Filling the Gaps in Latvian BLARK: Case of the Latvian IT Competence Centre

Inguna SKADIŅA, Ilze AUZIŅA, Daiga DEKSNE, Raivis SKADIŅŠ, Andrejs VASIĻJEVS, Madara GAIĻŪNA, Ieva PORTNAJA

Latvian Basic Language Resource Kit (BLARK)

CLARIN BLARK (2010)

Resource	Dutch/Flem. (CLARIN)[10]	Estonian (CLARIN/ other)	Finnish (CLARIN/ other)	Latvian (CLARIN/ other)	Basque (CLARIN) [11]	Catalan (CLARIN)[12]	Galician (CLARIN)[13]
Thesaurus	++	+	-/-	-/-	-	-	-
Unannotated corpus	+++	+++	+++/-	+/+	-	-	+
Multi–media corpus	-	-	+/-	-/-	+	+	+
Multi-lingual lexicon	++	++	+++/-	++/++	-	-	-
Annotated corpus	+++	+++	++/-	+/-	-	-	-
Speech corpus (sound)	++	++	+++/-	-/-	-	-	-
Speech corpus (transcript)	+++	++	+++/-	+/+	-	-	-
Mono-lingual lexicon	+++	++	+++/-	++/++	+	+	+
Terminology data bases	+	+	+/-	++/-	-	++	++
Multi–lingual corpus	+	-	+++/+	++/-	-	++	+
Multi-modal corpus	-	-	+++/-	-/-	-	-	-
Multi-lingual comparable corpora	+	++	+++/++	-/-	-	-	-

Level of language technology support for Latvian in 2012

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology: Tools, Technologies and Applications							
Speech Recognition	0	0	0	0	0	0	0
Speech Synthesis	2	3	4	3	4	3	4
Grammatical analysis	2.5	2	3	3.5	4	3	4
Semantic analysis	1	0	0	0	0	0	0
Text generation	1	2	1	2	2	1	2
Machine translation	3	4	3	3	4	3	4
Language Resources: Resources, Data and Knowledge Bases							
Text corpora	2	4	4	3	3	3	4.5
Speech corpora	1	0	1	1	1	1	3
Parallel corpora	1	3	2	2	3	4	4
Lexical resources	3	3.5	4	3	4.5	4.5	4.5
Grammars	2	1	3	2	3	4	3

Availability of language resources and tools for languages of Baltic and Nordic countries

(META-NET Whitepapers, 2012)

	Excellent	Good	Moderate	Fragmentary	Weak/None
Speech Processing	–	–	Finnish	Danish, Estonian, Norwegian, Swedish	Icelandic, **Latvian**, Lithuanian
Machine Translation	–	–	–	–	Danish, Estonian, Finnish, Icelandic, **Latvian**, Lithuanian, Norwegian, Swedish
Text Analysis	–	–	–	Danish, Finnish, Norwegian, Swedish	Estonian, Icelandic, **Latvian**, Lithuanian
Resources	–	–	Swedish	Danish, Estonian, Finnish, Norwegian	Icelandic, **Latvian**, Lithuanian

IT Competence Centre Programme

IEGULDĪJUMS TAVĀ NĀKOTNĒ

IT Competence Centre Programme

- In 2010, Latvian **research institutions** and major information technology **companies** founded the IT Competence Centre (IT CC)
- The **aim** of IT CC is to support a **long term cooperation** between
  - **research organisations** and **industry**
  - in order to **create innovative technologies** and **prototypes** of **internationally competitive IT products**
- Tilde, University of Latvia, LETA among 13 founders of the IT CC

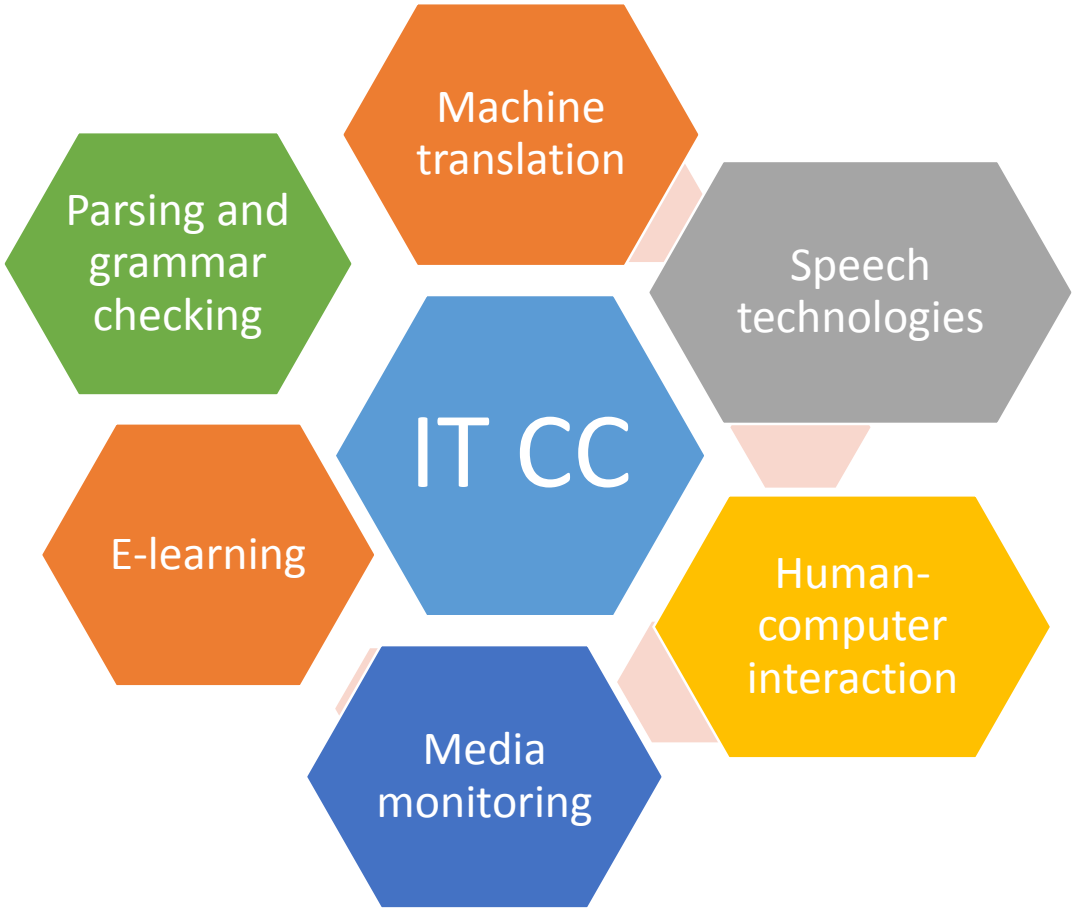
KOMPETENCES CENTRS

IT Competence Centre Programme

- IT Competence Centre has set two research directions as its main priorities:
  - business process analysis
  - language technologies
- Since 2013 **eleven** projects have been completed in different disciplines related to language technologies and natural language processing

