



CLARIN: A European Infrastructure for Language Resources and Technology

Erhard Hinrichs
University of Tübingen
CLARIN Board of Directors

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

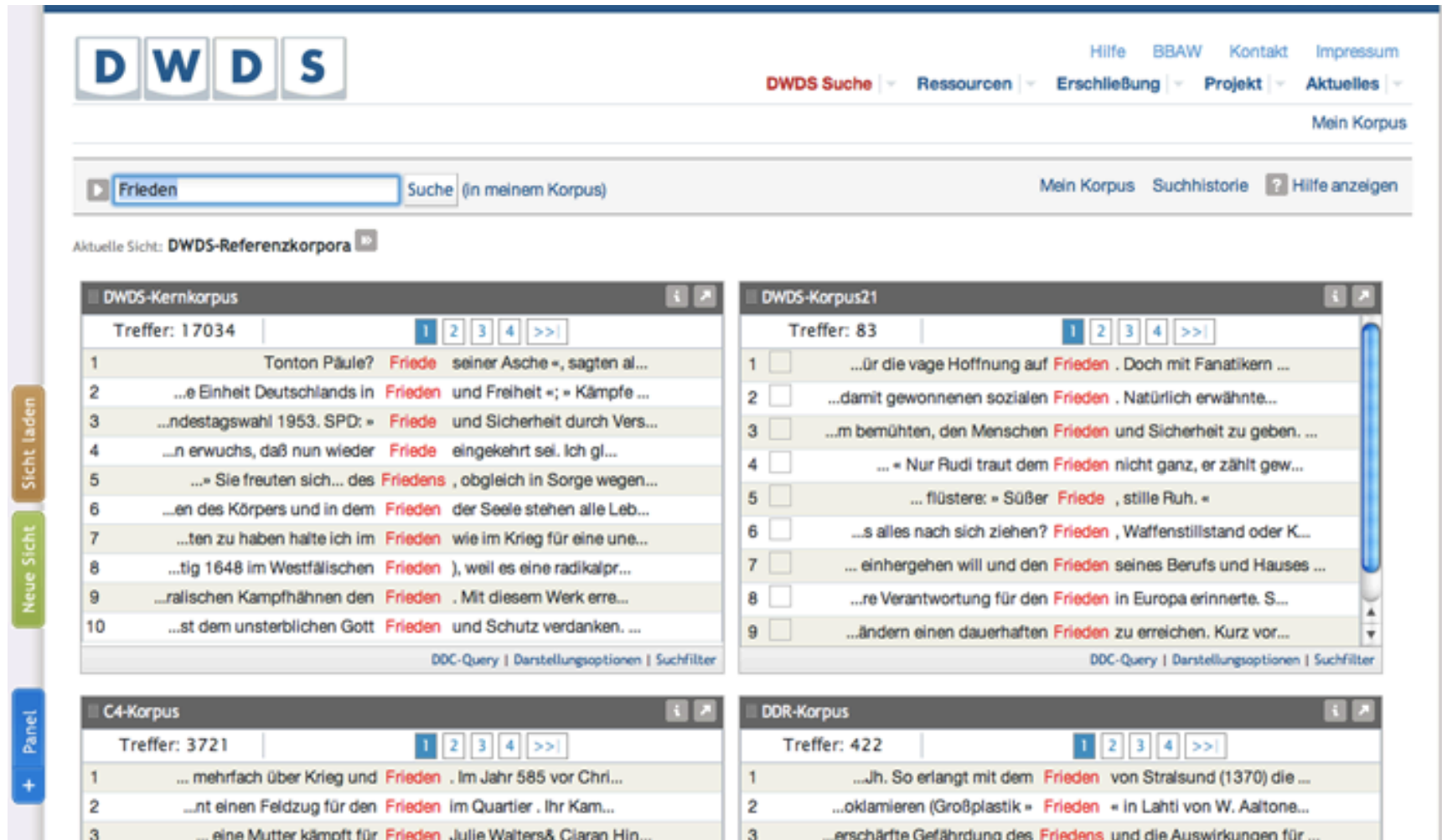
- CLARIN ERIC established by European Commission in February 2012
- 9 Founding Member States of CLARIN ERIC:
 - Austria, Bulgaria, Czech Republic, Denmark, Estonia, Germany, The Netherlands (Legal Seat), Nederlandse Taalunie (Flanders, the Netherlands), Poland
- Additional members are expected shortly: Croatia, Finland, France, Latvia, Lithuania, Norway, Slovenia

- 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



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

- 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



The screenshot displays the DWDS (Digitales Wörterbuch der Deutschen Sprache) interface. At the top, the logo 'DWDS' is visible, along with navigation links for 'Hilfe', 'BBAW', 'Kontakt', and 'Impressum'. Below the logo, there are dropdown menus for 'DWDS Suche', 'Ressourcen', 'Erschließung', 'Projekt', and 'Aktuelles'. A search bar contains the word 'Frieden' and a 'Suche' button. To the right of the search bar are links for 'Mein Korpus', 'Suchhistorie', and 'Hilfe anzeigen'. Below the search bar, the current view is identified as 'DWDS-Referenzkorpora'. The main content area is divided into four panels, each showing search results for 'Frieden' in a specific corpus:

- DWDS-Kernkorpus**: Treffer: 17034. Results 1-10 are shown, with 'Friede' and 'Friedens' highlighted in red.
- DWDS-Korpus21**: Treffer: 83. Results 1-9 are shown, with 'Frieden' highlighted in red.
- C4-Korpus**: Treffer: 3721. Results 1-3 are shown, with 'Frieden' highlighted in red.
- DDR-Korpus**: Treffer: 422. Results 1-3 are shown, with 'Frieden' and 'Friedens' highlighted in red.

On the left side of the interface, there is a vertical sidebar with buttons for 'Sicht laden', 'Neue Sicht', and 'Panel'.




Willkommen beim Wortschatz-Portal.

Wort:

Beachte Groß-/Kleinschreibung

Die Daten werden aus sorgfältig ausgewählten öffentlich zugänglichen Quellen automatisch erhoben. Die Beispielsätze werden automatisch ausgewählt und stellen keine Meinungsäußerung des Projektes Deutscher Wortschatz dar. Für die darin enthaltenen Inhalte und Meinungen sind ausschließlich die Autoren verantwortlich. Auch ohne besondere Kennzeichnung unterliegen im Wortschatz wiedergegebene Marken wie Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. den gesetzlichen Bestimmungen. Die synonyme Verwendung eines Trademarks beschreibt nicht notwendigerweise produktspezifische Eigenschaften sondern kennzeichnet stattdessen die Verwendung des Begriffs im allgemeinsprachlichen Kontext.

Wörter des Tages

Die tagesaktuellen Begriffe. Ausgewählt aus Tageszeitungen und Newsdiensten. Täglich um 7 Uhr früh. Seit April 2002 auf Deutsch – und seit März 2006 auch auf Norwegisch. Im Test: Jetzt auch als RSS 2.0! 

DIE WÖRTER DES TAGES
WORTSCHATZ.UNI-LEIPZIG.DE/WORT-DES-TAGES

Crawlen Sie mit!

Helfen Sie mit bei der Analyse des Internet! FindLinks nutzt freie Kapazitäten Ihres PCs, um das Internet zu analysieren und neue Daten für den Surfguide Nextlinks bereitzustellen. Seit August 2006. Zum Download ...



Webservices


Mit den Webservices ist ein direkter Zugriff auf die Daten des Projektes Deutscher Wortschatz aus einer beliebigen Software heraus möglich.



