DAS DEUTSCHE TEXTARCHIV

Deutsche Texte des 17.-19. Jahrhunderts

2424 Werke

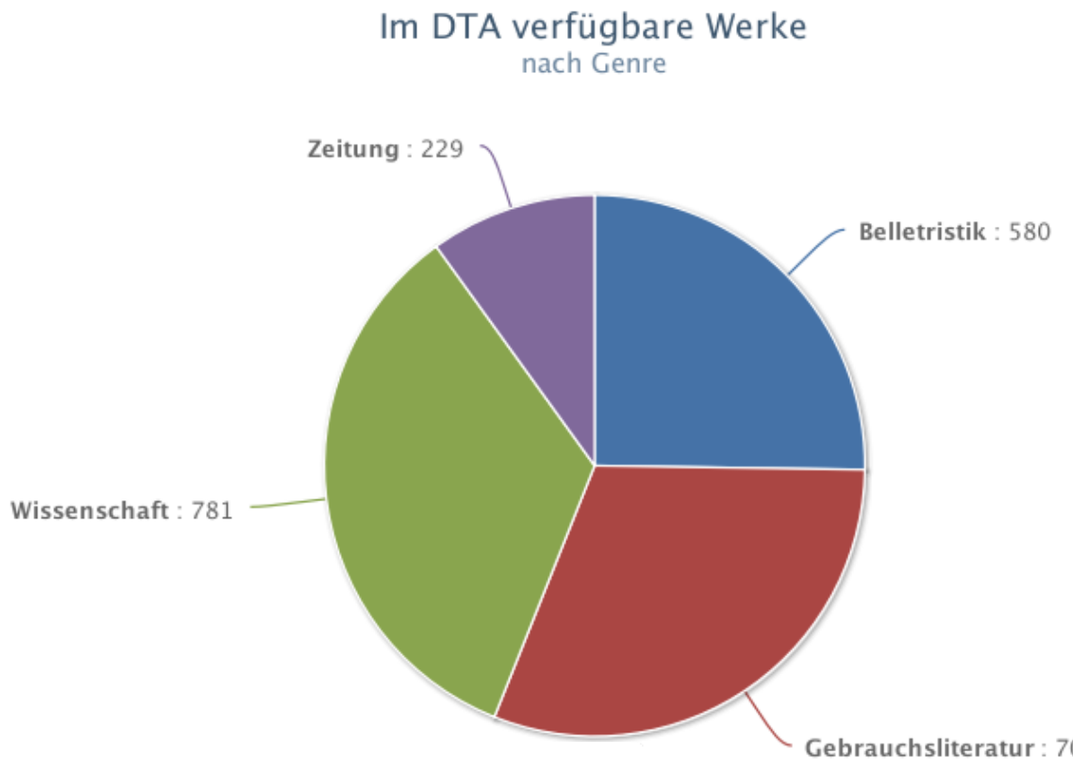
588507 digitalisierte Seiten

