



Member of the

# 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



Kupietz et al. (2017)

# 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**

# 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**

Thanks to Eliza Margaretha, Helge Stallkamp & Peter Harders!

## 2. VC References

# VC References

## Demo

Demo

Kupietz et al. (2018)

---



# 3. Plugins

[WiP]

# Plugins

## Motivation

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

The screenshot shows the KorAP web interface. At the top left is the KorAP logo. To its right is a search bar containing the text "Der alte Baum". Below the search bar is a query editor showing the filter "availability eq /CC-BY.\*/" with an edit icon. A "Statistics" button is located below the query editor. At the bottom of the header area, it says "in a virtual corpus" with an edit icon, "with Poliqarp" with a dropdown arrow, and a "Glimpse" button with a graduation cap icon. The main content area displays a table of search results. The table has two columns: an identifier and a snippet of text. The first four rows are visible, each containing an identifier and a snippet mentioning "Der alte Baum".

WPD15/168/56326	ehr. (92) <b>Der alte Baum</b> im Vordergrund des von van Gogh kopierten Pflaume
WPD15/N01/28099	m Satz „ <b>Der alte Baum</b> wurde vom Blitz getroffen.“ würde klassischerweise d
WPD15/J80/33968	us, 1946 <b>Der alte Baum</b> und andere Märchen, 1946 Vier Glöcklein, 1947 Es w
WPD15/K07/50372	epflanzt. <b>Der alte Baum</b> wurde 2001 aufwendig saniert und mit einer Kronens

At the bottom of the interface, there is a footer with the text "About KorAP - Imprint - Privacy - V 0.29.1" and a small icon.

Query

Result

# Plugins Demo

Demo

\* not very impressive ...

# Plugins

## Areas of Application



### Result plugins

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

# Plugins

## Areas of Application

eine Phrase (eine abgeschlossene syntaktische Einheit), deren Kern oder „Kopf“ ein Nomen (im Sinne von Substantiv) ist. Andere Formen wie Pronomina oder Substantivierungen von Adjektiven bilden Nominalphrasen, sofern sie der Wortart nach ebenfalls als nominal analysiert werden. In dem Satz „ **Der alte Baum** wurde vom Blitz getroffen.“ würde klassischerweise der Bestandteil „der alte Baum“ als eine solche Nominalphrase angesehen werden, da Adjektiv und Artikel hier als abhängige Begleiter des nominalen Kopfes "Baum" gesehen werden. Dieser Begriff der Nominalphrase ist in der allgemeinen Linguistik

Foundry	Layer	In	dem	Satz	Der	alte	Baum	wurde	vom	Blitz	getroffen
corenlp	p	APPR	ART	NN	ART	ADJA	NN	VAFIN	APPRART	NN	VVPP
opennlp	p	APPR	ART	NN	ART	ADJA	NN	VAFIN	APPRART	NN	VVPP

Metadata Tokens Relations

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


[WPD15/N01/28099]

## Match plugins

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

# Plugins

## Areas of Application

**availability** eq **/CC-BY.\*/** 

<b>documents:</b>	1.039.761	<b>paragraphs:</b>	10.199.791
<b>sentences:</b>	26.593.945	<b>tokens:</b>	475.442.767

Statistics

### 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
- ...

# Plugins

## Implementation

```
// 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'
```

# 4. Customizations

# Customizations Translations

**KorAP**

Der alte Baum

in a virtual corpus with Poliqarp

☒ Glimpse

ehr. (92) **Der alte Baum** im Vordergrund des von van Gogh kopierten Pfla

eine Phrase (eine abgeschlossene syntaktische Einheit), deren Kern oder „Kopf“ ein Nomen (im Sinne von Substantiv) ist. Andere Formen wie Pronomina oder Substantivierungen von Adjektiven bilden Nominalphrasen, sofern sie der Wortart nach ebenfalls als nominal analysiert werden. In dem Satz „ **Der alte Baum** wurde vom Blitz getroffen.“ würde klassischerweise der Bestandteil „der alte Baum“ als eine solche Nominalphrase angesehen werden, da Adjektiv und Artikel hier als abhängige Begleiter des nominalen Kopfes "Baum" gesehen werden. Dieser Begriff der Nominalphrase ist in der allgemeinen Linguistik

Metadata Tokens Relations **Nominalphrase** by Sokonbud, u.a. (2015-05-01) [WPD15/N01/28099]

us, 1946 **Der alte Baum** und andere Märchen, 1946 Vier Glöcklein, 1947 E  
epflanzt. **Der alte Baum** wurde 2001 aufwendig saniert und mit einer Kror

