

The Corpus Search Infrastructure at the IEAS

CSniper and ANNIS3

Niko Schenk

Institut für England- und Amerikastudien
Goethe-Universität Frankfurt am Main
Summer Term 2015

June 1, 2015

1 Background

2 Available Corpora

3 Web Interfaces

- Csniper
- ANNIS3

1 Background

2 Available Corpora

3 Web Interfaces

- Csniper
- ANNIS3

April 2013

- **My Task:** Set up an **infrastructure** for corpus search.

General Summary of the Required Components

① Data preprocessing (format conversion)

General Summary of the Required Components

- ① **Data preprocessing** (format conversion)
- ② Highly optimized **search tools** (for large amounts of textual data)

General Summary of the Required Components

- ① **Data preprocessing** (format conversion)
- ② Highly optimized **search tools** (for large amounts of textual data)
- ③ **Query language** (high-level access to the data, including annotations)

General Summary of the Required Components

- ① **Data preprocessing** (format conversion)
- ② Highly optimized **search tools** (for large amounts of textual data)
- ③ **Query language** (high-level access to the data, including annotations)
- ④ **Corpus storage** (data base)

General Summary of the Required Components

- ① **Data preprocessing** (format conversion)
- ② Highly optimized **search tools** (for large amounts of textual data)
- ③ **Query language** (high-level access to the data, including annotations)
- ④ **Corpus storage** (data base)
- ⑤ Web-based infrastructure (**web applications**/online user-interface, security restrictions, user management)

General Summary of the Required Components

- ① **Data preprocessing** (format conversion)
- ② Highly optimized **search tools** (for large amounts of textual data)
- ③ **Query language** (high-level access to the data, including annotations)
- ④ **Corpus storage** (data base)
- ⑤ Web-based infrastructure (**web applications**/online user-interface, security restrictions, user management)
- ⑥ **Export functionality** (to download results)

General Summary of the Required Components

- ① **Data preprocessing** (format conversion)
- ② Highly optimized **search tools** (for large amounts of textual data)
- ③ **Query language** (high-level access to the data, including annotations)
- ④ **Corpus storage** (data base)
- ⑤ Web-based infrastructure (**web applications**/online user-interface, security restrictions, user management)
- ⑥ **Export functionality** (to download results)
- ⑦ **Visualizations**

General Summary of my Work

- ① Set up of the **two** (existent) search interfaces.
 - ② Converted corpora into appropriate formats.
 - ③ Implemented download functionality.
 - ④ Configured data bases.
 - ⑤ Deployed web applications and registered user groups.
- **They are now ready for use.** (Effective: July 2014).

Two User Interfaces

Hosted at Rechenbereich Informatik (RBI):

- General Information:

<http://corpora.acoli.informatik.uni-frankfurt.de:8080/>

- Documentation:

<http://corpora.acoli.informatik.uni-frankfurt.de:8080/anglistik/>

Credentials/login data have been distributed and are available upon request!

Documentation on Csniper & ANNIS3

The screenshot shows a web page with a header containing logos for Goethe University Frankfurt am Main, Institut für Informatik, and ACoLi Applied Computational Linguistics. Below the header, there are navigation links for General Information, ANNIS3, CSNIPER, and specific links to Goethe-Uni Home, IEAS, ACoLi, and Informatik. The main content area is titled "Welcome!" and provides general information about the two interfaces. It includes a table with their URLs and a note about running linguistic queries. On the left side, there is a sidebar with links to General Information, Accessing the GUIs, A List of Available Corpora, FAQ, Contact Us, and icons for ANNIS3 and Csniper.

GENERAL INFORMATION ANNIS3 CSNIPER

Goethe-Uni Home IEAS ACoLi Informatik

General Information

Accessing the GUIs
A List of Available Corpora
FAQ
Contact Us




Welcome!

This web page provides general information on two powerful graphical user interfaces (GUIs) for annotation-based corpus search:

ANNIS3 & CSniper

In order to run linguistic queries on real corpus data you can easily access both user interfaces via the following two URLs:

Corpus Interface	Access
ANNIS3	http://corpora.acoli.informatik.uni-frankfurt.de:8080/annis
CSniper	http://corpora.acoli.informatik.uni-frankfurt.de:8080/csniper

More information on the interfaces is available here:

- <http://www.sfb632.uni-potsdam.de/annis/>
- <http://code.google.com/p/csniper/>

Use the navigation on the left side of this web page to obtain more specific information on how to use the GUIs, query data for linguistic information and for details on how to export analysis results.

