Morphology in CLARIN-D

Daniël de Kok

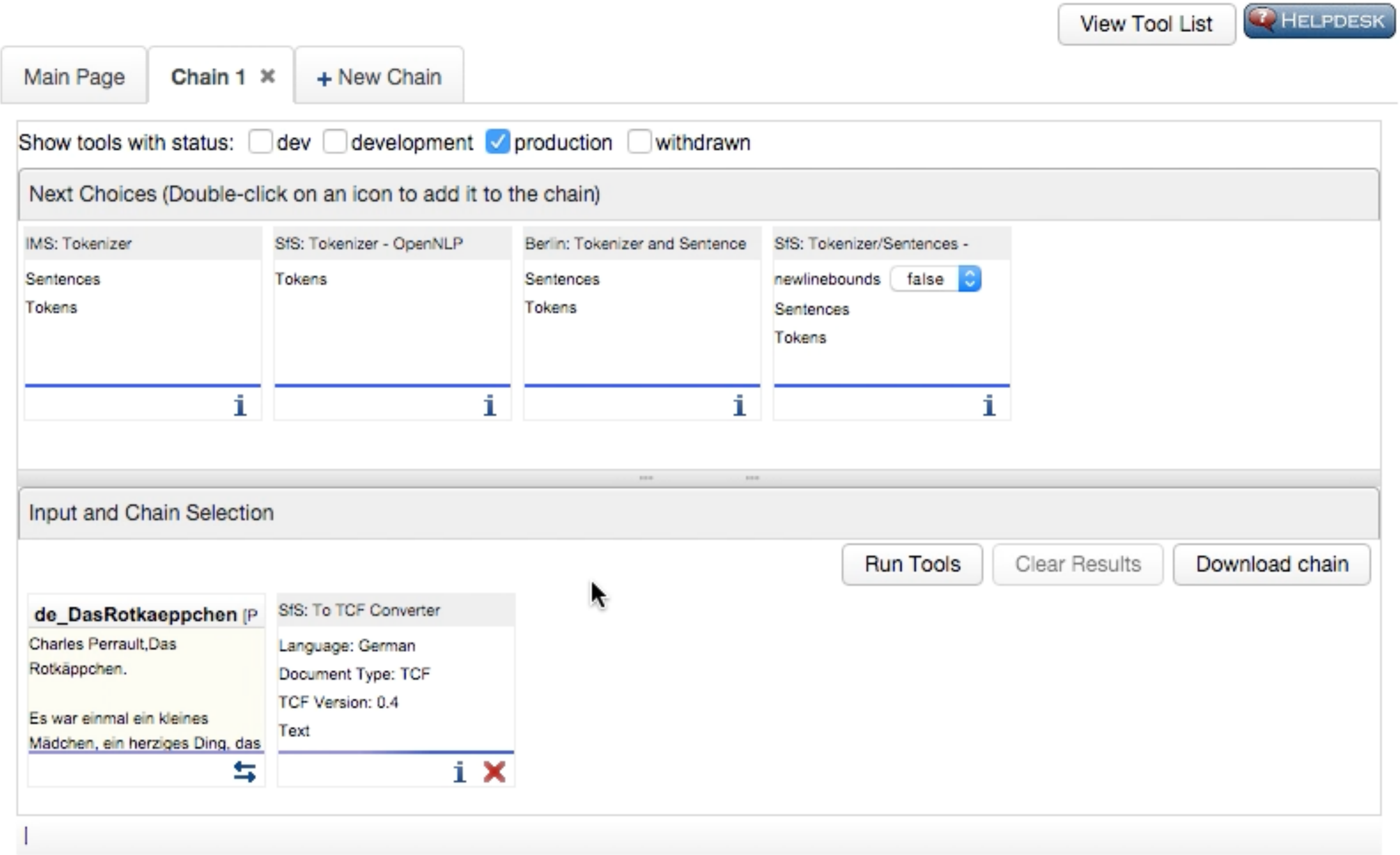
Introduction

A whirlwind introduction:

- CLARIN-D tools: WebLicht, TüNDRA
- Resources: corpora with morphology
- Mostly oriented towards inflectional morphology

