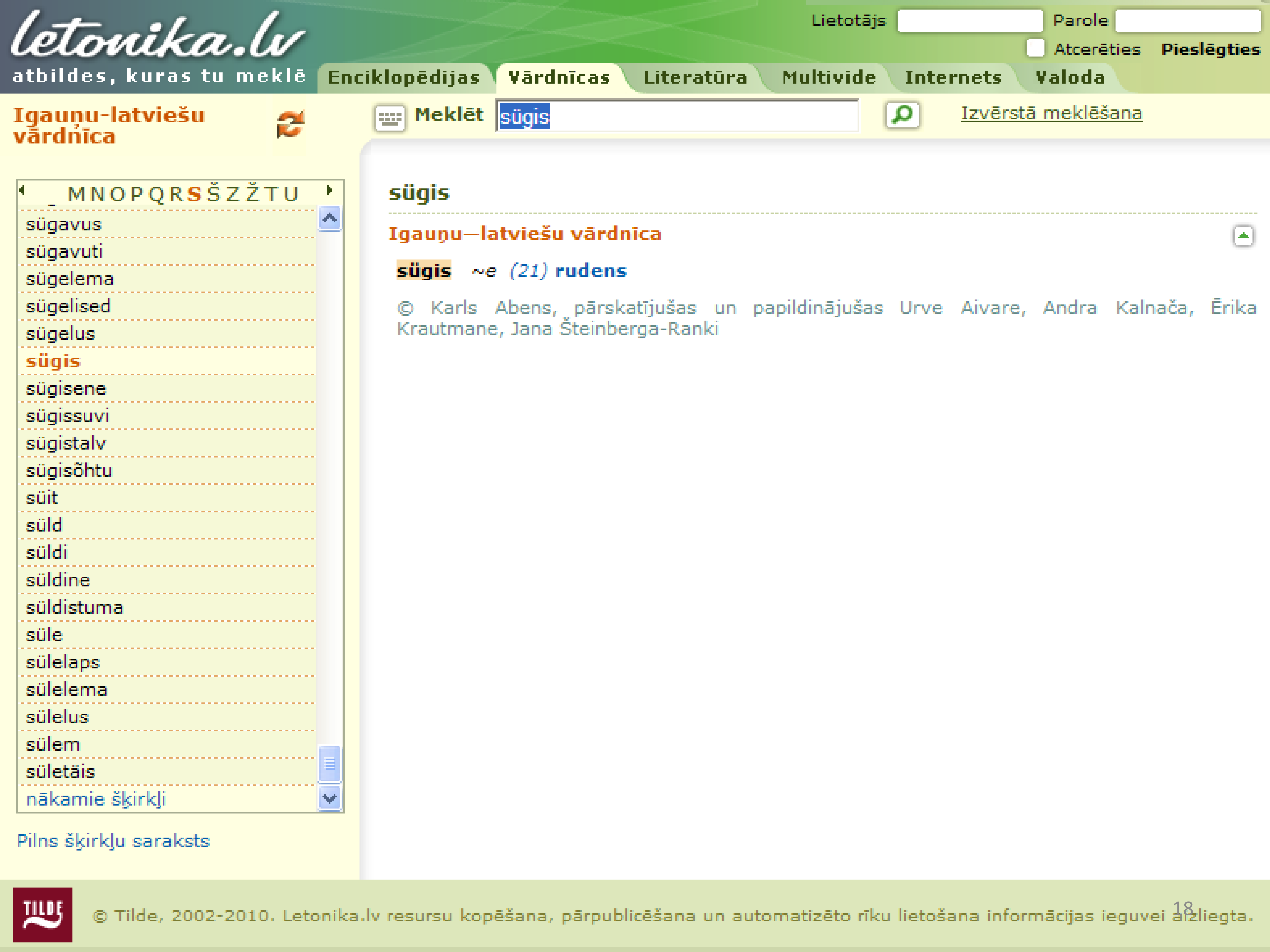
„Latvių-lietuvių kalbų žodynas" yra pirmasis nemokamas internetinis baltų kalbų žodynas, leidžiantis latviams ir lietuviams suprasti vienas kitą ir plėtoti kultūrinius bei ekonominius ryšius. Į šį išsamų ir šiuolaikinį kompiuterinį „Latvių-lietuvių kalbų žodyną" įtraukta ne tik daugiausia įrašų ir tikslių vartojimo reikšmių, bet ir atnaujintas informatikos bei gamtos mokslų žodynėlis ir naujausi terminai.

