



[Startseite](#) [Über dhmuc.](#) [Veranstaltungen](#) [Workshop: Digitale Editionen](#) [dhmuc. Site Visit](#) [Archiv](#) [Impressum](#)

Veröffentlicht am **29. September 2016** von **Eckhart Arnold**

[← Vorherige](#) [Weiter →](#)

## Ankündigung: Herbst-Workshop Digitale Editionen und Auszeichnungssprachen

Am 21./22. November veranstaltet dhmuc in der Bayerischen Akademie der Wissenschaften einen Workshop zum Thema „Digitale Editionen und Auszeichnungssprachen“.

# Digitale Editionen und Auszeichnungssprachen

## Computerlinguistische FinderApps mit Facsimile-Reader

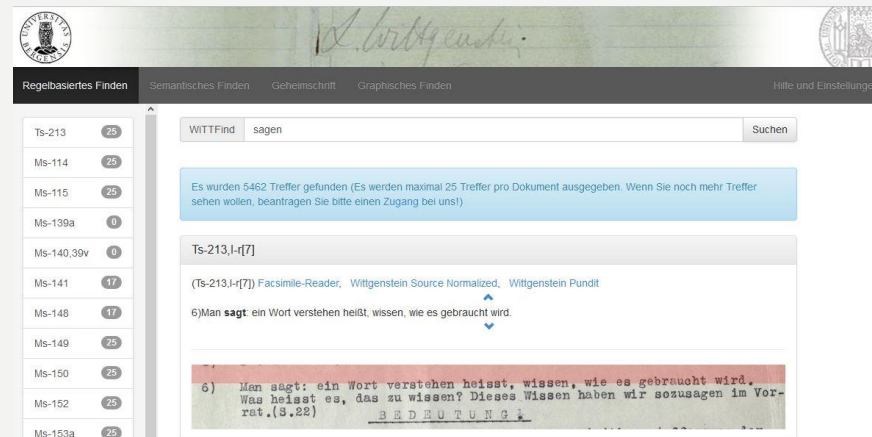
### Wittgenstein's Nachlass: WITTFind

### Goethe's Faust: GoetheFind

Hadersbeck M. et. al.

Centrum für Informations- und Sprachverarbeitung (CIS), LMU München

<http://wittfind.cis.uni-muenchen.de> and <http://goethefind.cis.uni-muenchen.de>





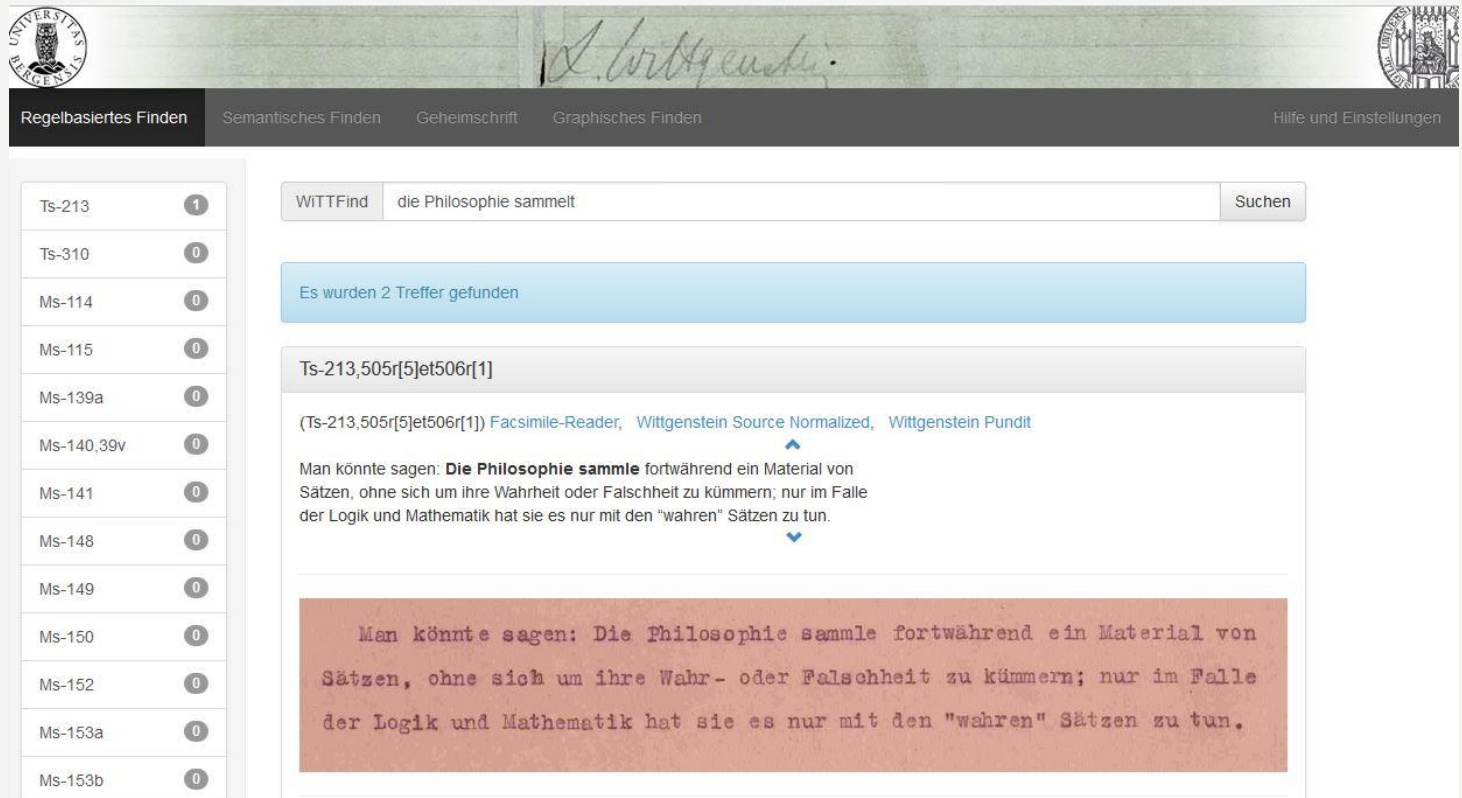
## Warum eine Suchmaschine neu entwickeln?