Int. Wortschatz Portal

Auf unserem internationalen Wortschatzportal in englischer Sprache können Sie derzeit in 17 verschiedenen Sprachen Wörter nachschlagen. Um das Suchen für Sie einfacher zu gestalten, werden Ihnen zu jeder Sprache zufällige Wörter vorgeschlagen.

CORPORA
CORPORA.INFORMATIK.UNI-LEIPZIG.DE



Wörterbuch

Über 100.000 Wörter und Wendungen auf Deutsch und Englisch. Die Besonderheit: Häufigkeitsangaben verraten Ihnen, wie oft die einzelnen Wörter verwendet werden. Seit Januar 2002.

WÖRTERBUCH
DICT.UNI-LEIPZIG.DE



Feedback

Feedback der Nutzer zu den verschiedenen Services des Wortschatz Portals.

- 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)



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



- **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

WordNet Search - 3.1

- [WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

Noun

- [S:](#) [\(n\)](#) **nail** (horny plate covering and protecting part of the dorsal surface of the digits)
- [S:](#) [\(n\)](#) **nail** (a thin pointed piece of metal that is hammered into materials as a fastener)
- [S:](#) [\(n\)](#) **nail** (a former unit of length for cloth equal to 1/16 of a yard)

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

WordNet Search - 3.1
- [WordNet home page](#) - [Glossary](#) - [Help](#)

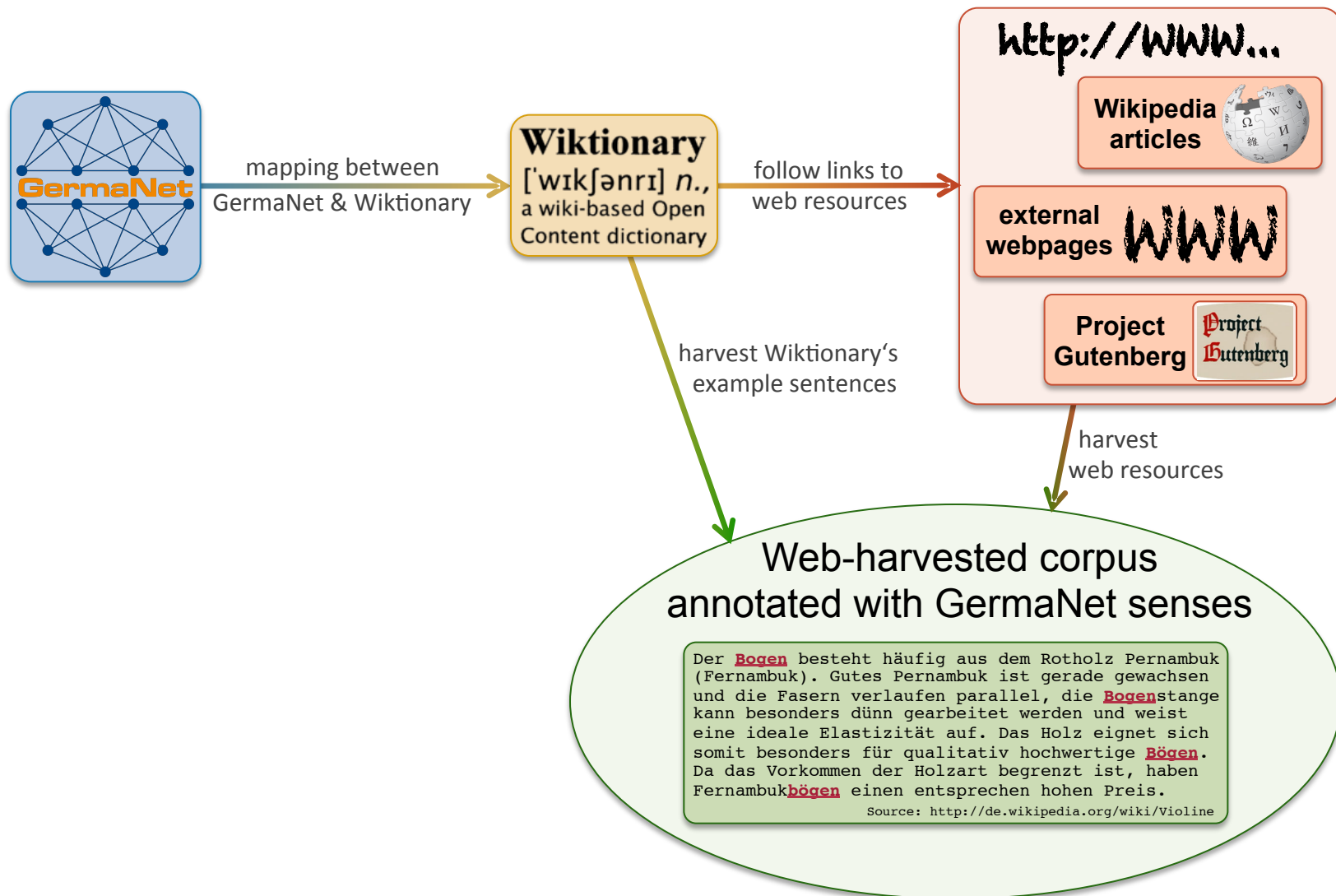
Word to search for:

Noun

- [S:](#) (n) nail
- [S:](#) (n) nail
- [S:](#) (n) nail

- 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



Wiktionary
[ˈwɪkʃənri] *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:

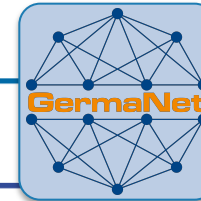
[1] Familienanzeige, Geburtsanzeige, Heiratsanzeige, Hochzeitsanzeige, Kontaktanzeige, Todesanzeige, Traueranzeige, Verlobungsanzeige