Daugiau informacijos apie žodyną

Darbas su žodynu

Latviešu-lietuviešu interneta vārdnīcas publicēšanu ir atbalstījusi Valsts valodas aģentūra.

„Latvių-lietuvių kalbų žodyną" sudaryti padėjo Latvijos valstybinės kalbos agentūra.



Igauņu-latviešu vārdnīca

Meklēt: sügis

sügavus
sügavuti
sügelema
sügelised
sügelus
sügis
sügisene
sügissuvi
sügistalv
sügisõhtu
süit
süld
süldi
süldine
süldistuma
süle
sülelaps
sülelema
sülelus
sülem
sületäis

sügis

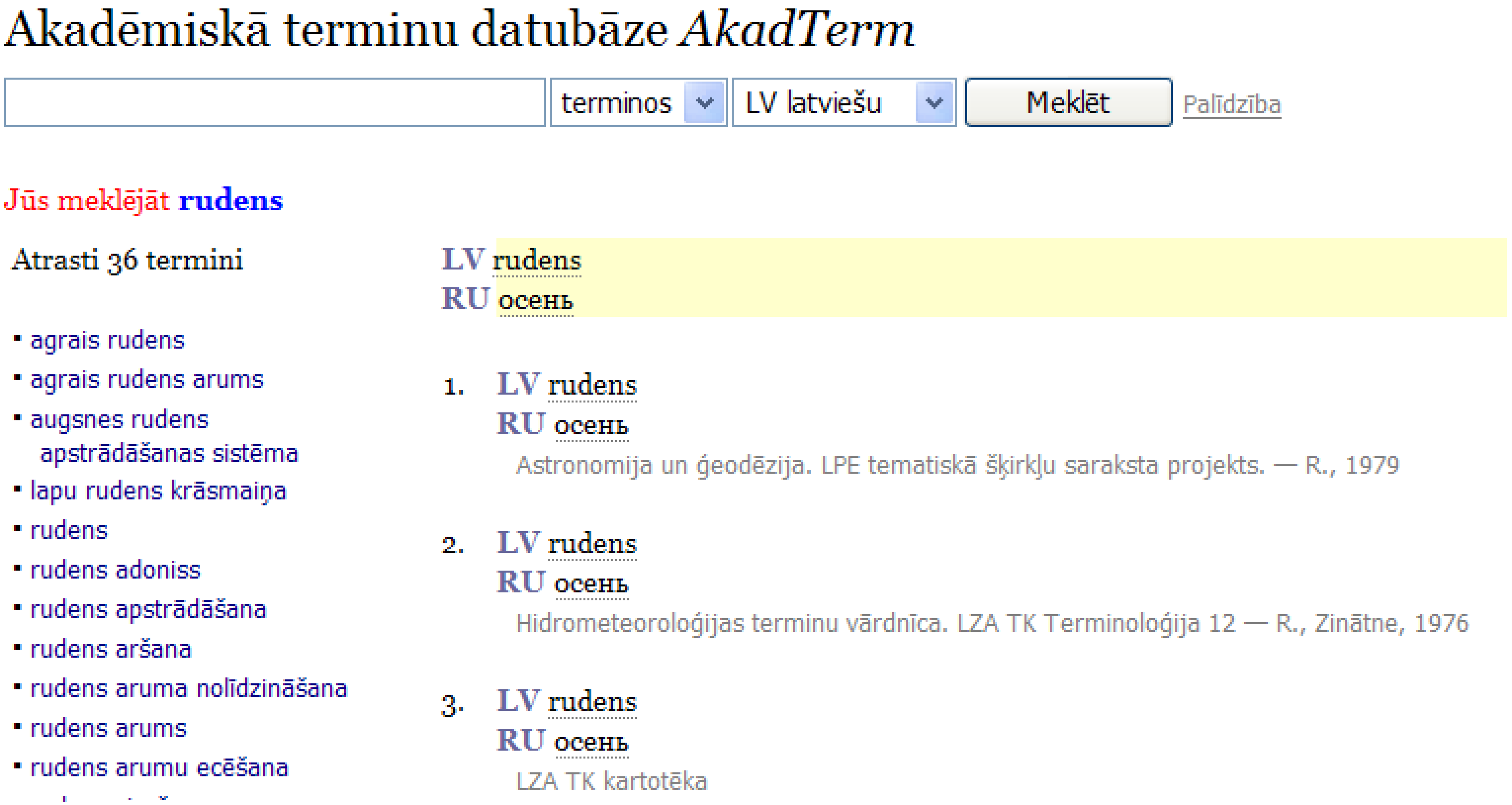
Igauņu—latviešu vārdnīca

sügis *~e (21)* **rudens**

© Karls Abens, pārskatījušas un papildinājušas Urve Aivare, Andra Kalnača, Ērika Krautmane, Jana Šteinberga-Ranki

Terminology Resources

Terminology Commission of the Latvian Academy of Sciences publishes official terminology in two large online databases:: www.termnet.lv and termini.lza.lv/akadterm



Akadēmiskā terminu datubāze *AkadTerm*

terminos | LV latviešu | Meklēt | Palīdzība

Jūs meklējāt **rudens**

Atrasti 36 termini

- agrais rudens
- agrais rudens arums
- augsnes rudens apstrādāšanas sistēma
- lapu rudens krāsmaiņa
- rudens
- rudens adoniss
- rudens apstrādāšana
- rudens aršana
- rudens aruma nolīdzināšana
- rudens arums
- rudens arumu ecēšana

LV rudens
RU осень

1. **LV** rudens
   RU осень
   Astronomija un ģeodēzija. LPE tematiskā šķirkļu saraksta projekts. — R., 1979