Figure: <http://corpora.acoli.informatik.uni-frankfurt.de:8080/anglistik/>

Documentation on Csniper & ANNIS3

The screenshot shows a web page for the English corpus (anglistik) at the Applied Computational Linguistics (ACoLi) website. At the top, there are logos for Goethe University Frankfurt am Main, Institut für Informatik, and ACoLi. Below the header, there are navigation links for GENERAL INFORMATION, ANNIS3, CSNIPER, and specific links to Goethe-Uni Home, IEAS, ACoLi, and Informatik. On the left, a sidebar contains links for General Information (Accessing the GUIs, A List of Available Corpora, FAQ, Contact Us), ANNIS3 (Csniper, Sample Usage, Token-based Search, Phrase-based Search, Exporting Results), and Csniper Sample Usage (1. Login). The main content area displays the Csniper Sample Usage interface, showing a login screen with fields for Username and Password.

Figure: <http://corpora.acoli.informatik.uni-frankfurt.de:8080/anglistik/>

Documentation on Csniper & ANNIS3



Figure: <http://corpora.acoli.informatik.uni-frankfurt.de:8080/anglistik/>

1 Background

2 Available Corpora

3 Web Interfaces

- Csniper
- ANNIS3

A List of Currently Available Corpora at the IEAS

Corpus	Properties			
	language	words	GUI	annotations
TüPP-D/Z	German (news)	200 million	Csniper	word, pos, lemma, (chunks?) automatically tagged & chunk annotated

A List of Currently Available Corpora at the IEAS

Corpus	Properties			
	language	words	GUI	annotations
TüPP-D/Z	German (news)	200 million	Csniper	word, pos, lemma, (chunks?) automatically tagged & chunk annotated
BNC	English (spoken, news, academic, fiction)	100 million	CSniper	word, pos, lemma, syntax automatically tagged & parsed (Darmstadt—RTF, Stanford)

A List of Currently Available Corpora at the IEAS

Corpus	Properties			
	language	words	GUI	annotations
TüPP-D/Z	German (news)	200 million	Csniper	word, pos, lemma, (chunks?) automatically tagged & chunk annotated
BNC	English (spoken, news, academic, fiction)	100 million	CSniper	word, pos, lemma, syntax automatically tagged & parsed (Darmstadt—RTF, Stanford)
deWaC	German (news)	85 million $(\approx \frac{1}{20})$	CSniper	word, pos, lemma, (syntax?) automatically tagged & parsed (Darmstadt—check availability?)

A List of Currently Available Corpora at the IEAS

Corpus	Properties			
	language	words	GUI	annotations
TüPP-D/Z	German (news)	200 million	Csniper	word, pos, lemma, (chunks?) automatically tagged & chunk annotated
BNC	English (spoken, news, academic, fiction)	100 million	CSniper	word, pos, lemma, syntax automatically tagged & parsed (Darmstadt—RTF, Stanford)
deWaC	German (news)	85 million $(\approx \frac{1}{20})$	CSniper	word, pos, lemma, (syntax?) automatically tagged & parsed (Darmstadt—check availability?)
Digitale Bibliothek	German (literature— old/contemporary)	2 million	Csniper	words, pos, lemma

A List of Currently Available Corpora at the IEAS

Corpus	Properties			
	language	words	GUI	annotations
TüPP-D/Z	German (news)	200 million	Csniper	word, pos, lemma, (chunks?) automatically tagged & chunk annotated
BNC	English (spoken, news, academic, fiction)	100 million	CSniper	word, pos, lemma, syntax automatically tagged & parsed (Darmstadt—RTF, Stanford)
deWaC	German (news)	85 million $(\approx \frac{1}{20})$	CSniper	word, pos, lemma, (syntax?) automatically tagged & parsed (Darmstadt—check availability?)
Digitale Bibliothek	German (literature— old/contemporary)	2 million	Csniper	words, pos, lemma
Tüba-D/Z	German (news)	1.3 million	ANNIS3	words, pos, lemma, syntax, dependencies, anaphora res., named-entities, coreferences (syntax manually annotated, new release 8.0)