... es gibt doch opensource – Suchmaschine: solr (lucene)

- Außerordentlich hoher Einsatz bei der Entwicklung von Editionen
- Sehr komplexe Editionen
- Sehr spezifische Fragestellungen der „Humanities“
- Software-Knowhow vorhanden:
  - Digitales Vollformenlexikon (CISLEX)
  - Lokale Grammatiken, OCR, IMPACT (EU-Projekt)
  - Finite-State Automaten, approximative Suche (Levenshtein)
  - Natural Language Processing Tools
  - WEB-Technologie
  - Software Versionierung mit GIT

## FinderApp WiTTFind (Entwicklung seit 2009)

<http://wittfind.cis.uni-muenchen.de>



UNIVERSITÄT MÜNCHEN

Regelbasiertes Finden Semantisches Finden Geheimschrift Graphisches Finden Hilfe und Einstellungen

WITTFind die Philosophie sammelt Suchen

Es wurden 2 Treffer gefunden

Ts-213,505r[5]et506r[1]

(Ts-213,505r[5]et506r[1]) [Facsimile-Reader](#), [Wittgenstein Source Normalized](#), [Wittgenstein Pundit](#)

Man könnte sagen: **Die Philosophie sammle** fortwährend ein Material von Sätzen, ohne sich um ihre Wahrheit oder Falschheit zu kümmern; nur im Falle der Logik und Mathematik hat sie es nur mit den "wahren" Sätzen zu tun.

Man könnte sagen: Die Philosophie sammle fortwährend ein Material von Sätzen, ohne sich um ihre Wahr- oder Falschheit zu kümmern; nur im Falle der Logik und Mathematik hat sie es nur mit den "wahren" Sätzen zu tun.