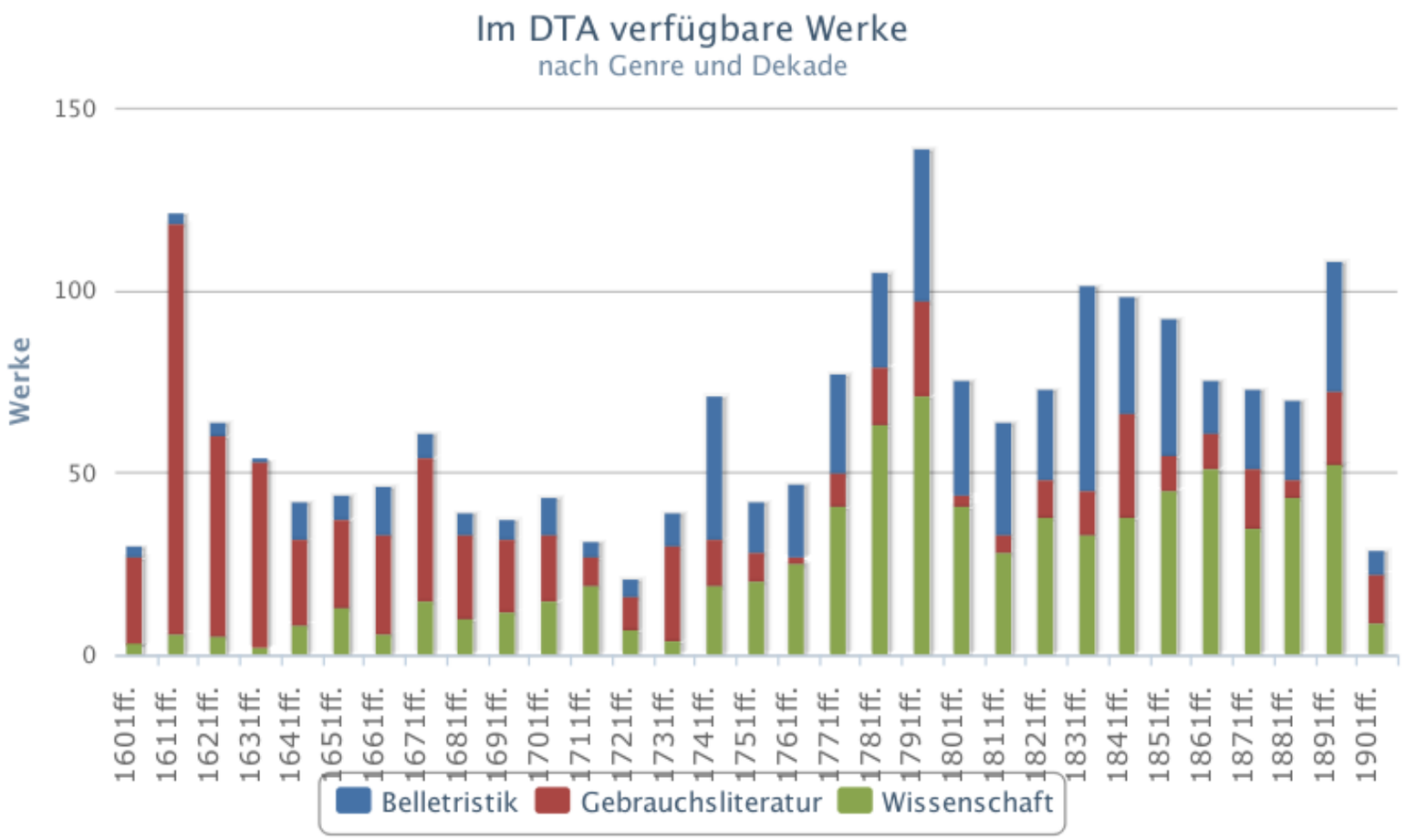
138658619 fortlaufende Wortformen

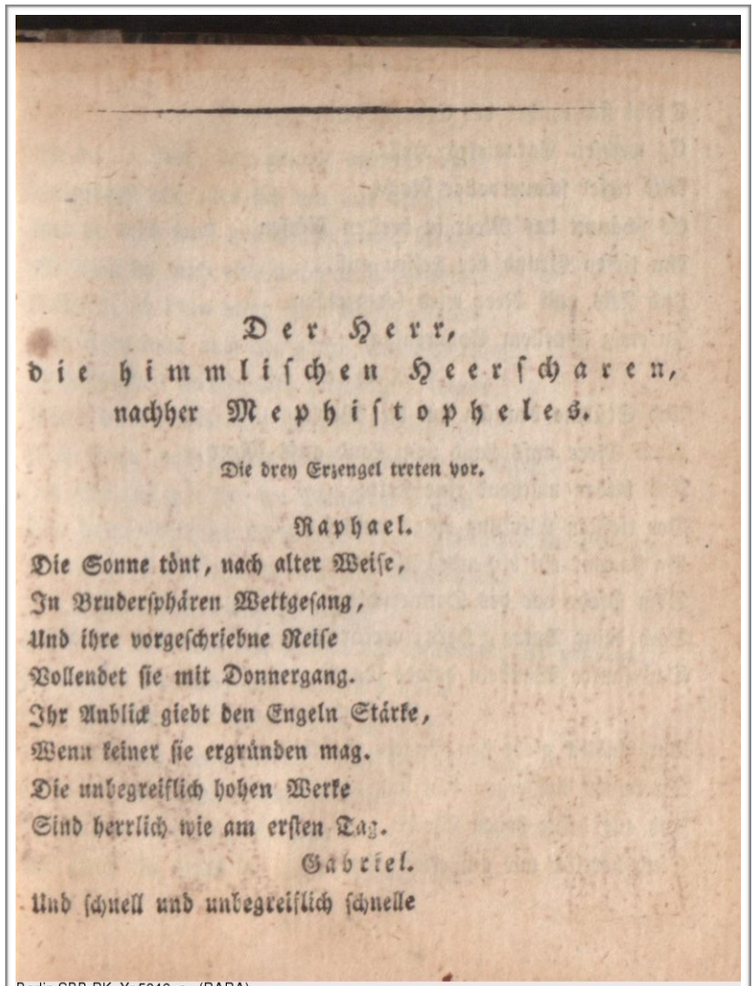
983530500 Zeichen (Unicode)

