CLARIN: A European Infrastructure for Language Resources and Technology

Erhard Hinrichs
University of Tübingen
CLARIN Board of Directors

Presentation Overview

CLARIN-D

- Introduction to CLARIN and CLARIN-D
- Survey of CLARIN-D Resources and Curation of Resources:
- CLARIN-D tools: WebLicht, WebMaus, CityViz, WhoIsInTheNews

- CLARIN ERIC established by European Commission in February 2012

- 9 Founding Member States of CLARIN ERIC:

  Austria, Bulgaria, Czech Republic, Denmark, Estonia, Germany, The Netherlands (Legal Seat), Nederlandse Taalunie (Flanders, the Netherlands), Poland

- Additional members are expected shortly: Croatia, Finland, France, Latvia, Lithuania, Norway, Slovenia

Main Goals of CLARIN

CLARIN-D

- Construction of an integrated, interoperable, and scalable research infrastructure **via a network of centers**
- Deployment of linguistic data, tools, and services
- Target group: researchers from the social sciences and humanities

CLARIN-D – Centers

BAS, University of Munich; Florian Schiel

BBAW, Berlin; Wolfgang Klein

IDS, Mannheim; Ludwig Eichinger

MPI, Nijmegen; Sebastian Drude

University of Hamburg; Kristin Bührig

University of Leipzig; Gerhard Heyer

Saarland University; Elke Teich

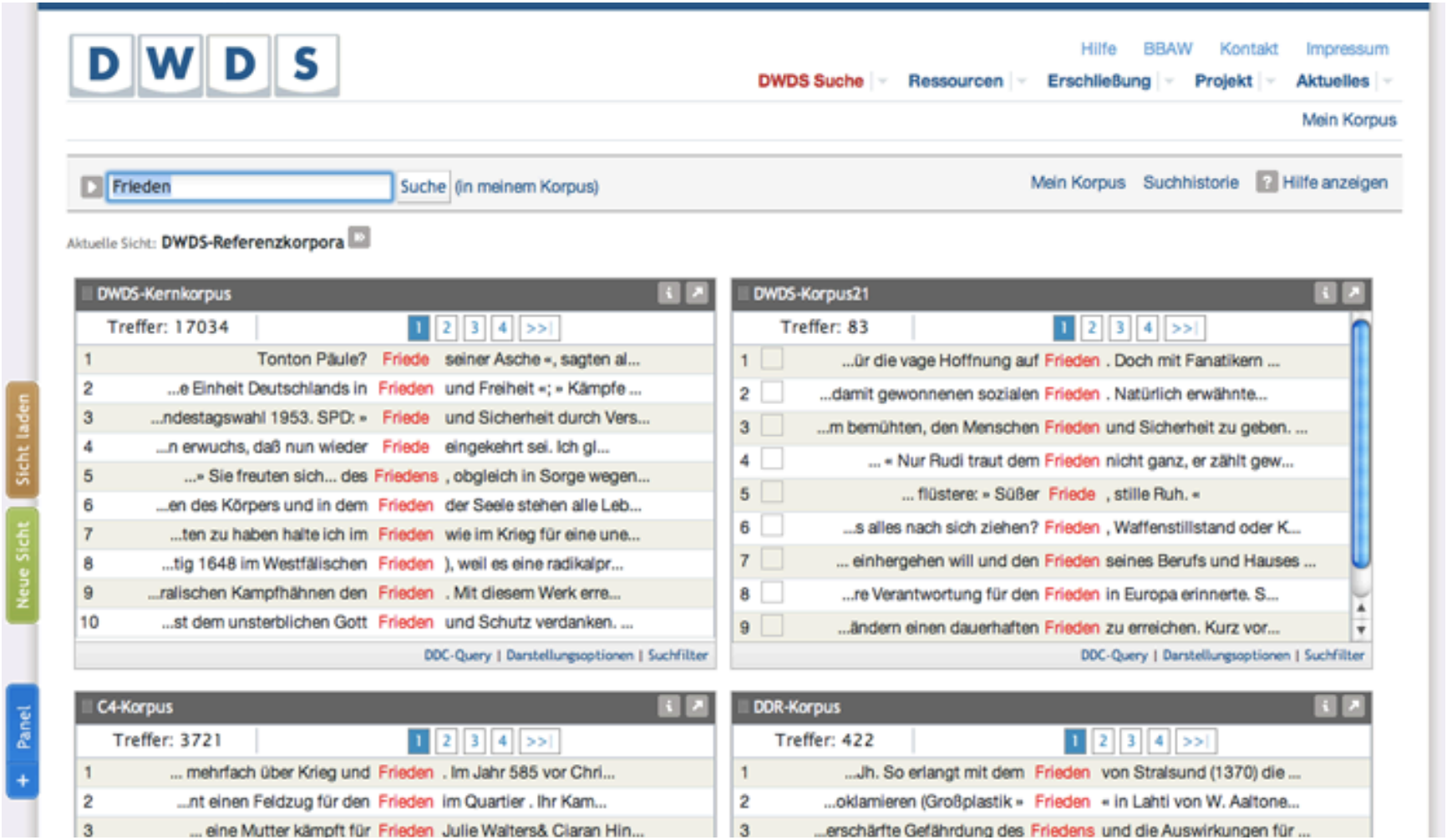
University of Stuttgart; Jonas Kuhn

University of Tübingen; Erhard Hinrichs

CLARIN-D – Integration

- Multitude of existing language resources and tools used by many researchers
- Seamless integration into the CLARIN-D infrastructure:
  - Corpora of spoken and written language, as well as multimedia corpora
  - Digital lexica
  - (Web-based) tools for the annotation and research of language resources

BBAW: Digitales Wörterbuch der Deutschen Sprache (DWDS)



Leipzig: Wortschatz-Portal

IDS: DeReKo and COSMAS

- Deutsches ReferenzKorpus (DeReKo)
  - largest collection of German electronic corpora (5.4 billion words)
  - Can be queried via COSMAS II (Corpus Search, Management and Analysis System)

BAS München: SIGNUM

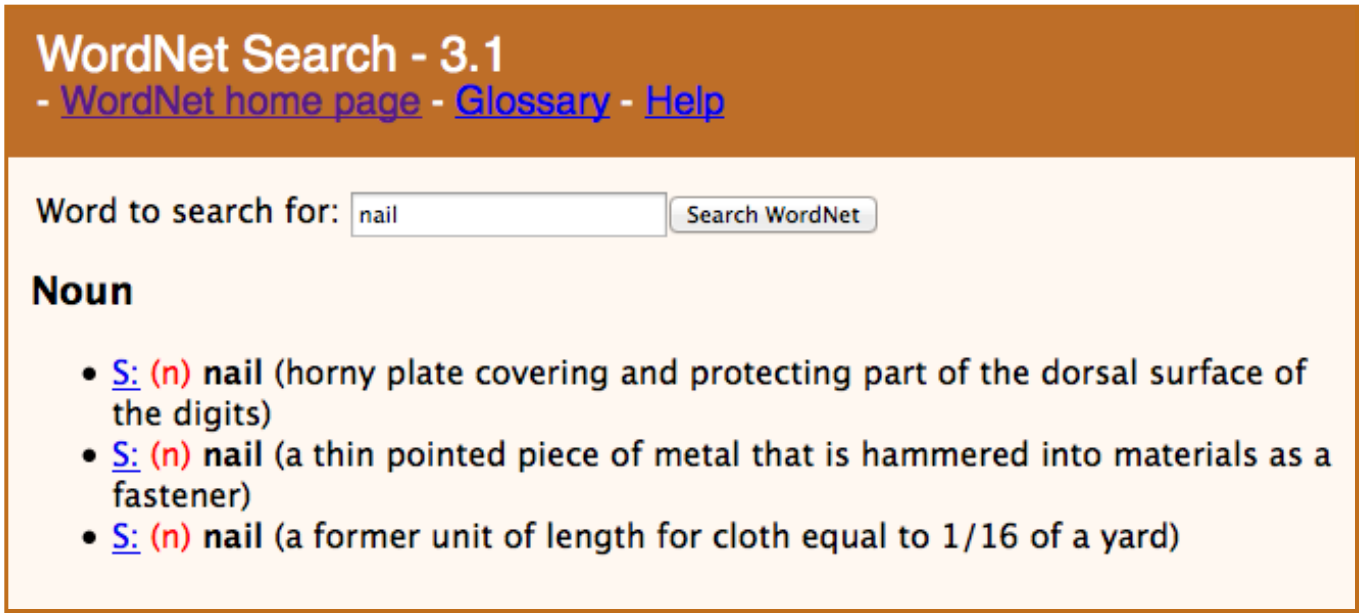
- First German sign language corpus
- Cooperation with University of Aachen
- 25 signers with 450 glosses and 780 ‚sentences‘
- 1 Terabyte of video data

GermaNet: Extend and Improve

CLARIN-D

- Sense descriptions
  - Crucial component for wordnets
  - Only 10% of all synsets in GermaNet had definitions
- Sense-annotated corpora
  - The comprehension of a word sense is much easier when its usages are illustrated by example sentences
  - Sense-annotated corpora are a prerequisite for word sense disambiguation
- Purely manual work would be arduous tasks
- The possibility of (semi-)automatic alternatives would be extremely valuable