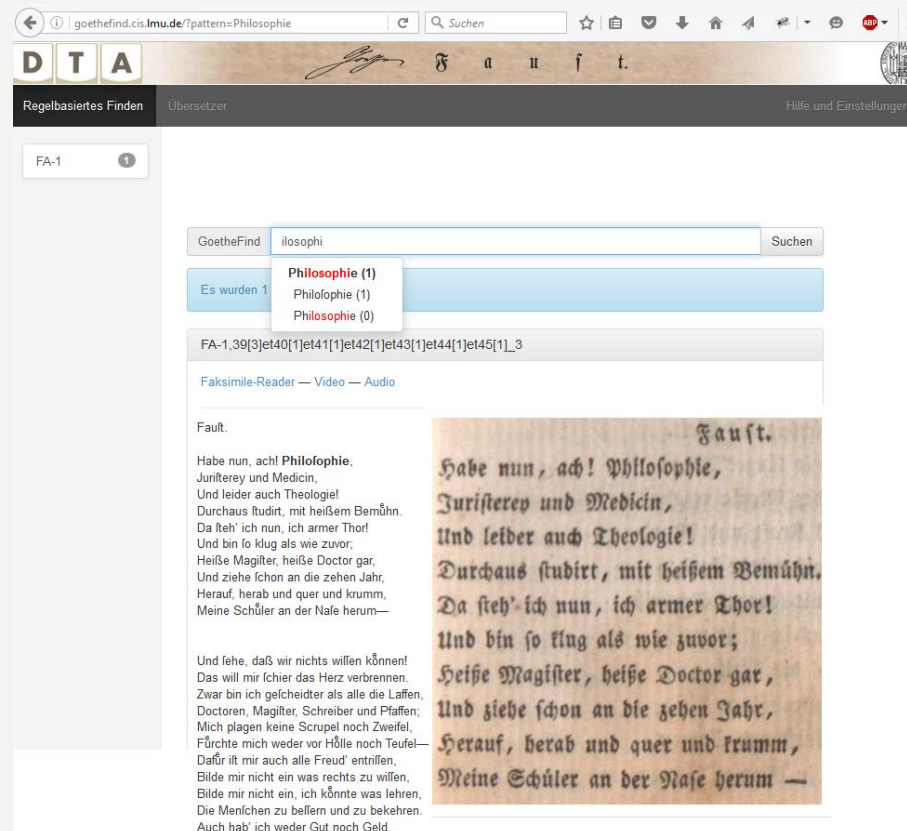
Ts-213 1  
Ts-310 0  
Ms-114 0  
Ms-115 0  
Ms-139a 0  
Ms-140,39v 0  
Ms-141 0  
Ms-148 0  
Ms-149 0  
Ms-150 0  
Ms-152 0  
Ms-153a 0  
Ms-153b 0

**Grundlagen von WiTTFind: sehr komplexe XML Edition**  
Bergen Electronic Edition, Wittgenstein Archiv in Bergen/Norwegen,  
5000 Seiten, öffentlich zugänglicher Teil des Nachlasses (1951)  
TEI-P5 konformes XML-Format, Prof. A. Pichler, Ø. Gjesdal  
<http://www.wittgensteinsource.org/>



## Multimediale FinderApp GoetheFind (Entwicklung seit 2015)

<http://goethefind.cis.uni-muenchen.de>



The screenshot shows the GoetheFind web application interface. At the top, there is a search bar with the URL `goethefind.cis.lmu.de/?pattern=Philosophie` and a search button labeled "Suchen". Below the search bar, there are navigation tabs for "D", "T", and "A". The main content area displays the search results for "Philosophie", showing a dropdown menu with "Philosophie (1)" and "Philosophie (0)". Below the search results, there is a section for "FA-1,39[3]et40[1]et41[1]et42[1]et43[1]et44[1]et45[1]\_3" with options for "Faksimile-Reader", "Video", and "Audio". The text of the search results is displayed in a digital edition format, showing the text of Faust's monologue "Habe nun, ach! Philosophie, Juristerey und Medicin, Und leider auch Theologie!".

GoetheFind  Suchen

Es wurden 1 **Philosophie (1)**  
Philosophie (1)  
Philosophie (0)

FA-1,39[3]et40[1]et41[1]et42[1]et43[1]et44[1]et45[1]\_3

Faksimile-Reader — Video — Audio

Fault.

Habe nun, ach! **Philosophie**,  
Juristerey und Medicin,  
Und leider auch Theologie!  
Durchaus studirt, mit heißem Bemühn.  
Da seh' ich nun, ich armer Thor!  
Und bin so klug als wie zuvor;  
Heiße Magister, heiße Doctor gar,  
Und ziehe schon an die zehen Jahr,  
Herauf, herab und quer und krumm,  
Meine Schüler an der Nase herum—

Und sehe, daß wir nichts wissen können!  
Das will mir schier das Herz verbrennen.  
Zwar bin ich gelcheider als alle die Laffen,  
Doctoren, Magister, Schreiber und Pfaffen;  
Mich plagen keine Scrupel noch Zweifel,  
Fürchte mich weder vor Hölle noch Teufel—  
Dafür ist mir auch alle Freud' entfallen,  
Bilde mir nicht ein was rechts zu wissen,  
Bilde mir nicht ein, ich könnte was lehren,  
Die Menschen zu bessern und zu bekehren.  
Auch hab' ich weder Gut noch Geld,

Faust.  
Habe nun, ach! Philosophie,  
Juristerey und Medicin,  
Und leider auch Theologie!  
Durchaus studirt, mit heißem Bemühn.  
Da seh' ich nun, ich armer Thor!  
Und bin so klug als wie zuvor;  
Heiße Magister, heiße Doctor gar,  
Und ziehe schon an die zehen Jahr,  
Herauf, herab und quer und krumm,  
Meine Schüler an der Nase herum —