KOMPETENCES CENTRS

Projects supported by IT CC



KOMPETENCES CENTRS

Basic Language Resources and Tools

Speech Resources and Technologies

- Until now, the major gap in BLARK for Latvian was the missing Automatic Speech Recognition (ASR) technology
- This important gap is now filled through the results of **four** IT CC projects

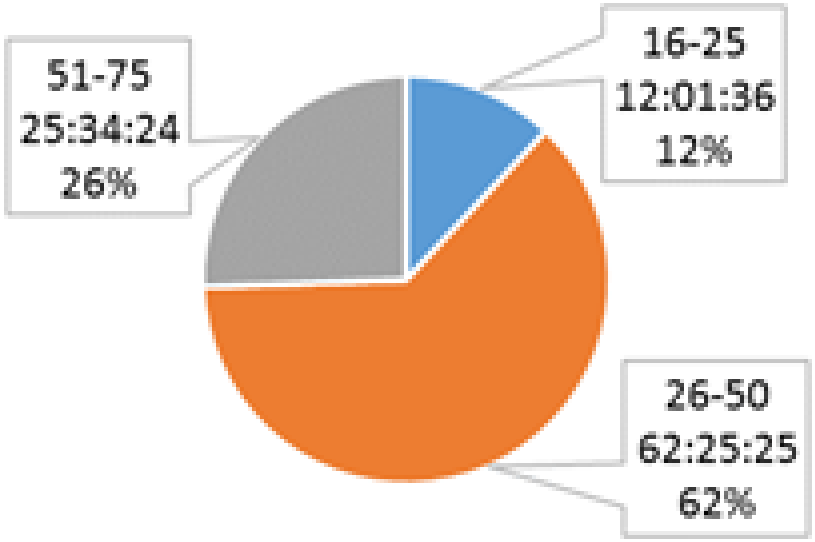
First Latvian speech corpus

- Specification and creation of the corpus took about a year
- 100-hour orthographically annotated and 4-hour phonetically annotated Latvian Speech Recognition Corpus

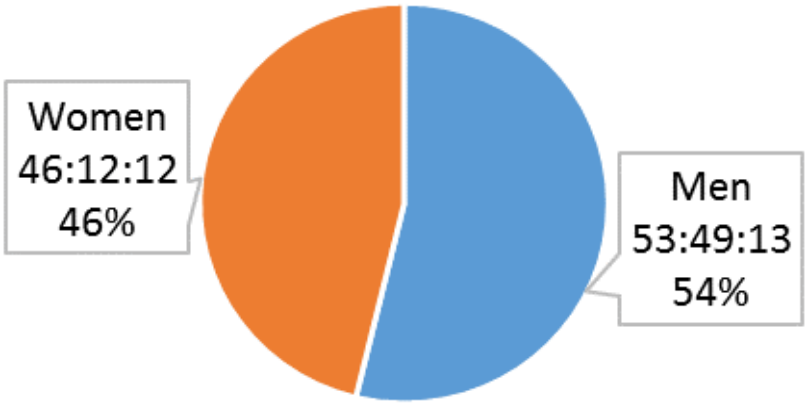
Number of unique words	~72.5 k
Number of running words	~837 k
Total number of speakers	1,851
Men	1,016
Women	835

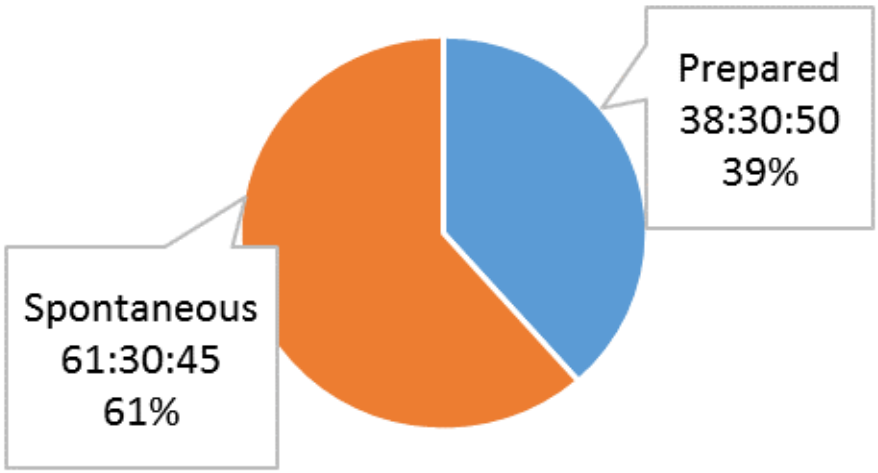
Type	Total length
Inhalation, exhalation	3 h 45 min (13,538 s)
Pauses	1 h 55 min (6,911 s)
Non-verbal segments	19 min (1,137 s)
Verbal segments	94 h 1 min (338,500 s)
The whole corpus	100 h 1 min (360,086 s)

