CLARIN 2021
Annotation and Acquisition Tools

Chair: António Branco
Day 3
Wednesday 29 September
12:30 - 13:10
A Method for Building Non-English Corpora for Abstractive Text Summarisation

Julius Monsen & Arne Jönsson

CLARIN 2021, 27–29 SEPTEMBER 2021
Overview and main findings

• A filtered corpus with properties similar to the CNN/Daily Mail corpus
• Fine-tuned models on four differently filtered corpora
• Evaluation on two test sets (filtered to different degrees)
• Highly filtered corpora → higher evaluation scores
• Results comparable to results on CNN/Daily Mail
Summarisation results with various filters

- Highly filtered testset (test-SN) - **most important factor**

- Best scores with the most filtered corpora (DN-SN)
  - **ROUGE-1**: 37.22 (39.89 on CNN/DailyMail)
  - **ROUGE-2**: 16.32 (18.18 on CNN/DailyMail)

<table>
<thead>
<tr>
<th></th>
<th>ROUGE-1</th>
<th></th>
<th>ROUGE-2</th>
<th></th>
<th>ROUGE-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN/DailyMail</td>
<td>39.89</td>
<td>18.18</td>
<td>27.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DN-LC</td>
<td>29.08</td>
<td>35.46</td>
<td>9.67</td>
<td>14.71</td>
<td>20.23</td>
</tr>
<tr>
<td>DN-S</td>
<td>28.44</td>
<td>36.97</td>
<td>8.98</td>
<td>15.79</td>
<td>19.48</td>
</tr>
<tr>
<td>DN-N</td>
<td>27.42</td>
<td>36.96</td>
<td>8.42</td>
<td>16.17</td>
<td>18.74</td>
</tr>
<tr>
<td>DN-SN</td>
<td>26.48</td>
<td><strong>37.22</strong></td>
<td>7.44</td>
<td><strong>16.32</strong></td>
<td>17.84</td>
</tr>
</tbody>
</table>
First filtering stage

- Filters
  - Length (in words)
  - Compression ratio (summary length/article length)
- Much larger than CNN/DailyMail
- Using preambles as summaries
  - Higher novelty
  - Less semantically similar

<table>
<thead>
<tr>
<th></th>
<th>DN-LC</th>
<th>CNN/DailyMail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus size</td>
<td>802,405</td>
<td>311,971</td>
</tr>
<tr>
<td>Vocabulary size</td>
<td>2,575,130</td>
<td>803,487</td>
</tr>
<tr>
<td>Occurring 10+ times</td>
<td>387,370</td>
<td>161,820</td>
</tr>
<tr>
<td>Article words</td>
<td>370.41</td>
<td>677.21</td>
</tr>
<tr>
<td>Article sentences</td>
<td>24.00</td>
<td>28.52</td>
</tr>
<tr>
<td>Summary words</td>
<td>29.31</td>
<td>48.34</td>
</tr>
<tr>
<td>Summary sentences</td>
<td>2.19</td>
<td>3.70</td>
</tr>
<tr>
<td>Compression ratio</td>
<td>0.13</td>
<td>0.09</td>
</tr>
<tr>
<td>Novelty (uni-gram)</td>
<td>0.42</td>
<td>0.14</td>
</tr>
<tr>
<td>Novelty (bi-gram)</td>
<td>0.80</td>
<td>0.57</td>
</tr>
<tr>
<td>Novelty (tri-gram)</td>
<td>0.93</td>
<td>0.77</td>
</tr>
<tr>
<td>Semantic similarity (doc/doc)</td>
<td>0.49</td>
<td>0.65</td>
</tr>
<tr>
<td>Semantic similarity (doc/sent)</td>
<td>0.52</td>
<td>0.67</td>
</tr>
</tbody>
</table>
Second filtering stage

- Filters
  - Novelty (N)
  - Semantic similarity (S)
- Previous methods
  - MLSUM