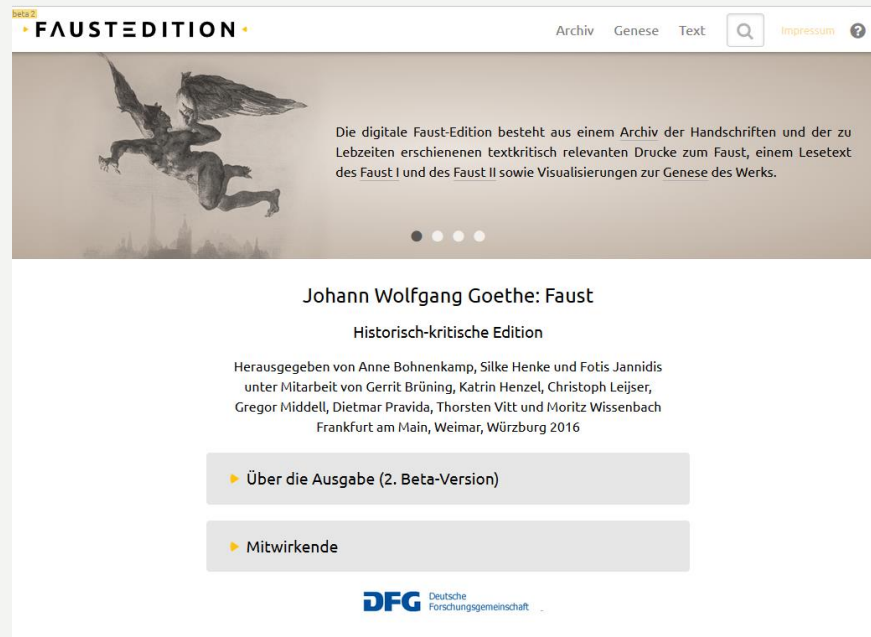
## Grundlagen von GoetheFind: einfachere XML Edition

J. W. Goethe, Deutsches Textarchiv: [www.deutschestextarchiv.de](http://www.deutschestextarchiv.de)

XML-TEI P5-based DTA-Basisformat (DTABf), HTML, Text, annotation Layer TCF

Historisch kritische Edition: Frankfurter Goethe-Haus

<http://www.goethehaus-frankfurt.de/forschung-und-editionen/digitale-faust-edition>



The screenshot shows the homepage of the 'FAUST EDITION' website. At the top, there is a navigation bar with links for 'Archiv', 'Genese', 'Text', a search icon, and 'Impressum'. Below the navigation bar is a large banner image of a winged figure (Faust) falling. To the right of the image, there is a text block: 'Die digitale Faust-Edition besteht aus einem Archiv der Handschriften und der zu Lebzeiten erschienenen textkritisch relevanten Drucke zum Faust, einem Lesetext des Faust I und des Faust II sowie Visualisierungen zur Genese des Werks.' Below the banner, the title 'Johann Wolfgang Goethe: Faust' is displayed, followed by 'Historisch-kritische Edition'. Below this, the editors are listed: 'Herausgegeben von Anne Bohnenkamp, Silke Henke und Fotis Jannidis unter Mitarbeit von Gerrit Brüning, Katrin Henzel, Christoph Leijser, Gregor Middell, Dietmar Pravida, Thorsten Vitt und Moritz Wissenbach Frankfurt am Main, Weimar, Würzburg 2016'. At the bottom, there are two buttons: 'Über die Ausgabe (2. Beta-Version)' and 'Mitwirkende'. The DFG logo is visible at the bottom right.

# Grundlagen der FinderApps: Faksimile der Editionen

Highlighting der Treffer im Faksimile (EU-Projekt Impact, Prof. Schulz)

Dazu WEB-basiertes OCR Korrektur-Tool: multiuserfähig, halbautomatisch

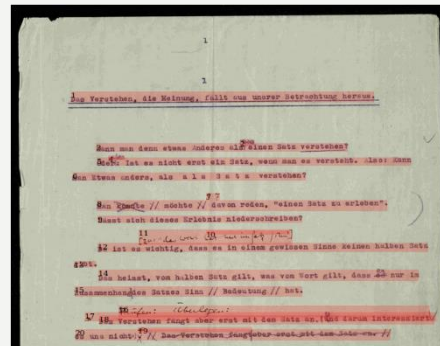
OCR mit tesseract

Typoskripte

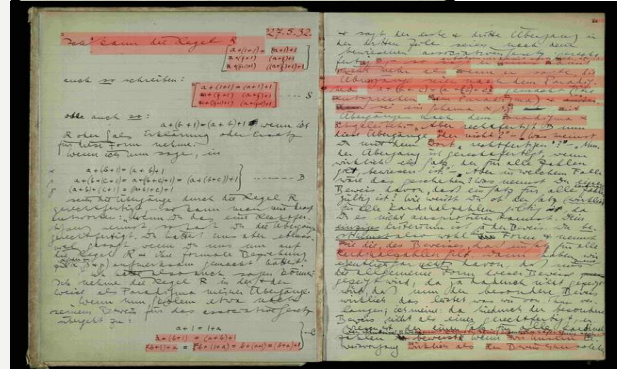
(recht gut)

