New features in view of EuReCo

Workshop: Corpus Harmonization and Language Comparison
2018-09-26 – Bucharest, Romania
Overview
New features in view of EuReCo

1. Introduction
2. Virtual Corpus References
3. Plugins
4. Customizations
5. Conclusion
Introduction
New features in view of EuReCo
Introduction
New features in view of EuReCo

While corpora can be made accessible with the same Corpus Platform and are comparable,
• the **underlying data** differs,
• the **user's language** differs,
• the **annotations** differ,
• the **research questions** may differ,
• ...
Introduction
New features in view of EuReCo

While corpora can be made accessible with the same Corpus Platform and are comparable,

- the underlying data differs,
- the user's language differs,
- the annotations differ,
- the research questions may differ,
- ...

Improvements for ...

- Comparable Corpora
- Diverse Research Interests
While corpora can be made accessible with the same Corpus Platform and are comparable,

- the **underlying data** differs,
- the **user's language** differs,
- the **annotations** differ,
- the **research questions** may differ,
- ...

**Improvements for ...**

- **Comparable Corpora**
- **Diverse Research Interests**

Thanks to Eliza Margaretha, Helge Stallkamp & Peter Harders!
2. VC References
Kupietz et al. (2018)
3. Plugins [WiP]
When tools or features ...  

- are very specific regarding ...  
  - a language  
  - the underlying data  
  - research questions or research projects

... they should **not be part of core KorAP**!
Plugins
Motivation

When tools or features ...

- are very specific regarding ...
  - a language
  - the underlying data
  - research questions or research projects

- are not of general interest

- need to be developed outside the KorAP project

- are not compatible with the BSD-2 license

- ...

... they should not be part of core KorAP!
Plugins
Panels and Views
* not very impressive ...
Plugins
Areas of Application

Result plugins
- e.g. export
- e.g. visualisations of match distributions
- ...

{...}
Plugins
Areas of Application

<table>
<thead>
<tr>
<th>Foundry</th>
<th>Layer</th>
<th>In</th>
<th>dem</th>
<th>Satz</th>
<th>Der</th>
<th>alte</th>
<th>Baum</th>
<th>wurde</th>
<th>vom</th>
<th>Blitz</th>
<th>getroffen</th>
</tr>
</thead>
<tbody>
<tr>
<td>corenlp</td>
<td>p</td>
<td>APPR</td>
<td>ART</td>
<td>NN</td>
<td>ART</td>
<td>ADJA</td>
<td>NN</td>
<td>VAFIN</td>
<td>APPRART</td>
<td>NN</td>
<td>VVPP</td>
</tr>
<tr>
<td>opennlp</td>
<td>p</td>
<td>APPR</td>
<td>ART</td>
<td>NN</td>
<td>ART</td>
<td>ADJA</td>
<td>NN</td>
<td>VAFIN</td>
<td>APPRART</td>
<td>NN</td>
<td>VVPP</td>
</tr>
</tbody>
</table>

Match plugins
- e.g. Translations, Mappings etc.
- e.g. embedding of external lexical resources
- ...

Nominalphrase by Sokonbud, u.a. (2015-05-01)
Plugins

Areas of Application

<table>
<thead>
<tr>
<th>availability eq /CC-BY.*|</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>documents:</th>
<th>1.039.761</th>
<th>paragraphs:</th>
<th>10.199.791</th>
</tr>
</thead>
<tbody>
<tr>
<td>sentences:</td>
<td>26.593.945</td>
<td>tokens:</td>
<td>475.442.767</td>
</tr>
</tbody>
</table>

Corpus plugins

- e.g. Visualisations of VC compilation
- ...
Plugins
Areas of Application

**Query plugins**
- e.g. including the visual query creator of Cosmas2win
- e.g. NLP-CQP (CoRoLa)
- e.g. Mapping of Tagsets
- ...
// Register plugin
KorAP.Plugin.register({
  'name': 'Catify',
  'desc': 'Add some cats to my matches',

  // Where to embed in the UI
  'embed': [{
    'panel': 'match',  // Add to the match panel
    'title': 'Catify',  // With a button called 'Catify'
    'classes': ['catify'],
    'onClick': {
      'action': 'addWidget',  // Add a new view to the panel
      'template': 'http://external-website.com/cat.html'
    }
  }]
});
Plugins
Implementation

- Registerable per instance and per user (wip)
- Embedded as sandboxed `<iframe>`s in the frontend (no OpenSocial-like Container)
- API communication via HTML5 Web Messaging API
- API will be defined „on demand“ (security challenge)
- Plugin services can be written in any programming language applicable
- Some plugins will require API access on behalf of the user (e.g. export)
Plugins
Oauth 2.0

- Open protocol for (Web-)API **Authorization**

- Enables **Third-party applications** to access data on behalf of a user
  - with the same (or less) permissions
  - but without giving away password information

- Not only relevant for plugins from third parties, but also for alternative clients
Plugins
Oauth 2.0

- Open protocol for (Web-)API **Authorization**
- Enables **Third-party applications** to access data on behalf of a user
  - with the same (or less) permissions
  - but without giving away password information
- Not only relevant for plugins from third parties, but also for alternative clients

```
$ curl -H 'Authorization: Bearer e739DQpDVChxFgu6e0zkwA' 'https://korap.ids-mannheim.de/api/v1.0/search?q=Wasser'
```

Kupietz, Diewald & Fankhauser (2018)
4. Customizations
Customizations
Translations


Metadaten Token Relationen Nominalphrase von Sokonbud, u.a. (2015-05-01)

Customizations
Annotations

Der alte Baum

Base Annotation
Structure

CoreNLP
Constituency, Named Entities, Part-of-Speech

DeReKo
Structure

Malt
Dependency

MarMoT
Morphology, Part-of-Speech

OpenNLP
Part-of-Speech

TreeTagger
Lemma, Part-of-Speech
Customizations
Example Corpora

and an autocompletion for closed annotations (type in foundry prefixes like cnx/).

Example Queries

**Poliqarp**: Find all occurrences of the lemma "baum" as annotated by the default foundry.

```
[base=Baum]
```

**Poliqarp**: Find all sequences of adjectives as annotated by Treetagger, that are repeated 3 to 5 times in a row.

```
[tt/p=ADJA]{3,5}
```

**Cosmas-II**: Find all occurrences of the words "der" and "Baum", in case they are in a maximum distance of 5 tokens. The order is not relevant.

```
der /w5 Baum
```
Customizations
Templates and Content Blocks

*KorAP* is a new Corpus Analysis Platform, suited for large, multiply annotated corpora and complex search queries independent of particular research questions.

**New to KorAP?** Please check out our [tutorial](#)!

*KorAP* is developed at the [Institute for the German Language](#), member of the [Leibniz Association](#). The separated modules are being published as open source at [GitHub](#).
5. Conclusion
Conclusion
KorAP in the view of EuReCo

Improvements for ...

- Comparable Corpora
  VC References + Statistics

- Diverse Research Interests
  Plugins + Oauth + Customizations
Thank you very much for your attention!


• Trawiński, Beta/Kupietz, Marc (2018): **Language Comparison and Corpus Data: From Monolingual through Parallel to Comparable Corpora,** Talk, ars grammatica 2018, IDS Mannheim