Latviešu valodas apguvēju korpusa izveide

Ilze Auziņa, Roberts Darģis,
Kristīne Levāne-Petrova

Latvijas Universitātes 76. zinātniskās konferences Datorlingvistikas sekcija
2018. gada 1. martā

Mērķis

- Izveidot valsts valodas prasmes pārbaudes darbu datubāzi – t.s. latviešu valodas apguvēju korpusu.
- Noskaidrot latviešu valodas prasmes kvalitāti, analizējot valsts valodas prasmes pārbaudes kārtotāju darbus un to rezultātus.

Korpusā iekļautie teksti

- 900 valsts valodas prasmes pārbaudes darbu rakstītprasmes testi: no katra valodas apguves līmeņa (A1, A2, B1, B2, C1, C2) datubāzē iekļauti 150 darbi.
- Tikai tie rakstītprasmes testa uzdevumi, kuros ir saistīts teksts.
- Apjoms: 1496 teksti, 142 684 teksta vienības.
- Kļūdu skaits tekstos: 22,2 %.
- Datubāzē iekļauti arī valsts valodas prasmes pārbaudes darbu runātprasmes testi (no katra valodas apguves līmeņa 17 audioieraksti).

Korpusa izveides posmi

- Datu digitalizācija
- Tekstu normalizēšana
- Automatizēta morfoloģiskā anotēšana, tostarp vārdšķiru noteikšana, lemmatizēšana
- Oriģinālā un normalizētā teksta sastatīšana
- Automatizēta kļūdu anotēšana un manuāla pārskatīšana

Datu digitalizācija un tekstu normalizēšana

Text

Labdien mana miļa draudzene Gunita. Es gribu ar tevi satakties. Es gribu tevi aicinat uz cirku. Mes kopa brauksim uz Rīga. Pastaigājsimies pa vecrigu un parunat. Biļeti cenu 12 eiro. Es zinu ka manu draudzenei interese cirk.
Gaidišu tavu atbildi
Visu labu.

Correct text

Labdien, mana mīļā draudzene Gunita! Es gribu ar Tevi satikties. Es gribu Tevi aicināt uz cirku. Mēs kopā brauksim uz Rīgu. Pastaigāsimies pa Vecrīgu un parunāsim. Biļetes cena ir 12 eiro. Es zinu, ka manai draudzenei interesē cirks.
Gaidīšu Tavu atbildi.
Visu labu!

Kļūdu tipi (1/3)

Tips	Apakštips	Piemēri
Pareizrakstības kļūdas	Lielie /mazie sākumburti	*latvija – Latvija; Kā tev iet? – Kā Tev iet?*
	Diakritiskās zīmes	*tēv – tev ; bēt – bet; bilete – biļete*
	Kopā / šķirti rakstāmi vārdi	*Uzredzēšanos! – Uz redzēšanos!* *vis lielākais – vislielākais*
	Izlaisti burti	*30. maja – 30. maijā*
	Lieki burti	*uotraja augustā – otrajā augustā*
	Citas pareizrakstības kļūdas	*lajme – laime* *Es gaigu atbildi! – Es gaidu atbildi!*
Formveidošanas kļūdas	Nepareiza vārda forma (piem., lietv. – locījums, skaitlis; īpaš. v. – dzimte, noteiktā /nenoteiktā galotne; verbiem – izteiksme, persona)	*Es dzīvoju Latvija – Es dzīvoju Latvijā* *Es gribas apmeklēt šo pasākumu –* *Man gribas apmeklēt šo pasākumu*
	Skaņu mijas esamība/ neesamība	*vīrieti – vīrieši; angla valoda – angļu valoda*

Kļūdu tipi (2/3)