<table>
<thead>
<tr>
<th></th>
<th>DN-S</th>
<th>DN-N</th>
<th>DN-SN</th>
<th>CNN/DailyMail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus size</td>
<td>122,419</td>
<td>124,105</td>
<td>38,151</td>
<td>311,971</td>
</tr>
<tr>
<td>Vocabulary size</td>
<td>727,406</td>
<td>1,070,351</td>
<td>435,412</td>
<td>803,487</td>
</tr>
<tr>
<td>Occurring 10+ times</td>
<td>118,661</td>
<td>171,100</td>
<td>70,450</td>
<td>161,820</td>
</tr>
<tr>
<td>Article words</td>
<td>362.52</td>
<td>630.36</td>
<td>512.43</td>
<td>677.21</td>
</tr>
<tr>
<td>Article sentences</td>
<td>22.51</td>
<td>40.27</td>
<td>31.71</td>
<td>28.52</td>
</tr>
<tr>
<td>Summary words</td>
<td>32.15</td>
<td>33.16</td>
<td>35.67</td>
<td>48.34</td>
</tr>
<tr>
<td>Summary sentences</td>
<td>2.38</td>
<td>2.51</td>
<td>2.62</td>
<td>3.70</td>
</tr>
<tr>
<td>Compression ratio</td>
<td>0.13</td>
<td>0.07</td>
<td>0.10</td>
<td>0.09</td>
</tr>
<tr>
<td>Novelty (uni-gram)</td>
<td>0.32</td>
<td>0.14</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>Novelty (bi-gram)</td>
<td>0.73</td>
<td>0.57</td>
<td>0.57</td>
<td>0.57</td>
</tr>
<tr>
<td>Novelty (tri-gram)</td>
<td>0.89</td>
<td>0.78</td>
<td>0.78</td>
<td>0.77</td>
</tr>
<tr>
<td>Semantic similarity (doc/doc)</td>
<td>0.65</td>
<td>0.52</td>
<td>0.65</td>
<td>0.65</td>
</tr>
<tr>
<td>Semantic similarity (doc/sent)</td>
<td>0.67</td>
<td>0.60</td>
<td>0.67</td>
<td>0.67</td>
</tr>
</tbody>
</table>
Conclusion

• Useful for building high-quality corpora
• Extension of the CNN/Daily Mail corpus
• The DN corpus will be freely available as a SweCLARIN resource
Enhancing CLARIN-DK Resources While Building the Danish ParlaMint Corpus

Bart Jongejan, Dorte Haltrup Hansen, Costanza Navarretta
University of Copenhagen, Denmark
2021.09.29
TASK: NLP processing TEI P5 input

- Enrich TEI P5 text with NER, morpho-syntactic descriptions, lemmas & dependency structures.

- Input: TEI P5 containing plain text within <seg> elements. To comply with the requirements in an earlier project, the NLP workflow must deal with and preserve <add>, <del>, <ex>, <lem>, <app> and <rdg> elements.

- Output: the same TEI P5, enriched with <s>, <w>, <pc>, <name>, <linkGrp> and <link> elements. @lemma and @msd in <w> tags.

- Everywhere following ParlaMint TEI P5 annotation schema.
SOLUTION: add/adapt NLP tools in Text Tonsorium

- Text Tonsorium (TT) is a workflow manager that automatically computes candidate workflows by selecting NLP tools from its toolbox.

- Handles many different file formats. New ones can be added very easily.

- Tool integration by wrapping in automatically generated PHP code.

- Handles stand-off as well as in-line annotation layers.

- Annotation layers can be, but don’t need to be, in different files.

- Already handles many languages. For Danish, also three time periods.

- Open source & still improving.

- Runs on CST’s servers. A complete TT instance can be run in WSL just fine.

- https://clarin.dk/clarindk/tools-texton.jsp
Enhancements to CLARIN-DK Resources

- CSTNER and dapipe (Danish UDPipe): capability to handle TEI P5 I/O.
- Conversion of TEI P5 to graphical presentation in HTML+SVG via CoNLL-U.
- Segmentation/Tokenisation: handling of sentence-final abbreviations. Recognition of 18 multicharacter name initials such as Chr. and Joh.
- Dapipe: the lemmas it produces do not agree with the PoS it also produces. Used CSTlemma instead.
- CSTlemma: added support of Universal PoS tags, as produced by dapipe.
- CSTNER: we made sure it recognised all names of political parties and PMs. Some bugs were found and fixed.
Pilot evaluation

5058 tokens from one randomly extracted debate.