[1] Werbeanzeige

[1] Kleinanzeige

[2] Selbstanzeige, Strafanzeige

[3] Temperaturanzeige, Wasserstandsanzeige



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’**

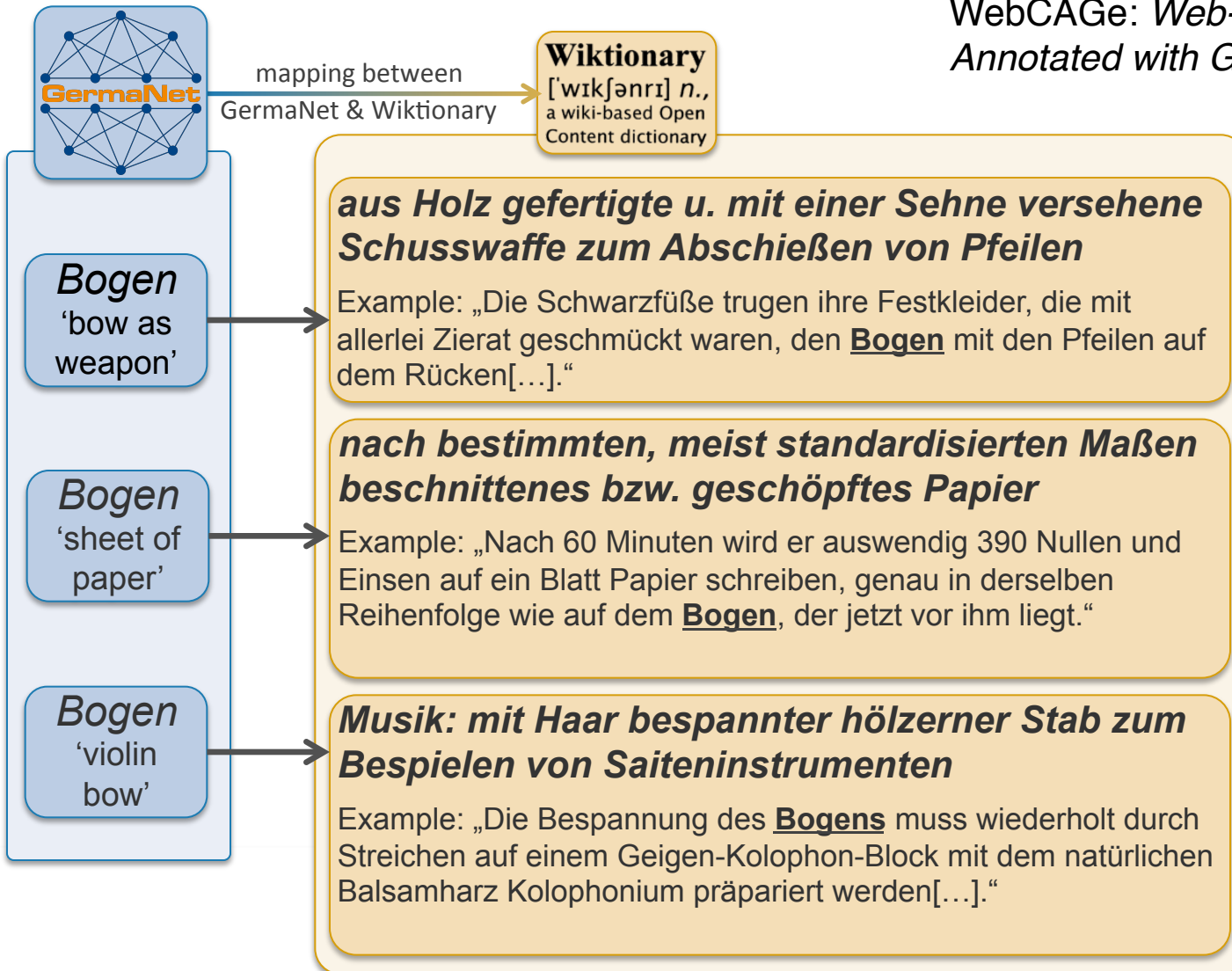
synonyms: [Display]

hypernyms: [Device, Gerät]

hyponyms: [Handydisplay] [Bildschirmseite]

- 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>

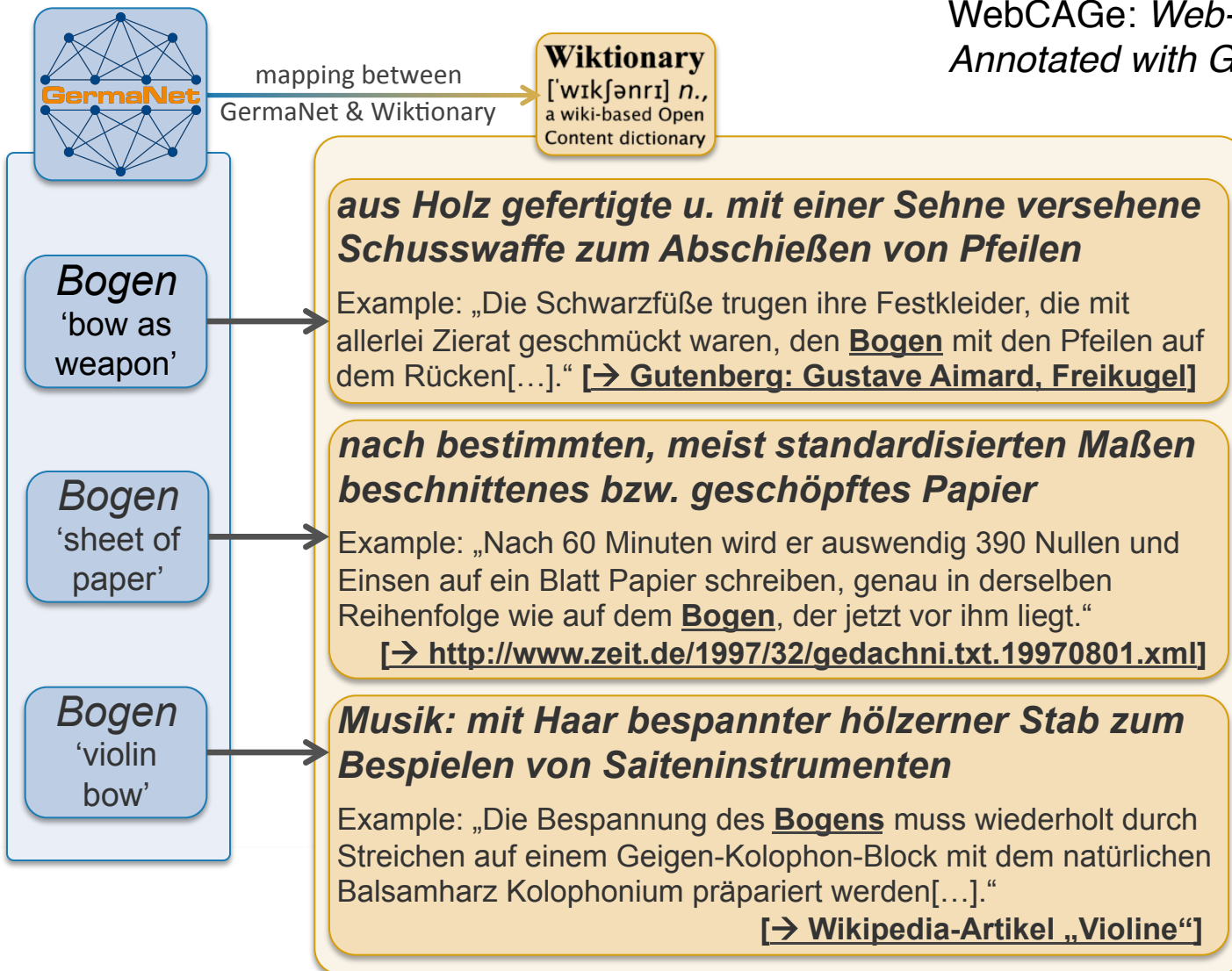
WebCAGe: *Web-Harvested Corpus Annotated with GermaNet Senses*



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



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



WebCAGe: Web-Harvested Corpus Annotated with GermaNet Senses



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




mapping between
GermaNet & Wiktionary

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

Bogen
'bow as
weapon'

aus Holz gefertigte u. mit einer Sehne versehene Schusswaffe zum Abschießen von Pfeilen

Example: „Die Schwarzfüße trugen ihre Festkleider, die mit allerlei Zierat geschmückt waren, den **Bogen** mit den Pfeilen auf dem Rücken[...].“ [\[→ Gutenberg: Gustave Aimard, Freikugel\]](#)

Project Gutenberg 

Bogen
'sheet of
paper'

nach bestimmten, meist standardisierten Maßen beschnittenes bzw. geschöpftes Papier

Example: „Nach 60 Minuten wird er auswendig 390 Nullen und Einsen auf ein Blatt Papier schreiben, genau in derselben Reihenfolge wie auf dem **Bogen**, der jetzt vor ihm liegt.“

[\[→ http://www.zeit.de/1997/32/gedachni.txt.19970801.xml\]](http://www.zeit.de/1997/32/gedachni.txt.19970801.xml)

external webpages 

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 

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 Pernambuco ist gerade gewachsen und die Fasern verlaufen parallel, die Bogenstange kann dann besonders dünn gearbeitet werden und weist eine ideale Elastizität auf. Pernambuco 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 (**Karbonfaser**) verwendet.

