Digital resources created by or related to the research teams of the applicant and partners

1. Estonian Literary Museum

A. Digital repository of the Estonian Literary Museum KIVIKE

(http://kivike.kirmus.ee/) has been created for the long-term preservation of all the digital collections of the three archives of the Estonian Literary Museum, including digital as well as digitized materials (manuscripts, sound and video recordings, photographs, multimedia files), along with metadata and deciphered texts if available. KIVIKE contains 862431 files with archival materials, 319476 metadata records of archival documents and 602423 metadata records of content units within the archival documents (mainly folklore items). Data records from various databases based on archival materials (see B. and C.) have been aggregated. Kivike enables users to browse, order and submit materials into the archives, and includes also the module for crowdsourcing activities (deciphering, commenting and tagging the archival records).

B. Research databases and text corpora in the field of folklore studies:

Berta: database of Estonian popular calendar holidays – http://www.folklore.ee/Berta/ introduces approx. 80 more or less known holidays. Compiled by M. Kõiva, L. Vesik, T. Särg.

Basic electronic dictionary of Estonian phraseologisms (FES) – http://www.folklore.ee/justkui/sonastik. Coontain 20,749 Estonian phraseologisms with complete semantic, syntactic and morphological description. Compiled by A. Õim and K. Õim.

JUSTKUI: **Database of Estonian sayings and phraseologisms (EKFA) –** http://www.folklore.ee/justkui. Comprised of approx. 160,000 entries. Compiled by A. Baran, A. Hussar, A. Õim, K. Õim.

Sayings and phraseologisms – http://www.folklore.ee/rl/date/robotid/leht3.html. Comprises approx. 25,400 texts of Estonian proverbs and phraseologisms, which originate from the Phraseological Dictionary by A. Õim (Tallinn, 1993), manuscripts from Estonian Folklore Archives and the dialect archive of the Institute of the Estonian Language. Compiled by A. Baran, Tiit Konsand, K. Õim, A. Krikmann.

Eesti vanasõnad [Estonian proverbs] – http://www.folklore.ee/rl/date/robotid/leht1.html. About 13,000 entries represent the original forms of all the known proverbs that were or are in use in Estonian tradition. The database is based on an academic publication: Estonian Proverbs I–IV (1980–1988). Compiled by A. Krikmann.

Eesti maskeerimiskombestik [Estonian masking tradition] – http://www.folklore.ee/Sanditajad/. Compiled by Ü. Tedre.
Estonian riddles – http://www.folklore.ee/moistatused. The database comprising of approx. 100,000 riddles uses for its source similar material that was used in the academic publications on Estonian riddles (2001–2002). Compiled by A. Krikmann.
Mõista, mõista, mõlle-rõlle (Riddle me this, riddle me that) – http://www.folklore.ee/~kriku/MOISTA/**.** Thematic database of classical riddles. Compiled by A. Krikmann.
Estonian conundrums – http://www.folklore.ee/Keerdkys**.** Contains approx. 25,000 conundrums. Compiled by P. Voolaid.
Estonian compound puns – http://www.folklore.ee/Sonamang**.** Comprised of approx. 5000 compound puns. Compiled by P. Voolaid.
Estonian acronyms – http://www.folklore.ee/Lyhendid. Consists of approx. 3000 acronyms. Compiled by P. Voolaid.
Eesti värssmõistatused [Estonian rhyming riddles] – http://www.folklore.ee/Varssmoistatused. Comprised of approx. 2000 texts. Compiled by P. Voolaid.
Eesti piltmõistatused (Estonian Droodles) – http://www.folklore.ee/Reebus **(Estonian Droodles** – http://www.folklore.ee/Droodles**).** Contains approx. 7500 droodles. Compiled by P. Voolaid.
Database of Children's Sayings – http://www.folklore.ee/Lapsesuu. Compiled by P. Voolaid.
Database of modern Estonian anecdotes – http://www.folklore.ee/~liisi/o2/**.** Contains approx. 30,000 Estonian Internet jokes as of 1996. Compiled by L. Laineste.
HERBA: Database of ethno-botany – http://herba.folklore.ee/. Comprises data from older manuscript collections, can be searched by keywords and botanical characteristics. Compiled by R. Sõukand and R. Kalle.
LEPP: South-Estonian tradition portal. Comprises 10,000 folklore texts from Võrumaa and Setumaa. Compiled by M. Kõiva.
RADAR: Map of Estonian cultural history – http://www.folklore.ee/radar. Interactive tradition-based educational material on Estonian cultural history. The primary material of the map is Estonian local folkloristic narratives. With the help of tradition – stories and legends of particular places on local landscape – the inhabitants have identified themselves throughout centuries in their close neighbourhood and established the feeling of "us – them", which is important for the preservation and self-definition of every small society. Compiled by T. Jonuks et al.
Rehepapp: Database of folk belief and legends – http://www.folklore.ee/rehepapp/. More than 10,000 texts on *hiis* sites, plague, underground creatures, malaria, lumbago, Tõnn (god of fertility), fairies, lakes, trees, werewolves, hoards, other mythological creatures, cosmology, spells, mediums. Compiled by M. Kõiva et al.