- Sentence segmentation errors (2 out of 229 sentences): full stop after numeral identified as part of number (ordinal).

- PoS errors, 4.4% of data: ambiguous words, e.g. auxiliary/main verb readings of *have* (have), *være* (be) and *blive* (be/become), *det* (it/the), *som* (like/which), *der* (there/that), abbreviations tagged as proper names and proper names tagged as substantives.

- Lemmatization: errors derived from wrong PoS, but also errors in some forms of common verbs, e.g. *være* (be) and *synes* (think) (independent from PoS errors 1.2%).

- NER (155 NER): four errors following PoS errors (proper names tagged as substantives).
Tools used in the Danish ParlaMint workflow.

- TEI tokenizer – creates <w> and <c> elements based on simple rules.
- Sentence extractor – creates <span> elements in a <spanGrp>, each <span> indicating start and end of a sentence.
- Token extractor – idem for tokens. A token can comprise several <w>s.
- TEI-segmenter – defines sentences in terms of ‘true’ tokens.
- Dapipe – the Danish version of UD-pipe. (To be replaced by UD-pipe.)
- Anno-splitter (three times) – utility to split dapipe’s output.
- CSTlemma – CST’s lemmatiser. Handles affixes, not just suffixes.
- CSTner – CST’s rule based NER software.
- TEI annotator – merges all annotations into input of the workflow.
Creating an Error Corpus: Annotation and Applicability

Þórunn Arnardóttir, Xindan Xu, Dagbjört Guðmundsdóttir, Lilja Björk Stefánsdóttir and Anton Karl Ingason

CLARIN 2021, 27–29 SEPTEMBER 2021
The Icelandic Error Corpus

• Modern Icelandic error corpus
• Roughly 57,000 categorized error instances
• Three text genres
• Published under a CC BY 4 license at the Icelandic CLARIN repository: http://hdl.handle.net/20.500.12537/105
• Created to guide the development of an open-source Icelandic spell and grammar checker, GreynirCorrect
Data

• Three text sources to reflect different styles of writing
  - Student essays
    • Students 16–20 years of age
    • Sentences shuffled to comply with original license
  - Online news
    • Written between 2004 and 2014, selected randomly
  - Wikipedia articles
    • Selected randomly
• Texts published as part of the Icelandic Gigaword Corpus
• All texts published anonymously

Annotation Process

- Five steps resulting in augmented TEI-format XML documents:
  1. Text cleanup
  2. Manual proofreading
  3. Conversion to TEI-format XML
     - Corrections explicitly marked using a revision span
  4. Manual error code labeling
     - All errors assigned an error code
     - Annotators separate to proofreaders
  5. Format checking
Annotation Scheme

- Descriptive scheme created for the error corpus
- Three levels:
  - Main categories
    - Six in total
  - Subcategories
    - 31 in total
  - Error codes
    - 253 in total
- Error codes used during annotation
- Subcategories reflect error types in general, e.g. agreement errors, typographical errors, etc.
- Revision of the scheme done throughout annotation
Statistical Information

- 4,044 files with 56,794 categorized error instances
- 45.76 errors per 1000 words
- Most common subcategories:
  - Punctuation
  - Wording
  - Spacing
  - Nonword
  - Typo