WebLicht

WebLicht is a web application for creating and running NLP pipelines



Show tools with status: □ dev □ development ☑ production □ withdrawn

Next Choices (Double-click on an icon to add it to the chain)

IMS: Tokenizer
Sentences
Tokens

SfS: Tokenizer - OpenNLP
Tokens

Berlin: Tokenizer and Sentence
Sentences
Tokens

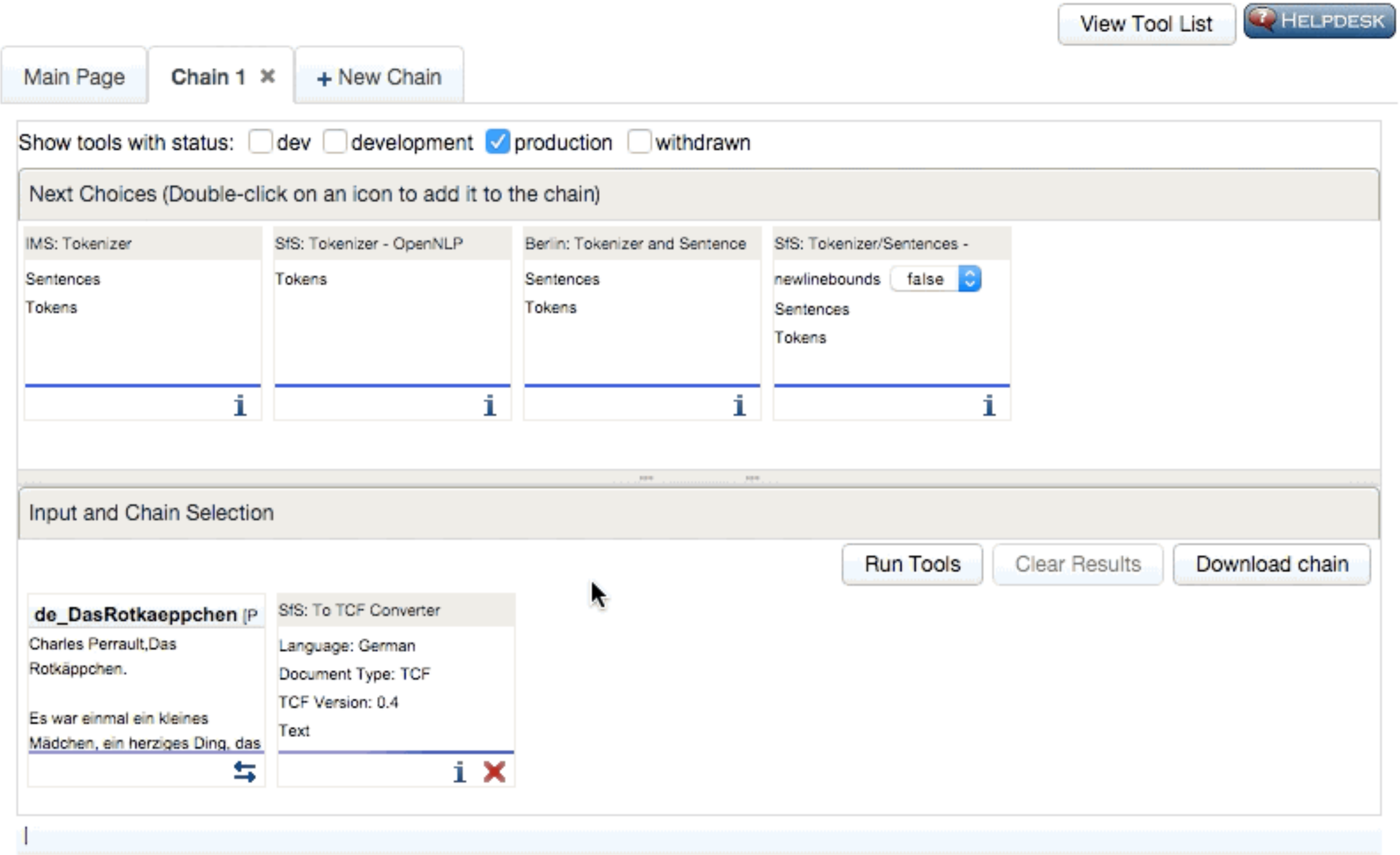
SfS: Tokenizer/Sentences -
newlinebounds false
Sentences
Tokens

Input and Chain Selection

de_DasRotkaeppchen [P
Charles Perrault,Das Rotkäppchen.