Graffiti database – http://www.folklore.ee/Graffiti. Typological database comprises approx. 600 paremiological graffiti, recorded within the Estonian-Polish joint project "Creativity and tradition in cultural communication" in 2010–2012. Compiled by P. Voolaid.
Estonian runic songs' database. http://www.folklore.ee/regilaul/. Includes all the runic song (regilaul) texts from older folklore collections. The database contains 83416 song texts (two thirds of all orginally collected runic songs).
Koobas. Estonian Place lore. The database of Estonian place lore. 20 000 texts http://galerii.kirmus.ee/koobas/ is connected with the map layer Estonian Memory Fields (http://www.maastikud.ee) and the aggregate database of the National Heritage Board.
Estonian folk melodies database on-site research database that enables to create automatic typologies of folk melodies. Contains 7000 folk melodies transcribed and analysed, together with rich metadata. User interface needs to be developed.
Estonian Fairy Tales Database (on-site database, web version is created during the year 2015). 10 000 texts (animal tales, tales of magic, religious tales, realistic tales, tales of the stupid ogre, formula tales).
Old and new games from Folklore Archives http://www.folklore.ee/ukauka/ selection of contributions to the children lore collection campaigns of Estonian Folklore Archives in 1935 and 2013.
Three Nations' Humour (Polish-Russian, Estonian-English)
The Ritual Year (English, http://www.folklore.ee/ritualyear/**)**
The European Incantations : Eesti loitsud/Estonian Incantations (Eesti/ English), http://www.folklore.ee/estonian_incantations/
Estonian Legends & Belief Narratives in English

C. Databases and text corpora in the field of cultural history and literary studies:

Archival Library participates in creation of the following united databases:
Estonian united library catalogue ESTER http://www.ester.ee/,
Bibliographic database of Estonian articles ISE http://ise.elnet.ee/,
Estonian national bibliographical database ERB http://erb.nlib.ee/,
Digitized Estonian newspapers database DEA http://dea.nlib.ee/.

There are several original databases created by the Archival Library:
Database of the analytical bibliography of the older periodical publications in the Estonian language - BIBIS http://www2.kirmus/biblioserver/,
Pseudonyms Database ISIK http://galerii.kirmus.ee/biblioserver/isik/
Catalogue of various cultural figures' personal collections
http://www2.kirmus.ee/memoriaal,
Database of older digitized books and other publications GRAFO
http://www2.kirmus.ee/grafo/ ,

Ellen http://galerii.kirmus.ee:8888/ellen/avalik.do - Estonian Cultural History Archive database, contains 223579 archival records.

Kreutzwald's Century: the Estonian Cultural History Web
http://kreutzwald.kirmus.ee/ - Estonian cultural history web portal that gives access to 268 author biographies, more than 10 000 photos and more than 2000 events descriptions based on newspapers material. Also are accessible more than 300 older fiction books in Estonian (in e-pub format), text corpora contains 13 808 pages (24 859 487 characters).

ERNI. Estonian Literature History in texts 1924-1925
http://galerii.kirmus.ee/erni/erni.html
Educational program on history of Estonian literature that contains materials illustrating the Estonian literary world in the 1920s. The database is comprised of library of literary texts, anthology of critical material, photo gallery, author biographies, and meanings of literary terms.

ELMA. Database of Estonian Biographies and Memoirs
http://galerii.kirmus.ee/elma/avaleht/? - gives an overview of Estonian autobiographical literature (memoirs, biographies, diaries, travelogues), its functionality allows to search the materials by their various attributes.