<table>
<thead>
<tr>
<th>Subcorpus</th>
<th>Files</th>
<th>Revisions</th>
<th>Categorized Errors</th>
<th>Errors/1000w</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Development corpus</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student essays</td>
<td>158</td>
<td>4,719</td>
<td>5,947</td>
<td>37.83</td>
</tr>
<tr>
<td>Online news</td>
<td>2,638</td>
<td>15,969</td>
<td>19,579</td>
<td>35.74</td>
</tr>
<tr>
<td>Wikipedia articles</td>
<td>881</td>
<td>20,216</td>
<td>26,786</td>
<td>62.03</td>
</tr>
<tr>
<td><strong>Test corpus</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student essays</td>
<td>18</td>
<td>645</td>
<td>828</td>
<td>43.30</td>
</tr>
<tr>
<td>Online news</td>
<td>267</td>
<td>1,334</td>
<td>1,663</td>
<td>32.74</td>
</tr>
<tr>
<td>Wikipedia articles</td>
<td>82</td>
<td>1,385</td>
<td>1,991</td>
<td>58.03</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>4,044</td>
<td>44,268</td>
<td>56,794</td>
<td>45.76</td>
</tr>
</tbody>
</table>
Conclusion

- The Icelandic Error Corpus
- 56,794 errors
- Three text genres
- Three-level hierarchical annotation scheme
- Used for improving a spell checker
Questions and Answers

CLARIN ANNUAL CONFERENCE
Five-minute paper presentations - chairs: A. Branco, R. Simov
28/09/2021 - 12:30

- ABSTRACTIVE TEXT SUMMARISATION
- ANNOTATION & ACQUISITION TOOLS
- ALEXIA
- PARLAMINT & DANISH CLARIN
- CLARIN-1S
- NATIONAL CLARIN CENTRES
- SWELL PORTAL
- ICELANDIC ERROR CORPUS
- CLARIN & BELARUSIAN
- CLARIN FLANDES
Reliability of automatic linguistic annotation: native vs non-native texts

Elena Volodina\textsuperscript{1}, David Alfter\textsuperscript{1}, Therese Lindström Tiedemann\textsuperscript{*}, Maisa Lauriala\textsuperscript{*}, Daniela Piipponen\textsuperscript{*}

\textsuperscript{*}=University of Helsinki, \textsuperscript{1}= University of Gothenburg
“L2 profiles”

Development of Lexical and Grammatical Competences in Immigrant Swedish

Project financed by
Riksbankens jubileumsfond (2018–2020 (2021))
P17-0716:1

Collaboration:
Swedish language, University of Gothenburg
“Nordica”, Dept of Finnish, Finnougrian and Scandinavian Studies, University of Helsinki

PI: Elena Volodina, Gothenburg

Homepage:
https://spraakbanken.gu.se/en/projects/l2profiles
Swedish (L2) Profile

– **lexical, grammatical, morphological.**

https://spraakbanken.gu.se/larkalabb/svlp
Manual checking

**Sparv pipeline** (Borin et al. 2016)
- Lemmatization
- Part-of-speech tagging
- Multi-word expression detection
- Word sense disambiguation
- Dependency parsing

- Course book texts
- Learner essays
- Corrected Learner essays

15 texts per dataset;
3 texts per proficiency level
Hypotheses

1. Pipelines trained on a standard language (L1) do not perform as well on non-standard language varieties such as learner language (L2)

2. Normalization of non-standard language, e.g. through error correction, improves tool performance

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Lemma</th>
<th>PoS</th>
<th>DepRel</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 Coctaill</td>
<td>0.93±0.0</td>
<td>0.98±0.0</td>
<td>74.89</td>
</tr>
<tr>
<td>L2 orig</td>
<td>0.90±0.02</td>
<td>0.95±0.0</td>
<td>63.01</td>
</tr>
<tr>
<td>L2 norm</td>
<td>0.93±0.02</td>
<td>0.97±0.0</td>
<td>69.02</td>
</tr>
</tbody>
</table>

Table 2: Lemmatization and PoS tagging: accuracy and standard deviation; Dependency: LAS

<table>
<thead>
<tr>
<th></th>
<th>#tokens excl punct</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>WSD Accuracy (corr/tot)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 Coctaill</td>
<td>1900</td>
<td>0.80</td>
<td>0.71</td>
<td>0.75</td>
<td>0.84±0.03</td>
</tr>
<tr>
<td>L2 orig</td>
<td>3635</td>
<td>0.90</td>
<td>0.72</td>
<td>0.80</td>
<td>0.82±0.07</td>
</tr>
<tr>
<td>L2 norm</td>
<td>3565</td>
<td>0.85</td>
<td>0.78</td>
<td>0.81</td>
<td>0.83±0.04</td>
</tr>
</tbody>
</table>

Table 3: Number of correctly identified MWEs including precision, recall and F1 score and automatic word sense disambiguation (WSD)
Conclusions

Reliable?