Am unteren Ende des Bogens befindet sich der sogenannte *Frosch* aus Ebenholz, meist verziert mit einer runden Perlmutter-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.: <i>violin</i> , ital.: <i>violino</i>	
	
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



Violine



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

Die **Violine** (**Geige**, Abk.: *Vi.*) 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 Pernambuco ist gerade gewachsen und die Fasern verlaufen parallel, die **Bogen**stange kann dann besonders dünn gearbeitet werden und weist eine ideale Elastizität auf. Pernambuco eignet sich somit besonders für qualitativ hochwertige **Bögen**. Da das Vorkommen der Holzart begrenzt ist, haben Pernambuco **bögen** einen hohen Preis. Einfachere Schüler **bögen** sind meist aus Brasilholz gefertigt. Heute werden, auch von Berufsgeigern, zunehmend **Bögen** aus Kohlefaser (**Karbonfaser**) verwendet.

Am unteren Ende des **Bogens** befindet sich der sogenannte *Frosch* aus Ebenholz, meist verziert mit einer runden Perlmutter-Einlage. Zwischen Frosch und **Bogens**spitze (Köpfchen) sind die **Bogen**haare 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.: <i>violin</i> , ital.: <i>violino</i>	
	
Klassifikation	Chordophon Streichinstrument
Tonumfang	

WebCAGe

“one sense per discourse” heuristic

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



		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	

- 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%

- 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

- 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>

- Status quo: Many language resources and tools are available to a large user community
- Goal: Integration of these resources and tools in a single, user-friendly CLARIN-D portal

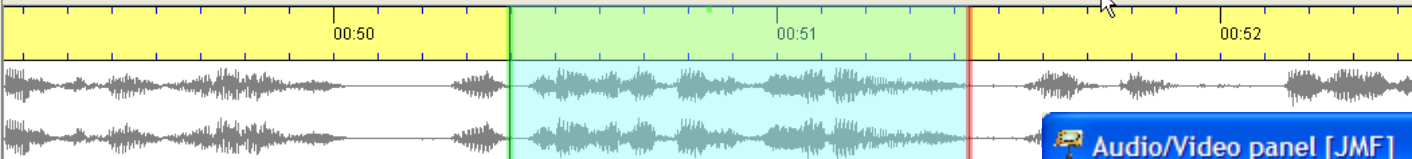
Hamburg: EXMARaLDA – Transcription and Annotation Tool for Video Data

EXMARaLDA Partitur-Editor 1.4.3 [C:\Dokumente und Einstellungen\thomas\Desktop\DEMO_KORPUS\Rudi\Rudi_Voeller_Wutausbruch.xml]

File Edit View Transcription Tier Event Timeline Format Help

Sind sie Tabellenführer oder nich?

00:50:39 1.037 00:51:43



Add event... Append interval

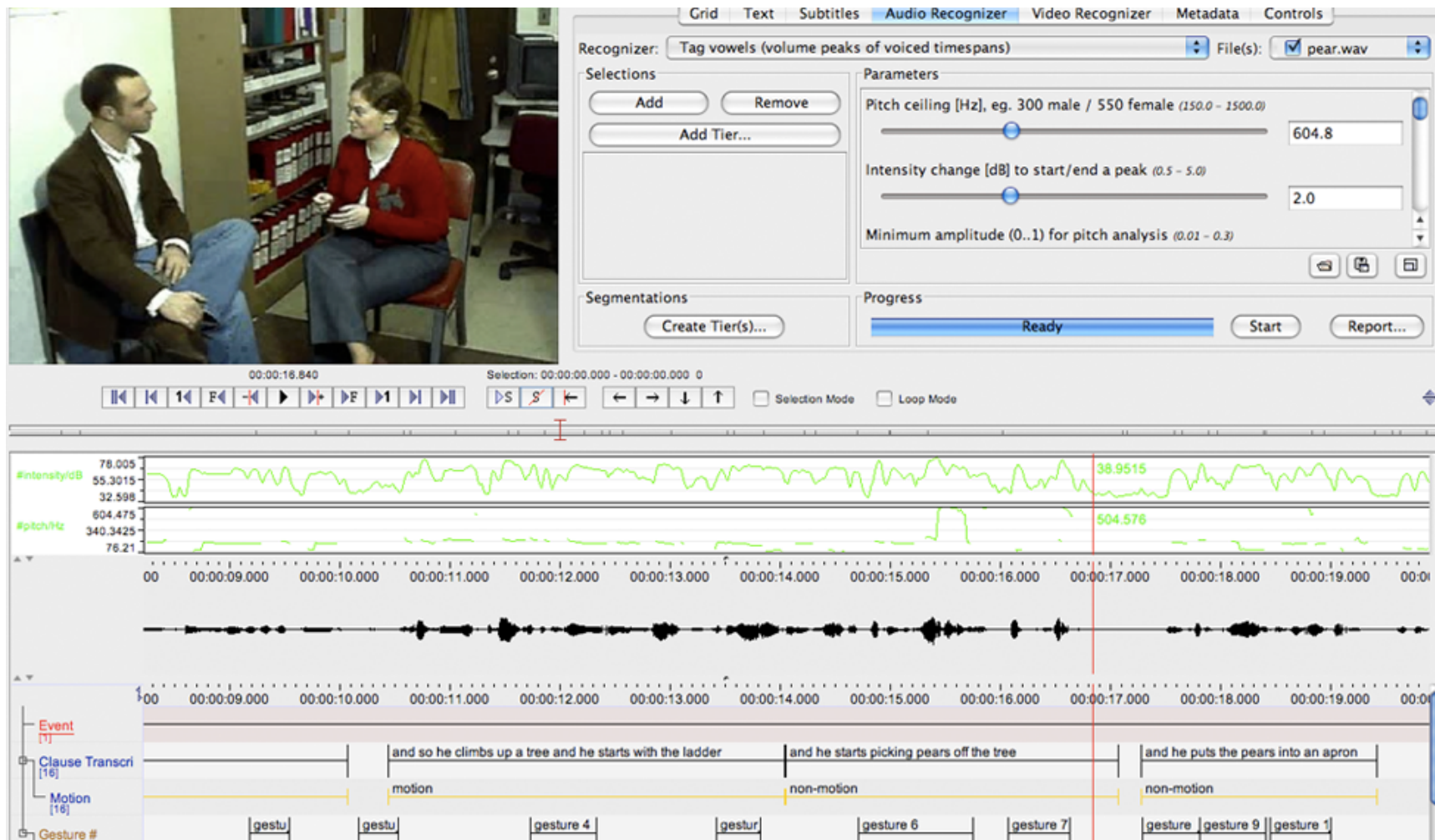
	37 [49.]	38 [49.1]	39 [49.4]	40 [49.6]	41 [49.]	42 [50.]	43 [50.3]		44 [51.4]	45 [51.9]
WH [sup]										
WH [v]				Ja.		• Ja.			Richtig.	
WH [en]				Yes.		Yes.			That's right.	
WH [nv]										
RV [sup]		betont								
RV [v]	. Das	weiß	du, Waldi!	Oder nic	ht?		Sind sie Tabellenführer oder nich?		(Ja) also!	
RV [en]	You know that, Waldi!			Don't you?			Are they group leaders or not?		Well then!	
RV [nv]										
WH [k]										
RV [k]							verärgerl, aggressiv		verärgerl, agg	

Done.