A List of Currently Available Corpora at the IEAS

Corpus	Properties			
	language	words	GUI	annotations
TüPP-D/Z	German (news)	200 million	Csniper	word, pos, lemma, (chunks?) automatically tagged & chunk annotated
BNC	English (spoken, news, academic, fiction)	100 million	CSniper	word, pos, lemma, syntax automatically tagged & parsed (Darmstadt—RTF, Stanford)
deWaC	German (news)	85 million $(\approx \frac{1}{20})$	CSniper	word, pos, lemma, (syntax?) automatically tagged & parsed (Darmstadt—check availability?)
Digitale Bibliothek	German (literature— old/contemporary)	2 million	Csniper	words, pos, lemma
Tüba-D/Z	German (news)	1.3 million	ANNIS3	words, pos, lemma, syntax, dependencies, anaphora res., named-entities, coreferences (syntax manually annotated, new release 8.0)
Brown Corpus	English (blogs)	1 million	(TODO)	words, pos, lemma

A List of Currently Available Corpora at the IEAS

Corpus	Properties			
	language	words	GUI	annotations
TüPP-D/Z	German (news)	200 million	Csniper	word, pos, lemma, (chunks?) automatically tagged & chunk annotated
BNC	English (spoken, news, academic, fiction)	100 million	CSniper	word, pos, lemma, syntax automatically tagged & parsed (Darmstadt—RTF, Stanford)
deWaC	German (news)	85 million $(\approx \frac{1}{20})$	CSniper	word, pos, lemma, (syntax?) automatically tagged & parsed (Darmstadt—check availability?)
Digitale Bibliothek	German (literature— old/contemporary)	2 million	Csniper	words, pos, lemma
Tüba-D/Z	German (news)	1.3 million	ANNIS3	words, pos, lemma, syntax, dependencies, anaphora res., named-entities, coreferences (syntax manually annotated, new release 8.0)
Brown Corpus	English (blogs)	1 million	(TODO)	words, pos, lemma
LOB Corpus	English	1 million	(TODO)	words, pos, lemma

A List of Currently Available Corpora at the IEAS

Corpus	Properties			
	language	words	GUI	annotations
Tüba-D/S	German (spoken— Verbmobil)	360,000	ANNIS3	words, pos, lemma, syntax, (syntax manually annotated)
Tüba-E/S	English (spoken— Verbmobil)	360,000	ANNIS3	words, pos, lemma, syntax, (syntax manually annotated)
English Web Treebank	English (blogs)	254,000	ANNIS3	words, pos, lemma, syntax, (syntax manually annotated)

Kolhapur Corpus

Australian Corpus of English

Wellington Corpus

London-Lund Corpus

Lancaster/IBM

COLT

Polytechnic of Wales Corpus

Helsinki Corpus of English Texts

FLOB

ICE

More Corpora... (potentially interesting)

Properties				
Corpus	language	words	GUI	annotations
Older Scots Corpus	Scottish			
Corpus of Early English Correspondence	early-modern-english			
Lampeter Corpus	early-modern-english			
The Newdigate Newsletters	early-modern-english			
20 News Groups	IR			
Reuters	IR			
Dortmunder Chat-Korpus	IR			
Own Social Media Corpora	IR			
Septembertestament	Informatik			
Gothic Bible	Informatik			
Penn Treebank (?)	Informatik			
Penn Discourse Treebank (?)	Informatik			
Gigaword Corpus (?)	Informatik	1 billion		
PukWaC (Open Source)	several billion			
ukWaC (Open Source)	2 billion			
WaCkypedia.EN (Open Source)	2 billion			

Our Available Corpora—Summary

- In principle, we can install any type of corpus (only with restrictions on the size).
 - Linguistic annotations can be added automatically (lemma, pos, coreference, discourse, information structure...)
 - Generalizations over large amounts of automatically (possibly erroneously) annotated data are better than only a handful of manually annotated (gold) texts.

1 Background

2 Available Corpora

3 Web Interfaces

- Csnicer
- ANNIS3

CSniper

The screenshot shows the 'Project Home' page of the CSniper project on Google Code. The header features the project logo (a blue house icon with a 'P') and the name 'csniper'. Below the logo, the tagline 'Combining search and annotation on large corpora' is displayed. The top navigation bar includes links for 'Project Home' (which is active), 'Wiki', 'Issues', and 'Source'. A secondary navigation bar below it shows 'Summary' and 'People'. The main content area is divided into several sections:

- Project Information:** Shows that the project is starred by 1 user and provides a link to 'Project feeds'.
- Code license:** Apache License 2.0
- Members:** Lists 'richard.eckart' and 'eriklan...@gmail.com'.
- Featured:** Includes links to 'Wiki pages' (ConversionGuide, InstallationTutorial, UsageGuide) and a 'Show all »' link.
- Links:** Includes 'Groups'.

