CLARIN 2021 PhD Student Session

Moderator: Darja Fišer

Day 1
Monday 27 September
13:30 - 14:30
What and How?

Aims of the Session

- CLARIN community learns about the next generation of research supported by or contributing to the infrastructure
- PhD students receive feedback on their work from CLARIN experts

Before the Session

- Abstracts with description of their PhD project published on the conference programme webpage
- Clearly articulate how the PhD work makes use of the CLARIN infrastructure or how the results will contribute to CLARIN

During the Session

- Presentations
- Plenary Q&A session
- In-depth discussions in breakout rooms

After the Session

- Invited to submit a full paper to be reviewed for the post-conference proceedings.
Programme

13:30-13:35 Wiiktoria Mieleszczenko-Kowszewicz
Reflecting Cognitive Processing of Trauma in Language - a Result of Trauma’s Experience Corpus Analysis with a Usage of Words’ Meanings Category

13:35-13:40 Antonio Pallotti
Machine Learning Applied to Voice Signal in Parkinson’s Disease

13:40-13:45 Zainab Almugbel
Building English-Arabic Parallel Medical Corpora

13:45-13:50 Fabio Ardolino
Attitudes and Language Acquisition: An Investigation on the Italian-French Community in Aix-Marseille

Towards the Precise Detection of Adverbial Roles in Hungarian – Manual Clustering of Adverbial Adjuncts

13:55-14:00 Alina Tsikulina
Tough Constructions and their Analogs in English, French and Russian: A Parallel Corpus Study Using the CLARIN VLO platform

14:00-14:05 Erica Biagetti
The RigVeda Goes ‘Universal’: Annotating Comparative Constructions in the Most Ancient Poetry of India

14:05-14:15 Plenary Q&A

14:15-14:30 Follow-Up Discussion in Breakout Rooms
Reflecting Cognitive Processing of Trauma in Language - a Result of Trauma’s Experience Corpus Analysis with a Usage of Words’ Meanings Categories

Wiktoria Mieleszczenko-Kowszewicz
SWPS University of Social Sciences and Humanities, Warsaw, Poland
Polish-Japanese Academy of Information Technology, Warsaw, Poland

Day 1
Monday 27 September
13:30 - 13:35
Insight and Causation, Blaming and Acceptance Categories as the Indicators of Cognitive Processing of Trauma
Definition of Words’ Meanings Categories Indicating Cognitive Processing of Trauma

**Acceptance**: Meanings indicating acceptance and coming to terms with the situation and the reality that followed, e.g. accept, reconcile, submit.

**Blaming**: Meanings indicating the attribution of blame to someone (oneself) for an event, e.g. blame.

**Insight**: Meanings indicating human intellectual activity related to understanding and finding the meaning of an event, e.g. “I think this event changed me.”

**Causation**: Meanings indicating the mental consequences of an event, e.g. “Since that event, I have changed my attitude toward life because I know that I cannot control everything.”
61 people who experienced a traumatic event were asked to describe it. Participants included 48 women and 13 men. Analysis of the corpora with CPT categories results in the following numbers:

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
<th>M</th>
<th>SD</th>
<th>Mdn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insight</td>
<td>1004</td>
<td>1.22</td>
<td>0.56</td>
<td>1.20</td>
</tr>
<tr>
<td>Causation</td>
<td>3611</td>
<td>4.22</td>
<td>0.85</td>
<td>4.13</td>
</tr>
<tr>
<td>Blaming</td>
<td>12</td>
<td>0.01</td>
<td>0.7</td>
<td>0</td>
</tr>
<tr>
<td>Acceptance</td>
<td>12</td>
<td>0.02</td>
<td>0.67</td>
<td>0</td>
</tr>
</tbody>
</table>
Machine Learning Applied to Voice and Handwriting Signal in Parkinson’s Disease