DTA

- Sorgfältige Auswahl von Texten
- Bevorzugt Erstausgaben
- Sehr hohe Erfassungsgenauigkeit (Fehlerrate 0,02%)
- Ermöglicht sprachhistorische Forschung







Der Herr,
die himmlischen Heerscharen,
nachher Mephistopheles.

Die drey Erzengel treten vor.

Raphael.

Die Sonne tönt, nach alter Weise,
In Brudersphären Wettgesang,
Und ihre vorgeschriebne Reise
Vollendet sie mit Donnergang.
Ihr Anblick giebt den Engeln Stärke,
Wenn keiner sie ergründen mag.
Die unbegreiflich hohen Werke
Sind herrlich wie am ersten Tag.

Gabriel.

Und schnell und unbegreiflich schnelle

Beispiel:
Faust I: Seite 23

Der Herr,
die himmlischen Heerscharen,
nachher Mephistopheles.

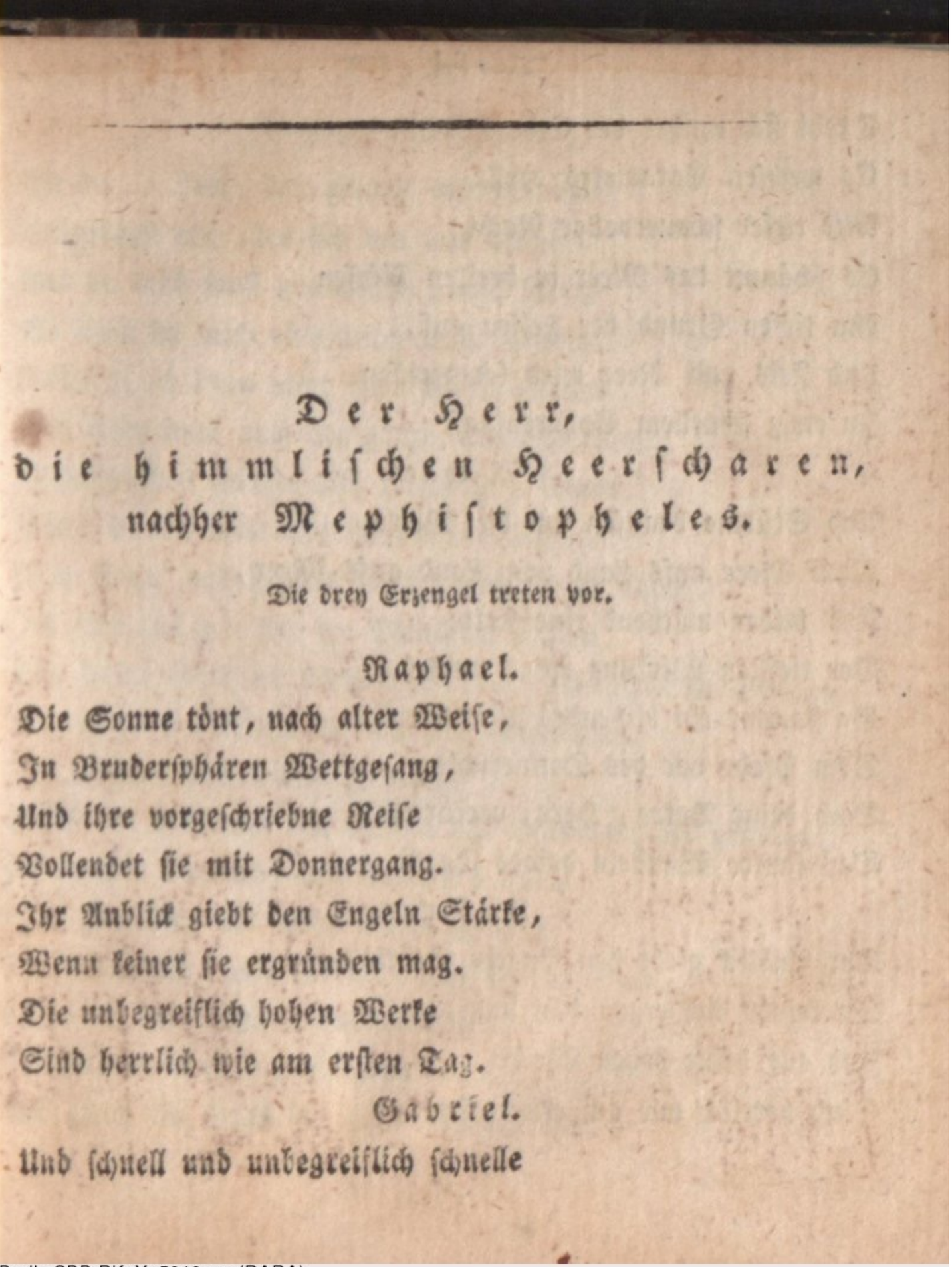
Die drey Erzengel treten vor.