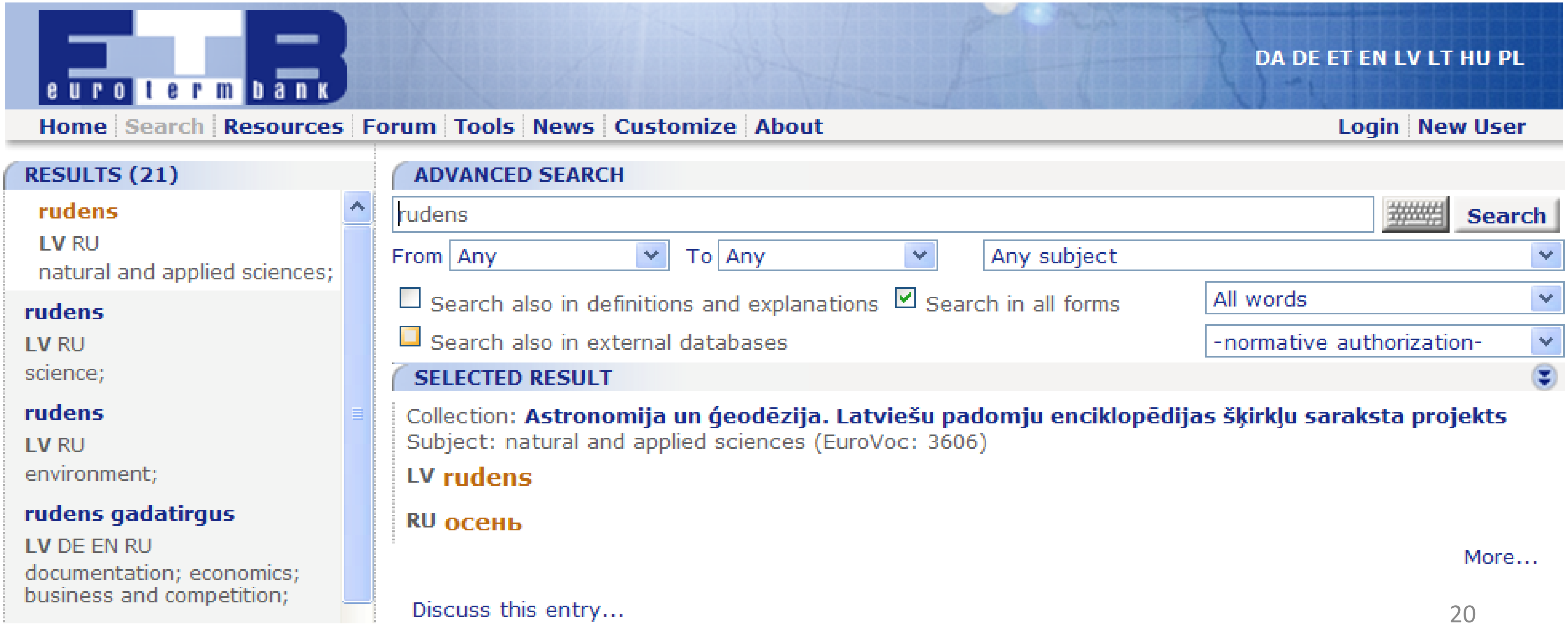
2. **LV** rudens
   RU осень
   Hidrometeoroloģijas terminu vārdnīca. LZA TK Terminoloģija 12 — R., Zinātne, 1976

3. **LV** rudens
   RU осень
   LZA TK kartotēka

LZA
Terminoloģijas
komisija

EuroTermBank portal

- Enables searching almost 2 million terms in over 25 languages
- Provides a single access point to interlinked term banks, such as IATE, WebTerm, Microsoft Terminology Collection, Terminology database of the Latvian Terminology Commission, and others



RESULTS (21)

rudens
LV RU
natural and applied sciences;

rudens
LV RU
science;

rudens
LV RU
environment;

rudens gadatirgus
LV DE EN RU
documentation; economics; business and competition;