Manuskripte

(sehr schlecht)



- 1: Das Verstehen die Meinung fällt aus unsrer Betrachtung heraus
- 2: iwü
- 3: Kann man dann etwas Anderes ul'sTeinen Satz verstggg
- 4: 1 634
- 5: Oedeff Ist es nicht erst ein Satz wenn man es versteht 150 Kann
- 6: män Etwas gnders. als a l s 3 a t z verstehen
- 7: S 7
- 8: Man gynäze möchte f davon reden einen Satz zu erleben
- 9: Lasst sich dieses Erlebnis niederschreiben
- 10: ., B h
- 11: fxti Luv LAauhAJLÄ thj
- 12: Da ist es wichtig dass es in einem gewissen Sinne Keinen halben Satz
- 13: gibt
- 14: Das heisst vom hxlben Satz gilt was vom hort gilt äassfü-nur im
- 15: zusammenhanäes Satzes Sinn Bedeutung hat
- 16: Qhagw am i Ä



- 1; a4 CH aFfvÖf
- 2; ßHfic»
- 3; w w
- 4; cuMj
- 5; z 7 552
- 6; auVI
- 7; a f I aaf ,L
- 8; a 43700 fdfdflljH
- 9; 4 aJrljIiY
- 10; cL a p



## Grundlagen der FinderApps:

## Linguistik

### Editionen sind keine „raw-texte“

- Strukturierte Edition: siglum Technik als feste Anker in der Edition
  - Dokument, Bemerkung/Abschnitt, Satz
- Textgenetische Information (Tabellen, XML)
- Semantische Informationen: Eigennamen, Orts- Zeitangaben
- Strukturierte Dokumente:
  - Linguistisch: Token, Wörter, Worttrennungen, Sätze
  - Physikalisch: Zeilen, Abschnitte, Paragraphen, Seiten, Dokumente

## Beispiel zum Wittgenstein Nachlass und Bergen Electronic Edition

[Ts-213] Viewer: Faksimile-Extrakt & Transkription für Ts-213,12r[2]

HTML

XML

Schließen

*eine sehr häufige Auffassung: ...*  
*geläufige*  
 Es ist ~~eine häufige~~ Auffassung, dass Einer ~~gleichsam~~ nur unvollkommen  
*(einen Satz, ein Zeichen (einen Befehl))*  
 zeigen kann, ob er ~~verstanden hat~~.  
*sein Verständnis nur unvollkommen*

Dass er gleichsam nur immer aus der Ferne darauf deuten, auch sich ihm  
 nähern, es aber nie mit der Hand berühren // ergreifen // kann. Und das  
 Letzte immer ungesagt bleiben muss.

```
<ab ana="abnr:216" n="Ts-213,12r[2]">
  <s ana="facs:Ts-213,12r abnr:216 satznr:538" n="Ts-213,12r[2]_1">Es ist <choice type="s"><seg n="s
  _alt1"><emph rend="usb">eine</emph><seg type="stripped"> geläufige</seg> Auffassung,</seg><seg n="s
  _alt2"> eine sehr häufige Auffassung:...</seg></choice> daß Einer <seg type="stripped"><seg type="st
  ripped"> sein Verständnis nur unvollkommen zeigen kann.</seg></seg></s>
  <s ana="facs:Ts-213,12r abnr:216 satznr:539" n="Ts-213,12r[2]_2">Daß er gleichsam nur immer aus de
  r Ferne <emph rend="uw1_h">darauf</emph> deuten, auch sich ihm<lb></lb> nähern, es aber nie mit der
  Hand <choice type="s"><seg n="s_alt1">berühren</seg><seg n="s_alt2"> ergreifen</seg></choice> ka
  nn.</s>
  <s ana="facs:Ts-213,12r abnr:216 satznr:540" n="Ts-213,12r[2]_3">Und das<lb></lb> Letzte immer ung
  esagt bleiben muß.</s>
</ab>
```