Es war einmal ein kleines Mädchen, ein herziges Ding, das

SfS: To TCF Converter
Language: German
Document Type: TCF
TCF Version: 0.4
Text



Show tools with status: □ dev □ development ☑ production □ withdrawn

IMS: Tokenizer
Sentences
Tokens

SfS: Tokenizer - OpenNLP
Tokens

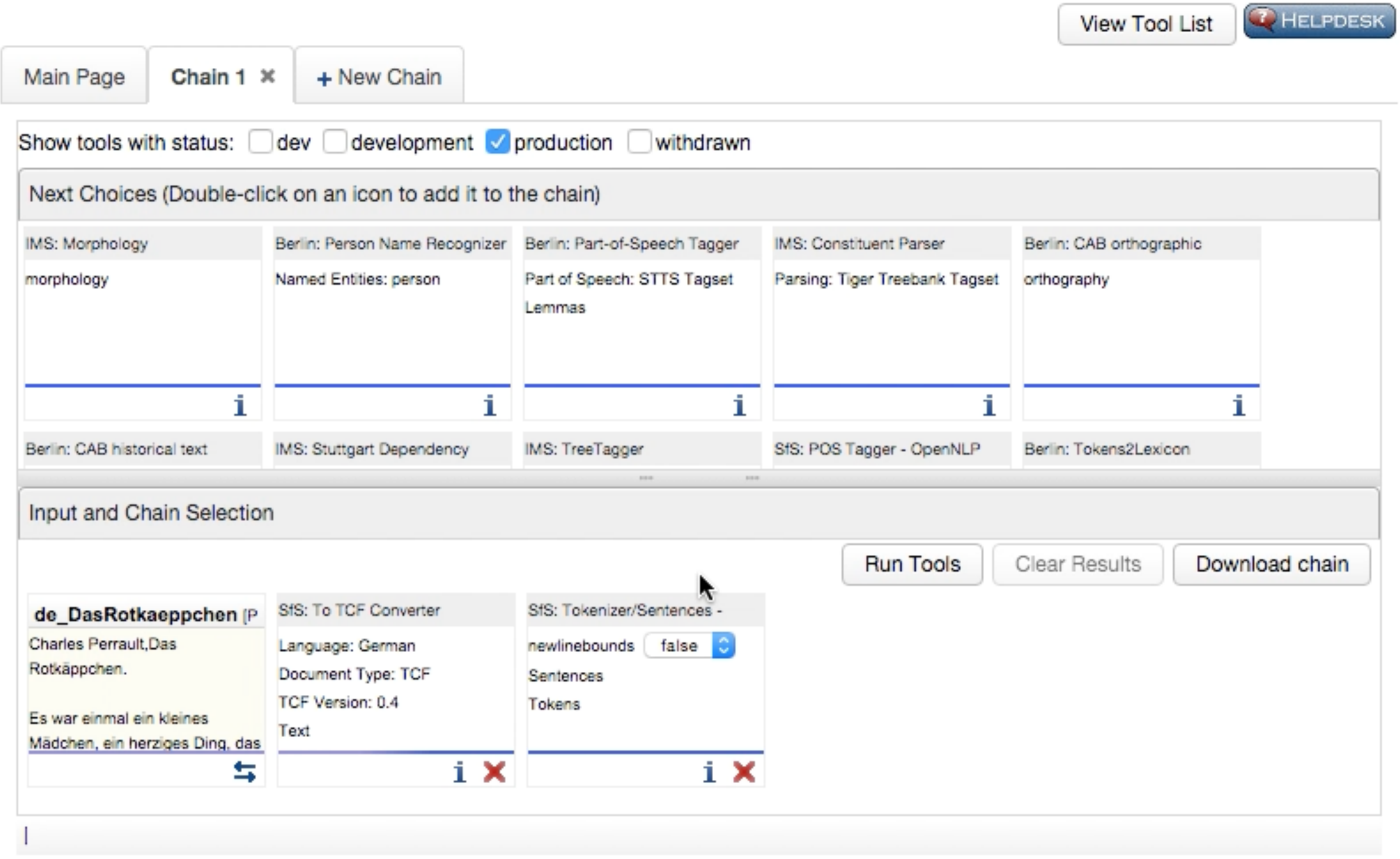
Berlin: Tokenizer and Sentence
Sentences
Tokens

SfS: Tokenizer/Sentences -
newlinebounds false
Sentences
Tokens

de_DasRotkaeppchen [P
Charles Perrault,Das Rotkäppchen.