Speech corpus: some statistics



Data distribution with respect to the age of speakers





Automatic Speech Recognition

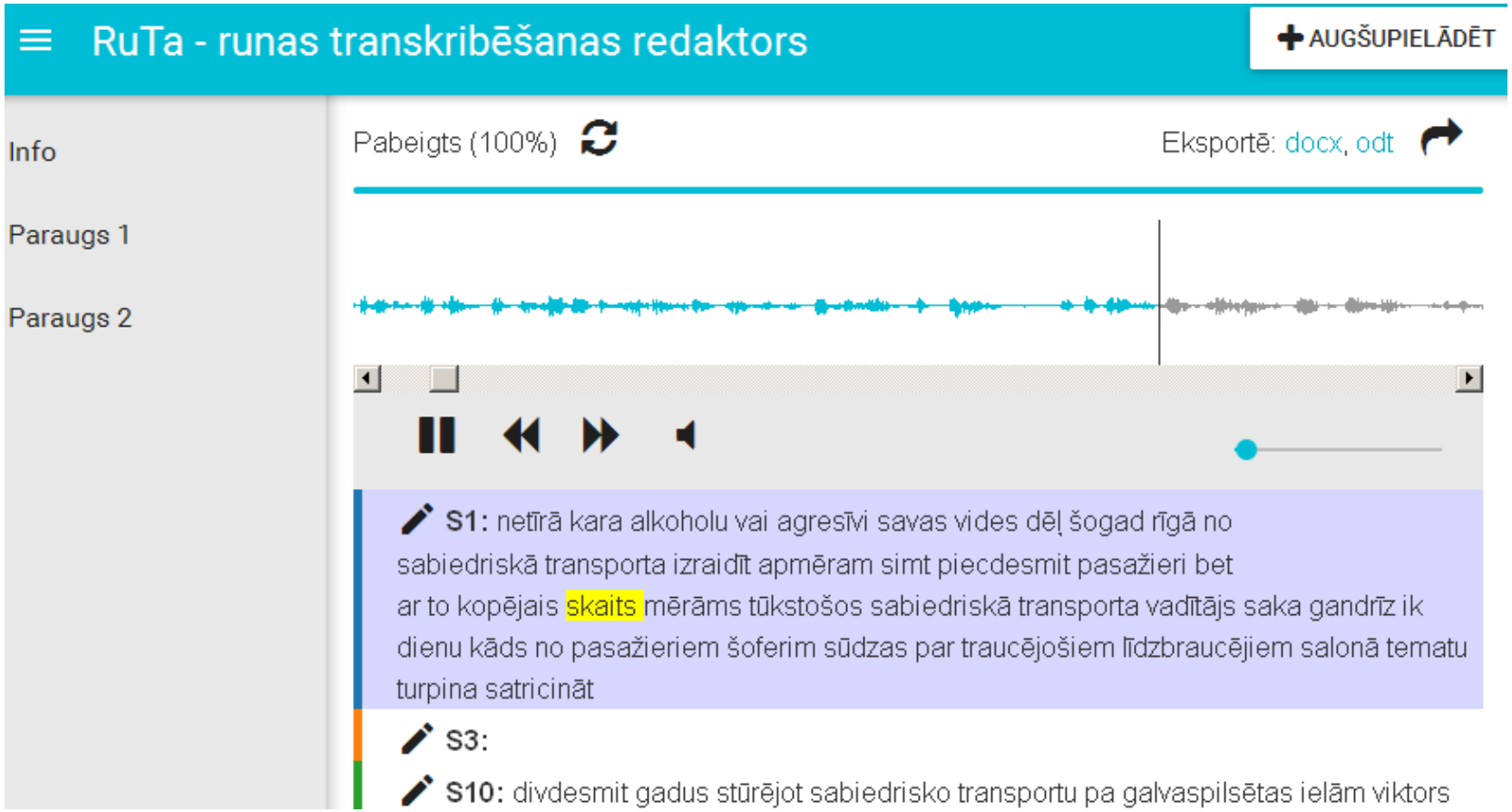
The Latvian Speech Recognition Corpus was used in two separate IT CC projects for building Latvian ASR systems:

- Language technology company Tilde created Latvian ASR systems for audio transcription and for text dictation
- The news agency LETA, together with IMCS, created an ASR system for keyword recognition in media monitoring

RuTa - a speech transcription editor

- Accumulated experience from IT CC projects about speech recognition and its application to media monitoring led to development of RuTa - a speech transcription editor
- RuTa includes: audio file uploading, automatic speech transcription and speaker segmentation, manual result editing and result exporting or sharing

RuTa - a speech transcription editor



Latvian ASR in real-world applications

Test Set	OOV (out of vocabulary)	WER
Lectures	3%	20.71%
News	6%	19.63%

The evaluation results encouraged to integrate Latvian ASR in real-world applications:

- Tilde created a public online service for audio transcription
- integrated Latvian speech recognition in the product *Tildes Birojs*
- developed a dictation system

Tava audiofaila saturs ir pārvērsts rakstītā tekstā

R1: Labvakar, pienācis laiks raidījumam aktuālais temats studijā Lauris zvejnieks, lai mudinātu iedzīvotājus vairāk izmantot ei pakalpojumus. Šonedēļ norit kampaņa dienas bez rindām.

R1: Šīs kampaņas laikā iecerēts cilvēkus informēt par tiem iestāžu pakalpojumiem, kurus var saņemt.

R1: Neizejot no mājas, tādējādi gan ietaupot naudu, gan laiku, lai runātu par kampaņu, kā arī par šiem AF pakalpojumiem uz sarunu aktuālo tematu studijā aicinājām vides aizsardzības un reģionālās attīstības ministrijas elektronisko pakalpojumu nodaļas vadītāju Gati Ozolu vakar novakarē.

R1: Tā kā jūs šobrīd raksturotu cik aktīvi iedzīvotāji izmanto interneta vidi dažādu pakalpojumu saņemšanai dažādu lietu kārtošanai.

Teksta dokuments (*.txt)