## Grundlagen der FinderApps:

## Linguistik

### elektronisches, multilinguales Vollformen- und Phrasenlexikon (WiTTLex)

- morphologische, syntaktische und semantische Kategorien
- Eigennamen
- Präfixverben
- Mehrwortlexeme

```
1 nachdenken, nachdenken.V+#4
2 farbgleiche, farbgleich.ADJ+COL+KOMP
3 part of the practice of communication, .X+z1
4 part of the time, .X+z1
5 nachdenkender, nachdenkend.ADJ+ER
6 rot, rot.ADJ+COL+Grundfarbe
7 Russell, ..EN:geM
8 Russells, Russell.EN:geM
9 Russellsche, Russell.EN
10 Russellschen, Russell.EN
11 Russen, Russe.N:aeM:amM:deM:dmM:geMS:gmM:nmM
```

## Grundlagen der FinderApps:

## Linguistik

### Lokale Grammatiken

Partikelverb Disambiguierung

Regelbasierter Zugriff

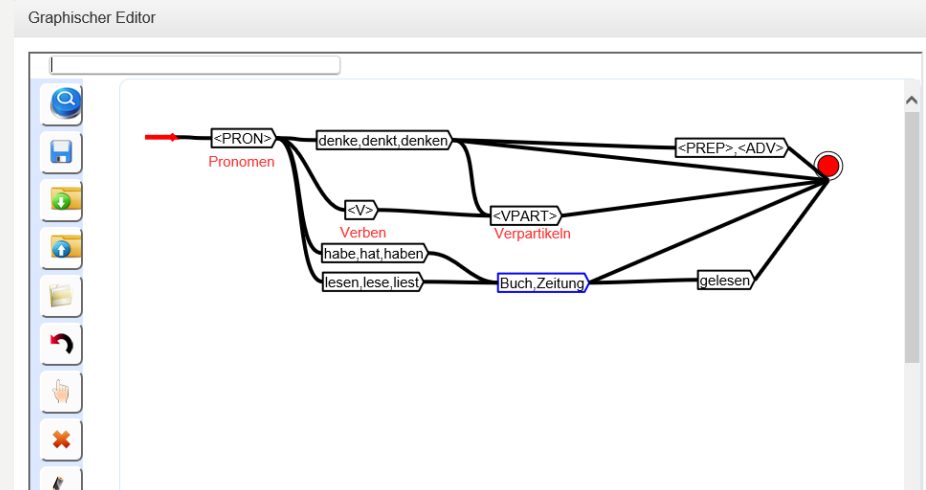
Semantische Kategorien

Phrasengrammatik

Satzebene

Basierend auf Ideen von

Unitex/GramLab, Paris-Est Marne-la-Vallée





## Grundlagen der FinderApps:

## Linguistik

### Part of Speech Tagger `treetagger`, Dr. Schmid (CIS)

- Problemlos einsetzbar bei “raw” Texten
- Problematisch bei Multilingualen Texten, XML Annotation
- Problematisch bei speziellen Texten da darauf nicht trainiert
  - historisches Deutsch, “seltene Wörter” (WiTTFind: 20000 Wörter mit Wortfrequenz=1)
- Ausweg: eigenes Lexikon
- Prä- und Postprocessing beim “taggen”

## Grundlagen der FinderApps:

## Informatik

### Annotierte Texte

- XML TEI-P5: (Reduziertes XML, Deutsches Text Archiv Basis Format (DTAbf))
- Problem: Überannotation, Überlappung
- Semantic WEB Techniken: Resource Description Framework Trippel ((rdf))
- Datenbanken mit OWL Information und mit Ankern in den Dokumenten
- WEB-Technologie: Bootstrap, Javascript und HTML5