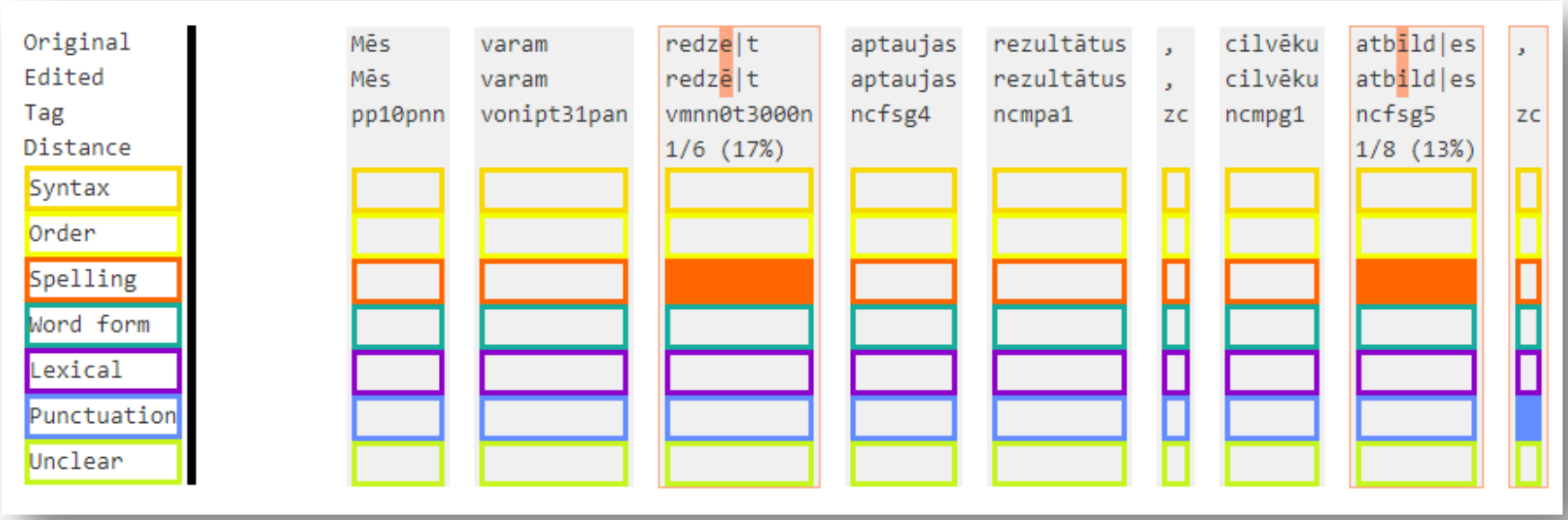
Tips	Apakštips	Piemēri
Leksikas kļūdas	Nozīmes ziņā neatbilstoša vārda lietojums	*Izeja ir bez maksas. – Ieeja ir bez maksas.* *Es tev pazvanīšu. – Es Tev piezvanīšu.* *Aptauja piedalījās Latvijās dzivotāji. –* *Aptaujā piedalījās Latvijas iedzīvotāji.*
		Ka klajas? – Kā klājas? *Nopirkt lūdzu mani 2 biļetus no cirku. –* *Nopirksi, lūdzu, man 2 biļetes uz cirku.* *Biļešu maksa 1,5 eiro, kāpēc es esmu pensioniere. –* *Biļešu maksa ir 1,5 eiro, tāpēc ka es esmu pensionāre.*
	Nesaprotams vārds / latv. val. neeksistējošs vārds	*Tas ir smuk pūtik un dažados krāsa. –* *Tās ir smukas puķes un dažadās krāsās.*
	Kalkēts vārds	*gimnasti – vingrotāji; nedārgs – nav dārgs;* *surpriziņš – pārsteigums; papraši – lūgt;* *importanta – svarīga*

Kļūdu tipi (3/3)