D. Interactive web applications:

Kratt (http://kratt.folklore.ee/) for the collection of folklore (thematic questionaries).
Kaardimasin http://www.folklore.ee/moistatused/kaardimasin/index.php

2. Institute of the Estonian Language

A. Databases, dictionaries, speech and text corpora in the field of linguistic studies:

Dictionary of Standard Estonian (ÕS 2013) http://www.eki.ee/dict/qs/ , compiled by M. Raadik, T. Leemets, S. Mäearu.
Multimodal Basic Estonian Dictionary http://www.eki.ee/dict/psv/, compiled by J. Kallas, K. Koppel, M. Tuulik, I. Hein.
Basic Estonian Dictionary http://www.eki.ee/dict/ekss/, compiled by M. Langemets, T. Valdre, M. Tiits.
Place Names Database (KNAB) http://www.eki.ee/knab/knab.htm, compiled by P. Päll.
Written Text Corpus http://portaal.eki.ee/corpus, compiled by I. Hein.
Estonian Emotional Speech Corpus http://peeter.eki.ee:5000/?lg=en, compiled by H. Pajupuu, R. Altrov, K.Tamuri, J. Pajupuu.

Valence Corpus http://peeter.eki.ee:5000/valence/paragraphsquery, compiled by H. Pajupuu, R. Altrov, K. Tamuri.

Speech Synthesis Corpus http://heli.eki.ee/koduleht/index.php/korpused, compiled by L. Piits, M. Mihkla.

Archive of Estonian dialects https://www.etis.ee/portaal/collection.aspx?kk=1&tab=des&VID=71, compiled by J. Viikberg, L. Raasik

Speech Styles Corpus (forthcoming). Modalities: Spoken Language, Voice, Written text. The texts will be chosen according to their intent: informative, persuasive, entertainment. The speech will be segmented and labelled for the features that are relevant for the research: formality (on the formal/informal dimension), expressiveness (on the dimension of activeness and valency), phrasal prominences and breaks. For each speaker the following data will be recorded: gender, age, education, and region of origin.

Historical Concordance of Estonian Bible Translation

http://portaal.eki.ee/piibel (Manuscripts and published versions of Estonian Bible translations of the 17th and 18h century; respective dictionaries and verse-concordance.)

Keskmurde rahvalaulude korpus (lemmatiseeritud ja morfoloogiliselt märgendatud)

17./18. sajandi põhjaeestikeelsete kirikulaulutõlgete korpus (lemmatiseeritud ja morfoloogiliselt märgendatud).

B. Software

Lexicographer's Workbench EELex http://eelex.eki.ee/ veebipõhiste töövahendite kompleks, mis ühendab sõnastike koostajatele ja toimetajatele vajaliku tarkvara ja keeleressursid, toetab rühmatööd ja pakub eesti keele tuge.

Emotion Detector https://github.com/EKT1/valence/ The Emotion Detector allows to identify the positivity, negativity and neutrality in paragraphs of written text.https://github.com/EKT1/valence/

Emotional speech recognition technology https://github.com/EKT1/emotional/

3. Tallinn University of Technology, Institute of Cybernetics

A. Corpora

BABEL Estonian Database developed within EU Copernicus project in 1995-1998; includes speech recordings of 70 subjects, in total 12 hours. Compiled by E.Meister and A.Eek.
https://metashare.ut.ee/repository/browse/babel-estonian-database/3efcb1ea6feb11e4a6e4005056b40024f9515fbdacc6461097c68580031940d6/

Estonian Speechdat-like Database based on SpeechDat format. Recorded via telephone channels by 1300 subjects, in total 70 hours. Compiled by E.Meister, J.Lasn, L.Meister.

https://metashare.ut.ee/repository/browse/estonian-speechdat-like-database/1ad05f22907711e4a6e4005056b40024615d6469dc2441c6b634c4674005c1d1/

Estonian Foreign Accent Corpus includes speech recordings of non-native Estonian speech. Among speakers (180 subjects) 18 different first language backgrounds are represented; as the reference material a subset of native Estonian speakers (10 male, 10 female) has been recorded. In total 80 hours of speech. Compiled by E.Meister and L.Meister.
https://metashare.ut.ee/repository/browse/estonian-foreign-accent-corpus/f262dc945a5f11e2a6e4005056b400244ec7fd71252140bda0ce6f23e829195e/

Corpus of Lecture Speech includes recordings of academic lectures and oral conference presentations in Estonian. Recorded ca 400 hours in total, ca 60 hours transcribed. Compiled by E.Meister.

https://metashare.ut.ee/repository/browse/corpus-of-lecture-speech/22be9758906311e4a6e4005056b4002401e84da293f14c70b7c8bffaed66676b/

Corpus of Radio News includes news recordings from the Estonian Public Broadcasting. In total ca 300 hours, transcribed ca 50 hours. Compiled by E.Meister.
https://metashare.ut.ee/repository/browse/corpus-of-radio-news/915f8262906d11e4a6e4005056b40024bcd9dc092e1a441aaf45fcfecc4d4277/

Corpus of Radio Interviews includes telephone interviews from different radio programs. In total ca 40 hours, all transcribed. Compiled by E.Meister.
https://metashare.ut.ee/repository/browse/corpus-of-radio-interviews/1dd0a34c906a11e4a6e4005056b40024436e769bc0c448768042d7788f88f592/

Corpus of Adolescent Speech includes speech recordings of 300 native Estonian subjects in the age range from 9 to 18 years; ca 100 hours in total. Compiled by E.Meister and L.Meister. (Link will be available soon).