ADVANCED SEARCH

rudens

From: Any To: Any Any subject

□ Search also in definitions and explanations ☑ Search in all forms All words

□ Search also in external databases -normative authorization-

SELECTED RESULT

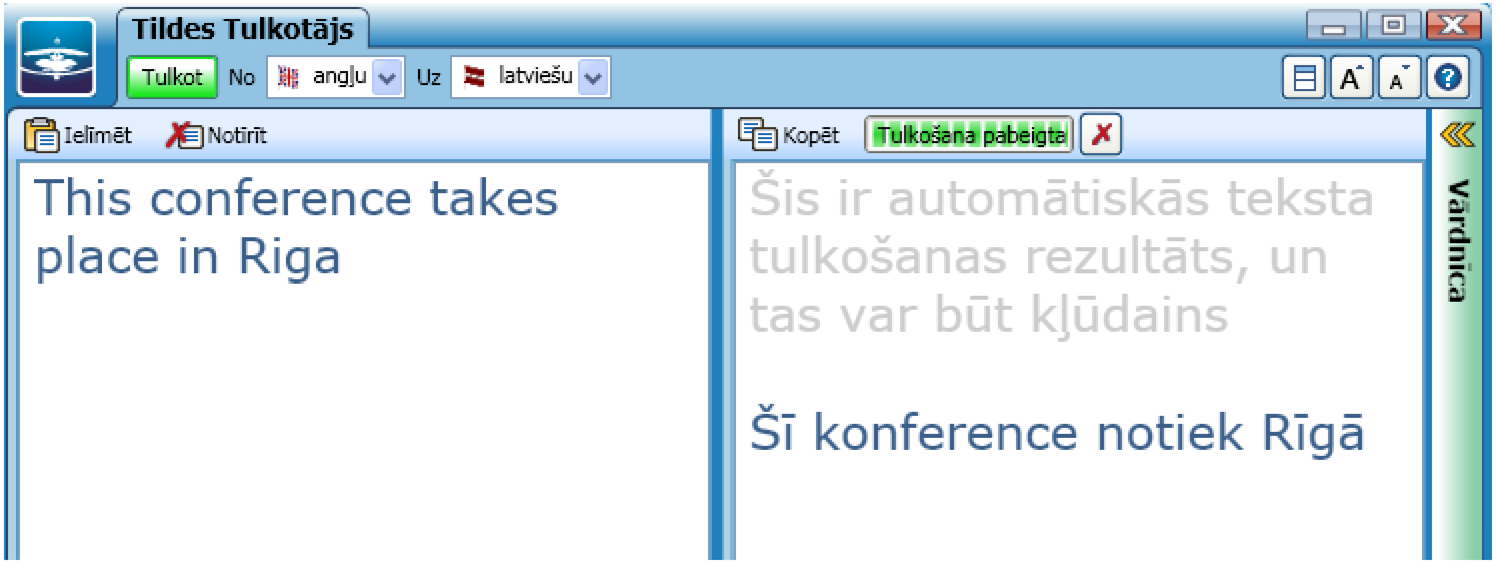
Collection: **Astronomija un ģeodēzija. Latviešu padomju enciklopēdijas šķirkļu saraksta projekts**
Subject: natural and applied sciences (EuroVoc: 3606)

LV **rudens**

RU **осень**

Machine Translation

- The rule-based approach to machine translation has been dominant in Latvia since mid-90-ies when the first version of the *LATRA* system (Latvian-English-Latvian) has been developed at IMCS
- The rule-based MT system *Tildes Tulkotājs* has been released in 2007 as part of *Tildes Birojs 2008,* the system translates texts from English into Latvian and from Latvian into Russian