Tips	Apakštips	Piemēri
Sintakses kļūdas	Vārdu secība	*cena ekskursija – ekskursijas cena* *svetki sporta – sporta svētki*
	Lieks vārds	*Apsveicu tevi ar svētku dienu. – Apsveicu Tevi svētku dienā!*
	Izlaists vārds	*Gaidīšu tevi informāciju – Gaidīšu no Tevis informāciju*
	Nepilna teikuma struktūra	*Rozes sievietēm.* *Es gribu pasākum. – Es gribu apmeklēt pasākumu.*
Interpunkcijas kļūdas	Pieturzīmes trūkums	*Sveiki Anna! – Sveiki, Anna!*
	Lieka pieturzīme	*Labdien, dārgais, draugs – Labdien, dārgais draugs!*
	Neatbilstoša pieturzīme	*Kā tev iet. – Kā tev iet?*
Neskaidrs teksts	Jālabo viss izteikums. Turklāt ir iespējami vairāki varianti.	*Olga tu gribet pasakumu man teātra.* *Es ceru apmeklēt akrobāti un dresēti dzīvnieki.*

Automatizēta datu apstrāde

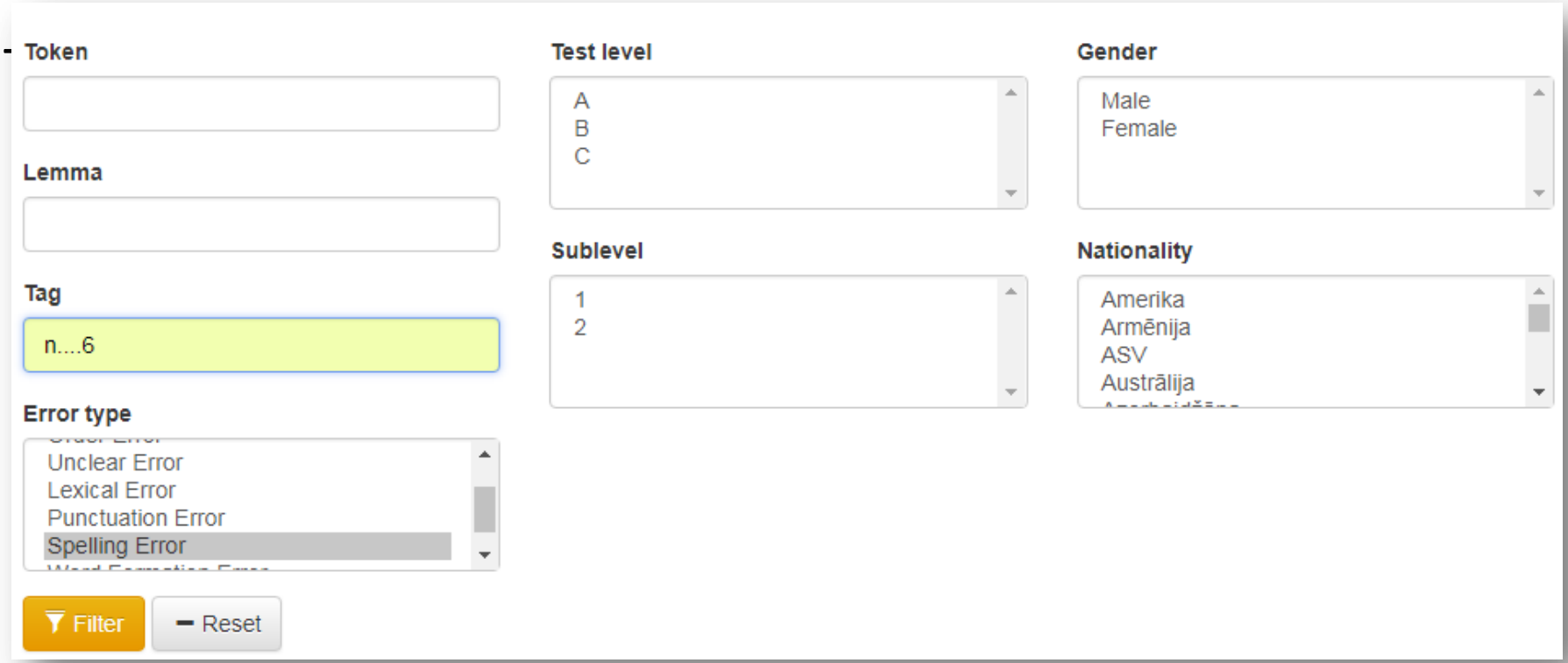
- Sadalījums tekstvienībās un morfoloģiskā analīze
- Tekstu sastatīšana (tostarp analīze tekstvienību un burtu līmenī)
- Pusautomātiska kļūdu marķēšana