Es war einmal ein kleines Mädchen, ein herziges Ding, das

SfS: To TCF Converter
Language: German
Document Type: TCF
TCF Version: 0.4
Text



Show tools with status: □ dev □ development ☑ production □ withdrawn

IMS: Morphology
morphology

Berlin: Person Name Recognizer
Named Entities: person

Berlin: Part-of-Speech Tagger
Part of Speech: STTS Tagset
Lemmas

IMS: Constituent Parser
Parsing: Tiger Treebank Tagset

Berlin: CAB orthographic
orthography

Berlin: CAB historical text

IMS: Stuttgart Dependency

IMS: TreeTagger

SfS: POS Tagger - OpenNLP

Berlin: Tokens2Lexicon

de_DasRotkaeppchen [P
Charles Perrault,Das Rotkäppchen.

Es war einmal ein kleines Mädchen, ein herziges Ding, das

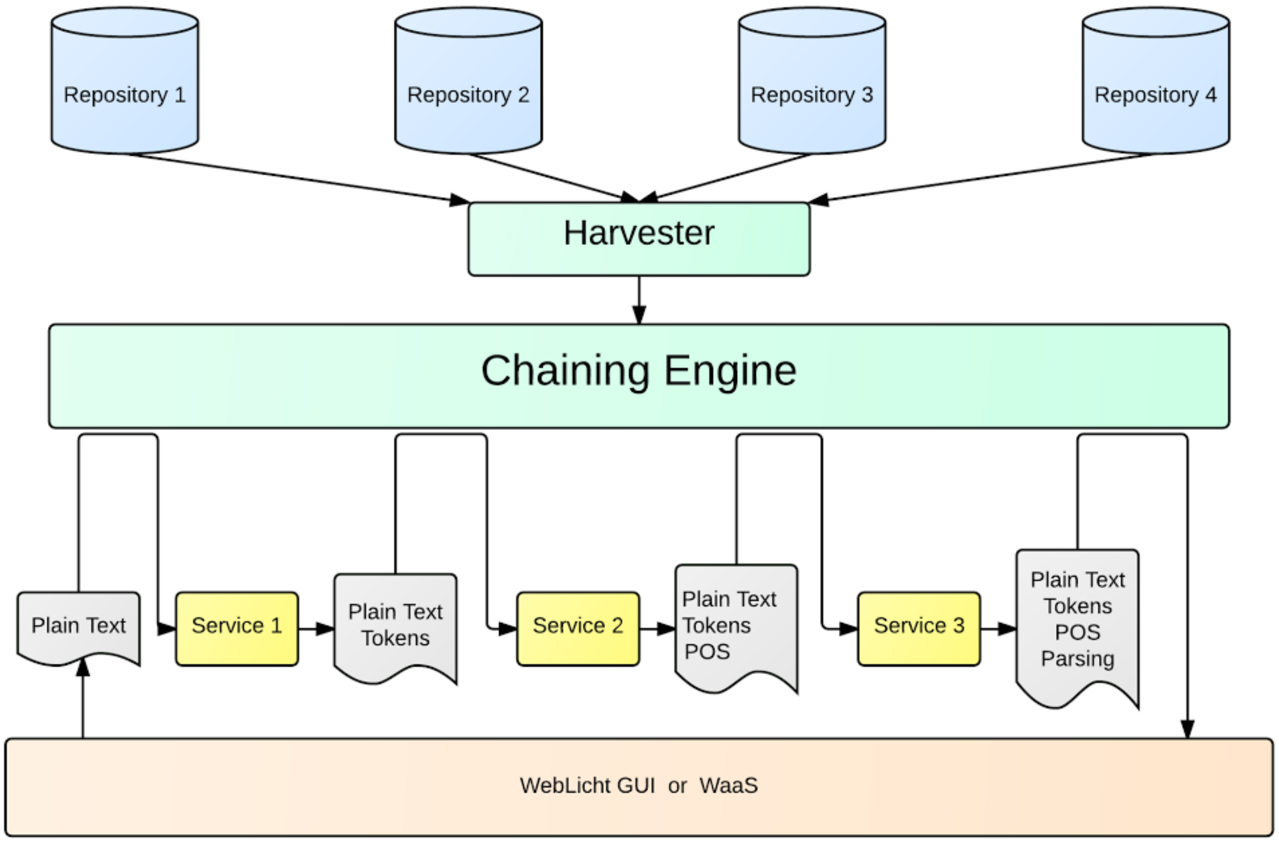
SfS: To TCF Converter
Language: German
Document Type: TCF
TCF Version: 0.4
Text