Transcription C:\Dokumente und Einstellungen\thomas\Desktop\DEMO_KORPUS\Rudi\Rudi_Voeller_Wutausbruch.xml opened



MPI: ELAN – Multimedia-Annotation Tool

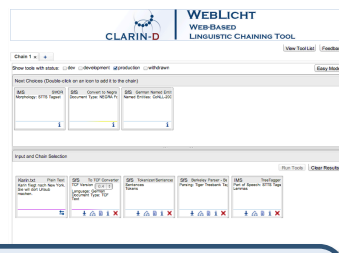


The screenshot displays the ELAN software interface. On the left, a video window shows a man and a woman in conversation. The main interface is divided into several sections:

- Top Panel:** Includes tabs for Grid, Text, Subtitles, Audio Recognizer (selected), Video Recognizer, Metadata, and Controls. The Recognizer is set to "Tag vowels (volume peaks of voiced timespans)" and the file is "pear.wav".
- Parameters:** Shows sliders for Pitch ceiling [Hz] (set to 604.8) and Intensity change [dB] to start/end a peak (set to 2.0). A "Minimum amplitude (0..1) for pitch analysis (0.01 - 0.3)" is also present.
- Progress:** A progress bar is labeled "Ready" with "Start" and "Report..." buttons.
- Timeline:** A central timeline shows the video and audio waveforms. A vertical red line marks the current time at 00:00:16.840.
- Analysis Tracks:** Below the timeline, there are tracks for #intensity/dB, #pitch/Hz, and a transcription track. The transcription track shows the text: "and so he climbs up a tree and he starts with the ladder | and he starts picking pears off the tree | and he puts the pears into an apron".
- Gesture Tracks:** The bottom track shows gesture annotations: [gestu] [gestu] [gesture 4] [gestur] [gesture 6] [gesture 7] [gesture | gesture 9 | gesture 1].

- 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

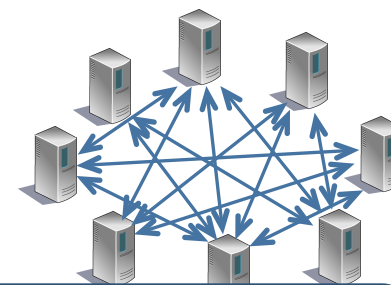
→ Solution: **WebLicht** (**Web**-Based **L**inguistic **C**haining **T**ool)



Web user interface:
interacts with the user



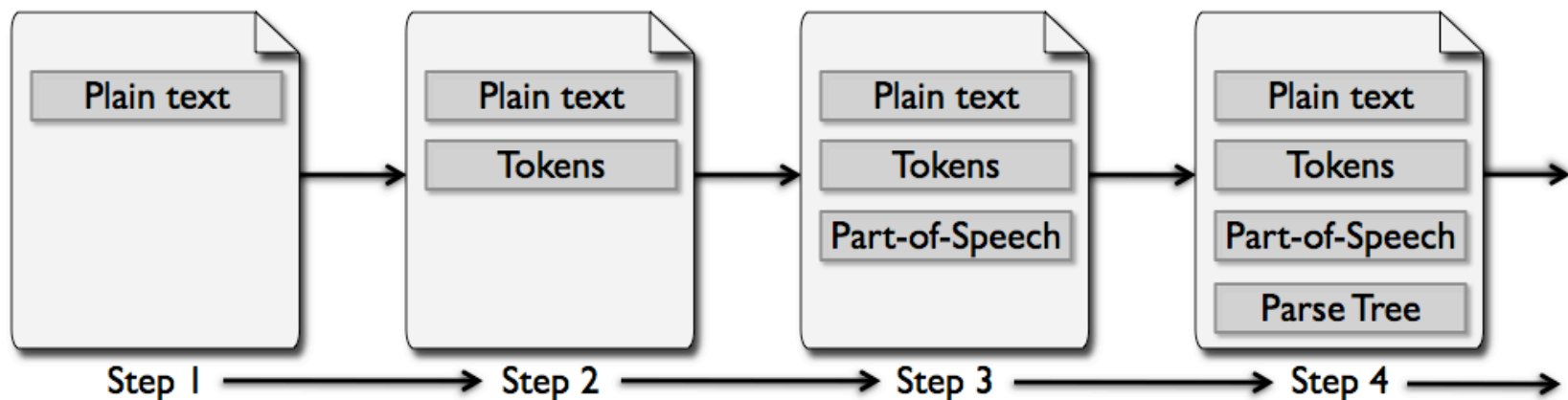
Registry: stores metadata
and technical information
about the web services

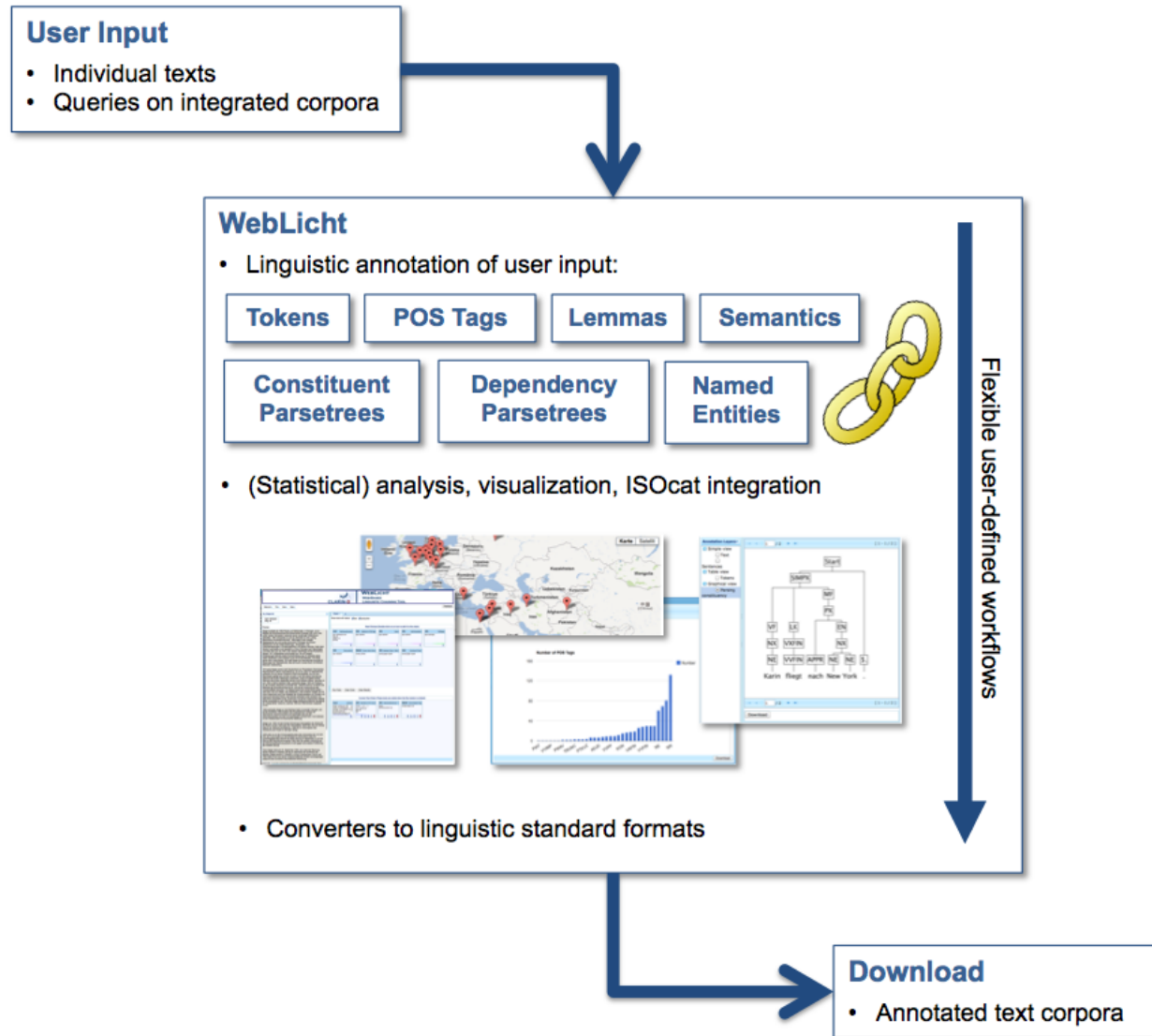


Distributed web services:
offer functionality over the
internet


- 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

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





WebLicht: User Interface



WEBLICHT




WEB-BASED
LINGUISTIC CHAINING TOOL

[View Tool List](#) [Feedback](#)