LEJUPIELĀDĒT

IESNIEGT JAUNU FAILU

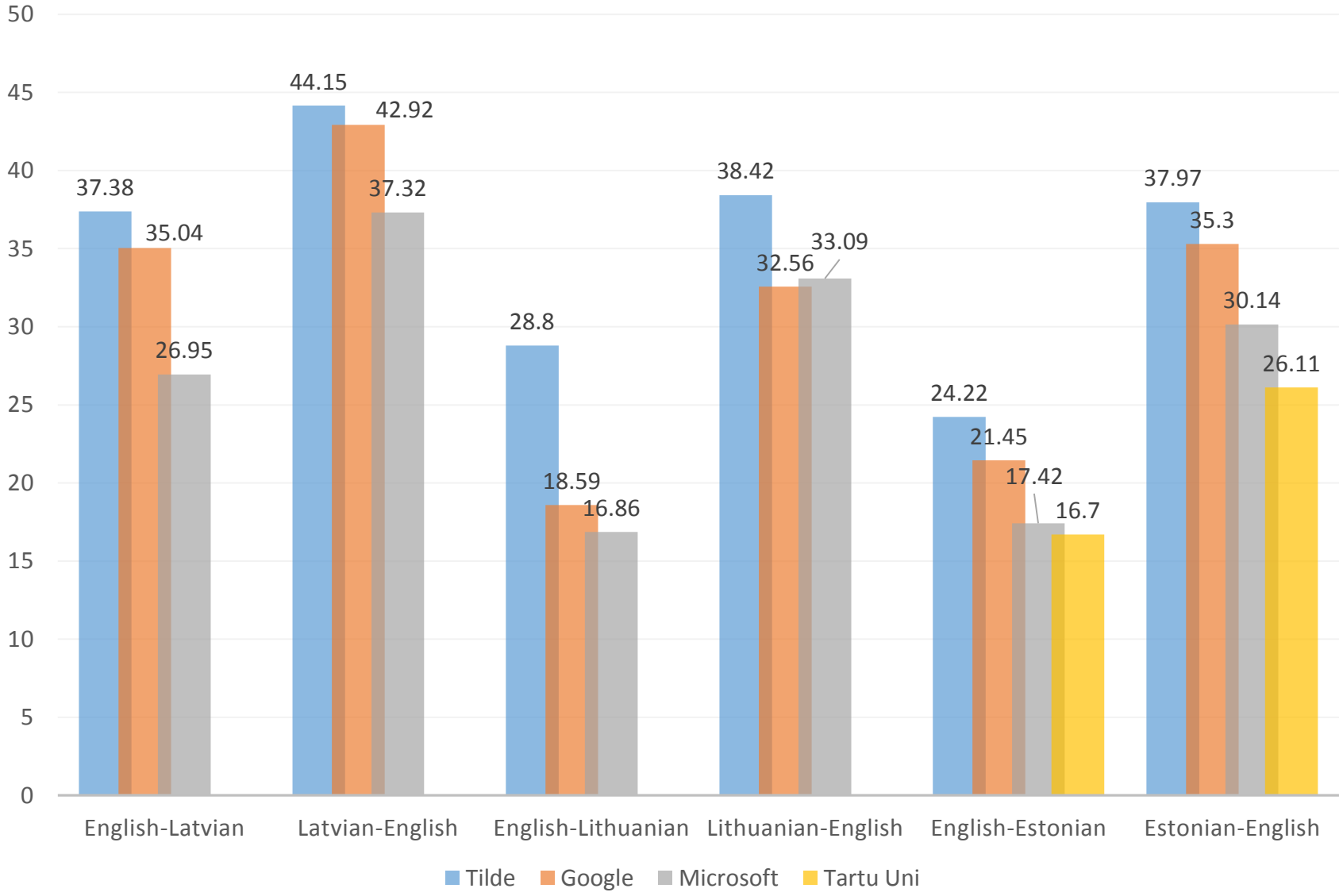
0:06 / 1:22

Machine Translation

- Focused on researching novel multilingual methods for Latvian and other under-resourced languages:
  - integration of knowledge about word morphology,
  - integration of translation dictionaries and term dictionaries,
  - treatment of special tokens (e.g. numbers, measurement units, file names, URLs etc.)
  - automation of corpora collection and processing to address the problem of data sparseness
- Validation in practical application for the localization industry

TILDE

Better than Google in MT quality



Productivity Increase in Localization

- Evaluation in software localization
  - Latvian, Lithuanian, Estonian, Polish, Hungarian, and Czech
  - Integration in CAT tools
  - Support of formatting tags
  - Document translation
- Translator productivity increase for all languages (15-32%)

Parsing and Grammar Checking

A rule-based solution

A framework for the parser and grammar checker development is derived from a context-free grammar (CFG) formalism, contains

- 580 rules describing the correct syntactic constructions
- 263 error rules that describe incorrect syntax
- 239 error rules that contain only terminal symbols

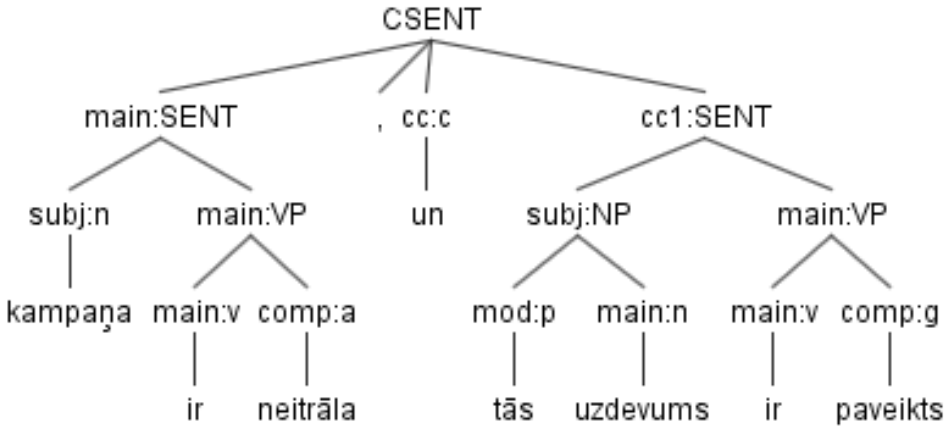
TILDE

- Parsing rule

CSENT -> main:SENT pm:T cc:C cc:SENT

cc:C.ConjType==Coord

pm:T.PunctType==comma



- Error rule

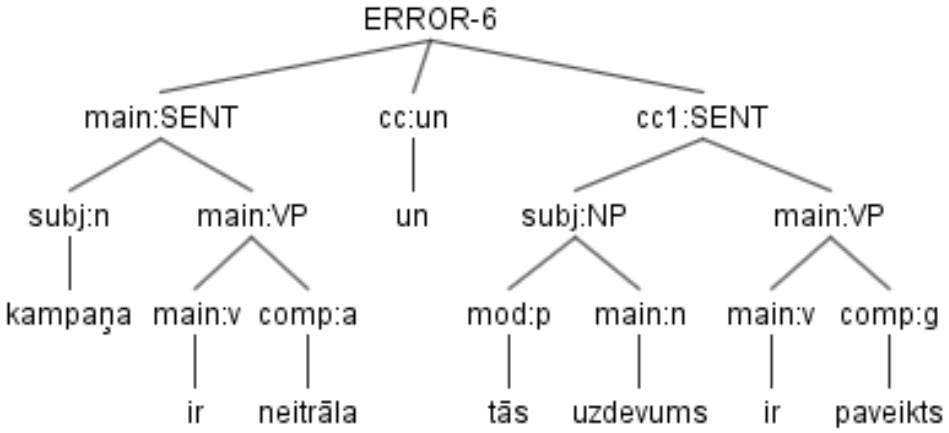
DESCR «Comma is missing"