SfS: Tokenizer/Sentences -
newlinebounds false
Sentences
Tokens

Services

- Centers provide RESTful annotation services
  - Input: Text Corpus Format (TCF)
  - Output: TCF with the added layers
- Centers create metadata for their annotations services and put them in their repository

WebLicht architecture



Morphology services

- Currently available (morphological tagging):
  - German: **Stuttgart Morphology (RFTagger)**, SMOR
  - Dutch: **Alpino**
  - English: MorphAdorner
- Adding new services for morphology:
  - Since WebLicht is decentralized, any CLARIN center could add additional morphology services.
  - If some interesting tool is missing, let us know!

Stuttgart morphology (German)

- HMM tagger specialized for large, feature-rich tag sets.
- Trained on the Tiger treebank.
- Uses a supplementary lexicon.
- Outputs morphological tags in the TIGER morphology scheme:
    - Part-of-speech
    - Gender
    - Case
    - Number
    - Degree
    - Person
    - Tense
    - Mood
    - Finiteness

Alpino (Dutch)

- Wide-coverage dependency parser for Dutch.
- But also has:
  - An extensive lexicon with subcategorization frames.
  - A guesser for unknown words.
- Eventual frames are decided by:
  - Filtering by n-best tagging.
  - The parse selected by the disambiguation model.

Resources for German

- Semi-automatically annotated
  - Tiger treebank
  - TüBa-D/Z
- Automatically annotated
  - TüBa-D/W

Tiger treebank

- ~50,000 sentences
- Newspaper text (Frankfurter Rundschau)
- Semi-automatically annotated
- Annotations:
    - STTS part-of-speech tags
    - Lemmas
    - Inflectional morphology
    - Constituency structure
    - Dependency conversion (subset hand-annotated)