- yes, for lemmatization, POS tagging, WSD & MWE detection
- not really for DepRel

Hypotheses?

1. yes, in general, but depend on linguistic features
2. yes, even if sometimes marginally
Thank you!

Questions? Comments? Suggestions?
Annotation Management Tool: A Requirement for Corpus Construction

Yousuf Ali Mohammed, Arild Matsson and Elena Volodina

CLARIN 2021, 27–29 SEPTEMBER 2021
Swell Infrastructure

Swedish Learner Language
https://spraakbanken.gu.se/en/projects/swell
Data collection and Data management

- Source
- Student
- Task
- Essay

- Data collection
- Import essay

- Metadata

- Task Manager
- Anonymization
- Normalization
- Correction Annotation

- Task
- Users
- Task State

- Export Data

Task Object
Data collection

- Source
- Student
- Task
- Essay

Metadata

Import essay

Data collection

Raw Text

XML
Data management

Task Manager

Users

Task State

Task

Anonymization

Normalization

Correction Annotation

Export Data
Export dataset

Task type
- Anonymized
- Normalized
- Correction anno

School
- A
- B
- C

L1
- English
- Spanish
- Arabic

Data type
- Source
- Target
- Both

Status
- Complete
- All

File format
- XML
- JSON
- Text
Thank you!

Comments? Questions?
ALEXIA
A Lexicon Acquisition Tool

Steinunn Rut Friðriksdóttir, Atli Jasonarson, Steinþór Steingrímsson and Einar Freyr Sigurðsson

CLARIN 2021, 27–29 SEPTEMBER 2021
Purpose

- A corpus tool that facilitates compilation of larger and better lexical resources, and research using the Icelandic Gigaword Corpus (IGC)\(^1\)

- Easier detection of data gaps in existing databases

- Useful in sociolinguistic research

---

Design

- Designed to be used with the IGC but is not limited to it
- Default settings involving either of two well-known Icelandic language resources, the Database of Icelandic Morphology or the Dictionary of Contemporary Icelandic
- User settings can include any word list and plain text corpus
- Applies POS-based filters
- Uses a stop-word list of 60k tokens, manually collected by us from the IGC
Candidate Lists

- Various options depending on the goal, e.g.
  - collocations show potential fixed expressions
  - text type frequencies to see if a word is a part of a specialized vocabulary
  - singular and plural frequencies for nouns indicates if a word only exists in one form

Please choose a type of output candidate list.
Enter the number corresponding to your choice and press ENTER.

1 Frequency list (lemma, freq) including information on whether a noun appears more frequently in the singular or plural.
2 Frequency list including all word forms appearing with the lemma (lemma, freq [word form])
3 Frequency list including individual frequencies from various text types (lemma, total freq: maths, freq, news, freq...)
4 Frequency lists including up to 5 collocations for each candidate (lemma : total freq [five word sentence example])

Enter number:
Conclusion

- The aim is to facilitate the creation and expansion of lexical resources
- Future work could expand the candidate options and include a GUI
- ALEXIA is available on CLARIN-IS and on Github with an Apache 2.0 license and we encourage anyone interested to use and modify it as they wish
CLARIN 2021
National CLARIN Centres

Chair: Kiril Simov

Day 3
Wednesday 29 September
13:10 - 13:25
National Language Technology Infrastructure Initiative on CLARIN-IS

CLARIN-IS is a fairly new CLARIN member currently building a broad collection of national language resources for use in language technology.

As a CLARIN C-centre, CLARIN-IS is hosting metadata for various text and speech corpora, lexical resources, software packages and models.