Articulatory Database includes multimodal articulatory data of two native Estonian subjects recorded with electropalatography (EPG), electroglottography (EGG) and electromagnetic articulography (EMA). Compiled by E.Meister. (Link will be available soon).

B. Software

Real-time speech recognition web service for Estonian
http://bark.phon.ioc.ee/speech-api/v1

Web service for transcribing long speech recordings in Estonian
http://bark.phon.ioc.ee/webtrans/

Kõnele - real-time speech recognition application for Android
https://play.google.com/store/apps/details?id=ee.ioc.phon.android.speak

Diktofon - voice recorder and transcriber for Android
https://play.google.com/store/apps/details?id=kaljurand_at_gmail_dot_com.diktofon

Forced alignment for Estonian speech https://phon.ioc.ee/dokuwiki/doku.php?id=projects:tuvastus:est-align.et

4. University of Tartu

A. Text corpora

Reference Corpus of Written Estonian, 260 million tokens, contains written present-day Estonian full texts. Non-balanced corpus, contains large amounts of newspaper texts but also ca 20 million tokens of Internet language. Annotated for text structure (texts, paragraphs, sentences, tables etc) according to TEI P5 XML standard, annotated files available under CC-BY license. The corpus has also a morphologically annotated version, available on request. The corpus has two user interfaces: http://www.cl.ut.ee/korpused/kasutajaliides/index.php?lang=et enables simple string queries; www.keeleveeb.ee by a close collaborator Filosoft Ltd. enables querying using also morphological information (POS and grammatical categories). Special subcorpus (balanced corpus) has been compiled in order to enable comparing three main text classes of the written language: newspaper texts, fiction and scientific texts.

Corpus of Written Estonian 1890-1990, 6,6 million tokens, a sample corpus containing mostly two text classes - fiction and newspaper texts. Available via user interface at http://www.cl.ut.ee/korpused/kasutajaliides/index.php?lang=et

Corpus of Old Literary Estonian contains written Estonian texts from the 16th to the 19th century. As for coverage, all texts from the 16th century are included, the subcorpus of 17.-18. centuries contains a selection of texts, representing more important registers, authors and works. The subcorpus of 19. century is also a selection representing more important authors; this subcorpus is in continous development. At the moment the Corpus of Old Literary Estonian contains ca 2,3 million tokens and is accessible at via web interface http://www.murre.ut.ee/vakkur/Korpused/

Corpus of Spoken Estonian consists of audio- and video recordings and transcriptions from real everyday and institutional, face-to-face and telephone dialogues. 2 million transcribed tokens (conversation analysis transcription). Metadata: all conversations are documented (information about participants, situation, recordings etc). Corpus has a search engine. Available via administrator for scientific purposes.

Corpus of Computer mediated Dialogues contains Instant Message and Chat room dialogues. Metadata: all conversations are documented (information about participants, situation, recordings etc). Available via administrator for scientific purposes.

Morphologically annotated corpus of Estonian texts (http://www.cl.ut.ee/korpused/morfkorpus/) contains 500 000 tokens, manually annotated and double-checked.

Estonian Dependency Treebank, contains 400 000 tokens. Manually annotated for syntactic roles and dependency relations.

Corpus of Estonian Dialects consists of sound recordings which are transcribed a) phonetically (using Finno-Ugric phonetic transcription system), b) in simplified transcription. The texts in simplified transcription have been morphologically annotated and (partially) syntactically parsed. The corpus includes also a database with metadata (information about informants and recordings). Data included to the corpus consists of texts from Estonian dialects (spoken data), Votic (spoken data + some published texts), Livonian (spoken data + published texts). The main part of the corpus (morphologically annotated texts) is available as a) xml files; b) online database: http://www.murre.ut.ee/mkweb/. No. of annotated words in the online database for Estonian dialects: 948180 text words, Livonian: 22224 text words, 34331 text words. Geographical information is added to the texts, can be used for measuring geographical distances, visualization (maps), etc.