Chain 1 x +






















Show tools with status: dev development production withdrawn [Easy Mode](#)

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

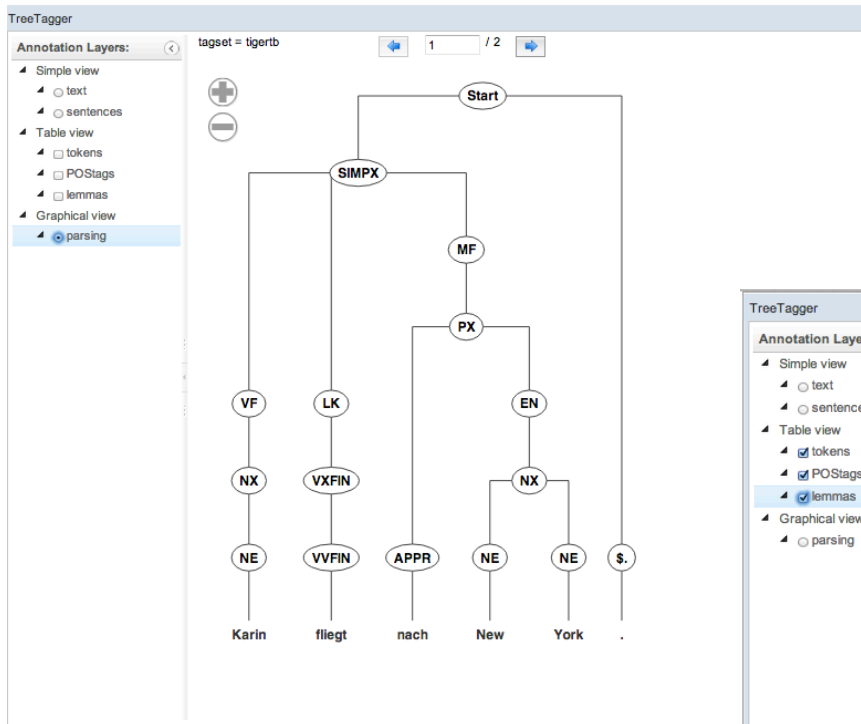
IMS SMOR Morphology: STTS Tagset 	SfS Convert to Negra Document Type: NEGRA Fc 	SfS German Named Entit Named Entities: CoNLL-200 
--	---	---

Input and Chain Selection

[Run Tools](#) [Clear Results](#)

Karin.txt Plain Text Karin fliegt nach New York. Sie will dort Urlaub machen. 	SfS To TCF Converter TCF Version: 0.4 Language: German Document Type: TCF Text     	SfS Tokenizer/Sentences: Sentences Tokens     	SfS Berkeley Parser - Be Parsing: Tiger Treebank Tag     	IMS TreeTagger Part of Speech: STTS Tags Lemmas     
---	---	---	---	---

| Done running tools.



TreeTagger

Annotation Layers: language = de

- Simple view
 - text
 - sentences
- Table view
 - tokens
 - POStags**
 - lemmas**
- Graphical view
 - parsing

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	.	\$.	.



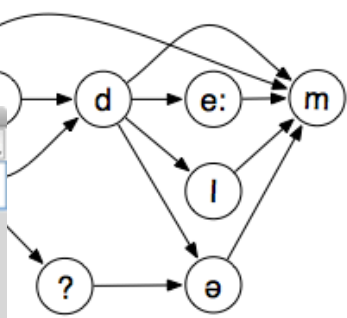
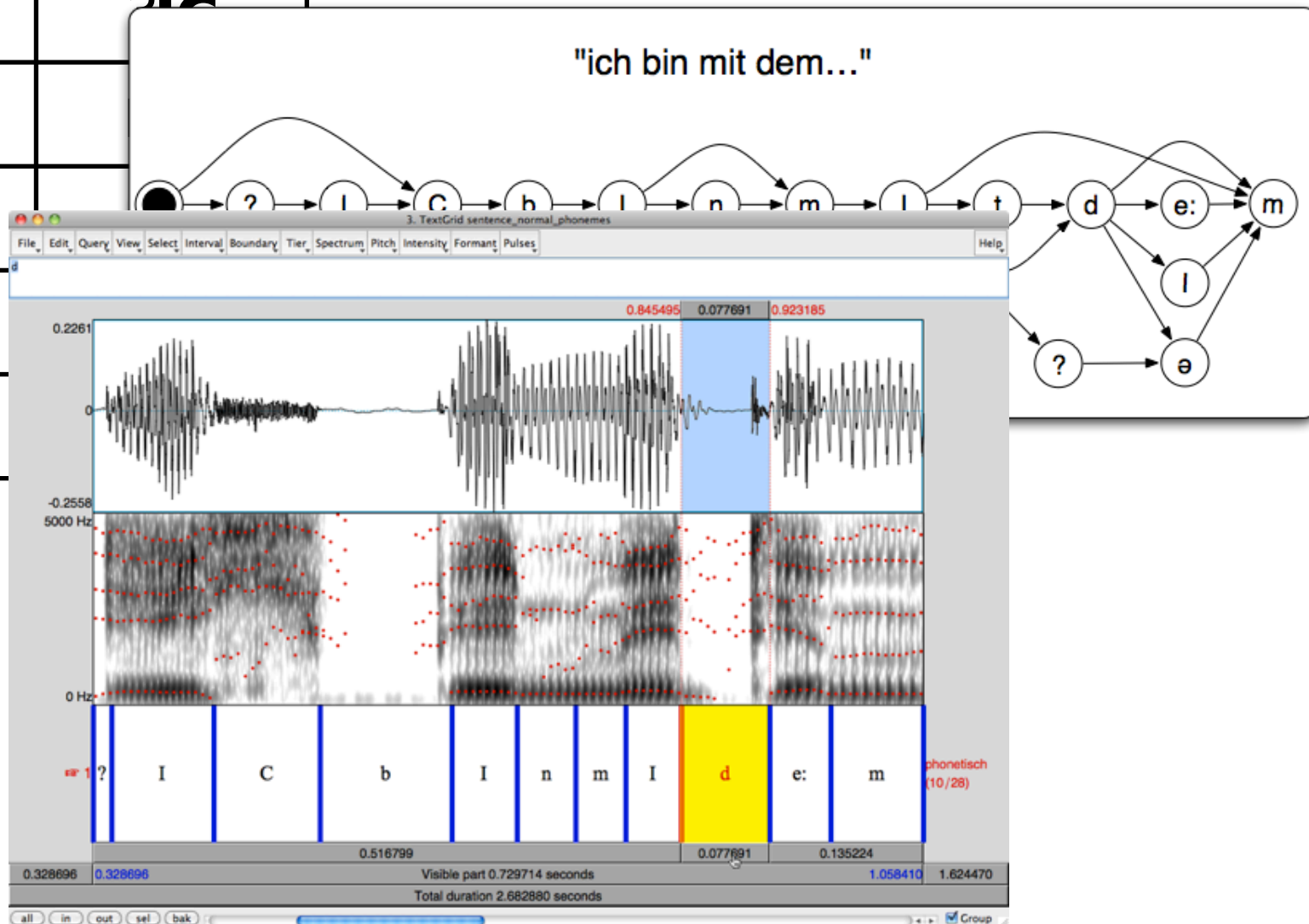
- Munich Automatic Segmentation
 - statistical modeling of pronunciation
 - orthographic transcript + lexicon
 - pronunciation generation for OOV words
- multilingual (D, Scottish E, HU...)
- 1st CLARIN-D result: WebMAUS