No angļu Uz latviešu

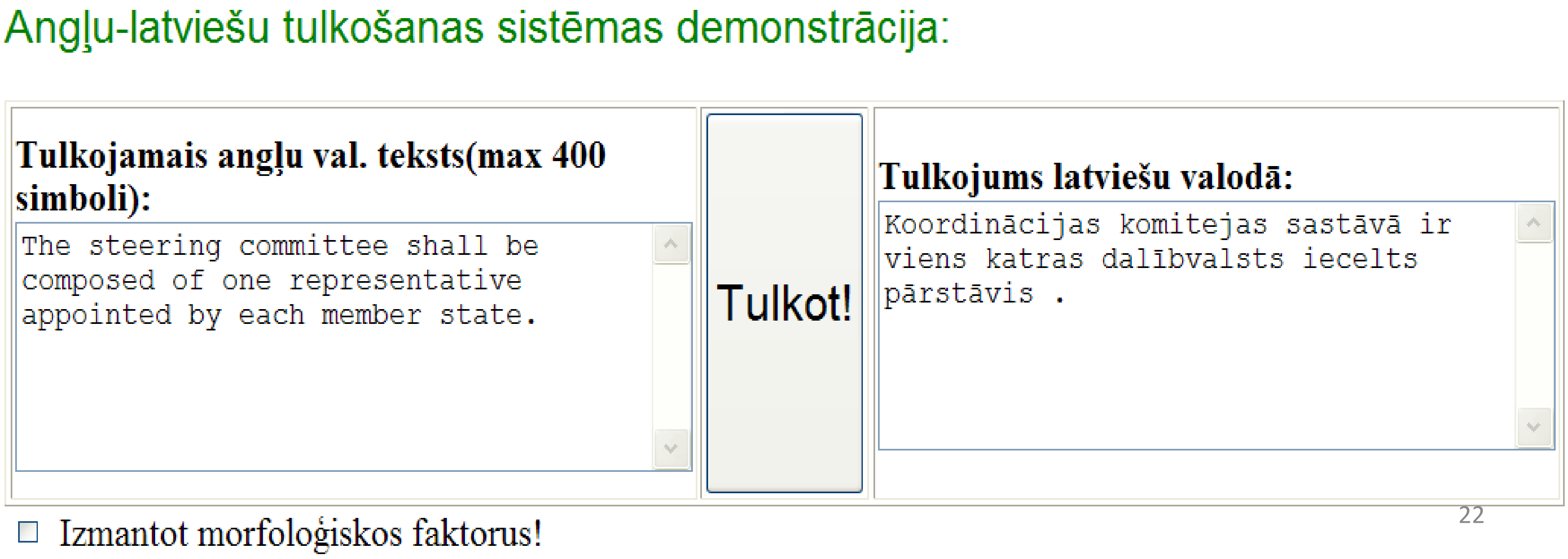
This conference takes place in Riga

Šis ir automātiskās teksta tulkošanas rezultāts, un tas var būt kļūdains

Šī konference notiek Rīgā

Statistical Machine Translation

- Research on Statistical Machine Translation (SMT) was started by IMCS in 2005 (eksperimenti.ailab.lv/smt)
  - *Evaluation of statistical Machine Translation methods for English Latvian translation system (2005-2008)*
  - *Application of Factored methods in English-Latvian Statistical Machine Translation System (2009-2012)*



Angļu-latviešu tulkošanas sistēmas demonstrācija:

Tulkojamais angļu val. teksts(max 400 simboli):

The steering committee shall be composed of one representative appointed by each member state.

Tulkojums latviešu valodā:

Koordinācijas komitejas sastāvā ir viens katras dalībvalsts iecelts pārstāvis .

□ Izmantot morfoloģiskos faktorus!

Statistical Machine Translation

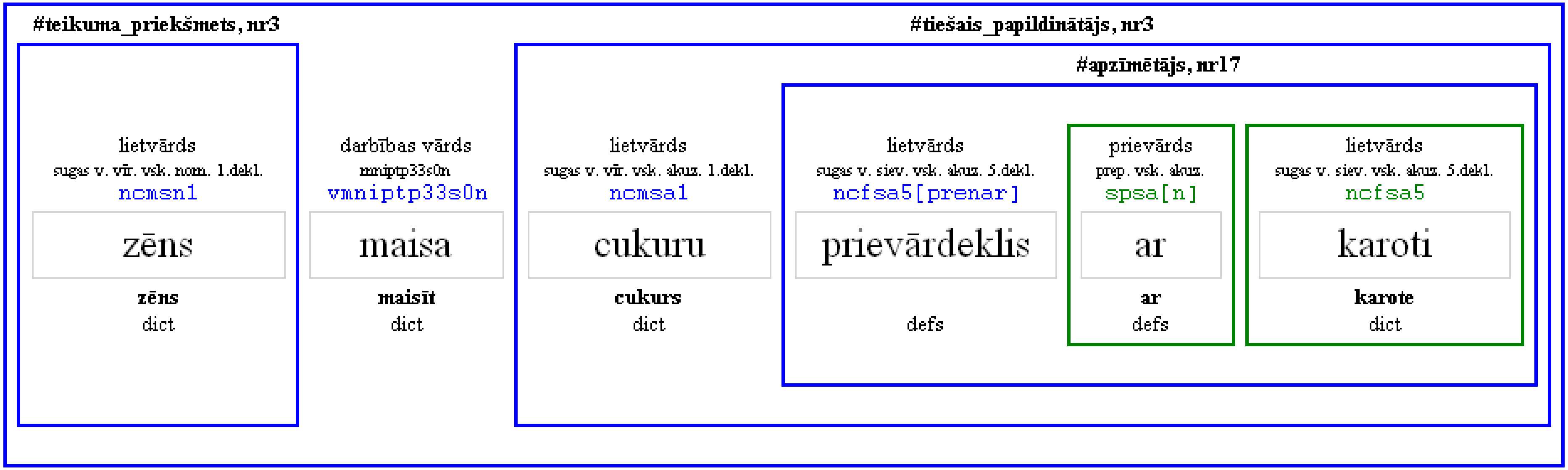
- In 2009/2010 Tilde released English-Latvian-English online SMT systems (translate.tilde.lv)
- Two SMT related EU projects coordinated by Tilde, have been started in 2010
  - the ICT PSP program project LetsMT!
  - the FP7 project ACCURAT