ERROR-6 -> main:SENT cc:C cc:SENT

cc:C.ConjType==Coord

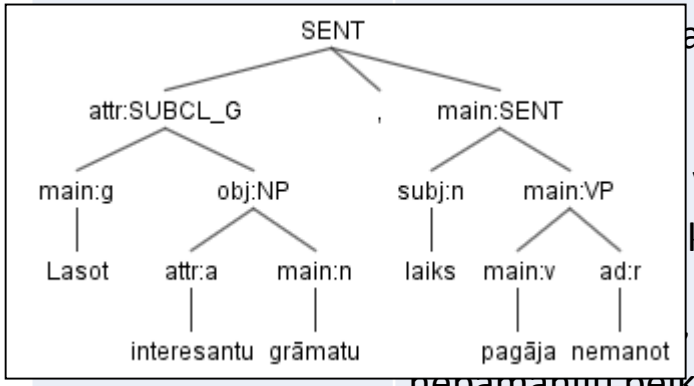
GRAMMCHECK MarkSpaceBefore(cc:C)

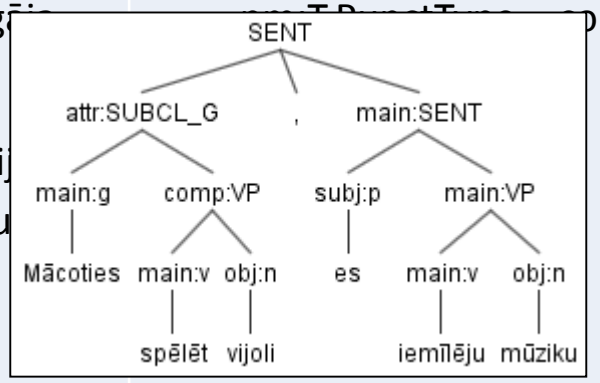
SUGGEST (","+cc:C)

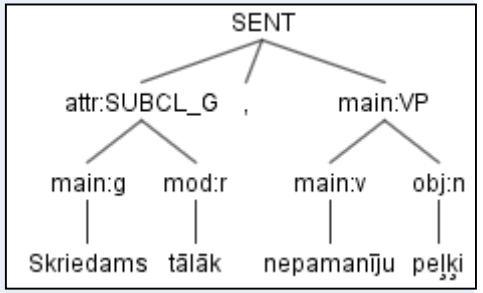


Treatment of participle phrase

Sentence type	Example	Rule
PP with an object	lasot interesantu grāmatu	SUBCL_G -> main:G obj:CNP obj:CNP.Case==accusative main:G.ParticipleType==undeclinable main:G.Voice==active
PP with the infinitive verb	Mācoties spēlēt vijoli	SUBCL_G -> main:G obj:CVP obj:CVP.Mode==infinitive main:G.ParticipleType==undeclinable main:G.Voice==active
PP with adverb	skriedams tālāk	SUBCL_G -> main:G advl:ADVP main:G.ParticipleType==partly
PP in sentence	Lasot interesantu	SENT -> attr:SUBCL_G pm:T main:SENT

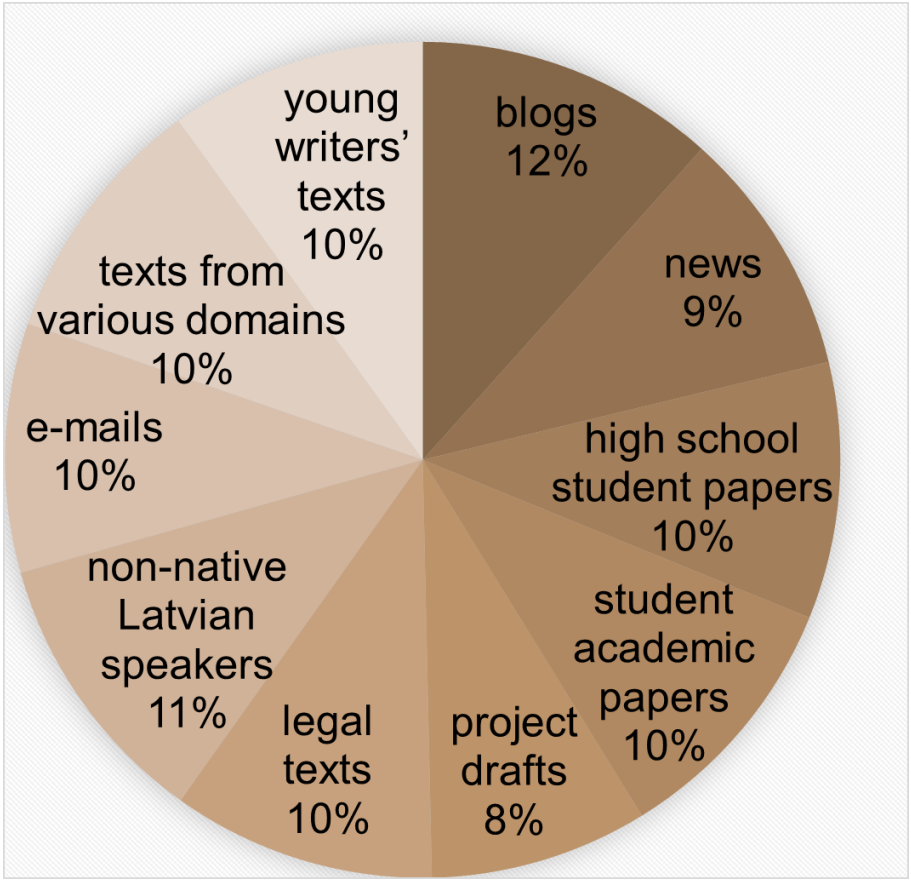






The balanced error annotated corpus

The balanced test corpus (**10,563 sentences**) was created with the aim to **assess the quality of the grammar checker**.