TüBa-D/Z

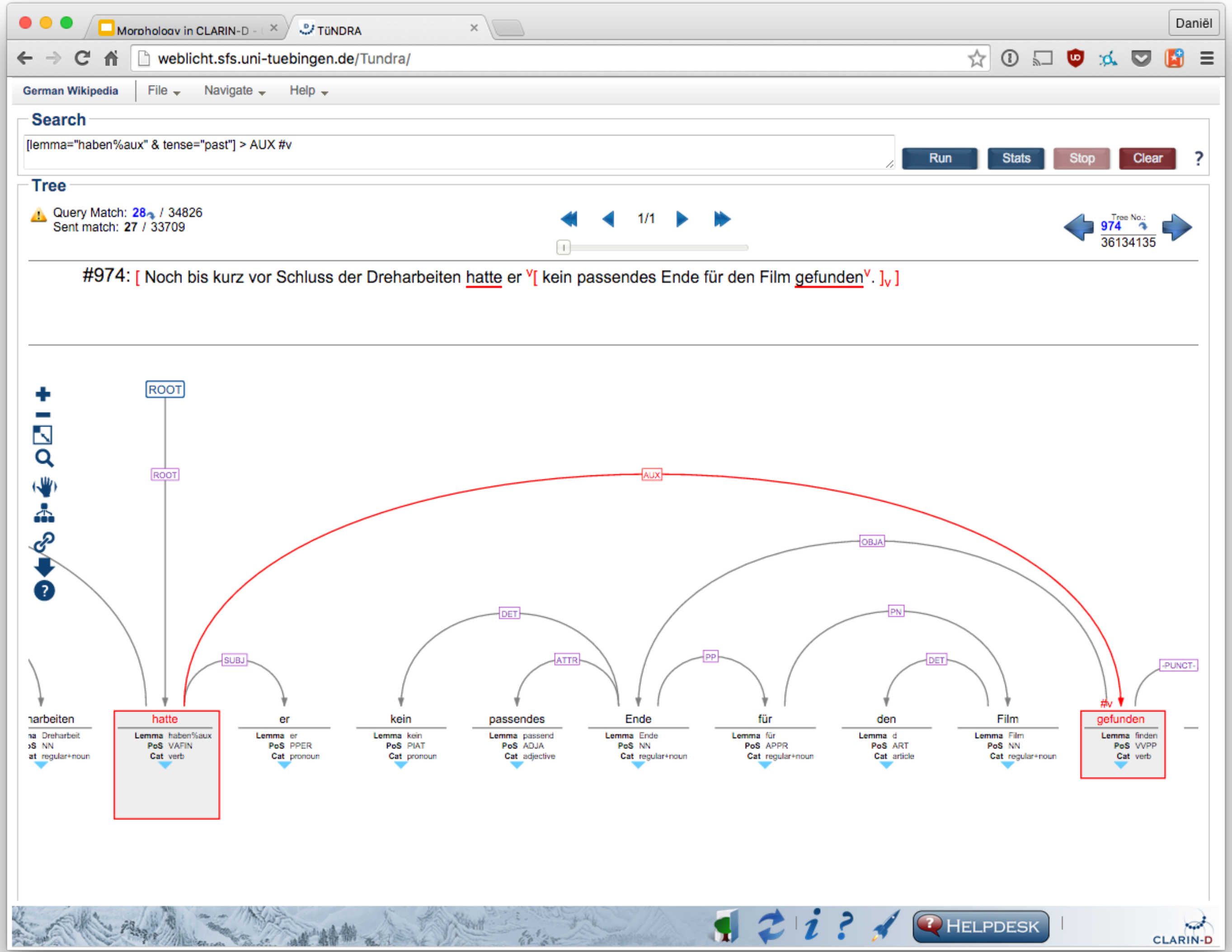
- ~95,500 sentences
- Newspaper text (*taz*)
- Semi-automatically annotated
- Annotations:
  - STTS part-of-speech tags
  - Lemmas
  - Inflectional morphology
  - Constituency structure
  - Dependency conversion
  - Anaphora and coreference relations
  - Subset with GermaNet word senses
  - Named entity class

TüBa-D/W

- 36.1 million sentences
- German Wikipedia
- Automatically annotated
- Annotations:
  - STTS part-of-speech tags
  - Lemmas
  - Inflectional morphology
  - Dependency structure
- Processed using WebLicht :-)