Antonio Pallotti

Department of Management and Law, University of Rome Tor Vergata

Day 1
Monday 27 September
13:35 - 13:40
Voice Signal

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Type</th>
<th>Accuracy (%)</th>
<th>TPR (%)</th>
<th>TNR (%)</th>
<th>F1 Score</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSL</td>
<td>KNN</td>
<td>97.3 ± 0.6</td>
<td>99.0 ± 0.6</td>
<td>93.8 ± 1.6</td>
<td>0.97 ± 0.006</td>
<td>0.97 ± 0.007</td>
</tr>
<tr>
<td>IBk</td>
<td>KNN</td>
<td>95.4 ± 0.6</td>
<td>98.2 ± 0.5</td>
<td>89.3 ± 1.7</td>
<td>0.94 ± 0.005</td>
<td>0.94 ± 0.009</td>
</tr>
<tr>
<td>MLP</td>
<td>ANN</td>
<td>87.6 ± 0.8</td>
<td>91.1 ± 1.1</td>
<td>74.5 ± 2.9</td>
<td>0.84 ± 0.011</td>
<td>0.91 ± 0.007</td>
</tr>
<tr>
<td>SMO</td>
<td>SVM</td>
<td>79.0 ± 0.8</td>
<td>89.4 ± 1.1</td>
<td>27.6 ± 4.5</td>
<td>0.68 ± 0.011</td>
<td>0.74 ± 0.009</td>
</tr>
<tr>
<td>Matlab SVM</td>
<td>SVM</td>
<td>91.2 ± 3.8</td>
<td>96.8 ± 3.9</td>
<td>80.8 ± 7.5</td>
<td>0.90 ± 0.039</td>
<td>0.95 ± 0.038</td>
</tr>
</tbody>
</table>

Handwriting Signal

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Specificity</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spiral</td>
<td>Linear SVM</td>
<td>74.2 %</td>
<td>82.4%</td>
</tr>
<tr>
<td>Ie</td>
<td>Fine KNN</td>
<td>80.6 %</td>
<td>82.4%</td>
</tr>
<tr>
<td>Spiral and Ie</td>
<td>Linear SVM</td>
<td>84.9 %</td>
<td>87.8 %</td>
</tr>
</tbody>
</table>


Future Developments

From unity APP INVENTOR Firebase MATLAB WEKA To CLARIN
From Sign language to CLARIN

Thank you for the attention!
Questions?

antonio.pallotti@uniroma2.it
Department of Management and Law – University of Rome Tor Vergata
antonio.pallotti@technoscience.it
Technoscience – University of Rome San Raffaele
Research Aims

The research aims to build an open-access parallel English-Arabic corpus of medical texts that is employed in improving specific domain word alignments.

In order to achieve this aim, we follow the following objectives:

• To collate a open-access parallel English-Arabic corpus of medical texts from reliable resources

• To align the corpus at document, sentence and word level

• To use the resulting aligned parallel corpus to extract a lexicon of English-Arabic medical terminology

• To publish the corpus as an open resource for research in AI for medicine into the CLARIN infrastructure that support re-using and easy finding of language data and tools for research.
Research Questions

• How much data I can collect for the corpus?

• What kind of annotations are required for this corpus?

• How do current alignment tools align the different types of terminologies (1-1, 1-M, M-1, M-M)?

• Is it possible to use linguistic features to improve the alignments?
Creating Corpus Steps

Identifying Multi-words Arabic Medical Terms

Algorithm 1: Identifying Multi-words Arabic Term Algorithm

Input: sentences: Arabic sentences

Output: sentences annotated with multi-words term

1. Step 1: Annotate sentences with the syntactic dependency tree.
2. Step 2: Retrieve the re-accruing fragment trees that have frequency equal or greater than 2 using FragmentSeeker [16].
3. Step 3: Cancel any fragments with length greater than 3.
4. Step 4: Calculate the Semantic Tightness Continuum (STC) [17] on fragmented units (fu) to annotate the ones with high values as candidate terms.

\[ STC(fu) = \frac{Freq(fu)}{\max(Freq(fu[0]|fu[1]|fu[2]), Freq(fu[0]|fu[1]|fu[2])) + \frac{1}{N}} \]

;where N is the number of medical articles.
5. Step 5: Filter the Arabic candidate terms: If the English version of the sentence has no medical term but there is an Arabic candidate term exists in the Arabic version, the term is canceled.
Attitudes and Language Acquisition: An Investigation on the Italian-French Community in Aix-Marseille

Fabio Ardolino
Università degli Studi di Siena

Day 1
Monday 27 September
13:45 - 13:50
Aims of the Project

CLARIN
Data Collection

**ATTITUDES**

- **Explicit attitudes**
  - Likert questionnaire

- **Implicit attitudes**
  - Implicit Association Test

- **Empathy**
  - Empathy Quotient Test

**SPEECH CORPUS**

- **French L2**
  - Picture description Test

- **Italian L1**
  - Modular Interview
Future Steps

Expansion of the corpus

- Additional recruitment & data collection campaigns

Full transcription of the corpus

- Literal & Phonetic (using Transcription chain)

Integration within the CLARIN infrastructure of the full dataset
Towards the Precise Detection of Adverbial Roles in Hungarian – Manual Clustering of Adverbial Adjuncts

Noémi Ligeti-Nagy

Faculty of Arts, Pázmány Péter Catholic University

Day 1
Monday 27 September
13:50 - 13:55
Motivation:
- QA system, training corpus (Novák et al., 2019)
  - (Hungarian: agglutinative language, stem+suffixes)
  - Formulate the relevant questions
  - Provide appropriate annotation

\[\text{London-ban} / \text{nadrág-ban}\]
London-Ine pants-Ine
‘in London’ ‘in pants’
Where? How? In what clothes?