**KorAP**

Der alte Baum

in einem virtuellen Korpus mit Poliqarp

☒ Glimpse

ehr. (92) **Der alte Baum** im Vordergrund des von van Gogh kopierten Pfla

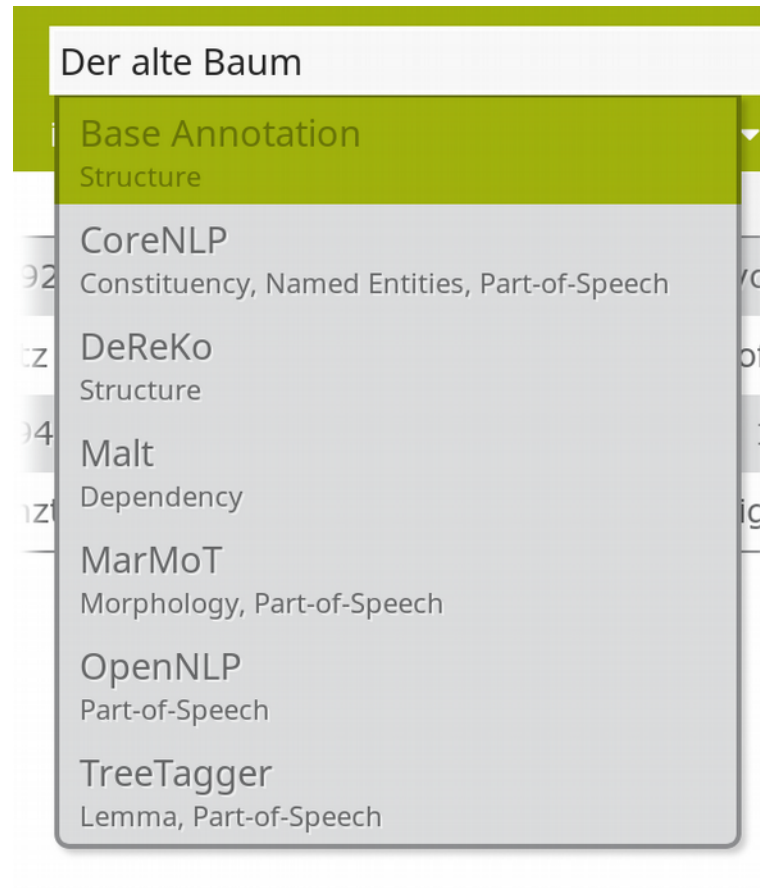
eine Phrase (eine abgeschlossene syntaktische Einheit), deren Kern oder „Kopf“ ein Nomen (im Sinne von Substantiv) ist. Andere Formen wie Pronomina oder Substantivierungen von Adjektiven bilden Nominalphrasen, sofern sie der Wortart nach ebenfalls als nominal analysiert werden. In dem Satz „ **Der alte Baum** wurde vom Blitz getroffen.“ würde klassischerweise der Bestandteil „der alte Baum“ als eine solche Nominalphrase angesehen werden, da Adjektiv und Artikel hier als abhängige Begleiter des nominalen Kopfes "Baum" gesehen werden. Dieser Begriff der Nominalphrase ist in der allgemeinen Linguistik

Metadaten Token Relationen **Nominalphrase** von Sokonbud, u.a. (2015-05-01) [WPD15/N01/28099]

us, 1946 **Der alte Baum** und andere Märchen, 1946 Vier Glöcklein, 1947 E  
epflanzt. **Der alte Baum** wurde 2001 aufwendig saniert und mit einer Kror

# Customizations

## Annotations





# Customizations

## Example Corpora

KorAP

Kalamar

Kustvakt

Koral

Krill

Karang

Query Languages

Cosmas II

Poliqarp+

Annis QL

CQL

FCSQL

Regular Expressions

Wildcards

Data

Corpora

Annotations

API

KoralQuery

Search API

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

**Cosmas-II**: Find all sequences of a word starting

# Customizations

## Templates and Content Blocks

The screenshot shows the KorAP website interface. On the left, there is a login section with fields for 'Username or Email' and 'Password', and a link 'Login with a registered Cosmas-II account'. The main content area is framed by a red border and contains the following text:

**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](#).

At the bottom right of the main content area, there is a logo for the 'INSTITUT FÜR DEUTSCHE SPRACHE' and a note 'Member of the Leibniz Association'.

The footer contains a navigation bar with links: 'About KorAP - Imprint - Privacy - V 0.29.1'.

# 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!**

- Anthony, Laurence (2013): **A critical look at software tools in corpus linguistics**. In: Linguistic Research 30, pp. 141-161.
- Kupietz, Marc/Diewald, Nils/Fankhauser, Peter (2018): **How to Get the Computation Near the Data: Improving data accessibility to, and reusability of analysis functions in corpus query platforms**. In: Bański, Piotr/Kupietz, Marc/Barbaredi, Adrien/Biber, Hanno/Breiteneder, Evelyn/Clematide, Simon/Witt, Andreas (Hrsg.): Proceedings of the LREC 2018 Workshop “Challenges in the Management of Large Corpora (CMLC-6)”. 07 May 2018 – Miyazaki, Japan. Paris: European language resources association (ELRA), 2018. S. 20-25
- Kupietz, Marc/Witt, Andreas/Bański, Piotr/Tuşiş, Dan/Cristea, Dan/Váradi, Tamás (2017): **EuReCo – Joining Forces for a European Reference Corpus as a sustainable base for cross-linguistic research**. In: Bański, Piotr et al. (eds.): Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017 including the papers from the Web-as-Corpus (WAC-XI) guest section. Birmingham, 24 July 2017. Mannheim: Institut für Deutsche Sprache, 2017. pp. 15-19.
- Kupietz, Marc/Cosma, Ruxandra/Cristea, Dan/Diewald, Nils/Trawiński, Beata/Tuşiş, Dan/Váradi, Tamás/Wöllstein, Angelika (2018): **Recent Developments in the European Reference Corpus EuReCo**. In: Using Corpora in Contrastive and Translation Studies – UCCTS (5th edition), 12.9.2018, Louvain-la-Neuve, Belgium.
- Trawiński, Beata/Kupietz, Marc (2018): **Language Comparison and Corpus Data: From Monolingual through Parallel to Comparable Corpora**, Talk, ars grammatica 2018, IDS Mannheim

## References