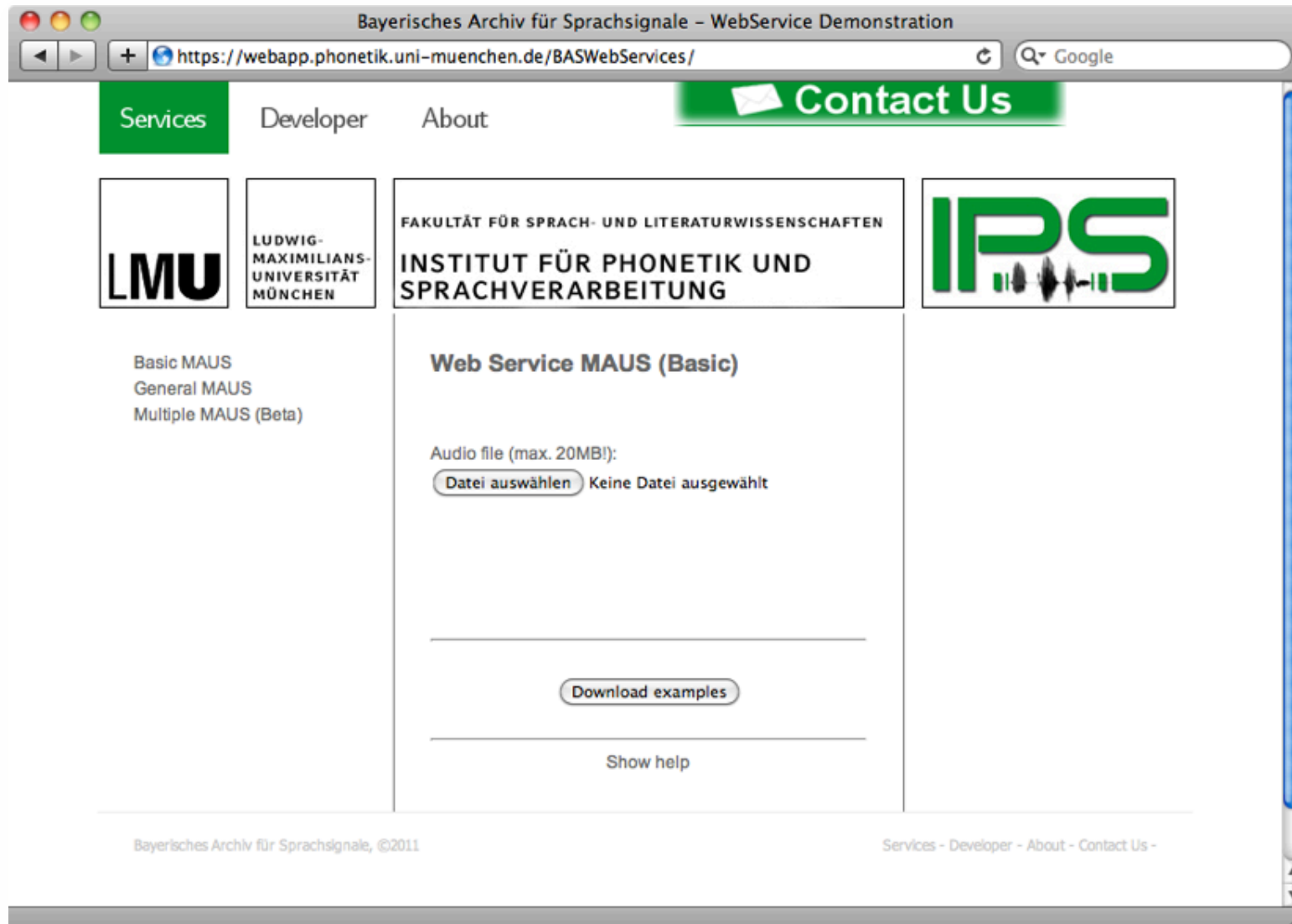
<https://webapp.phonetik.uni-muenchen.de/BASWebServices/>

ich
bin
mit
dem
Wagen

Ich bin mit dem Wagen nach Bonn gefahren.

"ich bin mit dem..."





The screenshot shows a web browser window titled "Bayerisches Archiv für Sprachsignale – Webservice Demonstration". The address bar contains the URL "https://webapp.phonetik.uni-muenchen.de/BASWebServices/". The page has a navigation menu with "Services", "Developer", "About", and "Contact Us" (highlighted with a green background and an envelope icon). The main content area is divided into four columns:

- Column 1:** LMU logo and text: "LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN". Below it, a list of services: "Basic MAUS", "General MAUS", and "Multiple MAUS (Beta)".
- Column 2:** Faculty and Institute information: "FAKULTÄT FÜR SPRACH- UND LITERATURWISSENSCHAFTEN" and "INSTITUT FÜR PHONETIK UND SPRACHVERARBEITUNG".
- Column 3:** "Web Service MAUS (Basic)" section. It includes an "Audio file (max. 20MB!)" upload area with a "Datei auswählen" button and the text "Keine Datei ausgewählt". Below this is a "Download examples" button and a "Show help" link.
- Column 4:** IPS logo.

At the bottom of the page, there is a footer with "Bayerisches Archiv für Sprachsignale, ©2011" on the left and "Services - Developer - About - Contact Us -" on the right.

Run WebLicht with External Data

Storage Usage: 3 MB out of 1 GB (0%)

Input Data

Text
ähm ich würde diesmal sagen Theater das ist dann immer so aufge in letzter Minute so was spielt man heute abend ich wäre eher dafür daß wir vielleicht ins Kino gehen und nachher irgendwo in eine nette Kneipe