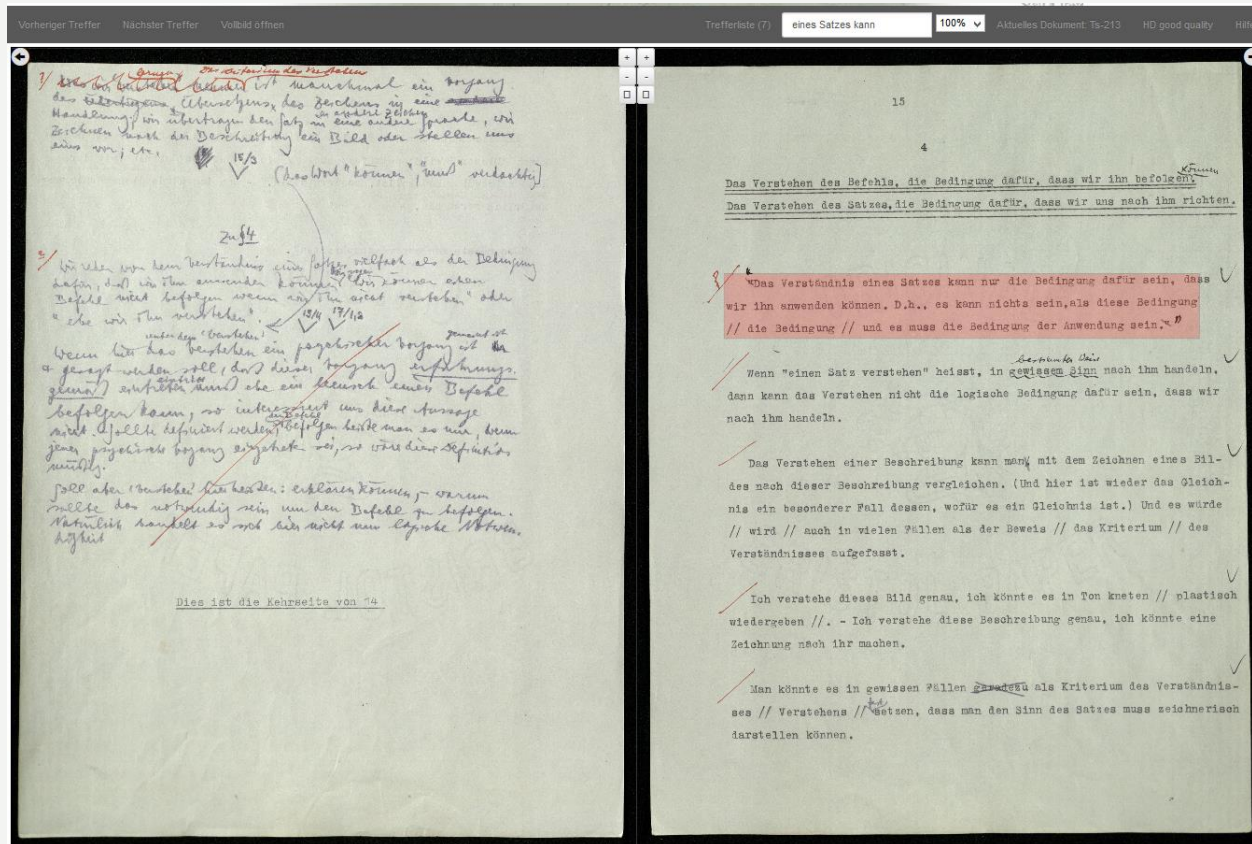
## Grundlagen der FinderApps:

## Informatik

### Annotierte Texte mit Ankeren

- Neue Finder Editions-Basis: Tokensegmentierung und Anker:
  - Anchored XML unter <http://dev.wittfind.cis.uni-muenchen.de/>
- Derzeit zwei Phd Projekte in Cambridge: „Neue Annotationsmethoden für Editionen“