B. Digital collections and databases

The online database of the University of Tartu Archives of Estonian Dialects and Kindred Languages, http://www.murre.ut.ee/arhiiv/. The archives is a digital online collection of Finno-Ugric material at University of Tartu. The archives include 1) (digitalized) sound recordings from Estonian dialects and other Finno-Ugric languages, 2) unpublished manuscripts (transcriptions, written notes on Estonian and other languages, student reports, fieldwork diaries, etc), 3) photos, 4) videos. The archives is accessible online, including recordings and manuscripts.

Digital Text Repository for Older Estonian Literature EEVA (http://www.utlib.ee/ekollekt/eeva/), started 2002, is a joint project of the University of Tartu Library and the Department of Literature and Folklore of the University of Tartu, partnered also by the Estonian Literary Museum. The main goal of EEVA is to make the old rare texts that have played an important role in Estonian cultural history accessible to the users in the most sparing way that would not harm the original copies. EEVA contains texts starting from the 13th century up to the mid-19th century in Estonian, German, Russian, French, Latin, Greek, Latvian, and is mainly engaged in scanning printed texts., trying to contribute to the comparative study of the multilingual Baltic cultural space. As of 17/08/15 EEVA contains 6515 texts (6418 digitised) from 889 authors, the total of 76014 scanned text files.

EWOD (Estonian Writers Online Dictionary) will present concise bio-bibliographical data of Estonian literature published in foreign languages, to fill a void that has until now considerably restrained the projection of Estonian literature outside Estonian-language cultural space and in world literature.

"Estonian verse" (http://www.ut.ee/verse) is a growing environment for presenting information on poetry and theory of verse. There are four statistical modules added to this site at this point, but there are several other modules already prepared and ready to be uploaded to the site (for instance, the database of Estonian Aeolic verse, the database of 17th century hexameters written in Tartu, the database of Uku Masing's metrical and rhythmical indices). These modules are so far the most comprehensive databases on Estonian verse. Together the modules form the first Internet database on Estonian verse culture, being at the same time the world's first verse theoretical page of such scope, containing real data and the possibilities to process the data online.

Terminological database of narratology and text theory contains 140 final records, with another 100-150 records potentially sieved from collected material. All records are relatively long (approximately 500-600 words). The structure of a terminological entry: definition(s) of the term, matches in foreign languages, interpretation by different authors in different traditions, examples and practical application of the term, list of bibliographical sources. Cross-referencing is actively implemented: connected terms form semantic nests or clasters.

Estonian Wordnet (http://www.cl.ut.ee/ressursid/teksaurus) is a digital thesaurus; the basic unit of a wordnet-type thesaurus is a synonym set (also called a synset), which is a set containing all the synonymous words or multi-word units that express

the same concept. The synsets are connected by links which correspond to semantic or lexical relations between concepts. The most important relations are hyponymy and hypernymy, but also meronymy, holonymy, antonymy, cause, role, derivational relations are marked. At the moment EWN contains ca 65 000 concepts and is a work in progress.

Database of Estonian Multiword Expressions (http://www.cl.ut.ee/ressursid/pysiyhendid) contains a subtype of multi-word expressions, namely those consisting of a verb and a particle or a verb and its complements. The present version of the database contains ca 13 000 expressions. It is available for download and also has a simple web interface.

Frequency lists of word-forms, lemmas, n-grams and collocations, also of parts of speech and grammatical categories derived from the balanced subpart of the Estonian Reference Corpus (http://www.cl.ut.ee/ressursid/). All frequencies are calculated for the whole subcorpus (15 million words) and separately for fiction (5 million), newspaper (5 million) and scientific (5 million) texts.

C. Software applications

Constraint Grammar Dependency Parser for Estonian is a rule-based software for syntactic (dependency) analysis of Estonian texts, performs also morphological disambiguation. Performance: LA 80.7% (syntactic label attachment score), LAS 69.4% (syntactic label plus dependency relation attachment score), UAS 77.2% (dependency relation attachment score). Web interface https://korpused.keeleressursid.ee/syntaks/

MaltParser version trained for Estonian. A version of MaltParser (http://www.maltparser.org/), a machine learning based software for syntactic (dependency) analysis trained on the Estonian Dependency Treebank. Performance: LA (syntactic label attachment score) 84.6%, LAS (syntactic label plus dependency relation attachment score) 76.6%, UAS (dependency relation attachment score) 81.0%. Web interface https://korpused.keeleressursid.ee/syntaks/