GermaNet: Necessity of Sense Descriptions

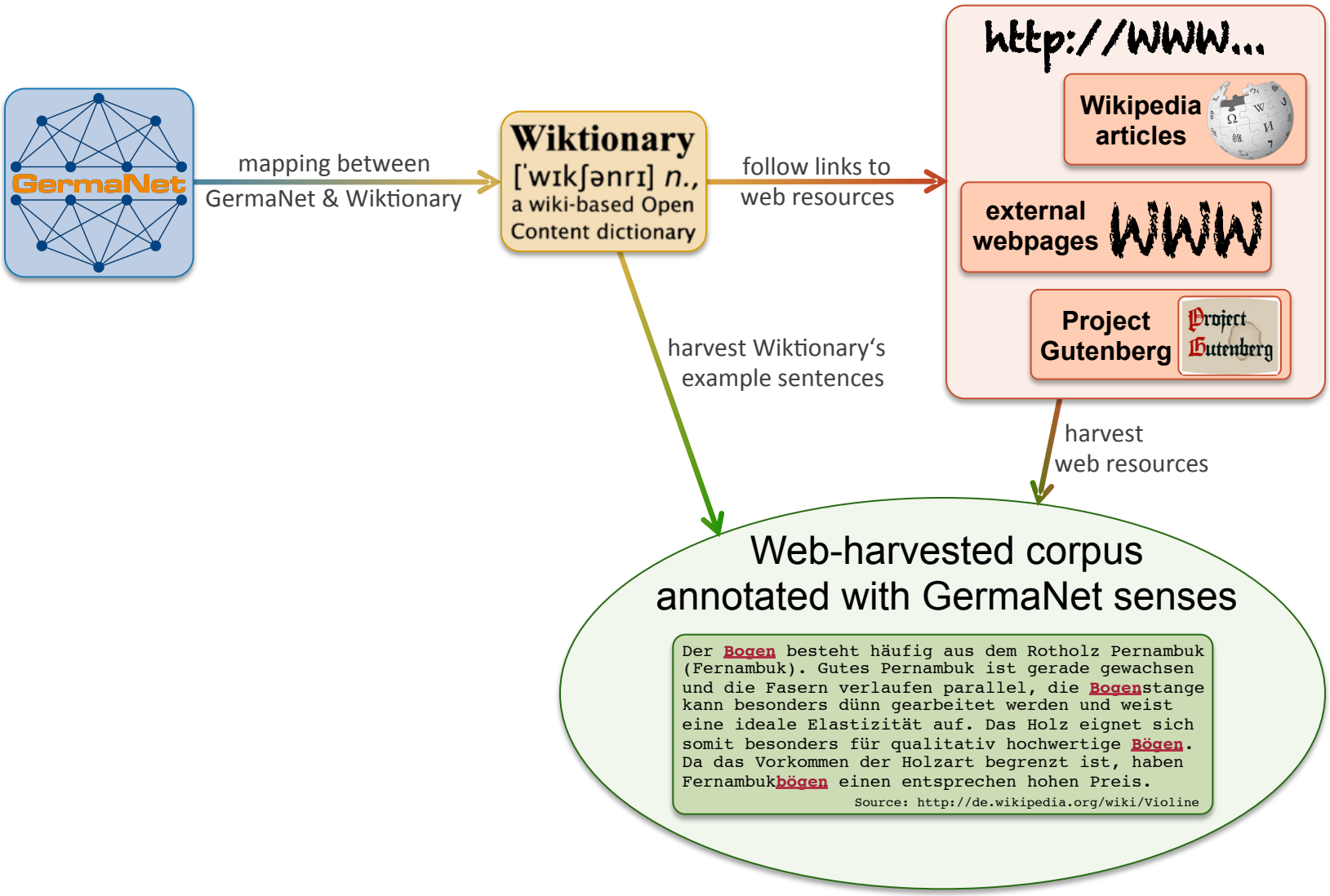


- Descriptions illustrate individual word senses in dictionaries
- For example: Princeton WordNet contains 3 senses for *nail*

GermaNet: Necessity of Sense Descriptions

- Descriptions illustrate individual word senses in dictionaries
- For example: Princeton WordNet contains 3 senses for *nail*
- Without definitions it is not easy to distinguish senses

GermaNet: The Big Picture



GermaNet-Wiktionary Mapping

CLARIN-D

Wiktionary
[ˈwɪkʃənrɪ] *n.*,
a wiki-based Open Content dictionary

Anzeige

Anzeige (Deutsch) [Bearbeiten]

Substantiv, *f* [Bearbeiten]

Bedeutungen:

[1] kurze Mitteilungen in den Medien, die der Bekanntmachung oder Werbung dienen

[2] *Recht:* Bekanntgabe einer Straftat bei einer Behörde

[3] *Technik:* eine Vorrichtung zur Signalisierung von Zuständen und Werten

[4] *veraltend:* ein sichtbarer Hinweis auf etwas Zukünftiges

Synonyme:

[1] Annonce; Inserat

[4] Anzeichen

Sinnverwandte Wörter:

[3] Display

Unterbegriffe:

[1] Familienanzeige, Geburtsanzeige, Heiratsanzeige, Hochzeitsanzeige, Kontaktanzeige, Todesanzeige, Traueranzeige, Verlobungsanzeige

[1] Werbeanzeige

[1] Kleinanzeige

[2] Selbstanzeige, Strafanzeige

[3] Temperaturanzeige, Wasserstandsanzeige

GermaNet

Anzeige

1. (n) [Anzeige] ‘advertisement’

synonyms: [Annonce, Inserat]

hypernyms: [Ausschreibung]

hyponyms: [Versandanzeige] [Kaufgesuch] [Verkaufsangebot] [Familienanzeige] [Partnergesuch, Kontaktanzeige] [Stellenanzeige, Stellenangebot, Stellenannonce] [Stellengesuch] [Kleinanzeige] [Großanzeige] [Zeitungsanzeige]

is part of : [Zeitung, Blatt, Gazette]

2. (n) [Anzeige] ‘complaint’

synonyms: [Strafanzeige]

hypernyms: [juristischer Text] [Klage, Anklage]

hyponyms: [Denunziation]

3. (n) [Anzeige] ‘display’

synonyms: [Display]

hypernyms: [Device, Gerät]

hyponyms: [Handydisplay] [Bildschirmseite]

GermaNet: "Sense Alignment“ Conclusion

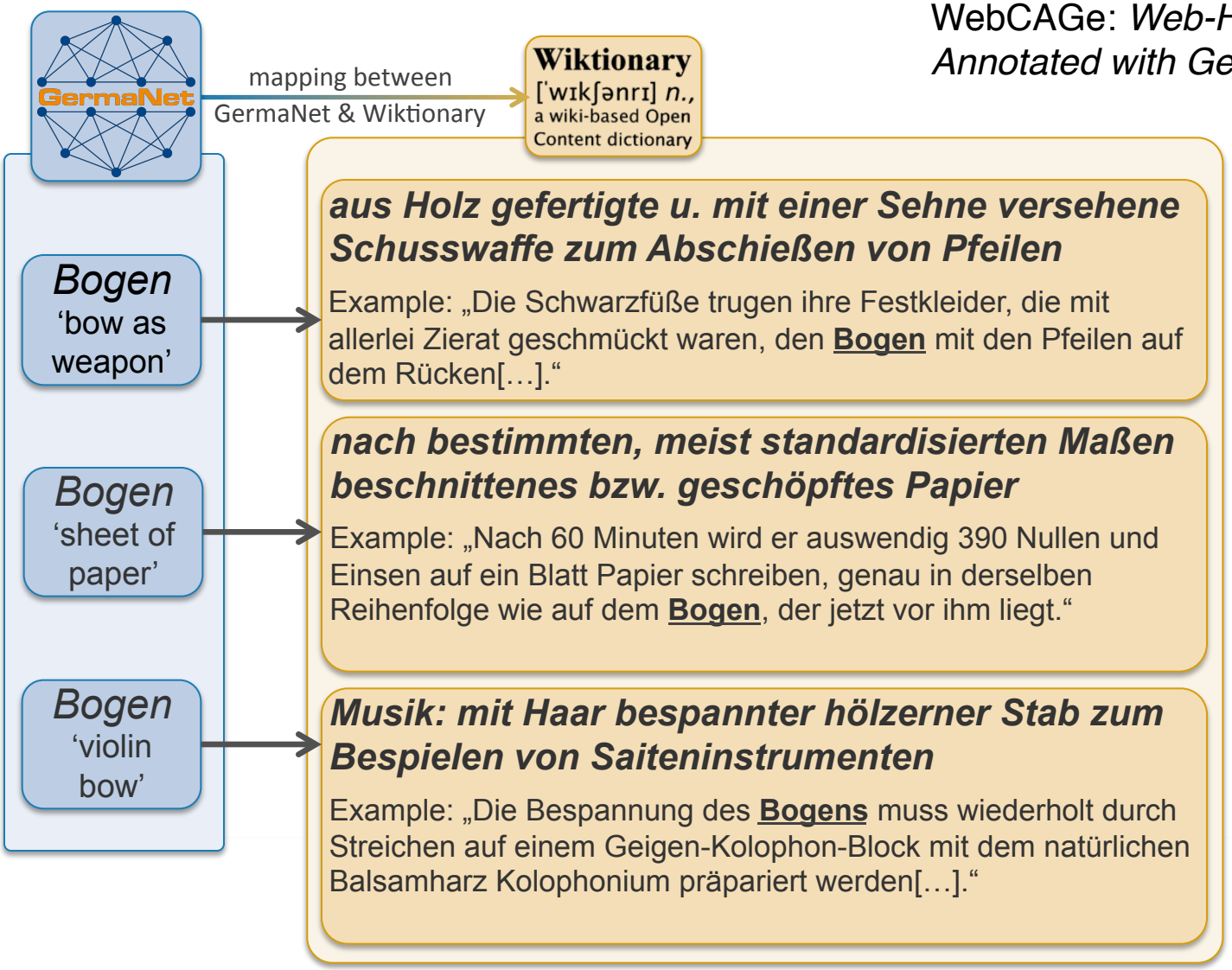
CLARIN-D

- We have presented a method for semi-automatically enriching GermaNet with sense definitions from Wiktionary
- Evaluation results underscore the feasibility of the approach
  - Accuracy: 91.9%
  - F1: 84.3
- After this alignment, 22296 synsets (32%) in GermaNet have sense definitions from Wiktionary
- All mapped definitions were manually post-corrected
- Extension of GermaNet is freely available
  - Included in GermaNet release
  - Online: http://www.sfs.uni-tuebingen.de/GermaNet/wiktionary.shtml