In the center, there is a detailed description of CSniper's purpose: 'CSniper (Corpus Sniper) is a tool that implements (i) a web-based multi-user interface for constructing queries in large corpora based on linguistic queries and (ii) an annotation-by-query approach efficiently harnesses expert knowledge to identify means of existing automatic annotation tools.'

A 'Cite CSniper' section provides citation information: 'If you use CSniper in scientific work, please cite' followed by a reference: 'Eckart de Castilho, R., Bartsch, S., and Gurevych, I. (2012). **CSniper - large corpora**. In Proceedings of the ACL 2012 System Demonstrations, pages 1–6. Association for Computational Linguistics. ([pdf](#), [bib](#))'

Figure: Csniper Google Code Project by Eckart de Castilho & Erik Lan, TU Darmstadt.

CSniper

Advantages:

- Active support / Wiki / Programmers contribute to the project.

CSniper

Advantages:

- Active support / Wiki / Programers contribute to the project.
- **Extremely efficient** for large amounts of textual data
(handles \approx 300 million tokens easily) / MySQL data base.

CSniper

Advantages:

- Active support / Wiki / Programers contribute to the project.
- **Extremely efficient** for large amounts of textual data
(handles \approx 300 million tokens easily) / MySQL data base.
- Provides users with easy-to-use search for annotations (flat & syntax).
 - *Tgrep2* integration possible.

CSniper

Advantages:

- Active support / Wiki / Programers contribute to the project.
- **Extremely efficient** for large amounts of textual data
(handles \approx 300 million tokens easily) / MySQL data base.
- Provides users with easy-to-use search for annotations (flat & syntax).
 - *Tgrep2* integration possible.
- Offers novel functionality (“annotation by query”, ML, prediction of likely annotations & **statistical evaluation**).

CSniper

Advantages:

- Active support / Wiki / Programers contribute to the project.
- **Extremely efficient** for large amounts of textual data
(handles \approx 300 million tokens easily) / MySQL data base.
- Provides users with easy-to-use search for annotations (flat & syntax).
 - *Tgrep2* integration possible.
- Offers novel functionality (“annotation by query”, ML, prediction of likely annotations & **statistical evaluation**).
- Secured/restricted user access possible.

CSniper

Advantages:

- Active support / Wiki / Programers contribute to the project.
- **Extremely efficient** for large amounts of textual data
(handles \approx 300 million tokens easily) / MySQL data base.
- Provides users with easy-to-use search for annotations (flat & syntax).
 - *Tgrep2* integration possible.
- Offers novel functionality (“annotation by query”, ML, prediction of likely annotations & **statistical evaluation**).
- Secured/restricted user access possible.
- New corpus data (own corpora) can be easily imported (*UIMA* pipeline).

CSniper

Disadvantages:

- Only standard visualizations (syntax trees, POS, context).

CSniper

Disadvantages:

- Only standard visualizations (syntax trees, POS, context).
- Requires *UIMA* pipeline to import corpus data.

CSniper

Disadvantages:

- Only standard visualizations (syntax trees, POS, context).
- Requires *UIMA* pipeline to import corpus data.
 - Difficult (impossible?) to import, e.g., TüPP-D/Z's "topologische Felder" annotations.

CSniper

Disadvantages:

- Only standard visualizations (syntax trees, POS, context).
- Requires *UIMA* pipeline to import corpus data.
 - Difficult (impossible?) to import, e.g., TüPP-D/Z's "topologische Felder" annotations.
- More or less limited to flat annotations. (word, lemma, pos, syntactic categories **but** no dependencies, no discourse, no ...)

CSniper

Disadvantages:

- Only standard visualizations (syntax trees, POS, context).
- Requires *UIMA* pipeline to import corpus data.
 - Difficult (impossible?) to import, e.g., TüPP-D/Z's "topologische Felder" annotations.
- More or less limited to flat annotations. (word, lemma, pos, syntactic categories **but** no dependencies, no discourse, no ...)
- I had to add download functionality for standard search (with context provided).