The providers of the resources are universities, institutions and private companies working on a national Icelandic LT infrastructure initiative.
Core Projects – Language Resources

Text corpora
Parallel text corpora
Lexicographic resources
Speech corpora – crowdsourced
Speech corpora – high quality recordings
Core Projects – NLP tools

Tokeniser
Part-of-speech taggers
Parsers
Lemmatiser
Named Entity Recogniser
Semantic Analysis
Core Projects – Infrastructure Software

Speech recognition – Speech synthesis – Machine Translation – Spell and grammar checking

For all infrastructure projects, the aim is to publish open environments for adaption and further research and development, as well as to deliver software packages that can be integrated into other products. The delivered software will be designed for defined domains, with instructions for adaption to other domains.
Standards and Deliverables

Aim is to publish the majority of the language resources under the CC BY 4.0 license.

All delivered software will be published under the Apache 2.0 license, or a comparable license.

Data and software development guidelines and quality standards are coordinated across all projects.

All deliverables will be stored and distributed by CLARIN.
CLARIN Knowledge Centre for Belarusian Text and Speech Processing (K-BLP)

Yuras Hetsevich, Jauheniya Zianouka, David Latyshevich, Mikita Suprchuk, Valery Varanovich and Katerina Lomat

Minsk, Belarus
kcenter@clarin-belarus.corpus.by
Main aim – is to collect and develop the text and speech processing tools in all levels of the Belarusian language

Main web page of Belarusian CLARIN K-center is https://clarin-belarus.corpus.by/.

Certificate of recognition issued on 10 February 2020.
The Clarin Knowledge Centre for Belarusian text and speech processing provides users knowledge for

- text processing (corpora, telemetry, udc, health and etc data),
- speech processing (real speech, emotions, intonations, pathological speech etc)
- localization and modernization of web-portals and mobile systems
- computer-assisted rehabilitation and educational systems
- other data processing for languages, especially for the Belarusian language
www.corpus.by – the main core web platform of Clarin Knowledge Centre for Belarusian text and speech processing

Basic Principles

- “1 click on start – 1 instant result”
- “ready to use tool”
- “everything has been saved”
- “towards to open source tools”
58 tools and services are provided for CLARIN VLO

https://vlo.clarin.eu/search?0&q=uiip
Example service: Text-to-Speech Synthesizer

Example service: Voiced Electronic Grammatical Dictionary

- spelling
- transcription (classical and IPA format)
- part of speech
- how it sounds (using Text-to-speech)

Example service : Pitch Plotter

Pitch Plotter service allows a user to get graphical image of the pitch frequency contour of a speech phrase online.

Groups of www.corpus.by services

- Proofreading
- UDC
- Writer
- Linguist
- Programmer
- Other

- <1%
- All
70 web services
for Belarusian, Russian and English processing (and growing!)

Allophone Frequency Counter
Allophone Plotter
Allophonic Phrase Plotter
Alphabetical Subject Index Generator
Alphabetizer
Character Frequency Counter
Character Information Generator
Common Slavic Linguistic Atlas Generator
Dialectological Maps
Electronic Keyboard
Emotion Speech Analyzer
Grammatical Dictionary Processor
Homograph Identifier
Intonation Recognizer
Intonational Processor
Intonational Processor Py
Keyboard Layout Converter
Language Identifier
Language Tutor
Lemmatizer
Morse Code Converter
Morse Code Recognizer
Music Recognizer
N-gram Frequency Counter
Nearest Words Finder
Numeric Expressions Processor
Orthoepic Dictionary Generator
Part-of-Speech Tagger
Phonetic Minimizer
Phonetic Phenomena Searcher
Pitch Plotter
Publication Reference Generator
QR Code Generator
RSS Reader
Romanizator
Service Demonstration
Service Demonstration Py
Service Demonstration With Authorization
Short U Spell Checker
Sound Recorder
Speaker Recognizer
Spreadsheet Dictionary
Speech Duration Predictor
Speech Recognition Trainer
Speech Segmentation Checker
Spell Checker
Syllabifier
Table Processor
Tag Identifier
Talking Head Synthesizer
Text Summarizer
Text-to-Speech Synthesizer
Thematic Lists Collector
Thematic Speech Recognizer
Tokenizer
Transcription Generator
Translator
TranslatorPy
UDC Code Finder
UDC Decoder
Unknown Words Processor
Voice Activity Detector
Voice Pathology Detector
Voiced Electronic Grammatical Dictionary
Web Page Content Saver
Word Frequency Counter
Word Paradigm Generator
...
1) Speech Synthesis and Recognition Laboratory UIIP NAS Belarus (K-BLP Clarin center) + Belarusian State University:
   - provides courses on computer linguistics, programming, data processing
   - provides students’ internship
2) Lab 0: How to be acquainted with text and speech processing services in 10 days? (free available resource about text and speech processing on Corpus.by)
3) Nooj corpus manager for Belarusian: Video tutorial on the Youtube
   (https://www.youtube.com/playlist?list=PLtc_R9i0zr6QiyLk5_Vn_9F4balHK42XW)
MAIN PLANS