Meklēšana korpusā (1/2)

Korpusā ir iespējams meklēt, izmantojot dažādas pazīmes:

- vārds / vārdforma
- pamatforma
- morfoloģiskās pazīmes (*Tag*)
- valodas prasmes līmenis
- dzimums
- pilsonība



Meklēšana korpusā (2/2)

Vaicājums: 6. deklinācijas lietvārdi, pareizrakstības kļūdas

Candidate 2132 (MALE, 42, Krievija) , 3. uzdevums, A līmenis

ar	jums	braukt	ekskursijā	.	Ekskursija	būs	14.	maijā	uz	Vinspili	vai	Liepaāju	.	Cena	ekskursijas	būs	60	eiro	vai	75	eiro
ar	Jums	braukt	ekskursijā	.	Ekskursija	būs	14.	maijā	uz	Ventspili	vai	Liepāju	.	Cena	ekskursijas	būs	60	eiro	vai	75	eiro

Candidate 3112 (FEMALE, 41, Krievija) , 3. uzdevums, B līmenis

Iesim	kopā	uz	džeza	mūzikas	festivālu	?	festivāls	būs	Cesu	koncertzāli	no	17.10	.	līdz	19.11.2015	.	Biļetes	maksā	55	eiro		
Iesim	kopā	uz	džeza	mūzikas	festivālu	?	Festivāls	būs	Cēsu	koncertzāli	no	17.10	.	līdz	19.11.2015	.	Biļetes	maksā	55	eiro		

Candidate 9212 (MALE, 30, Latvija) , 2. uzdevums, C līmenis

Latvija	ir	tik	bagata	ar	dzintaru	ka	viena		Valsta	.	Turistiem	ļoti	interesanti	atvest	to	,	ko	nav	nevienam			
Latvija	ir	tik	bagāta	ar	dzintaru	, kā	ne	neviena	valsts	.	Tūristiem	ļoti	interesanti	un	atvest	to	,	kā	nav	nevienam		

Candidate 9261 (MALE, 25, Latvija) , 2. uzdevums, C līmenis

labs	pamatēdiens	.	Labāk	pamatēdienā	dārzeņus	aizvietot	ar	gaļu	vai	zivim	.											
labs	pamatēdiens	.	Labāk	pamatēdienā	dārzeņus	aizvietot	ar	gaļu	vai	zivīm	.											

Candidate 9454 (MALE, 25, Krievija) , 3. uzdevums, C līmenis

es	nezināju		ko	tieši		par	mani		šie	ļaudi	,	ja	viņi	visi	smaida	.	Tādējādi		teiciens	par	smaidu	
es	nezināju	,	ko	tieši	domā	par	mani	domā	šie	ļaudis	,	ja	viņi	visi	smaida	.	Tādējādi	,	teiciens	par	smaidu	

Kļūdu analīze

Kļūdu tips	Skaits	Kļūdainās teksta vienības (%)	Kļūdu izplatība korpusā (%)
Pareizrakstības kļūdas	14956	47,13%	10,48%
Formveidošanas kļūdas	8075	25,45%	5,66%
Interpunkcijas kļūdas	5857	18,46%	4,10%
Leksikas kļūdas	1756	5,53%	1,23%
Sintakses kļūdas (vārdu secība)	1703	5,37%	1,19%
Nesaprotams teksts	1546	4,87%	1,08%
Sintakses kļūdas	1321	4,16%	0,93%