Demo

1 Background

2 Available Corpora

3 Web Interfaces

- CsniPer
- ANNIS3

ANNIS3

The screenshot shows the ANNIS3 web interface. At the top, there's a logo with three stylized orange flowers and the text "ANNIS". Below the logo, the title "ANNIS: Search and Visualization in Multilayer Linguistic Corpora" is displayed. A navigation bar with tabs for "Project", "ANNIS Query Language", "Visualizations", "Download", and "Access" is visible. A red banner at the top states: "New: stable version 3.0.1 (hotfix) has replaced 3.0.0 (What else is new in ANNIS3?)".

The main area is titled "ANNIS3" and contains a search interface. On the left, a "Search Form" panel shows a query: "AnniQL: <Jugendliche> & <pos=>/V.FIN/ & K2 ->dep[func='adv'] #1 & cat='S' & S1 & sentence #1". It also displays "Status: 1 match" and "1 document". To the right, the "Query Result" panel shows the search results for the query. A video player window displays a frame from a video showing a person writing the sentence "Jugendliche wollen und brauchen ohne auf die Idee" on a whiteboard. Below the video, the sentence is shown in German: "Jugendliche wollen und brauchen ohne auf die Idee". The interface then breaks down the sentence into its morphological components: "Jugendliche", "wollen", "und", "brauchen", "ohne", "auf", "die", "Idee". Each word is annotated with its part of speech (POS), such as "NOUN", "VERB", "CONJ", "VERB", "ADP", "DET", and "NOUN". A dependency tree diagram illustrates the grammatical structure, showing how each word depends on others. The root node is labeled "S". Other nodes include "NP" (NP), "VP" (VP), and "PP" (PP). The tree shows dependencies between "wollen" and "brauchen" (both labeled "VP"), "wollen" and "ohne" (labeled "PP"), and "ohne" and "auf" (labeled "PP"). The "PP" node "auf" is further connected to the noun "Idee".

Figure: ANNIS3—Zeldes, Zipser TU Berlin/Potsdam

ANNIS3

Advantages:

- Active online support + member contribution to the project.

ANNIS3

Advantages:

- Active online support + member contribution to the project.
- Integrates lot's of annotations.

ANNIS3

Advantages:

- Active online support + member contribution to the project.
- Integrates lot's of annotations.
- Very advanced query language.

ANNIS3

Advantages:

- Active online support + member contribution to the project.
- Integrates lot's of annotations.
- Very advanced query language.
- Download functionality (including context + export options).

ANNIS3

Advantages:

- Active online support + member contribution to the project.
- Integrates lot's of annotations.
- Very advanced query language.
- Download functionality (including context + export options).
- Frequency analysis (since updated version from last month).

ANNIS3

Advantages:

- Active online support + member contribution to the project.
- Integrates lot's of annotations.
- Very advanced query language.
- Download functionality (including context + export options).
- Frequency analysis (since updated version from last month).
- **Very (!) nice visualizations.**

ANNIS3

Disadvantages:

- Query language offers lot's of functionality → quite difficult.

ANNIS3

Disadvantages:

- Query language offers lot's of functionality → quite difficult.
- Limited to 2 million tokens per imported corpus.

ANNIS3

Disadvantages:

- Query language offers lot's of functionality → quite difficult.
- Limited to 2 million tokens per imported corpus.
 - The size of the Tüba-D/Z almost exceeds the limit.

ANNIS3

Disadvantages:

- Query language offers lot's of functionality → quite difficult.
- Limited to 2 million tokens per imported corpus.
 - The size of the Tüba-D/Z almost exceeds the limit.
 - Requires data base fine-tuning.

ANNIS3

Disadvantages:

- Query language offers lot's of functionality → quite difficult.
- Limited to 2 million tokens per imported corpus.
 - The size of the Tüba-D/Z almost exceeds the limit.
 - Requires data base fine-tuning.
- Difficult to import data (cf. *PAULA XML*, *relAnnis*, cascade of annotation scripts).

Online Demo:

<https://korpling.german.hu-berlin.de/annis3/>

Summary

- Two ready-to-use corpus UIs exist at the IEAS:
 - **Csnicper & ANNIS3**
- We encourage you to make use of their **full functionality** (in-class use/research purposes, etc.)
- More corpora can be imported upon request.