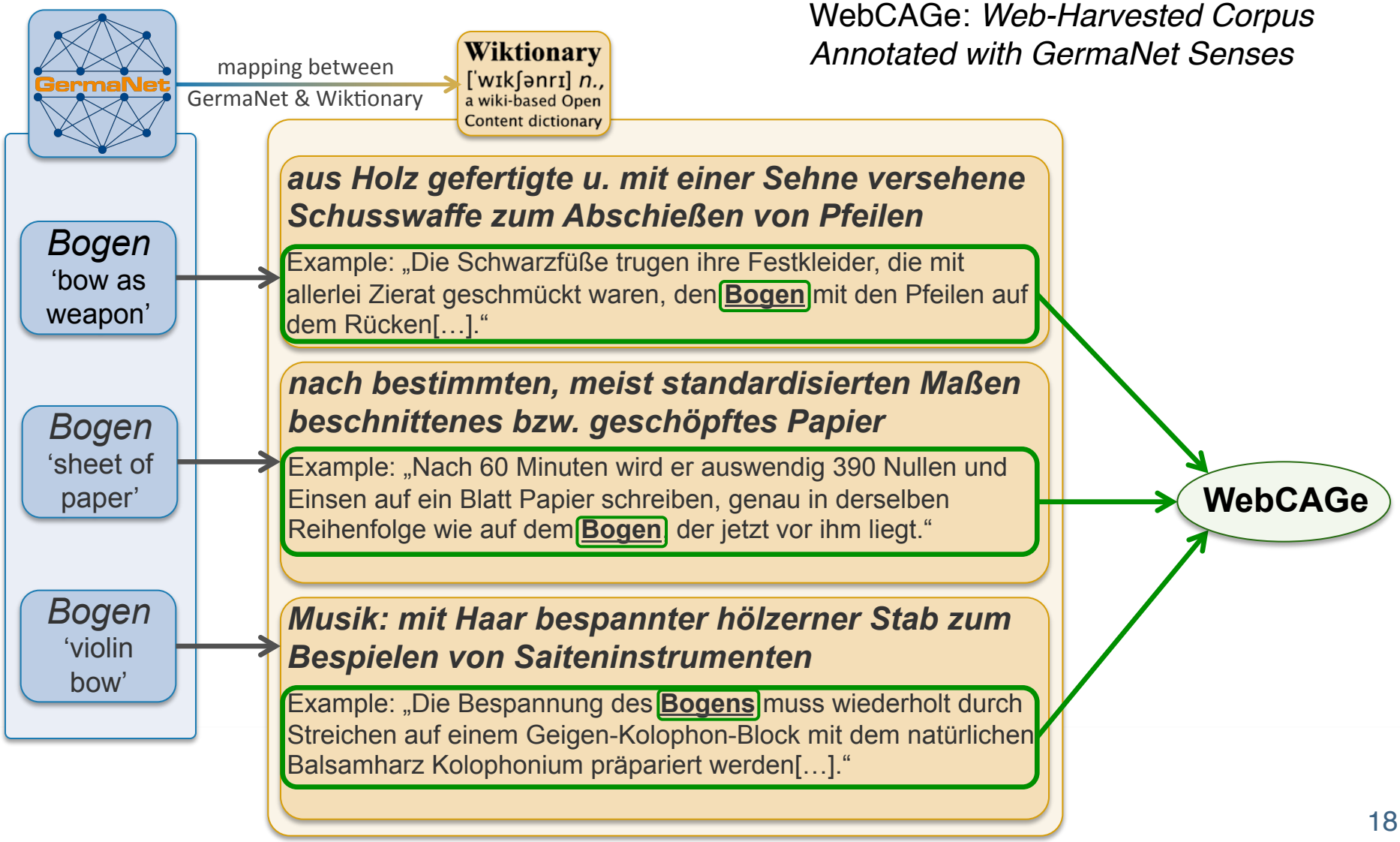
GermaNet: WebCAGe

CLARIN-D

WebCAGe: *Web-Harvested Corpus Annotated with GermaNet Senses*



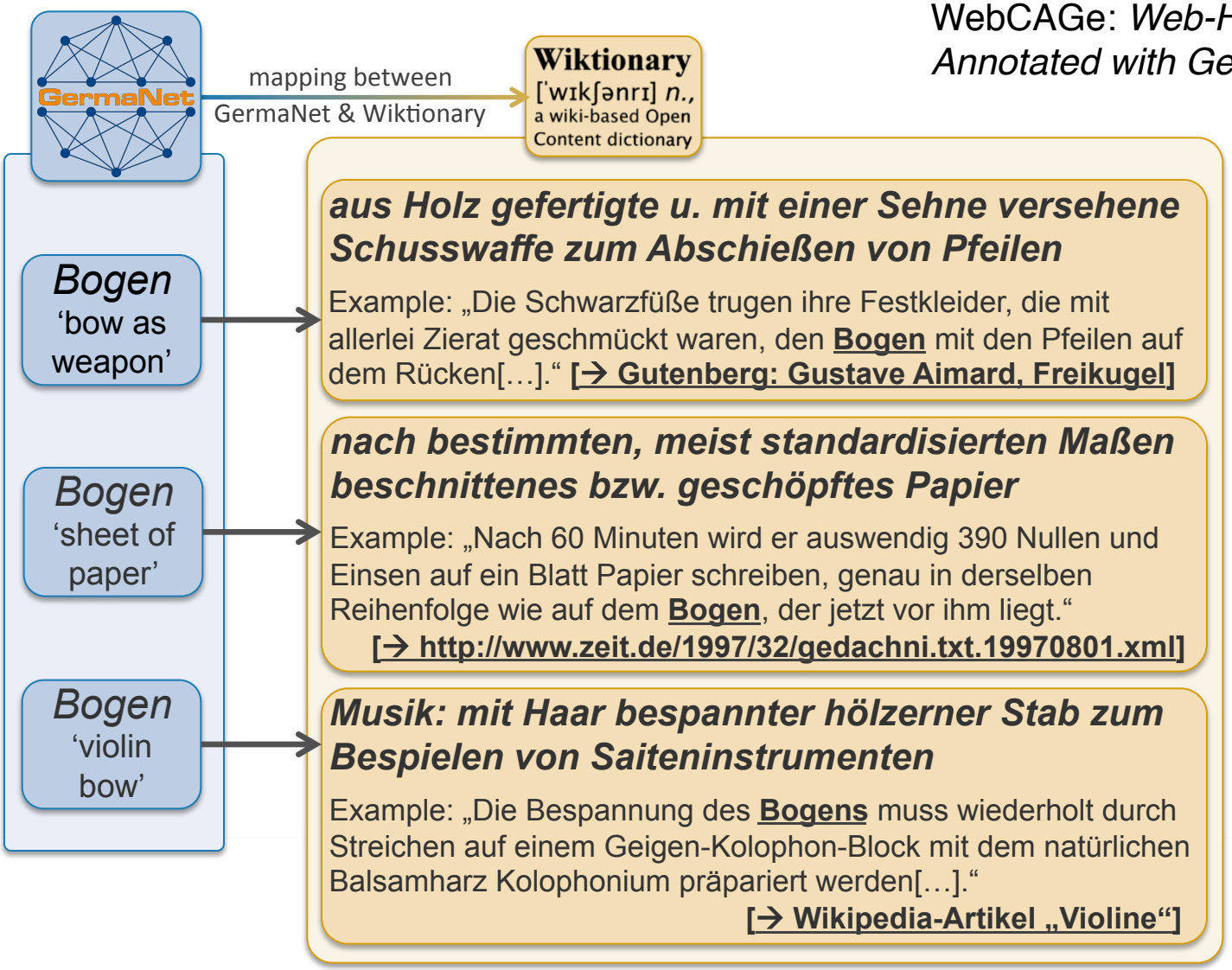
GermaNet: WebCAGe



GermaNet: WebCAGe

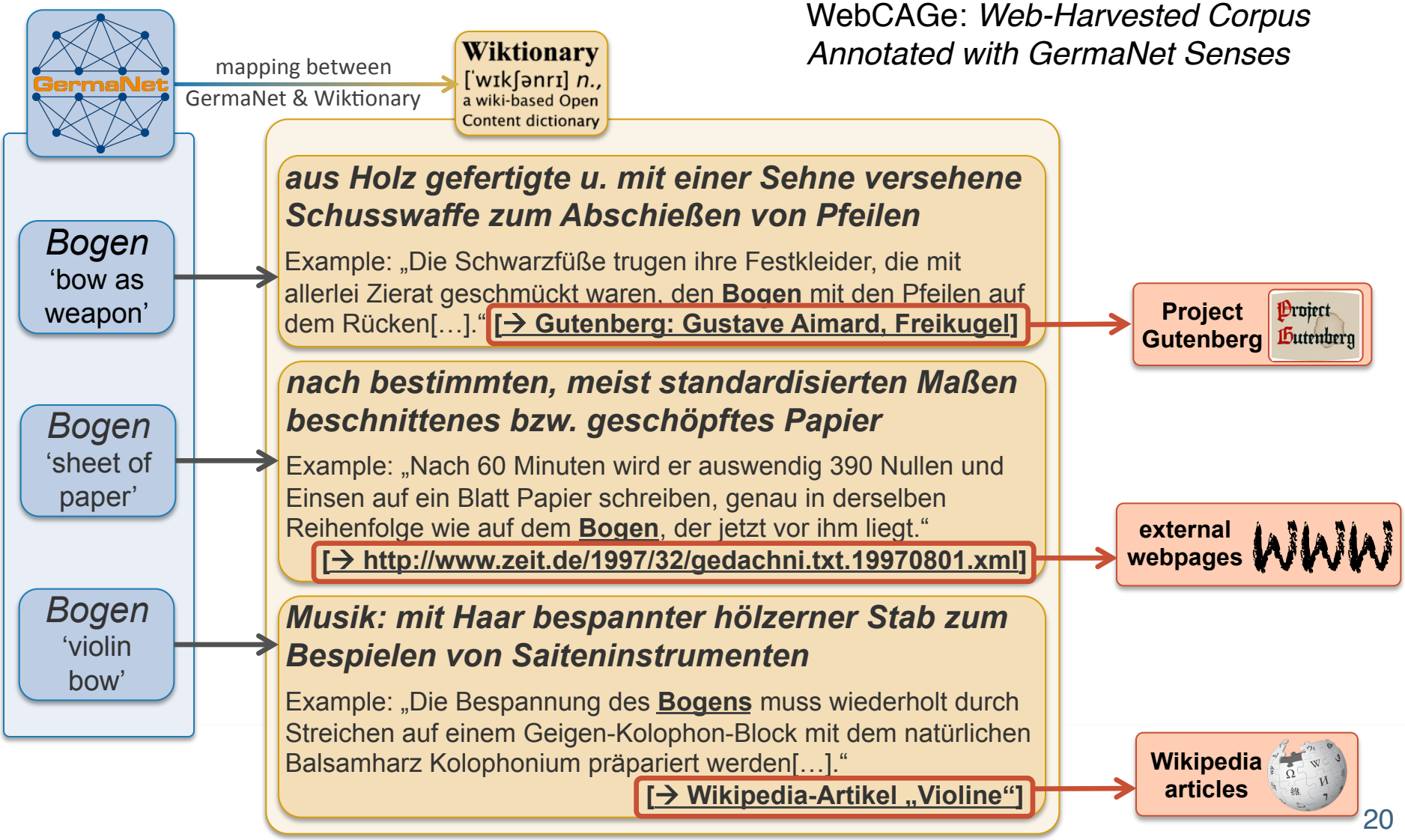
CLARIN-D

WebCAGe: *Web-Harvested Corpus Annotated with GermaNet Senses*



GermaNet: WebCAGe

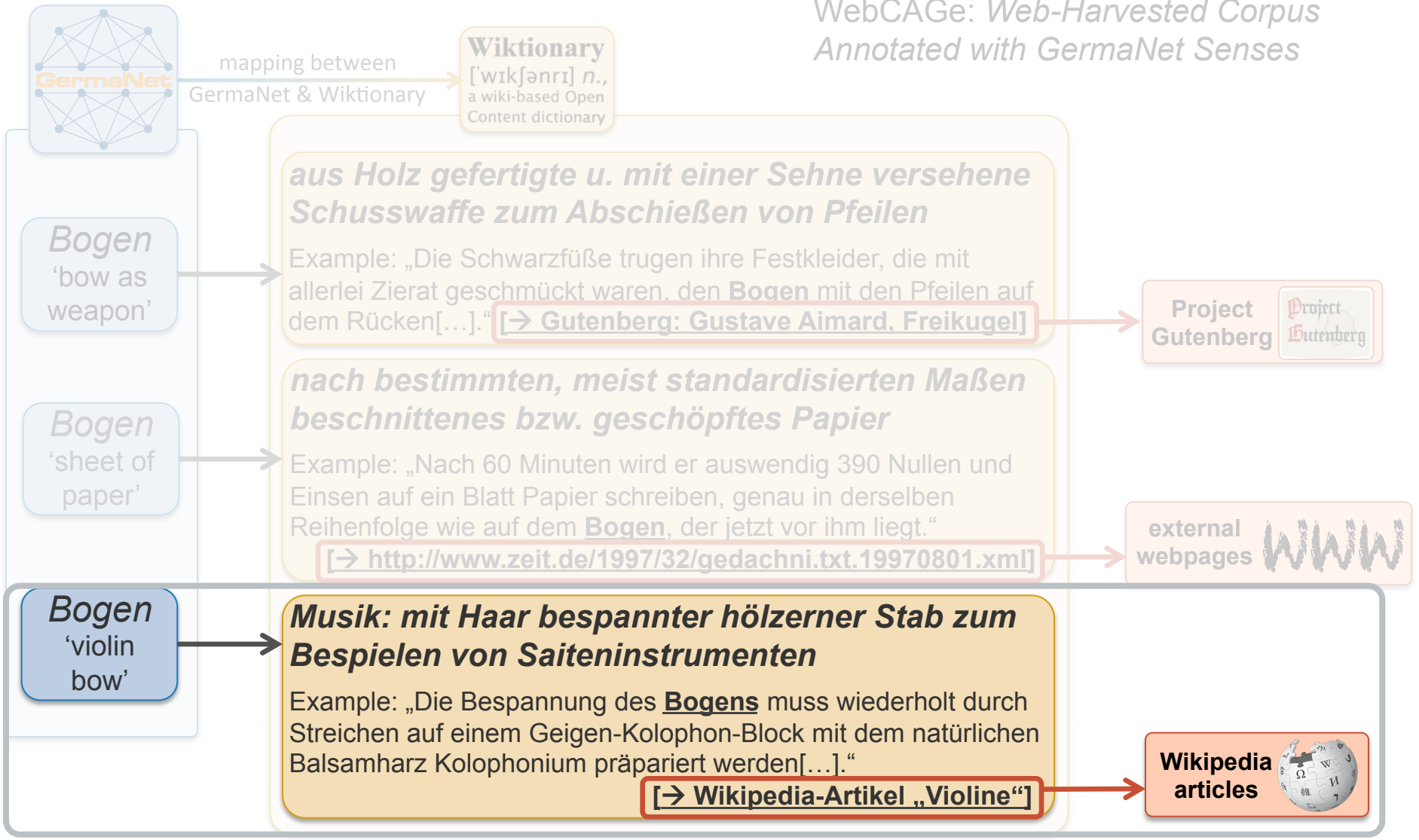
CLARIN-D



WebCAGe: *Web-Harvested Corpus Annotated with GermaNet Senses*

GermaNet: WebCAGe

CLARIN-D



Cropped Wikipedia article from:
http://de.wikipedia.org/wiki/Violine

Violine

Die **Violine** (**Geige**, Abk.: *Vl.*) ist ein Streichinstrument aus verschiedenen Hölzern. Ihre vier Saiten (g – d^1 – a^1 – e^2) werden mit einem Bogen gestrichen. In der Tradition der klassischen europäischen Musik spielt die Violine eine wichtige Rolle – viele große Komponisten haben ihr bedeutende Teile ihres Schaffens gewidmet. Violinen werden von Geigenbauern hergestellt.

Der Bogen [Bearbeiten]

Der *Bogen* besteht häufig aus dem Rotholz Pernambuk . Gutes Pernambuk ist gerade gewachsen und die Fasern verlaufen parallel, die Bogenstange kann dann besonders dünn gearbeitet werden und weist eine ideale Elastizität auf. Pernambuk eignet sich somit besonders für qualitativ hochwertige Bögen. Da das Vorkommen der Holzart begrenzt ist, haben Pernambukbögen einen hohen Preis. Einfachere Schülerbögen sind meist aus Brasilholz gefertigt. Heute werden, auch von Berufsgeigern, zunehmend Bögen aus Kohlefaser (Karbonfiber) verwendet.

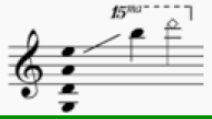
Am unteren Ende des Bogens befindet sich der sogenannte *Frosch* aus Ebenholz, meist verziert mit einer runden Perlmutt-Einlage. Zwischen Frosch und Bogenspitze (Köpfchen) sind die Bogenhaare eingespannt. Dies sind ca. 180 bis 250 Haare vom Hengstschweif[3] bestimmter Pferderassen. Durch das Drehen einer Schraube (Beinchen) wird der Bogen in Spannung versetzt (die Spannung muss nach dem Spiel jeweils wieder gelöst werden). Die Haare verfügen über feine Widerhaken, welche die Saiten beim Darüberstreichen in Schwingung bringen. Dafür müssen die Haare aber zuvor mit Kolophonium (natürliches Balsamharz) präpariert werden. Das erreicht man durch mehrfaches Streichen des Bogens über einen Kolophonium-Block.

Violine

engl.: *violin*, ital.: *violino*

Klassifikation Chordophon
Streichinstrument

Tonumfang



Bogen
'violin bow'

Musik: mit Haar bespannter hölzerner Stab zum Bespielen von Saiteninstrumenten

Example: „Die Bespannung des **Bogens** muss wiederholt durch Streichen auf einem Geigen-Kolophon-Block mit dem natürlichen Balsamharz Kolophonium präpariert werden[…]."

[→ Wikipedia-Artikel „Violine"]

Wikipedia articles

GermaNet: WebCAGe

CLARIN-D

WebCAGe

"one sense per discourse" heuristic

Violine

Cropped Wikipedia article from: http://de.wikipedia.org/wiki/Violine

Die **Violine** (**Geige**, Abk.: *Vl.*) ist ein Streichinstrument aus verschiedenen Hölzern. Ihre vier Saiten (g – d^1 – a^1 – e^2) werden mit einem Bogen gestrichen. In der Tradition der klassischen europäischen Musik spielt die Violine eine wichtige Rolle – viele große Komponisten haben ihr bedeutende Teile ihres Schaffens gewidmet. Violinen werden von Geigenbauern hergestellt.

Der Bogen [Bearbeiten]

Der *Bogen* besteht häufig aus dem Rotholz Pernambuk . Gutes Pernambuk ist gerade gewachsen und die Fasern verlaufen parallel, die Bogenstange kann dann besonders dünn gearbeitet werden und weist eine ideale Elastizität auf. Pernambuk eignet sich somit besonders für qualitativ hochwertige Bögen. Da das Vorkommen der Holzart begrenzt ist, haben Pernambukbögen einen hohen Preis. Einfachere Schülerbögen sind meist aus Brasilholz gefertigt. Heute werden, auch von Berufsgeigern, zunehmend Bögen aus Kohlefaser (Karbonfiber) verwendet.