Evaluation results

Parser

Results (using PARSEVAL measures) for the 484 sentences, first parsed by a parser, then corrected by a human editor:

- recall – 89.1%
- precision – 89.13%
- brackets do not cross – 87.6%

Grammar checker

Results for the student paper test corpus:

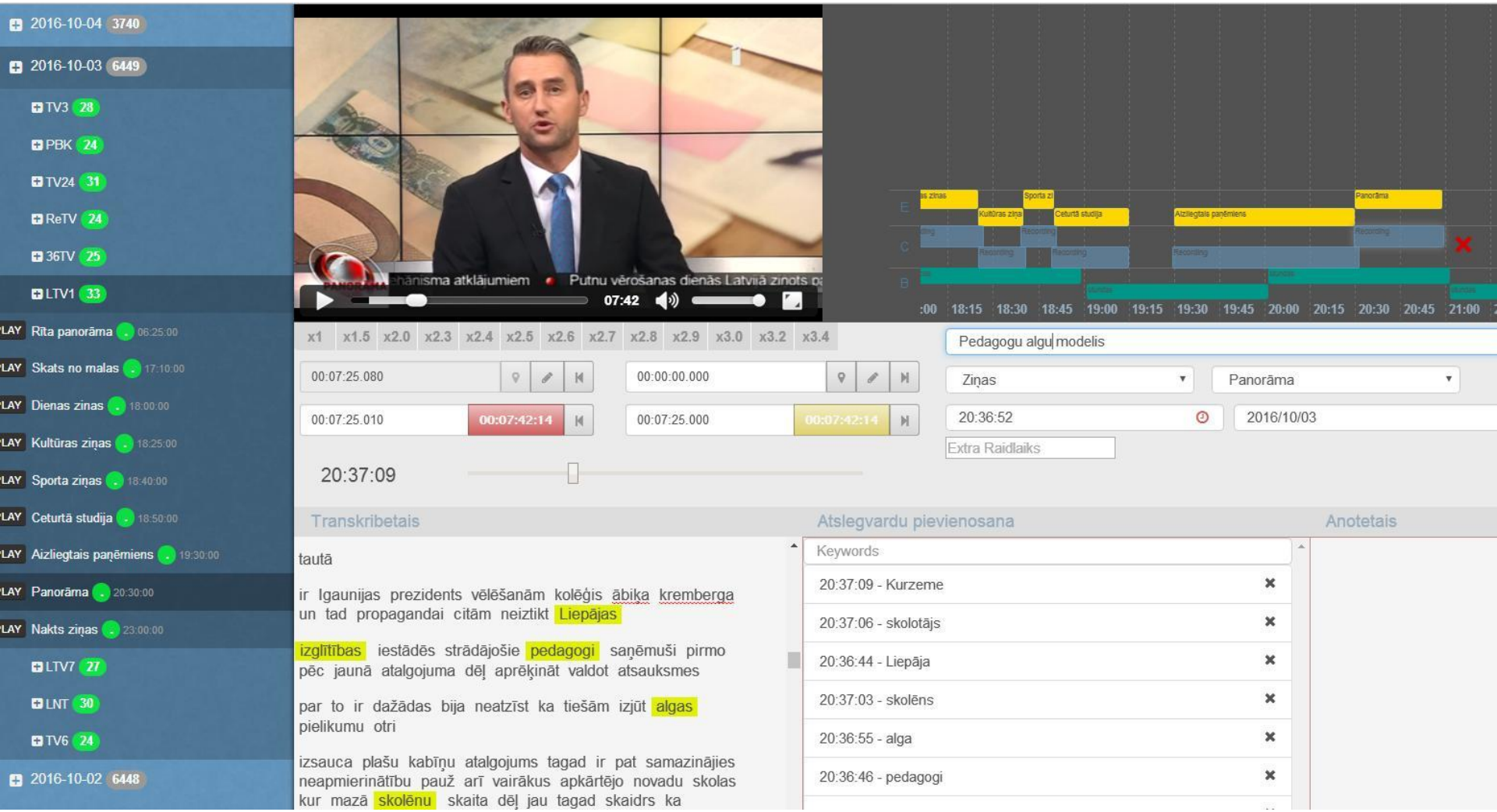
- recall – 18.7%
- precision – 65%

Applications and Products

System for Monitoring of Latvian Radio and Television Broadcasts

- Monitors more than 200 Latvian TV and radio programs
- Spots more than 5000 keywords
- Speech recognition accuracy: 62%
- Keyword spotting accuracy: 78%