Goal:
- Categorise the roles of adverbial adjuncts
- 9 locative case suffixes in Hungarian
  - inessive, adessive, superessive ...

Data:
- Szeged Dependency Treebank (HunCLARIN corpus) (Vincze et al., 2010)
- Tokens with OBL edge, with one of the nine locative case suffixes
Results

28 adverbial roles
1097 lemmas sorted into the categories
Default categories + additional roles

<table>
<thead>
<tr>
<th>Word</th>
<th>Default Category</th>
<th>Adverbial Roles with the Suffixes + Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>inessive</td>
</tr>
<tr>
<td><em>kör</em> ‘circle’</td>
<td>mode - <em>How?</em></td>
<td>loc - <em>Where?</em></td>
</tr>
<tr>
<td><em>gondozás</em> ‘caretaking’</td>
<td>thing - <em>In what?</em>, <em>At what?</em> etc.</td>
<td>form - <em>In what form?</em></td>
</tr>
</tbody>
</table>
Discussion

- Anything universal?
  - *in pants* = *nadrág-ban*, pants.Ine
  - other languages?

- Other suffixes and adverbial roles
- annotation for the QA system?


Tough Constructions and their Analogs in English, French and Russian: A Parallel Corpus Study Using the CLARIN VLO platform

Alina Tsikulina
UMR «Savoir, Textes & Language» Lab, University of Lille

Day 1
Monday 27 September
13:55 - 14:00
Tough constructions

1a. This road is difficult to cross.
1b. Cette route est difficile à traverser.

2a. This is a difficult road to cross.
2b. *C’est une difficile route à traverser.

3a. Ètu dorogu trudno pereseč’.
DEM-F.SG.ACC road-F.SG.ACC difficult-impers cross-INF-PERF
‘This road (is) difficult to cross.’

3b. Èta doroga trudna dlja peresečenija.
DEM-F.SG.NOM road-F.SG.NOM difficult-F.SG.NOM for crossing-N
‘This road (is) difficult for crossing.’

3c. Èta doroga trudno pereseekaetsja.
DEM-F.SG.NOM road-F.SG.NOM difficult-impers cross.REFL
‘This road with difficulty (can) be crossed’

3d. Èta doroga trudno peresekeaema.
DEM-F.SG.NOM road-F.SG.NOM difficult-impers cross.PR.PART.PASS-F.SG
‘This road (is) with difficulty crossable’

Case Marking (CM)
(Comrie & Matthews, 1990)

Deverbal (DEV)

Passive Reflexive verb (PASS)

Present Passive Participle (PASS)
(Paykin & Van Peterghem, 2020)
Aims and Corpus

1. Correspondence between EN, FR and RU
2. Commonalities and differences between these three languages
3. Accuracy of previous classifications

Platform: CLARIN VLO infrastructure
Opus corpus -> OpenSubtitles2018 multilingual corpus
(Lison and al., 2018)

Source language English (EN) ->
Target languages: French, Russian
EN: total 6530 (1827 ‘difficult’ and 4703 ‘easy’)

87% - Non-TC
13% - Extraposed constructions
7% - Tough-constructions + Tough-Nut constructions

377 target segments (125 – difficult, 252 - easy)
Main results and discussion

- English and French offer TCs
- French allows for more variability as Russian
- Various types of analogues in French and Russian

Applications - Where is it helpful?
- Annotation practices for TCs and crosslinguistic comparisons (contribution to the CLARIN)
- Translation Studies
- Foreign Language Teaching & Learning
- Cognitive Sciences

Future perspectives:
1. Source language Russian and French
2. Deeper corpus investigations (each language separately)
3. Experimental research: monolingual, bilingual and trilingual speakers.
The *RigVeda* Goes ‘Universal’: Annotating Comparative Constructions in the Most Ancient Poetry of India

Erica Biagetti
University of Pavia, Bergamo

Day 1
Monday 27 September
14:00 - 14:05
• **PhD Project:** Annotation of the *RigVeda* (RV); equative and similative constructions

• **RV:** 1028 hymns, second half of II millennium BCE: oldest layer of Vedic literature

❖ **Relevance for CLARIN (1):** Annotation according to UD scheme
  
  Annotation via DCS interface (Hellwig 2010-2021)
  
  → publication in UD
  
  (v. 1.0 released in May 2020; v. 2.0 and 3.0 in the next release?)