Text File
Selected file name

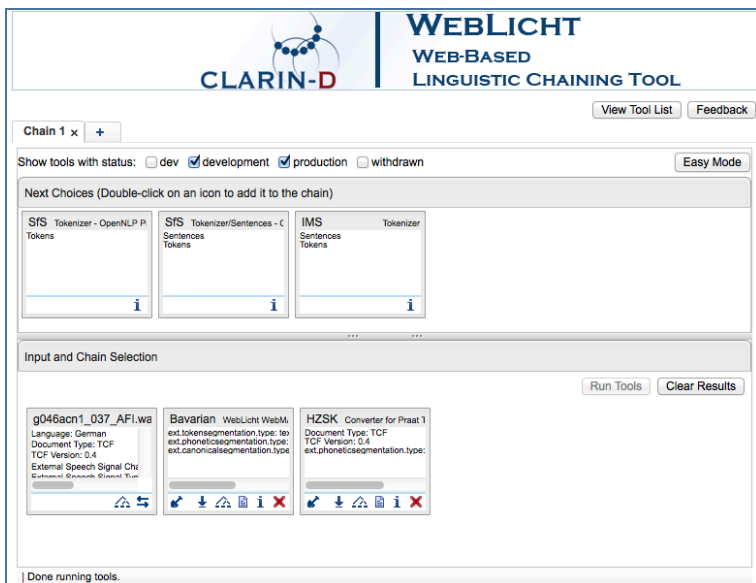
Language
German

Audio Format
audio/wav

Channels
1

Audio File
g046acn1_037_AFI.wav

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



CLARIN-D | **WEBLICHT**
WEB-BASED LINGUISTIC CHAINING TOOL

Chain 1 x

Show tools with status: dev development production withdrawn

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

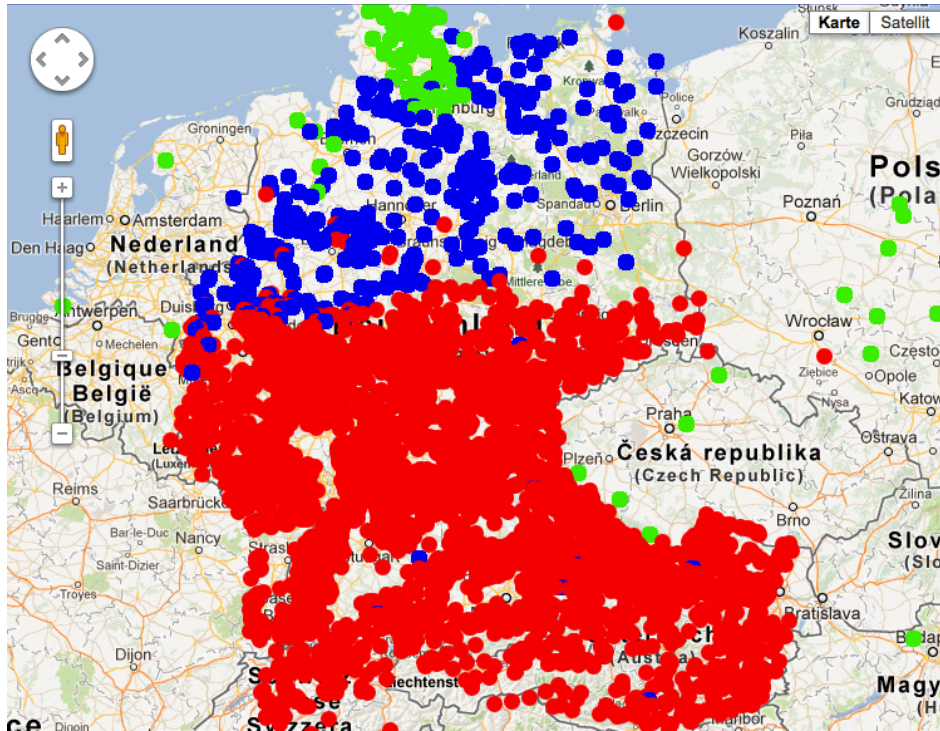
SIS Tokenizer - OpenNLP P Tokens	SIS Tokenizer/Sentences - C Sentences Tokens	IMS Sentences Tokens
-------------------------------------	--	----------------------------

Input and Chain Selection

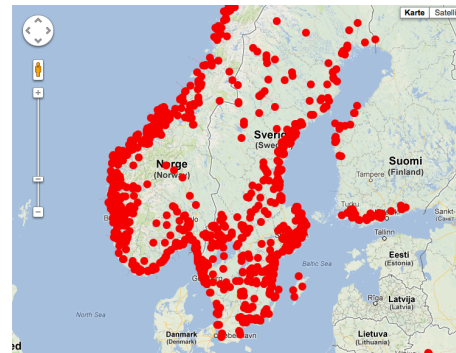
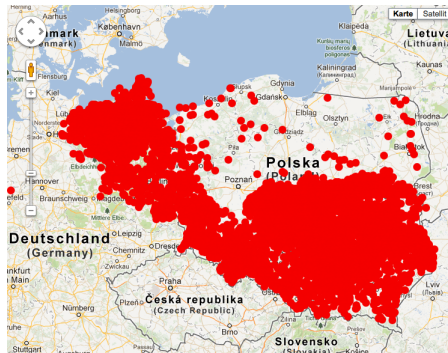
g046acn1_037_AFI.wa Language: German Document Type: TCF TCF Version: 0.4 External Speech Signal Ch: External Speech Signal Tm:	Bavarian WebLicht Webl ext.tokenizesegmentation.type: 1a ext.phonemealignment.type: 2 ext.canonicalsegmentation.type:	HZSK Converter for Praat 1 Document Type: TCF TCF Version: 0.4 ext.phonemealignment.type:
---	--	--

Done running tools.

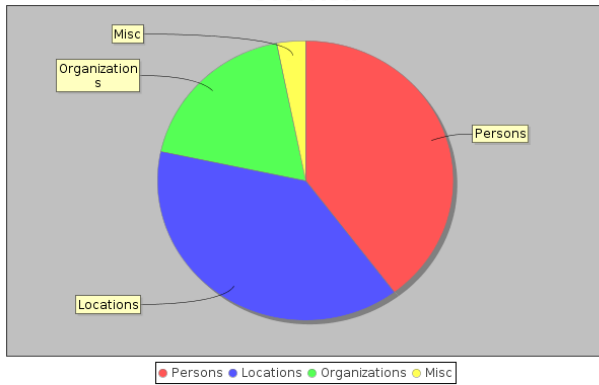
- Use WebMAUS webservice to do segmentation and labeling
- Generate ELAN format with another webservice (HZSK)



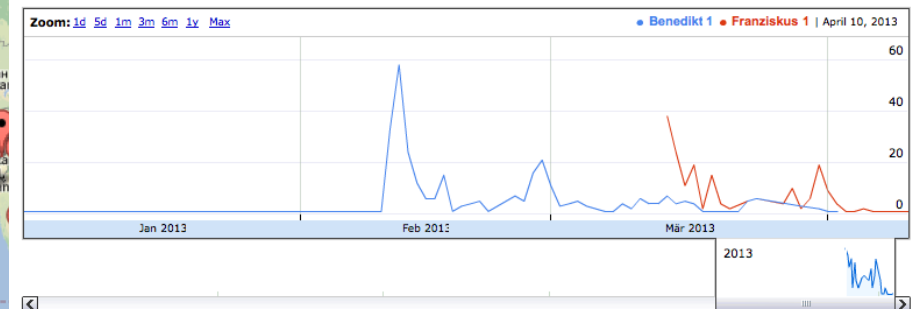
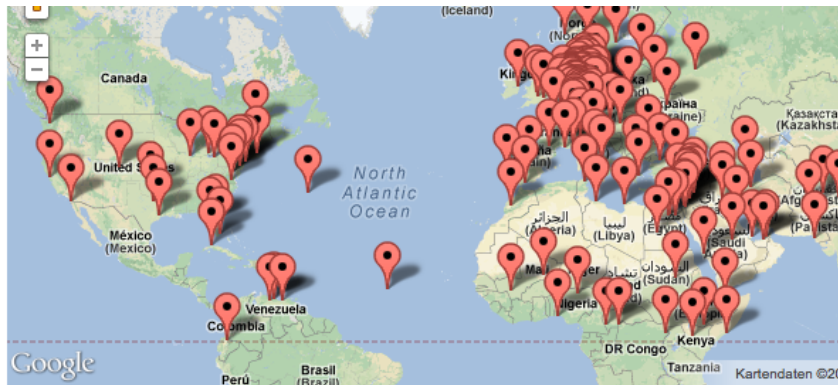
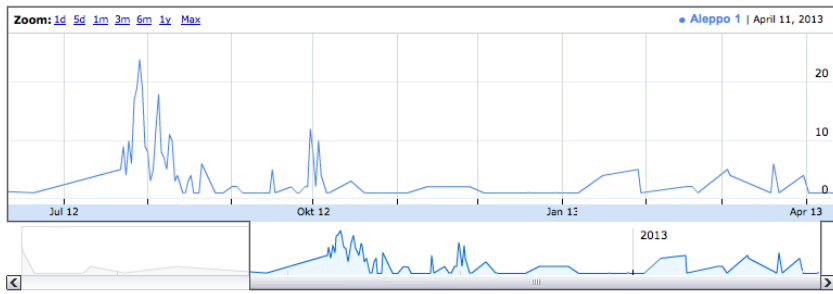
- Top: **-bach, -beck, -bek**



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



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!