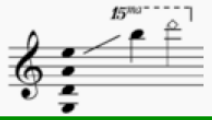
Am unteren Ende des Bogens befindet sich der sogenannte *Frosch* aus Ebenholz, meist verziert mit einer runden Perlmutt-Einlage. Zwischen Frosch und Bogenspitze (Köpfchen) sind die Bogenhaare eingespannt. Dies sind ca. 180 bis 250 Haare vom Hengstschweif[3] bestimmter Pferderassen. Durch das Drehen einer Schraube (Beinchen) wird der Bogen in Spannung versetzt (die Spannung muss nach dem Spiel jeweils wieder gelöst werden). Die Haare verfügen über feine Widerhaken, welche die Saiten beim Darüberstreichen in Schwingung bringen. Dafür müssen die Haare aber zuvor mit Kolophonium (natürliches Balsamharz) präpariert werden. Das erreicht man durch mehrfaches Streichen des Bogens über einen Kolophonium-Block.

Violine

engl.: *violin*, ital.: *violino*

Klassifikation Chordophon, Streichinstrument

Tonumfang



Bogen 'violin bow'

Musik: mit Haar bespannter hölzerner Stab zum Bespielen von Saiteninstrumenten

Example: „Die Bespannung des **Bogens** muss wiederholt durch Streichen auf einem Geigen-Kolophon-Block mit dem natürlichen Balsamharz Kolophonium präpariert werden[…]."

[→ Wikipedia-Artikel „Violine"]

Wikipedia articles

GermaNet: Size of WebCAGe (Version 1.0)

CLARIN-D

		WebCAGe	Further information
Number of tagged words	adjectives	217	2.5 senses per word
	nouns	1539	2.8 senses per word
	verbs	962	3.6 senses per word
	all word classes	**2718**	3.0 senses per word
Number of tagged word tokens	Wiktionary examples	7648	7378 examples
	Wikipedia articles	1470	56 articles
	Gutenberg texts	757	101 texts
	external webpages	566	252 pages
	all texts	**10441**	
Domain independent		yes	

GermaNet: Evaluation of the Algorithm

CLARIN-D

- Precision and recall above 94%
- One deviation: precision for verbs is slightly lower
- The average precision is of sufficient quality to be used as is if approximately 2-5% noise is acceptable
- In order to assure good quality, all annotations have been manually verified
- Heuristic “one sense per discourse” was highly reliable:
  - 99.96% Wiktionary example sentences
  - 95.62% Wikipedia articles
  - 96.75% external webpages

		All texts
Precision	adjectives	98.21%
	nouns	96.18%
	verbs	89.80%
	all word classes	94.30%
Recall	adjectives	97.54%
	nouns	96.10%
	verbs	96.58%
	all word classes	96.01%

GermaNet Related Work: WikiCAGe

- WikiCAGe: short for *Wikipedia-Harvested Corpus Annotated with GermaNet Senses*
- Similar approach to the already presented one: create a mapping of GermaNet with Wikipedia to automatically harvest sense-annotated Wikipedia texts
- The mapping is based on a combination of variants of the Lesk and PageRank algorithms
- The algorithm for the target word identification has been reused on Wikipedia texts
- Comparison to WebCAGe:
  - WikiCAGe has more occurrences per target word
  - WebCAGe covers adjectives, nouns, and verbs, whereas WikiCAGe covers nouns only

GermaNet: "Sense-Annotated Corpus" Conclusion

CLARIN-D

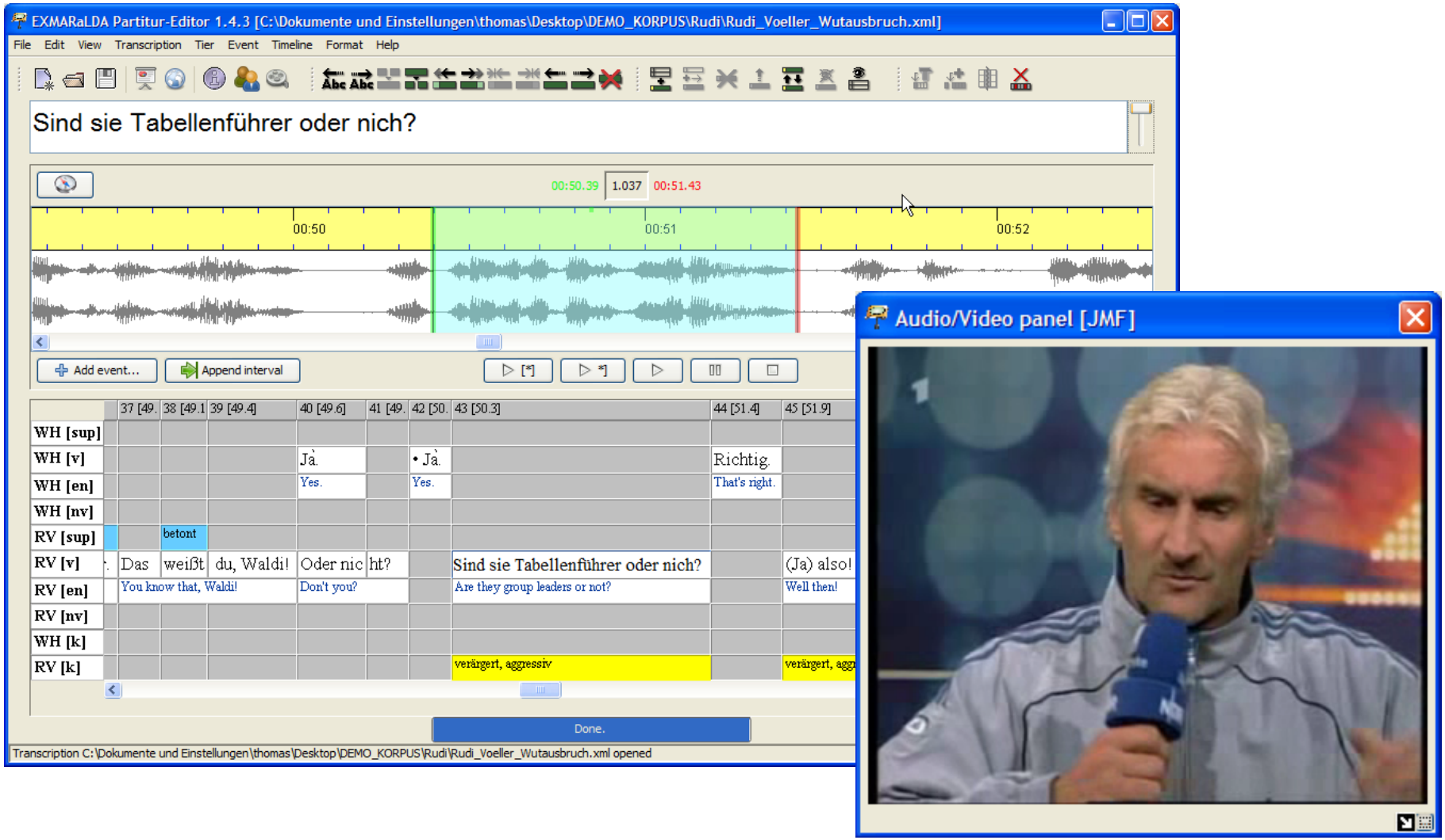
- The results prove the viability of the proposed method for automatically creating a sense-annotated corpus
- WebCAGe currently represents the largest sense-annotated corpus available for German
- WebCAGe (Wiktionary and Wikipedia parts for now) is freely available at: http://www.sfs.uni-tuebingen.de/en/webcage.shtml