• Before the project: treebank of Classical Sanskrit, pilot project on 250 sentences from the *Pañcatantra* (Dwivedi and Zeman 2018; Biagetti 2018)

• Project VTB (Zürich) + Project ChronBMM (Düsseldorf):
  24109 tokens in 3092 sentences from the RV
  63664 tokens in 8646 sentences from Vedic literature (Hellwig and Sellmer forth.)
Relevance for CLARIN (2): Hands-On Comparative Constructions

<table>
<thead>
<tr>
<th>CONSTRUCTION</th>
<th>EXAMPLE</th>
<th>ANNOTATION (dependent → relation → head)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREDICATIVE SIM.</td>
<td>‘Agni is like the sun.’</td>
<td>like → case:sim → sun</td>
</tr>
<tr>
<td></td>
<td>Agni → nsubj → sun</td>
<td></td>
</tr>
<tr>
<td>SIM. WITH ELLIPSIS</td>
<td>‘Agni shines like the sun.’</td>
<td>like → case:sim → sun → obl:grad → shines</td>
</tr>
<tr>
<td></td>
<td>‘The lightning bellows like a cow.’</td>
<td>like → case:sim → cow → obl:manner → bellow</td>
</tr>
<tr>
<td>SIM. WITH GAPPING</td>
<td>‘Thoughts lick Indra like mothers a calf.’</td>
<td>like → mark:sim → mothers → advcl:manner → lick,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>calf → orphan → mothers</td>
</tr>
<tr>
<td>CLAUSAL SIM.</td>
<td>‘Just as you drank the previous soma drinks, so take a drink today.’</td>
<td>as → mark → drank → advcl:manner → drink;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>previous drinks → obj → drank;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>so → advmod → drink</td>
</tr>
</tbody>
</table>

- Equative and simulative constructions marked by *ná, iva, and yáthā* ‘like’
- Research questions: origin, synchronic syntax, further developments
- NB: comparison expressed by several other different strategies in Vedic; *ná, iva, and yáthā* have other functions too
- Annotation according to UD scheme + sub-relations
  NB: not meant to be language-specific!
• **Research Outputs**

More granular and informative annotation scheme returned new insights on:

a. Origin of the three main comparative constructions

b. Synchronic syntax: order of STAND and PAR determined by a series of factors
   WO typology, Behaghel’s Law of Increasing Constituents, preference of RW gapping

c. Development (1): *iva* develops into an adaptor in Vedic prose

• **Drawbacks**

d. Development (2): expansion of *iva* at the expense of *ná*
   Distribution based on case ending of the preceding word
   What about semantics? (“family resemblance”; Viti 2002)

☞ *Post-PhD*: synsets from the Sanskrit WordNet (Biagetti et al. 2021) in CoNLL-U Plus
Plenary Q&A

Day 1
Monday 27 September
14:05 - 14:15
Breakout Room Discussions

Day 1
Monday 27 September
14:15 - 14:30

Room 1: Wiktoria Mielesczenko-Kowszewicz
Reflecting Cognitive Processing of Trauma in Language – a Result of Trauma’s Experience Corpus Analysis with a Usage of Word’s Meanings Category

Room 2: Antonio Pallotti
Machine Learning Applied to Voice Signal in Parkinson’s Disease

Room 3: Zainab Almugbel
Building English-Arabic Parallel Medical Corpora

Room 4: Fabio Ardolino
Attitudes and Language Acquisition: An Investigation on the Italian-French Community in Aix-Marseille

Room 5: Noémi Ligeti-Nagy
Towards the Precise Detection of Adverbial Roles in Hungarian – Manual Clustering of Adverbial Adjuncts

Room 6: Alina Tsikulina
Tough Constructions and Their Analogs in English, French and Russian: A Parallel Corpus Study Using the CLARIN VLO Platform

Room 7: Erica Biagetti
The RigVeda Goes ‘Universal’: Annotating Comparative Constructions in the Most Ancient Poetry of India
Thank you for contributing!

See you at
CLARIN 2022
PhD Student Session

Comments and suggestions welcome at events@clarin.eu

Coming up:
10 min break + 20 min social session with Ben Verhoeven

Do not miss:
Tue 13.30-14.30 Session on Teaching with CLARIN
Wed 10.15-11.00 Panel on The Role of Corpora for the Study of Language Use and Mental Health Conditions