# Spezielle Features der Finder: mit Treffer-highlighting, Doppelseitig blättern, Downloadmode

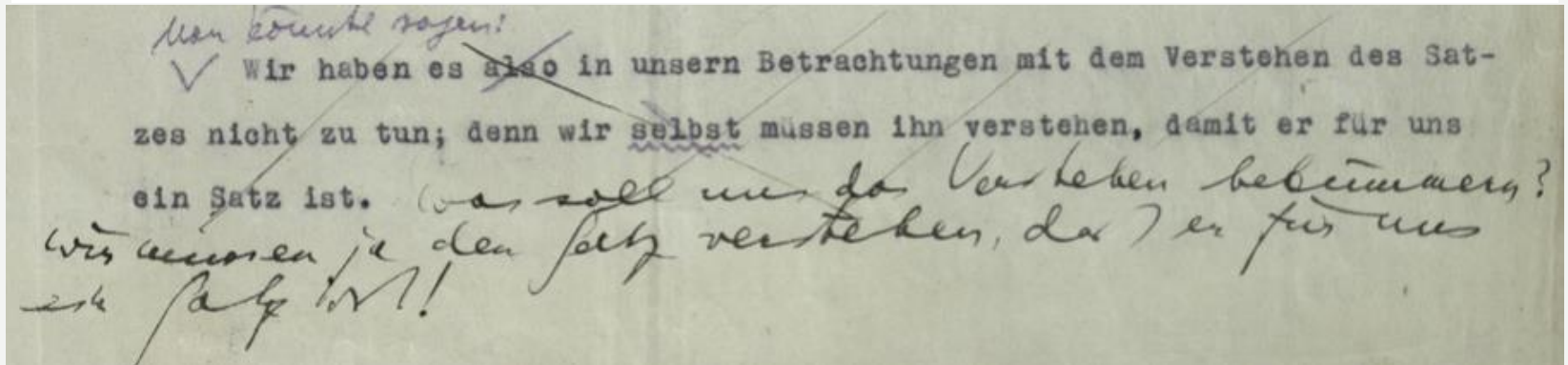


## “Grundtenor” der Humanists: Wer traut schon einer Edition?

### VOM DATENMODELL ZUR FINDERAPP

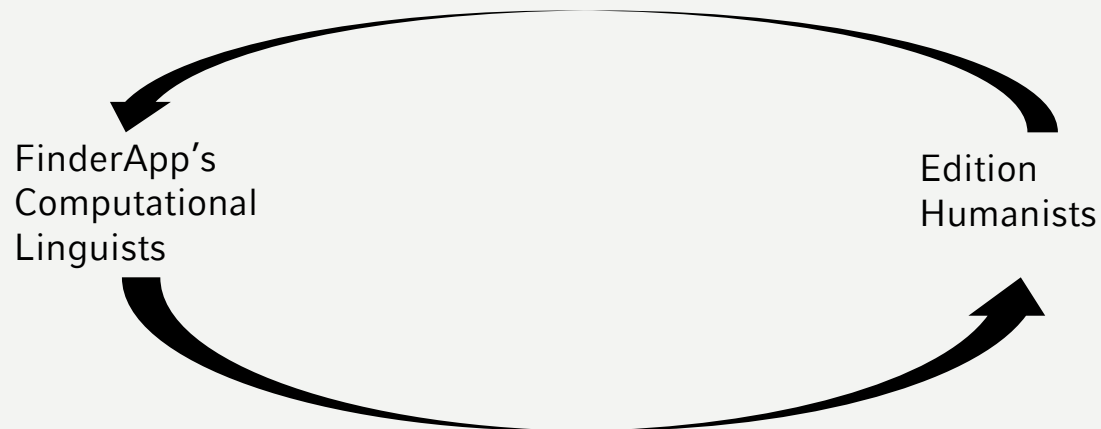
**Man könnte sagen:** Was soll uns das Verstehen kümmern?

Wir müssen ja den Satz verstehen, daß er für uns  
ein Satz ist!



## Erfolgreiche Kooperation “Humanists” und Computerlinguistik

(high Precision, high Recall)



# Kooperation über die Feedback-App

Vom hervorgehobenen Treffer zum Faksimile zur Edition zum Editor

Investigate Mode: [ Ts-213,217r[1] ] Faksimile-Extrakt & Transkription

HTML XML-norm XML-orig

50

Ist die Vorstellung das Porträt rar exzellente, also grund-  
verschieden, etwa, von einem gemalten Bild und durch ein solches oder  
etwas Ähnliches nicht ersetzbar? Ist sie das, was eigentlich eine be-  
stimmte Wirklichkeit darstellt, - zugleich Bild und Meinung?

Denn ein Wunderding, schenken, brau-  
che wir.  
Und die Vorstellung scheint die ganze  
Denn wir können uns nicht diese  
eigene Vorstellung von diesem Fleische  
möglich, die vor dem Auge liegen  
trouwe, aber sie sind für einen Augen der  
Zeit nicht.

```

<ab ana="abnr:1101" n="Ts-213,217r[1]">
<s ana="facs:Ts-213,217r abnr:1101 satznr:3499" n="Ts-213,217r[1]_1">Aber ist nicht der Satz diese
s
  <lb></lb> Wunderding —
  <seg part="N" type="edcom"></seg> der
  <emph rend="usl_h">sagt</emph>, was er
  <lb></lb>
  <emph rend="usl_h">meint</emph>
</s>
<lb rend="hl"></lb>
<s ana="facs:Ts-213,217r abnr:1101 satznr:3500" n="Ts-213,217r[1]_2">Denn so ein Wunderding, schen-
nt es, brau
  <lb rend="stypchen"></lb>chen wir.
</s>
<s ana="facs:Ts-213,217r abnr:1101 satznr:3501" n="Ts-213,217r[1]_3">Und die Vorstellung scheint
  <choice type="s">
    <seg n="s_alti">yes</seg>
    <seg n="s_alt2">dies</seg>
  </choice> zu sein
  <seg type="stripped">
    <seg type="stripped"> Denn</seg>
  
```

Feedback

Email-Adresse

Fehlerbeschreibung

Zum Abschicken der Fehlermeldung bitte die Figur in der Box nachzeichnen

Abschicken



## Vielen Dank für Ihre Aufmerksamkeit

- an A. Pichler (WAB, Bergen), A. Geyken (Deutsches Textarchiv, Berlin), G. Brüning (Freies Deutsches Hochstift, Frankfurt)
- an die Studierenden: Faridis Alberteris Azar, Matthias Lindinger, Stefan Schweter, Michael Wöß
- und dem LEHRE@LMU Team für die freundliche Förderung von studentischen Projekten