The CLARIN-D Vision

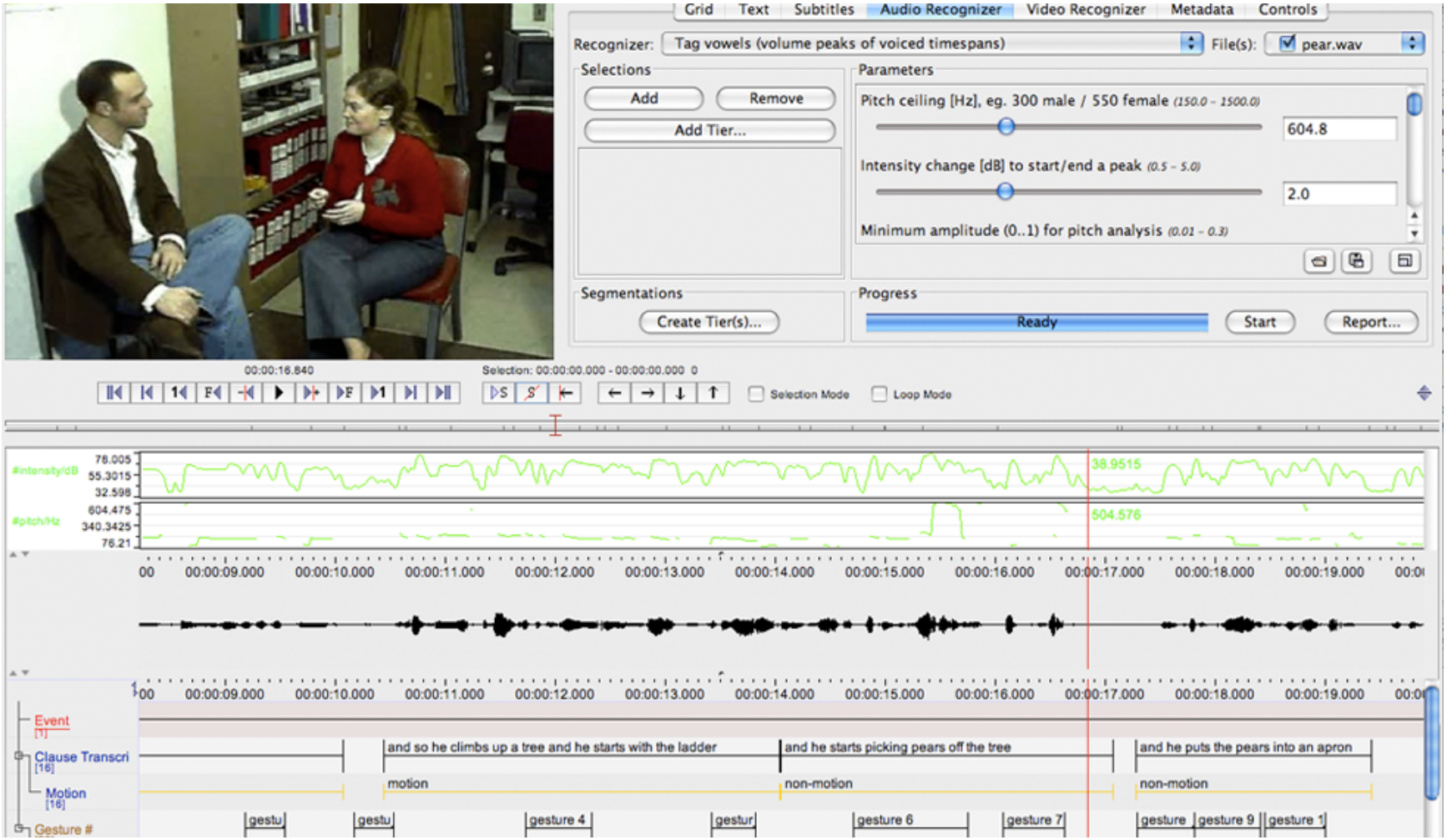
- Status quo: Many language resources and tools are available to a large user community

- Goal: Integration of these resources und tools in a single, user-friendly CLARIN-D portal

Hamburg: EXMARaLDA – Transcription and Annotation Tool for Video Data



MPI: ELAN – Multimedia-Annotation Tool



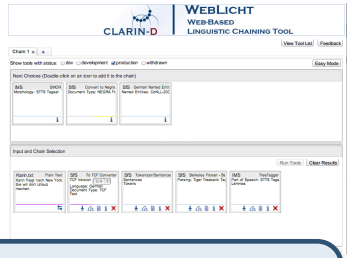
WebLicht - Motivation

CLARIN-D

- Many linguistic resources (corpora, dictionaries, …) and tools (tokenizer, tagger, parser, …) are available
- Most of them are implemented to run on local machines (inconvenient and time-consuming)
- Requirement: avoid “download-first” paradigm and make linguistic resources and tools available on the web

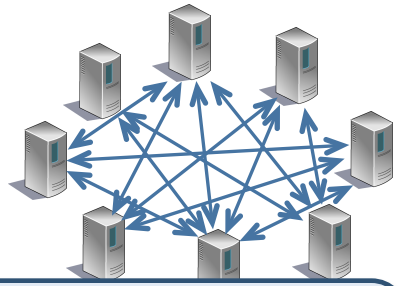
→ Solution: **WebLicht** (**Web**-Based **L**inguistic **Ch**aining **T**ool)

WebLicht - Architecture



Web user interface: interacts with the user

Registry: stores metadata and technical information about the web services



Distributed web services: offer functionality over the internet

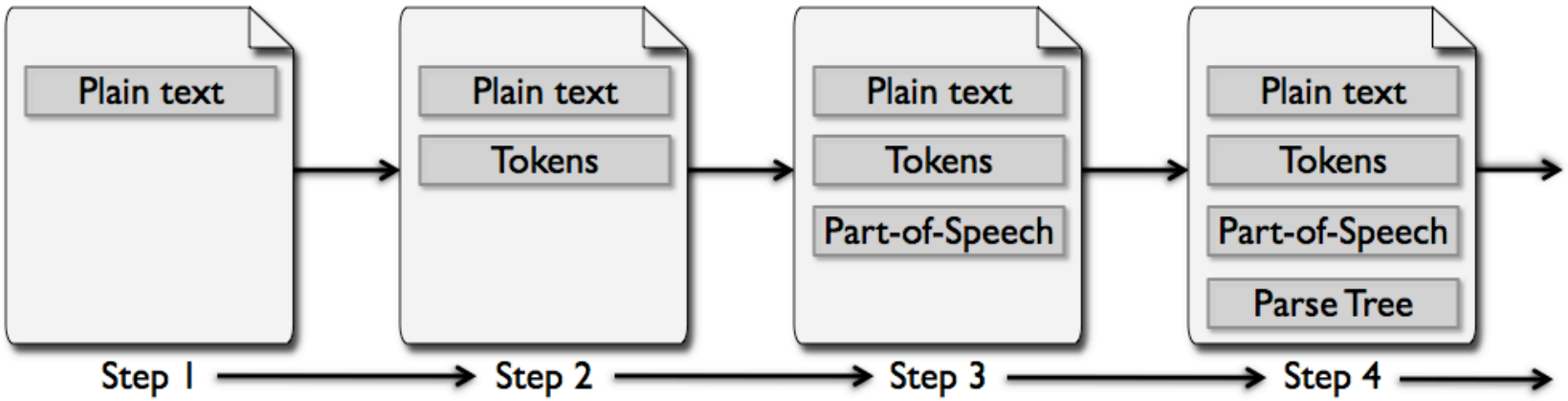
WebLicht – Tool Chains

CLARIN-D

- The chaining algorithm ensures that the list of possible next web services only contains web services that form a valid next step in the chain
  - For example: a part-of-speech tagger can only be added to a chain after a tokenizer was chosen
- Communication between the web services is based on the TCF data format