Raphael.
Die Sonne tönt, nach alter Weise,
In Brudersphären Wettgesang,
Und ihre vorgeschriebne Reise
Vollendet sie mit Donnergang.
Ihr Anblick giebt den Engeln Stärke,
Wenn keiner sie ergründen mag.
Die unbegreiflich hohen Werke
Sind herrlich wie am ersten Tag.

Gabriel.
Und schnell und unbegreiflich schnelle

VOM FAKSIMILE ZUM ANNOTIERTEN DIGITALISAT

GENUTZTE TECHNIKEN ZUR ERSTELLUNG DES ELEKTRONISCHEN VOLLTEXTES



Der Herr,
die himmlischen Heerscharen,
nachher Mephistopheles.

Die drey Erzengel treten vor.

Raphael.

Die Sonne tönt, nach alter Weise,
In Brudersphären Wettgesang,
Und ihre vorgeschriebne Reise
Vollendet sie mit Donnergang.
Ihr Anblick giebt den Engeln Stärke,
Wenn keiner sie ergründen mag.
Die unbegreiflich hohen Werke
Sind herrlich wie am ersten Tag.

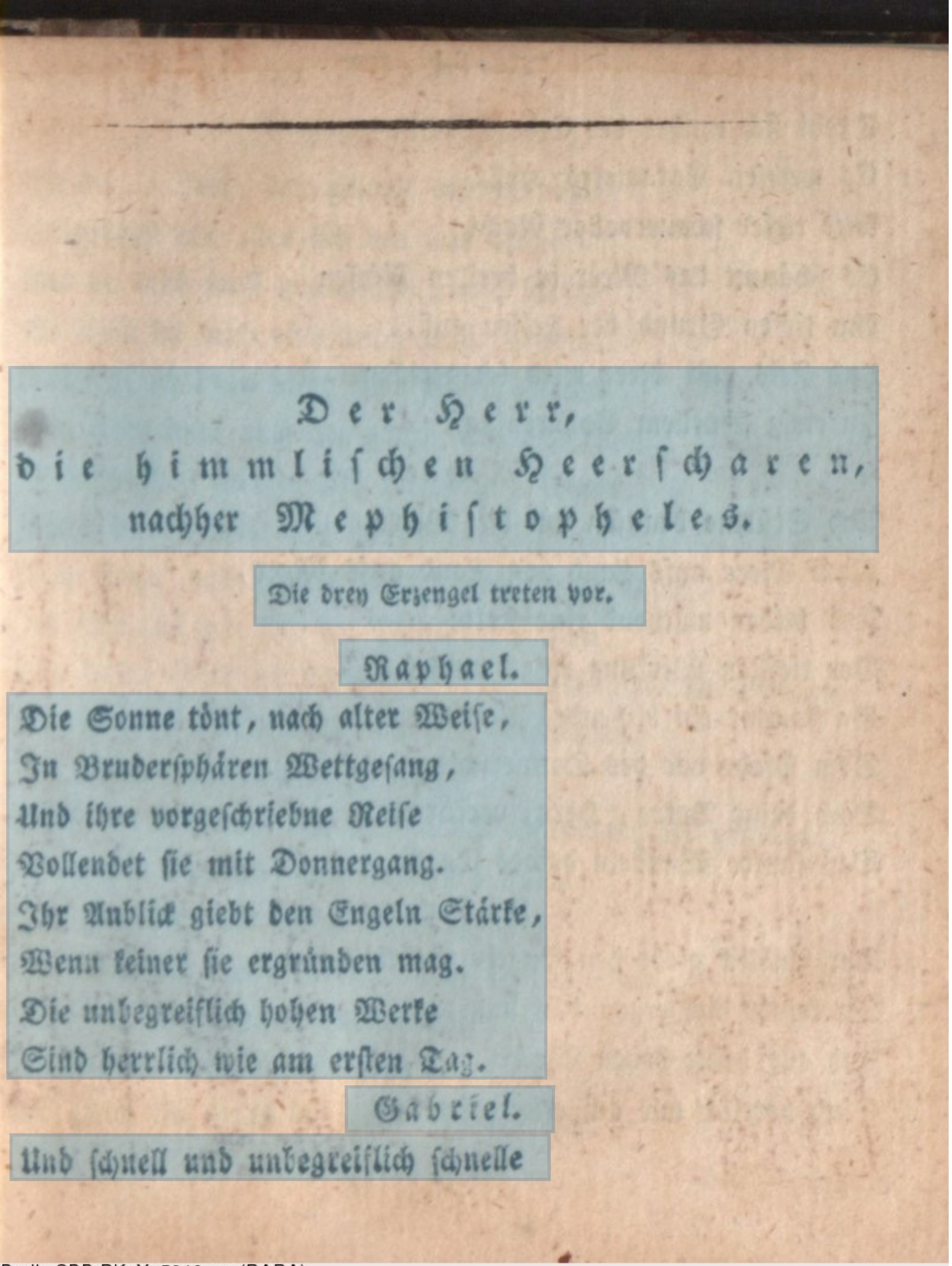
Gabriel.

Und schnell und unbegreiflich schnelle

PROBLEME DER FORMATIERUNG

Unterschiedliche Textpassagen

- „Kapitelüberschrift“
- „Untertitel“
- „Seitenzahl"
- „Fußnote"
- usw.



Der Herr,
die himmlischen Heerscharen,
nachher Mephistopheles.

Die drey Erzengel treten vor.