1. To attract other scientific organizations and institutes with research centers for computer processing of the Belarusian language to widen K-BLP.
2. To expand K-BLP with new resources: corpora, dictionaries, processing services, etc.
3. To optimize existing resources and tools in K-BLP according to CLARIN classification of resources.
4. To organize the overviews of developed Belarusian tools.
5. To provide a user-friendly overview of the available Belarusian language tools in the CLARIN infrastructure.
CLARIN-Flanders

New prospects

Vincent Vandeghinste
Els Lefever
Walter Daelemans
Tim Van de Cruys
Sally Chambers
A new member to CLARIN.eu: CLARIN-BE

Given that:

- CLARIN members need to be countries or international organisations
- Flanders is not a country

Ergo:

- Flanders can not become a member of CLARIN.eu

Solution:

- Apply for political support from Flanders (no funding) for the formal founding of CLARIN-BE and the membership fee
- Support was granted, membership is now a fact (since September 2021)
CLARIN-VL Consortium

- CLARIN-B centre for CLARIN-BE and CLARIN-VL
- CLARIN-K Centre K-Dutch for the Dutch language
- Vincent Vandeghinste, national coordinator CLARIN-BE

- Language and Translation Technology Team (LT3), Els Lefever, CLARIN-VL principal investigator
- Internet Technology and Data Science (imec-IDLab)

- Centre for Computational Linguistics and Psycholinguistics (CLIPS), Walter Daelemans
- Centre for Computational Linguistics (CCL), Tim Van de Cruys
- Language Intelligence and Information Retrieval Lab (LIIR), Sien Moens
- ESAT Processing Speech and Images (ESAT-PSI), Patrick Wambacq

OPEN FOR NEW RESEARCH GROUPS
CLARIAH-VL: Plans

- **Digital Text Analysis Dashboard and Pipeline for**
  - Dutch texts
  - Parallel texts
  - re-using existing CLARIN tools
  - extended with variation of models and NLU analysis layers

- **Benchmarking existing tools and new models**
  - Linguistic processing tools (pos, lemma, ner, ...): LeTs Preprocess, Frog, DeepFrog, SpaCy, Stanza
  - Natural Language Understanding: development of a suite of NLU evaluation tasks (cf. GLUE) for Dutch
  - Explore application of neural network models for search and extraction of linguistics structures

- **Integrate NLU tools in pipeline**
  - Sentiment analysis, Emotion detection, Document similarity clustering, Topic modelling, Stylometry

- **Integrate Tools for multilingual or parallel data**
  - Sentence alignment, Word alignment

- **Setup of Use Cases**

- **Setup of a Help Desk**
Breakout Rooms

• Room 1: Annotation and Acquisition Tools (Antonio Branco)
• Room 2: Repositories and National CLARIN Centres (Neeme Kahusk and Kiril Simov)
• Room 3: Research Data Management, Metadata and Curation (Part 2) (Jurgita Vaičenonienė)
• Room 4: Legal Issues Related to the Use of LRs in Research (Part 2) (Krister Lindén)
• Room 5: Breakout Room for CLARIN Committees (Franciska de Jong)