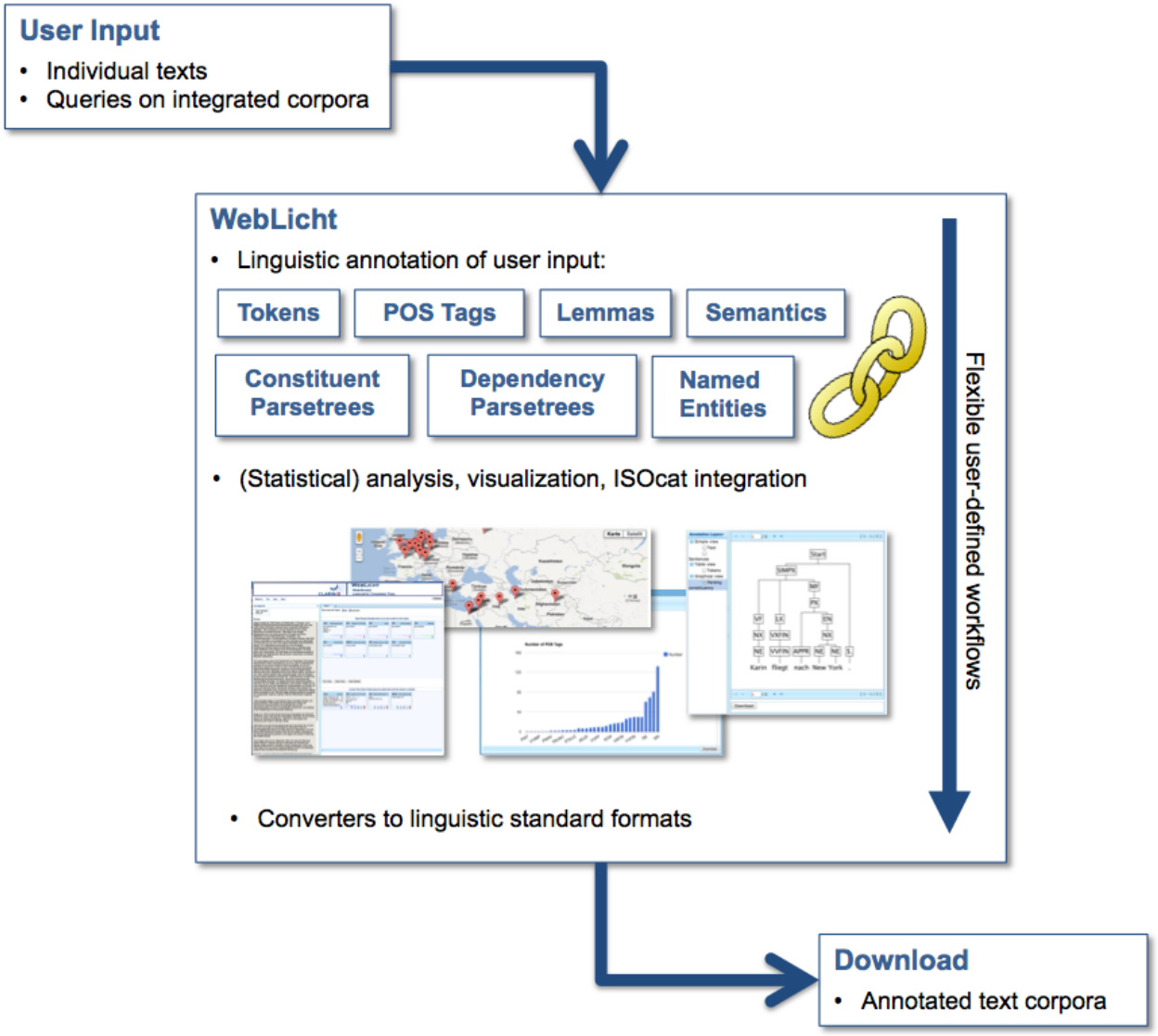
WebLicht – Text Corpus Format

- Common data format for in- and output of web services
- Each WS incrementally adds an annotation layer to TCF

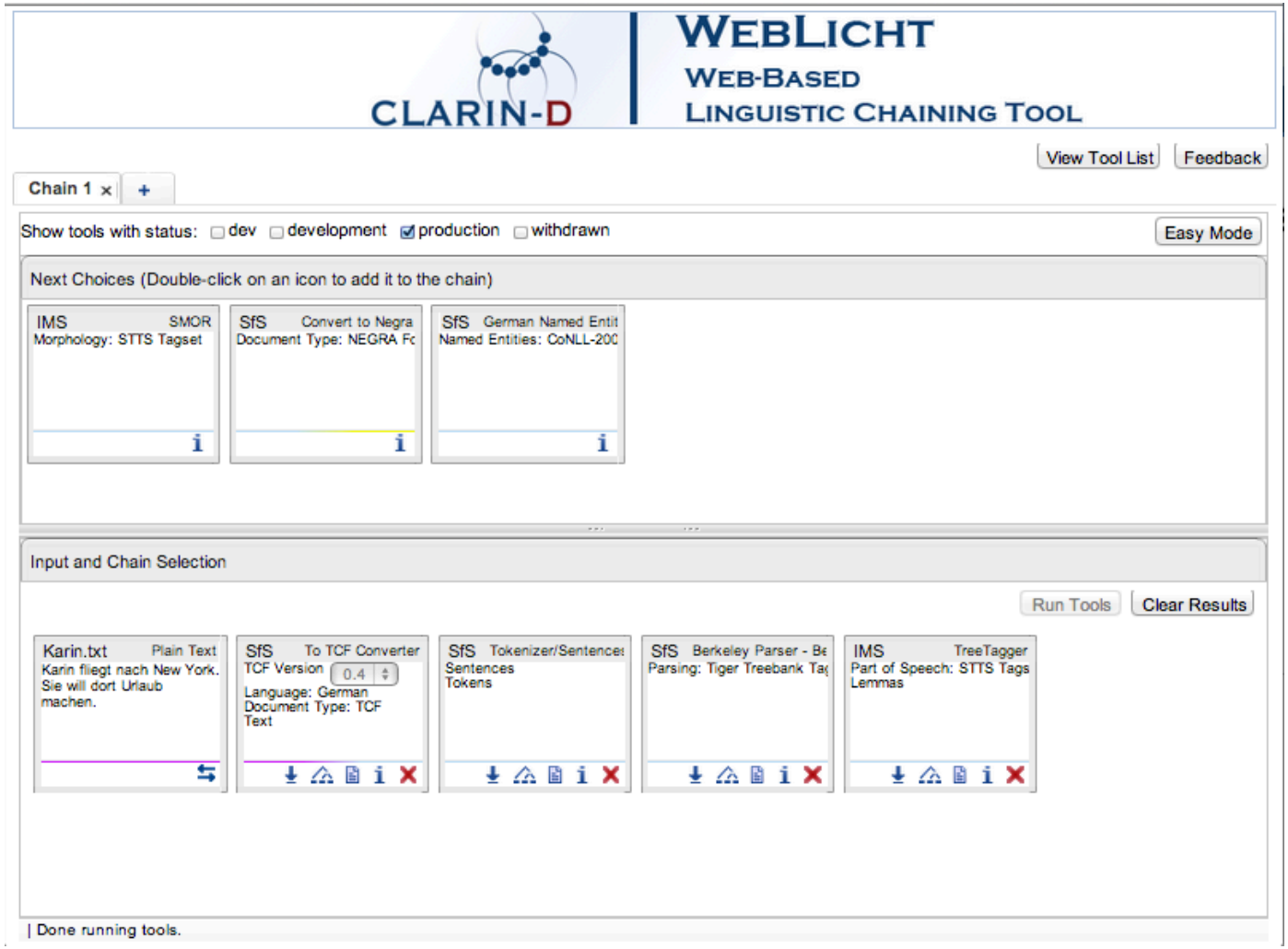


WebLicht: Integrated Webservice-Architecture

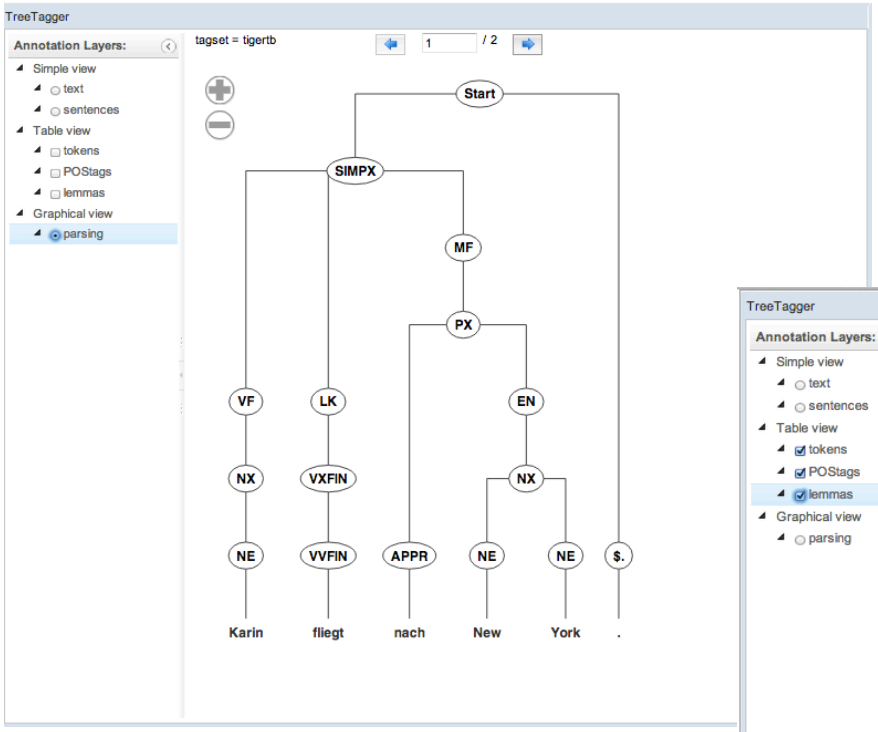
CLARIN-D

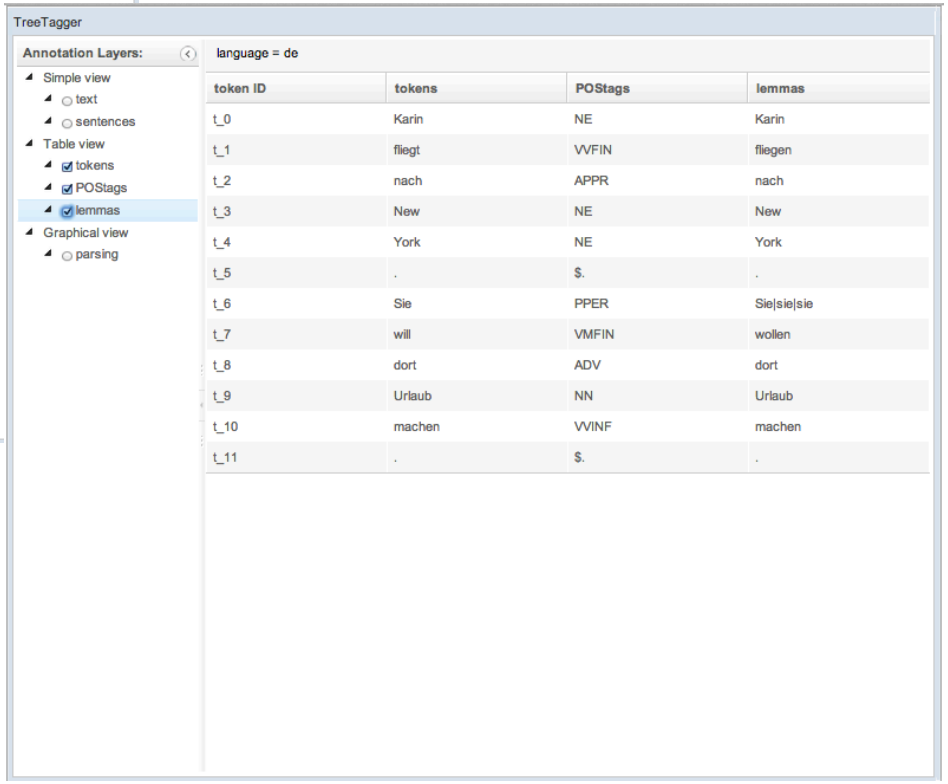


WebLicht: User Interface



WebLicht: Visualization





token ID	tokens	POStags	lemmas
t_0	Karin	NE	Karin
t_1	fliegt	VVFIN	fliegen
t_2	nach	APPR	nach
t_3	New	NE	New
t_4	York	NE	York
t_5	.	$.	.
t_6	Sie	PPER	Sie|sie|sie
t_7	will	VMFIN	wollen
t_8	dort	ADV	dort
t_9	Urlaub	NN	Urlaub
t_10	machen	VVINF	machen
t_11	.	$.	.