Speech Technologies

- IMCS had several projects devoted to experimental TTS and speech recognition systems
- In 2005 Tilde together with The Association of Blind People started a project to develop a Latvian text-to-speech (TTS) system
- Three speech synthesis systems have achieved the level of practical usability: *Visvaris* (Tilde), *T2S* (IMCS) and *Balss* (SIA Rubuls & Co).
- There has not been any serious research in Latvian language speech recognition, which could result in a practically usable speech recognition system

Tools for Natural Language Processing

- Morphology Tools: analysers and synthesizers, taggers
- Syntactic Parsers
  - dependency-based syntactic representation and a corresponding rule-based parser were created in the *SemTi-Kamols* project



  - Latvian shallow syntactic parser was built by Tilde in 2007. The formal grammar is derived from the unification grammar

CLARIN in Latvia

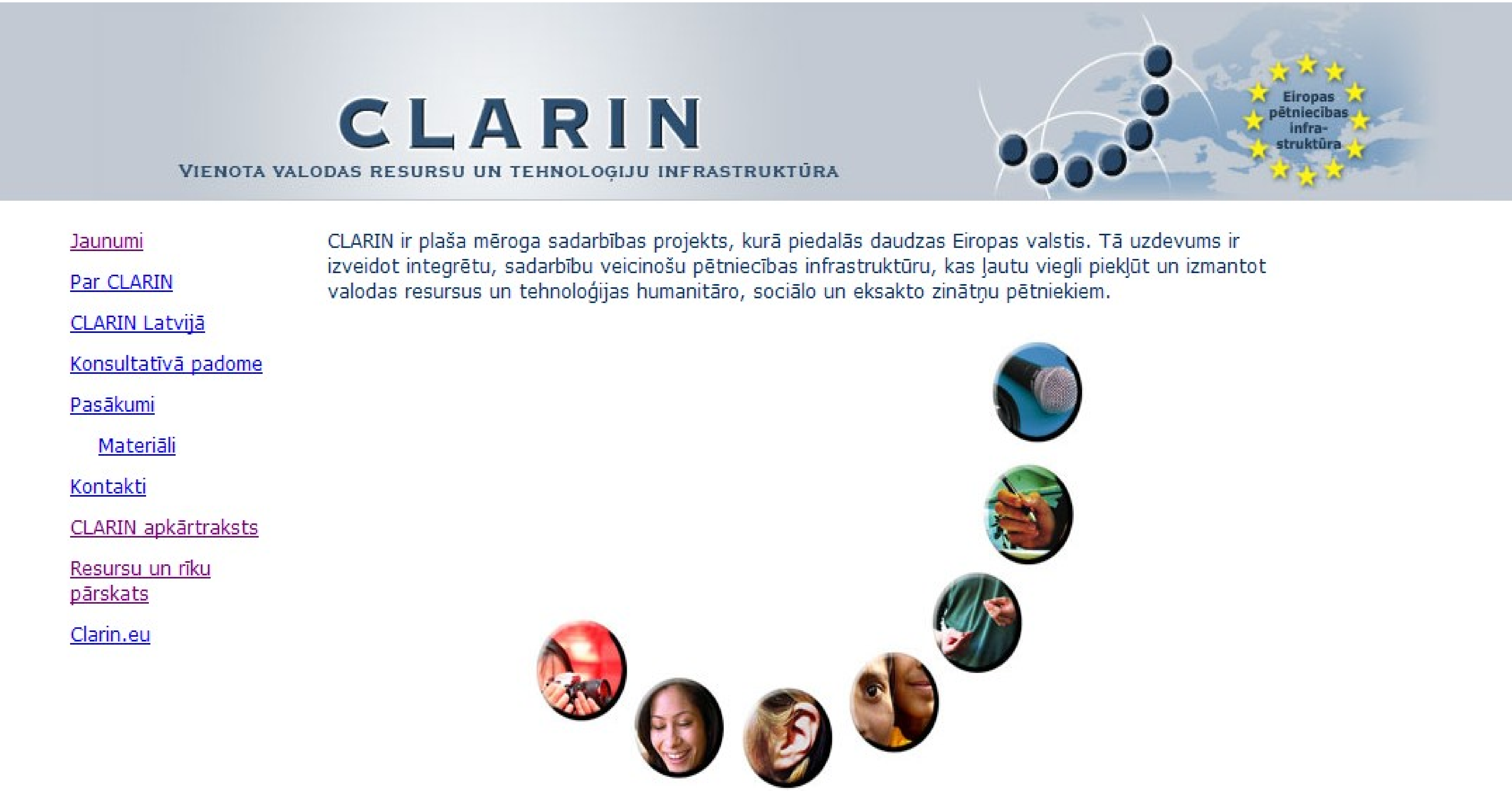
- Although the CLARIN initiative has been started only recently, the IMCS has been contributing to CLARIN aims already before by
  - collecting, preserving and making public available linguistic resources
  - development the Latvian language tools
  - co-operating with other research organizations in resource creation
  - by being Web publisher and maintainer of resources created in other research institutions