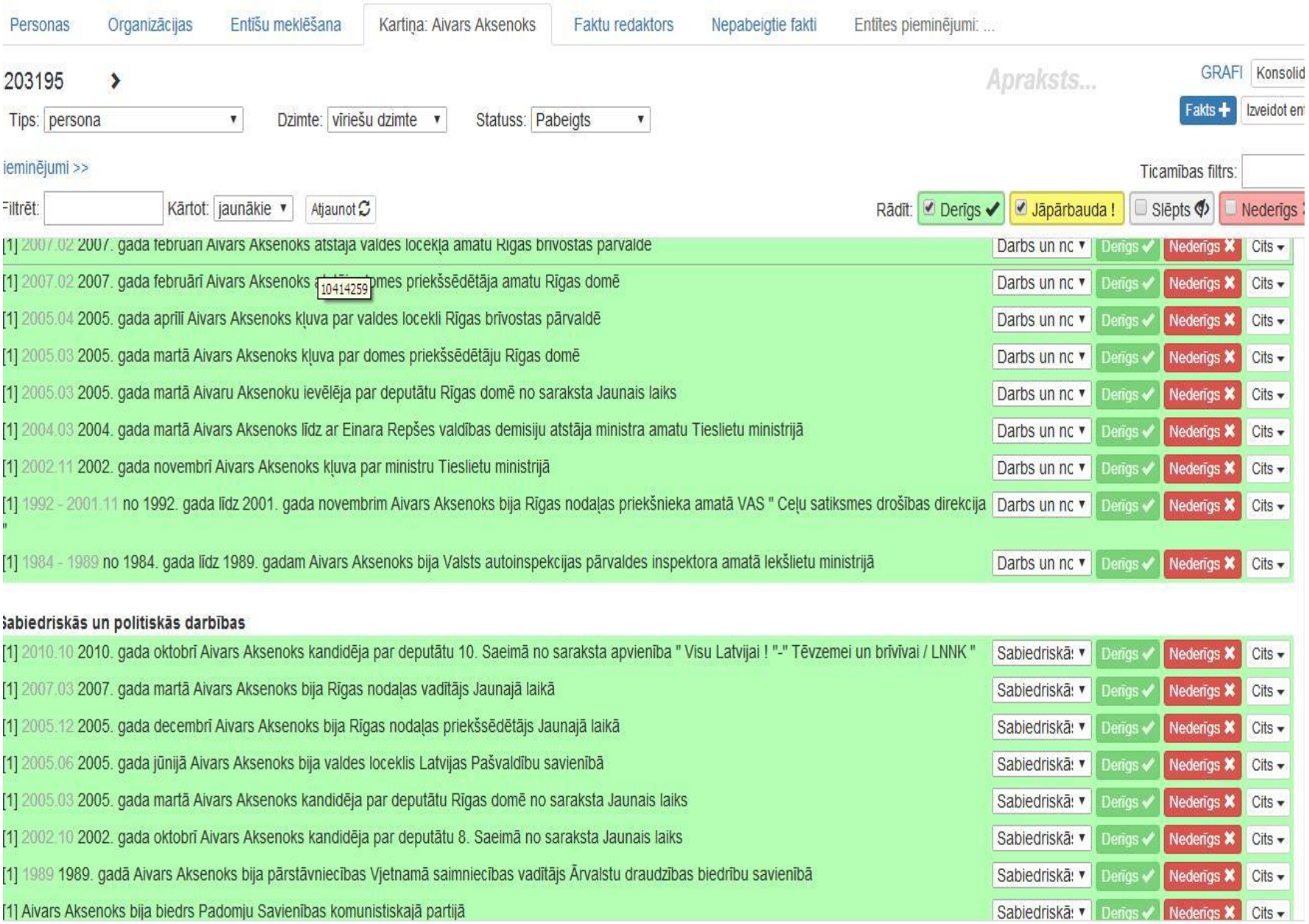
LETA
nacionālā informācijas aģentūra


2016-10-04 3740
2016-10-03 6449
TV3 28
PBK 24
TV24 31
ReTV 24
36TV 25
LTV1 33
PLAY Rīta panorāma 06:25:00
PLAY Skats no malas 17:10:00
PLAY Dienas zinas 18:00:00
PLAY Kultūras ziņas 18:25:00
PLAY Sporta ziņas 18:40:00
PLAY Ceturtā studija 18:50:00
PLAY Aizliegtais paņēmiens 19:30:00
PLAY Panorāma 20:30:00
PLAY Nakts ziņas 23:00:00
LTV7 27
LNT 30
TV6 24
2016-10-02 6448
07:42
x1 x1.5 x2.0 x2.3 x2.4 x2.5 x2.6 x2.7 x2.8 x2.9 x3.0 x3.2 x3.4
00:07:25.080
00:00:00.000
00:07:25.010
00:07:42:14
00:07:25.000
00:07:42:14
20:37:09
Pedagogu algu modelis
Ziņas
Panorāma
20:36:52
2016/10/03
Extra Raidlaiks
:00 18:15 18:30 18:45 19:00 19:15 19:30 19:45 20:00 20:15 20:30 20:45 21:00
Transkribetais
Atslegvardu pievienosana
Anotetais
tautā
ir Igaunijas prezidents vēlēšanām kolēģis ābika kremberga un tad propagandai citām neiztikt Liepājas
izglītības iestādēs strādājošie pedagogi saņēmuši pirmo pēc jaunā atalgojuma dēļ aprēķināt valdot atsauksmes
par to ir dažādas bija neatzīst ka tiešām izjūt algas pielikumu otri
izsauca plašu kabīņu atalgojums tagad ir pat samazinājies neapmierinātību pauž arī vairākus apkārtējo novadu skolas kur mazā skolēnu skaita dēļ jau tagad skaidrs ka
Keywords
20:37:09 - Kurzeme
20:37:06 - skolotājs
20:36:44 - Liepāja
20:37:03 - skolēns
20:36:55 - alga
20:36:46 - pedagogi

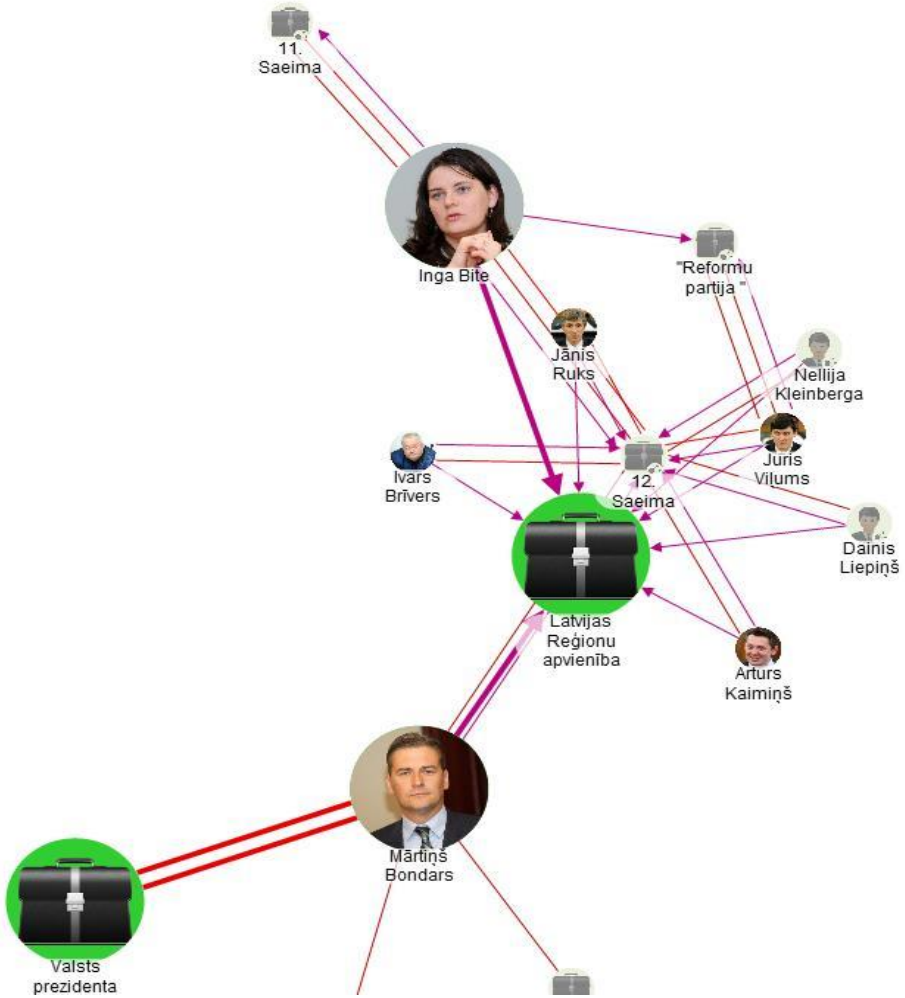
System of Text Summarisation and Information Extraction

- Automatically processes, analyses and extracts information about persons and organisations from news articles
- Extracts facts according to 27 frames (e.g. Name, age, relationship, etc.)
- Daily processes ~ 500 news articles
- LETA news archive collects and stores publically available information of more than 50,000 persons and 5,000 companies.