MAUS and WebMAUS

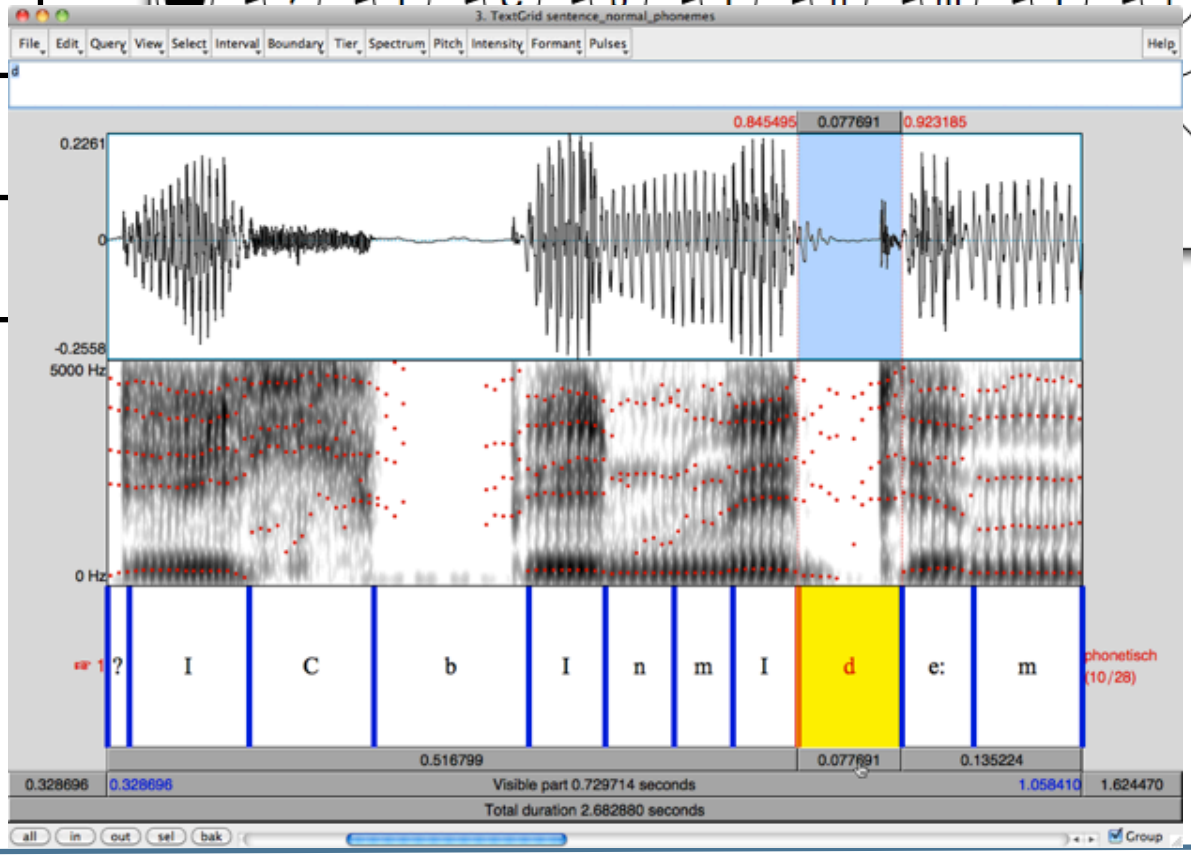
- Munich Automatic Segmentation
  - statistical modeling of pronunciation
  - orthographic transcript + lexicon
  - pronunciation generation for OOV words
- multilingual (D, Scottish E, HU…)
- 1st CLARIN-D result: WebMAUS

https://webapp.phonetik.uni-muenchen.de/BASWebServices/

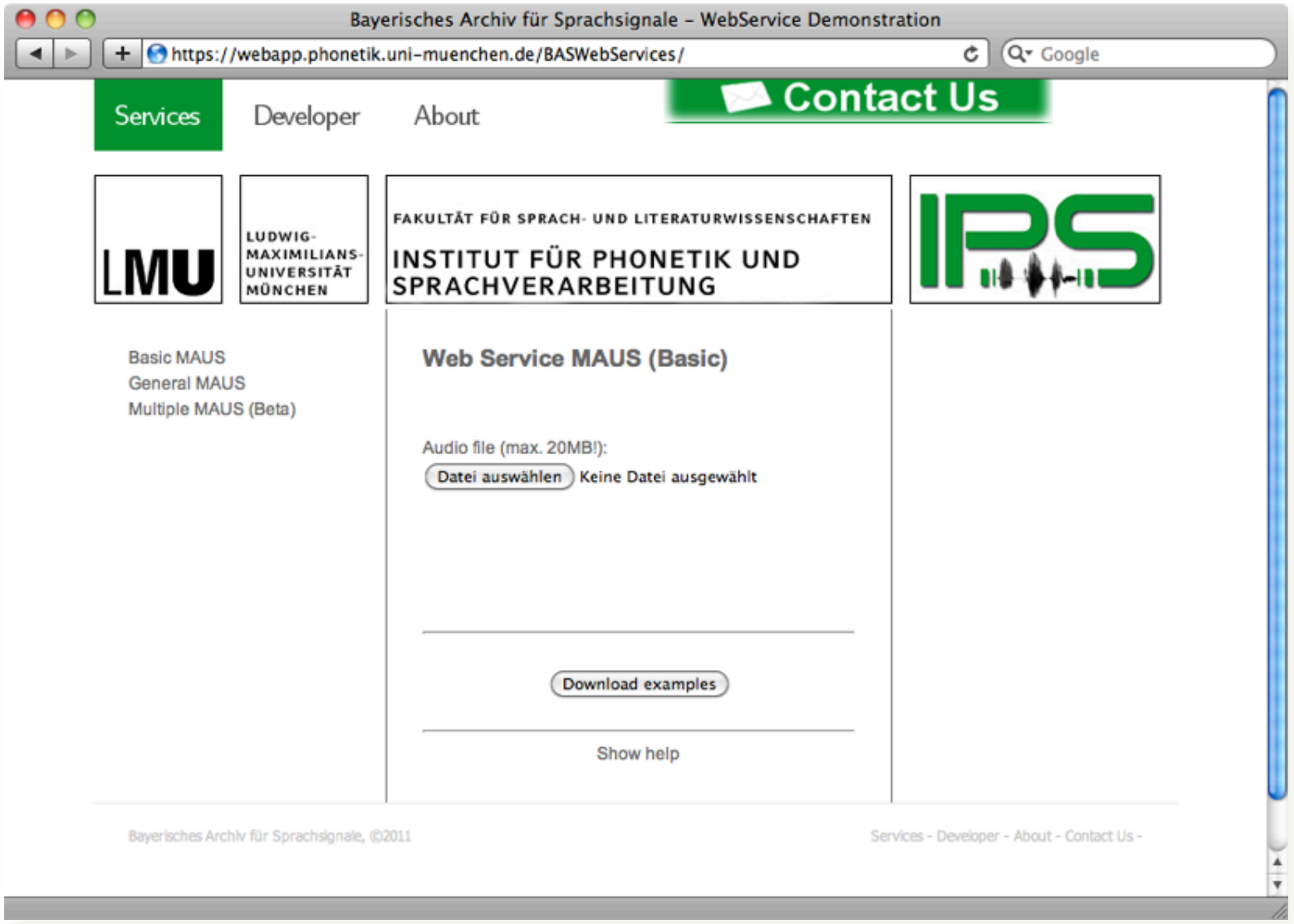
WebMAUS

ich	
bin	
mit	
dem	
Wagen	

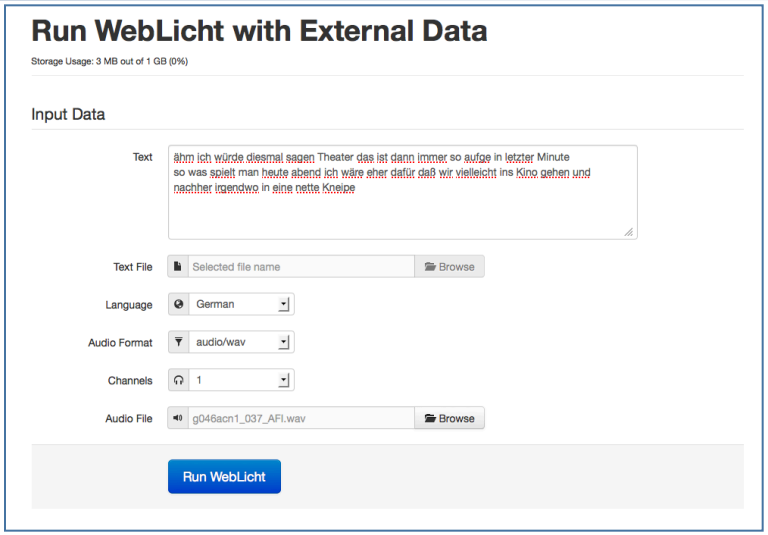
Ich bin mit dem Wagen nach Bonn gefahren.



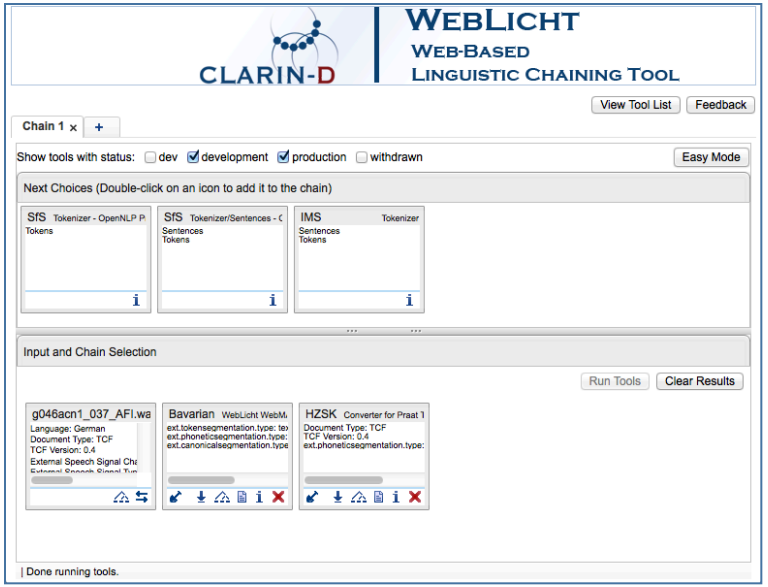
WebMAUS


Bayerisches Archiv für Sprachsignale – WebService Demonstration
https://webapp.phonetik.uni-muenchen.de/BASWebServices/
Google
Services
Developer
About
Contact Us
LMU
LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
FAKULTÄT FÜR SPRACH- UND LITERATURWISSENSCHAFTEN
INSTITUT FÜR PHONETIK UND SPRACHVERARBEITUNG
IPS
Basic MAUS
General MAUS
Multiple MAUS (Beta)
Web Service MAUS (Basic)
Audio file (max. 20MB!):
Datei auswählen
Keine Datei ausgewählt
Download examples
Show help
Bayerisches Archiv für Sprachsignale, ©2011
Services - Developer - About - Contact Us -