TüBa-D/W

TüBa-D/W is fully searchable using the TüNDRA treebank viewer



Search

[lemma="haben%aux" & tense="past"] > AUX #v

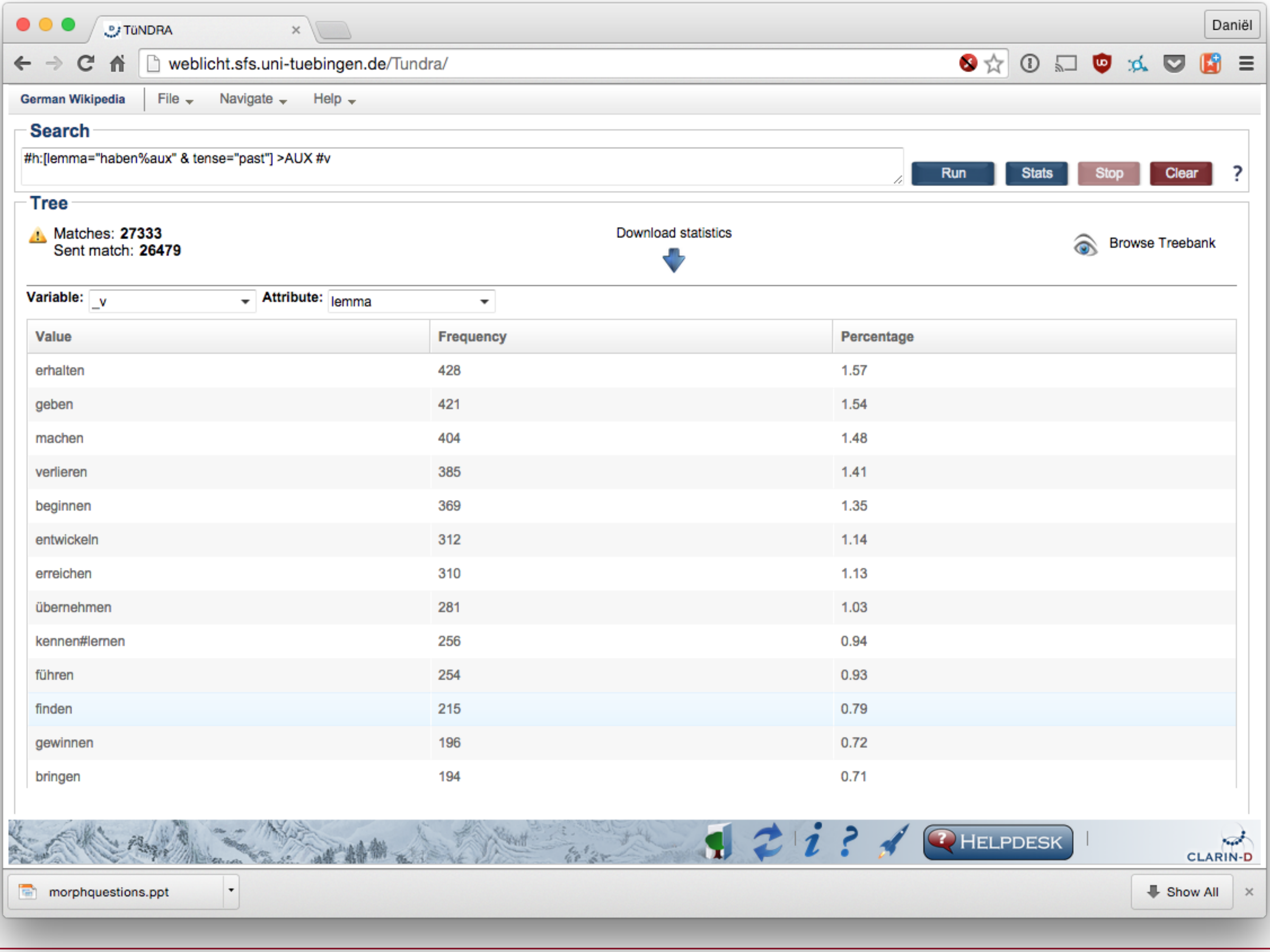
Tree

Query Match: 28 / 34826
Sent match: 27 / 33709

974
36134135

#974: [Noch bis kurz vor Schluss der Dreharbeiten hatte er ^{v}[kein passendes Ende für den Film gefunden^{v}.]$_{v}$]

	ıarbeiten	hatte	er	kein	passendes	Ende	für	den	Film	gefunden
Lemma	Dreharbeit	haben%aux	er	kein	passend	Ende	für	d	Film	finden
PoS	NN	VAFIN	PPER	PIAT	ADJA	NN	APPR	ART	NN	VVPP
Cat	regular+noun	verb	pronoun	pronoun	adjective	regular+noun	regular+noun	article	regular+noun	verb



```
#h:[lemma="haben%aux" & tense="past"] >AUX #v
```

Matches: **27333**
Sent match: **26479**

Variable: _v Attribute: lemma

Value	Frequency	Percentage
erhalten	428	1.57
geben	421	1.54
machen	404	1.48
verlieren	385	1.41
beginnen	369	1.35
entwickeln	312	1.14
erreichen	310	1.13
übernehmen	281	1.03
kennen#lernen	256	0.94
führen	254	0.93
finden	215	0.79
gewinnen	196	0.72
bringen	194	0.71

Links

WebLicht:

https://weblicht.sfs.uni-tuebingen.de/

TüNDRA:

https://weblicht.sfs.uni-tuebingen.de/Tundra/