LETA
nacionālā informācijas aģentūra


Personas
Organizācijas
Entīšu meklēšana
Kartiņa: Aivars Aksenoks
Faktu redaktors
Nepabeigtie fakti
Entītes pieminējumi: ...
203195
Apraksts...
GRAFI
Konsolid
Fakts +
Izveidot en
Tips: persona
Dzimte: vīriešu dzimte
Statuss: Pabeigts
ieminējumi >>
Ticamības filtrs:
Filtrēt:
Kārtot: jaunākie
Atjaunot
Rādīt: Derīgs
Jāpārbauda !
Slēpts
Nederīgs
10414259
[1] 2007.02 2007. gada februari Aivars Aksenoks atstaja valdes locekļa amatu Rīgas brīvostas pārvalde
[1] 2007.02 2007. gada februārī Aivars Aksenoks
omes priekšsēdētāja amatu Rīgas domē
[1] 2005.04 2005. gada aprīlī Aivars Aksenoks kļuva par valdes locekli Rīgas brīvostas pārvaldē
[1] 2005.03 2005. gada martā Aivars Aksenoks kļuva par domes priekšsēdētāju Rīgas domē
[1] 2005.03 2005. gada martā Aivaru Aksenoku ievēlēja par deputātu Rīgas domē no saraksta Jaunais laiks
[1] 2004.03 2004. gada martā Aivars Aksenoks līdz ar Einara Repšes valdības demisiju atstāja ministra amatu Tieslietu ministrijā
[1] 2002.11 2002. gada novembrī Aivars Aksenoks kļuva par ministru Tieslietu ministrijā
[1] 1992 - 2001.11 no 1992. gada līdz 2001. gada novembrim Aivars Aksenoks bija Rīgas nodaļas priekšnieka amatā VAS " Ceļu satiksmes drošības direkcija "
[1] 1984 - 1989 no 1984. gada līdz 1989. gadam Aivars Aksenoks bija Valsts autoinspekcijas pārvaldes inspektora amatā Iekšlietu ministrijā
Darbs un nc
Derīgs
Nederīgs
Cits
abiedriskās un politiskās darbības
[1] 2010.10 2010. gada oktobrī Aivars Aksenoks kandidēja par deputātu 10. Saeimā no saraksta apvienība " Visu Latvijai ! "-" Tēvzemei un brīvīvai / LNNK "
[1] 2007.03 2007. gada martā Aivars Aksenoks bija Rīgas nodaļas vadītājs Jaunajā laikā
[1] 2005.12 2005. gada decembrī Aivars Aksenoks bija Rīgas nodaļas priekšsēdētājs Jaunajā laikā
[1] 2005.06 2005. gada jūnijā Aivars Aksenoks bija valdes loceklis Latvijas Pašvaldību savienībā
[1] 2005.03 2005. gada martā Aivars Aksenoks kandidēja par deputātu Rīgas domē no saraksta Jaunais laiks
[1] 2002.10 2002. gada oktobrī Aivars Aksenoks kandidēja par deputātu 8. Saeimā no saraksta Jaunais laiks
[1] 1989 1989. gadā Aivars Aksenoks bija pārstāvniecības Vjetnamā saimniecības vadītājs Ārvalstu draudzības biedrību savienībā
[1] Aivars Aksenoks bija biedrs Padomju Savienības komunistiskajā partijā
Sabiedriskā:

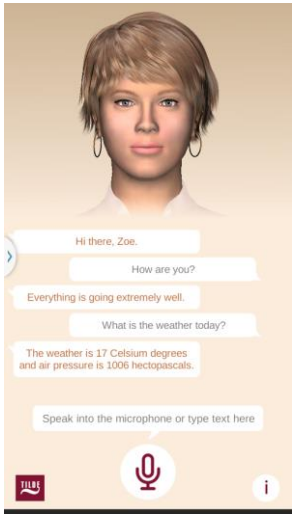
Gads: 1937 - 2016 | Iztīrīt visu | Paslēpt vājas | Skats | Iziet | Ieva Portnaja



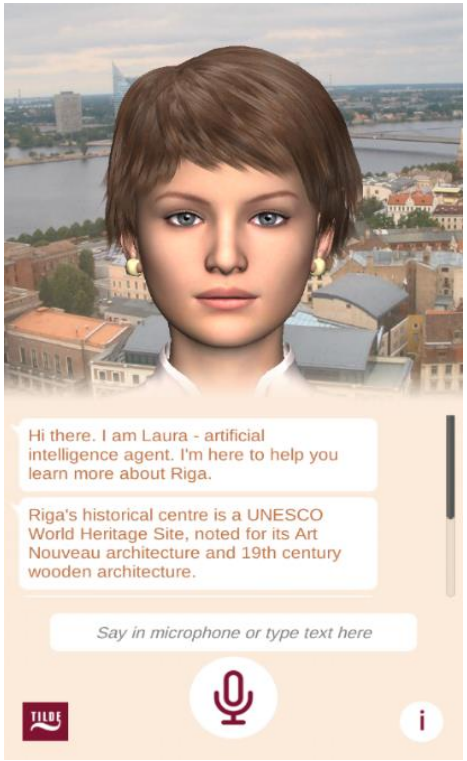
Multimodal interaction

The goal was to create animated virtual agents and research their usability in different applications:

- simple conversation
- city guide
- guide in library



Laura can hold conversations in English and provide information about several topics



Latvian-speaking virtual assistants

Two use scenarios were investigated:
- teaching multiplication to kids
- asking/telling facts about Latvia

Dialogue system	Speech recognition			Correctness of dialogue	
	Incorrect	Partly correct	Correct	Correct	Incorrect
Multiplication	9%	22%	69%	87%	13%
Latvian facts	14%	26%	14%	74%	26%

Synergy with EU level projects

QT21
QUALITY TRANSLATION

FREME

SUMMA

European Language
Resource Coordination

Conclusions and next steps

- The CC programme helped to advance Latvian language technologies and fill major gaps in the Latvian BLARK
- A particularly important achievement is the creation of a large annotated Latvian speech corpora and the first speech recognition systems for Latvian – the components that were completely missing in the Latvian BLARK.
- IT CC will continue research and development of the Latvian language technologies in the Competence Centre programme 2016-2021