WebMAUS



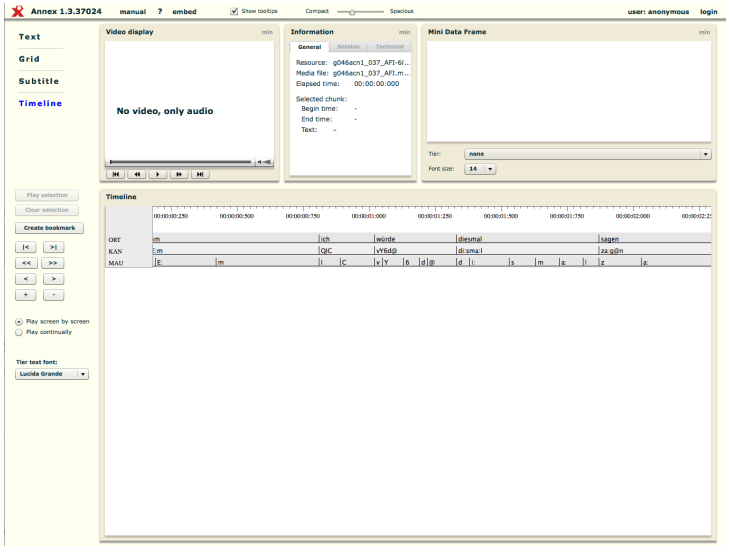
- Upload text and audio files
- Create TCF and run WebLicht

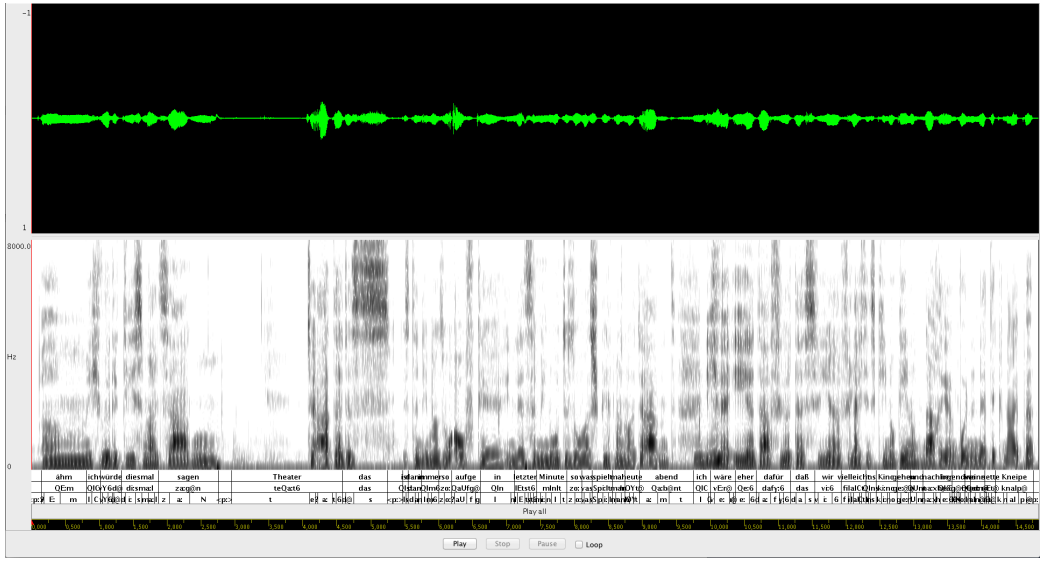


- Use WebMAUS webservice to do segmentation and labeling
- Generate ELAN format with another webservice (HZSK)

CLARIN-D

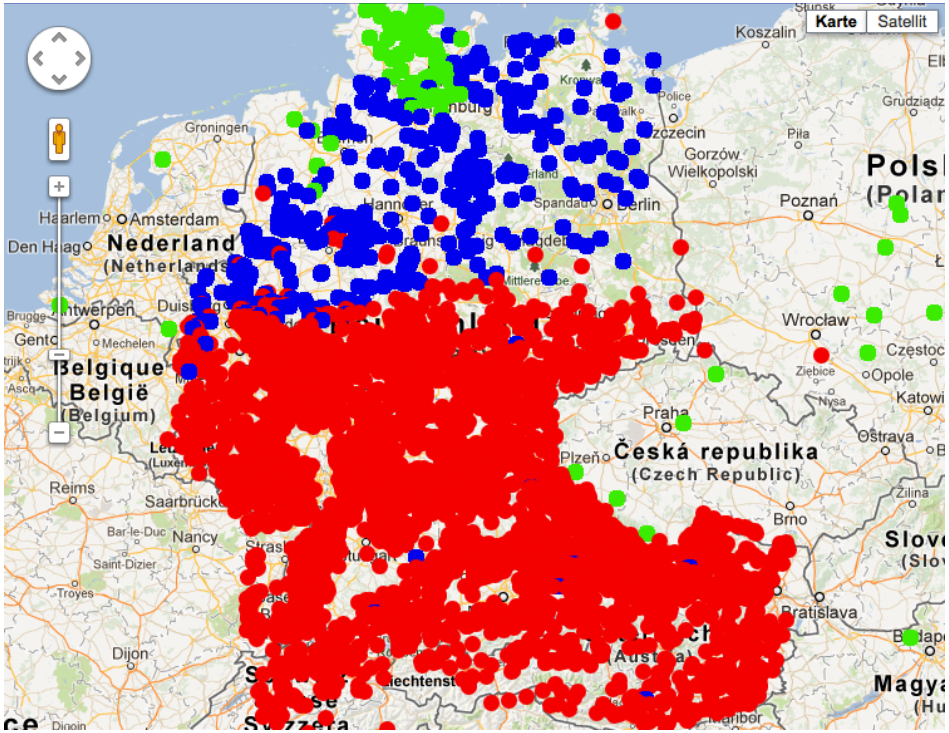
Annex Tool



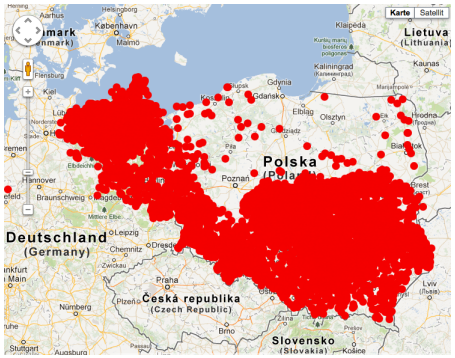


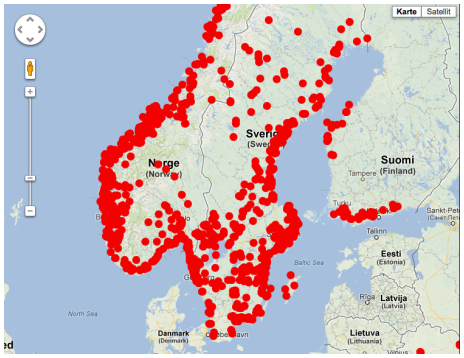
BAS Praat TextGrid Viewer

CiNaViz



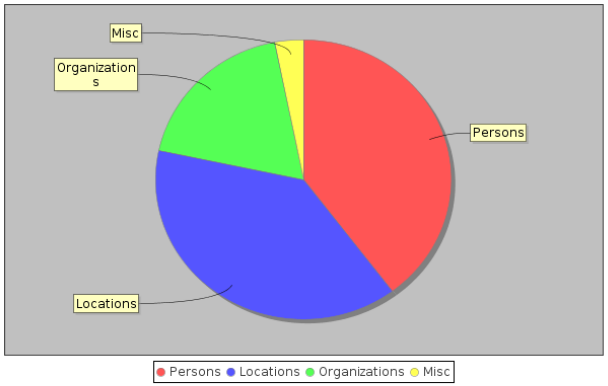
- Top: **-bach, -beck, -bek**



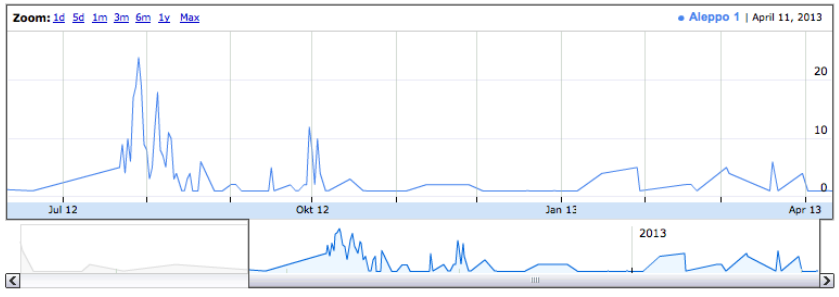


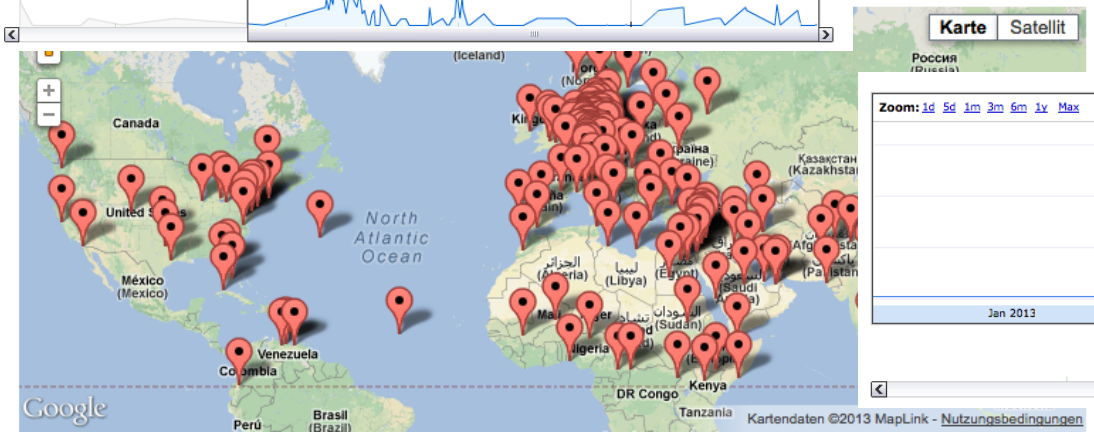
- Bottom left: **-ow**
- Bottom right: **-vik**

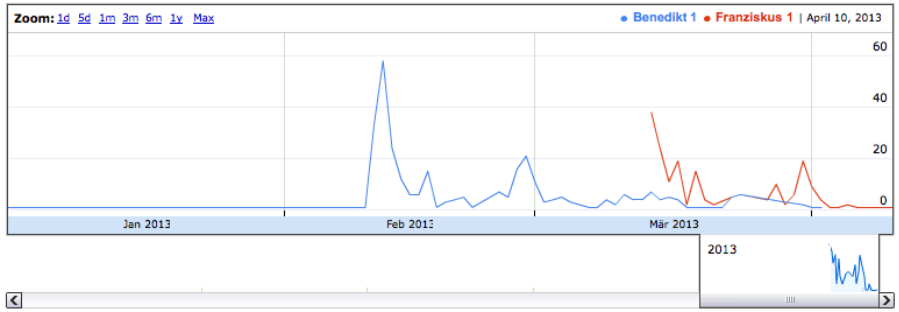
WhoIsInTheNews



1. Every day at 3 p.m., newsfeeds from German magazines are collected and named entity recognition is applied with the help of WebLicht web services
2. With a graphical user interface, these named entities can be analyzed and visualized







Thank You!