CLARIN in Latvia

- In 2006 IMCS and Tilde company have been invited to join CLARIN initiative
- IMCS has signed agreement to join CLAIN consortium starting form April 1, 2009
- Participation of Latvia in the CLARIN project is supported by the Ministry of Education and Science of the Republic of Latvia
- Recently the Cabinet of Ministers has approved "*Action Plan for Implementation of Guidelines for Science and Technology Development*"*.* One of the subtasks of the Action Plan is to ensure the participation of research institutions in the CLARIN project

CLARIN in Latvia

- IMCS has been appointed as the CLARIN National Contact Point (www.clarin.lv) by the Ministry of Education and Science
- Long term intention of IMCS is to become a CLARIN conformant national-level service and metadata providing centre
- To prioritize goals and tasks of the CLARIN project in Latvia and to facilitate the creation of the CLARIN infrastructure, the CLARIN National Advisory Board was established and approved by the Ministry of Education of Science
- Some important contributions:
  - Latvian resources in CLARIN LRT inventory
  - contribution to CLARIN BLARK
  - work on creation of reliable identity federation



CLARIN

VIENOTA VALODAS RESURSU UN TEHNOLOĢIJU INFRASTRUKTŪRA

- Jaunumi
- Par CLARIN
- CLARIN Latvijā
- Konsultatīvā padome
- Pasākumi
  - Materiāli
- Kontakti
- CLARIN apkārtraksts
- Resursu un rīku pārskats
- Clarin.eu

CLARIN ir plaša mēroga sadarbības projekts, kurā piedalās daudzas Eiropas valstis. Tā uzdevums ir izveidot integrētu, sadarbību veicinošu pētniecības infrastruktūru, kas ļautu viegli piekļūt un izmantot valodas resursus un tehnoloģijas humanitāro, sociālo un eksakto zinātņu pētniekiem.

Conclusions

- The last six years have been an active period in HLT development in Latvia
- Basic elements for research infrastructure of language resources and technology have been established in Latvia
- Urgent problem is the lack of programmes on computational linguistic at Latvian universities
- Targeted national research and development activities are urgently needed to fill these gaps in HLT development in Latvia
- New initiatives to support resource sharing and development of HLT products has been recently initiated in Baltic and Nordic countries