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

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

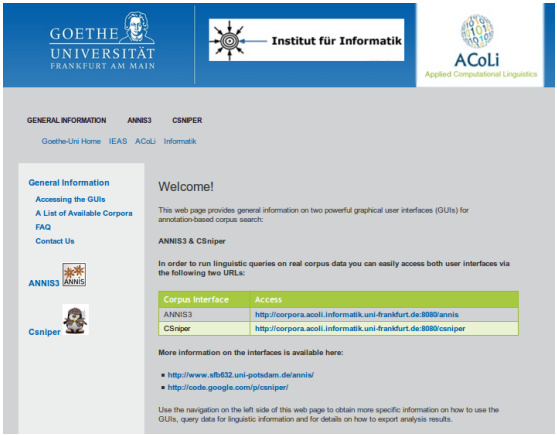

Figure: `http://corpora.acoli.informatik.uni-frankfurt.de:8080/anglistik/`

Documentation on Csniper & ANNIS3

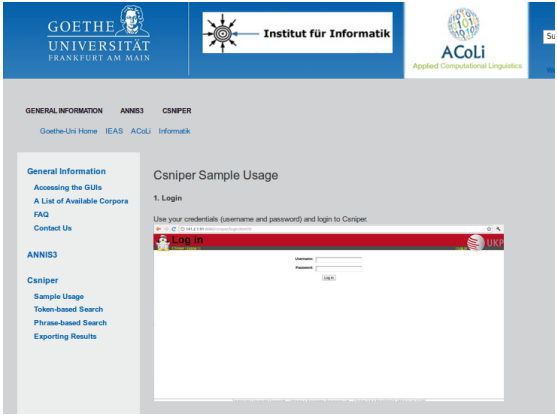

Figure: `http://corpora.acoli.informatik.uni-frankfurt.de:8080/anglistik/`

Documentation on Csniper & ANNIS3

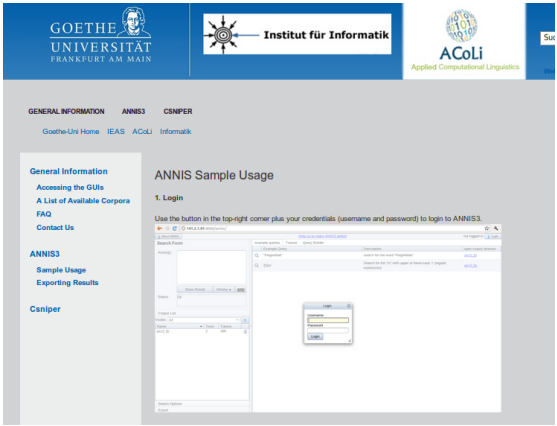

Figure: `http://corpora.acoli.informatik.uni-frankfurt.de:8080/anglistik/`

A List of Currently Available Corpora at the IEAS

	Properties			
Corpus	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

	Properties			
Corpus	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

	Properties			
Corpus	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

	Properties			
Corpus	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

	Properties			
Corpus	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

	Properties			
Corpus	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

	Properties			
Corpus	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

	Properties			
Corpus	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				
Lancester/IBM				
COLT				
Polytechnic of Walses 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**		tok, lem, pos, fully dependency-parsed
ukWaC (Open Source)		**2 billion**		tok, lem, pos, fully dependency-parsed
WaCkypedia_EN (Open Source)		**2 billion**		tok, lem, pos, fully dependency-parsed

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.

CSniper

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

CSniper

Advantages:

- Active support / Wiki / Programers 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

ANNIS3

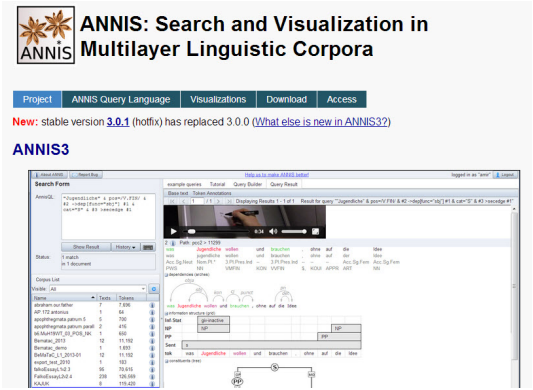

Figure: ANNIS3—Zeldes, Zipser TU Berlin/Potsdam

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 $\rightarrow$ quite difficult.

ANNIS3

Disadvantages:

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

ANNIS3

Disadvantages:

- Query language offers lot's of functionality $\rightarrow$ 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 $\rightarrow$ 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 $\rightarrow$ 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:
  - **Csniper & ANNIS3**
- We encourage you to make use of their **full functionality** (in-class use/research purposes, etc.)
- More corpora can be imported upon request.