Raphael.

Die Sonne tönt, nach alter Weise,
In Brudersphären Wettgesang,
Und ihre vorgeschriebne Reise
Vollendet sie mit Donnergang.
Ihr Anblick giebt den Engeln Stärke,
Wenn keiner sie ergründen mag.
Die unbegreiflich hohen Werke
Sind herrlich wie am ersten Tag.

Gabriel.

Und schnell und unbegreiflich schnelle

Berlin SBB-PK, Yr 5046<a> (RARA)

MAKROSTRUKTURIERUNG DER BILDDIGITALISATE

- Definieren verschiedener Textzonen
- Speichern der Strukturinformationen
- Vermeiden von Uneinheitlichkeit bei der digitalen Transkription

DOUBLE KEYING

Texte werden von zwei verschiedenen Personen erfasst

Dritte Person kontrolliert Fehler

Sehr hoher Aufwand - Fehlerrate von ~0.02%

OPTICAL CHARACTER RECOGNITION

Automatisierte Texterkennung innerhalb von Bildern

Deutlich höhere Fehlerrate

Hoher Aufwand der manuellen Nachkorrektur

SCHWIERIGKEITEN BEI DER ERKENNUNG VON WORT UND SATZGRENZEN

Typische Zeichen müssen nicht am Satzende stehen

- 1.000 Liter Wasser entsprechen einem Kubikmeter.
- E. T. A. Hoffmann starb 1822 in Berlin.
- „Warum nicht?“, rief er.
- http://www.google.de

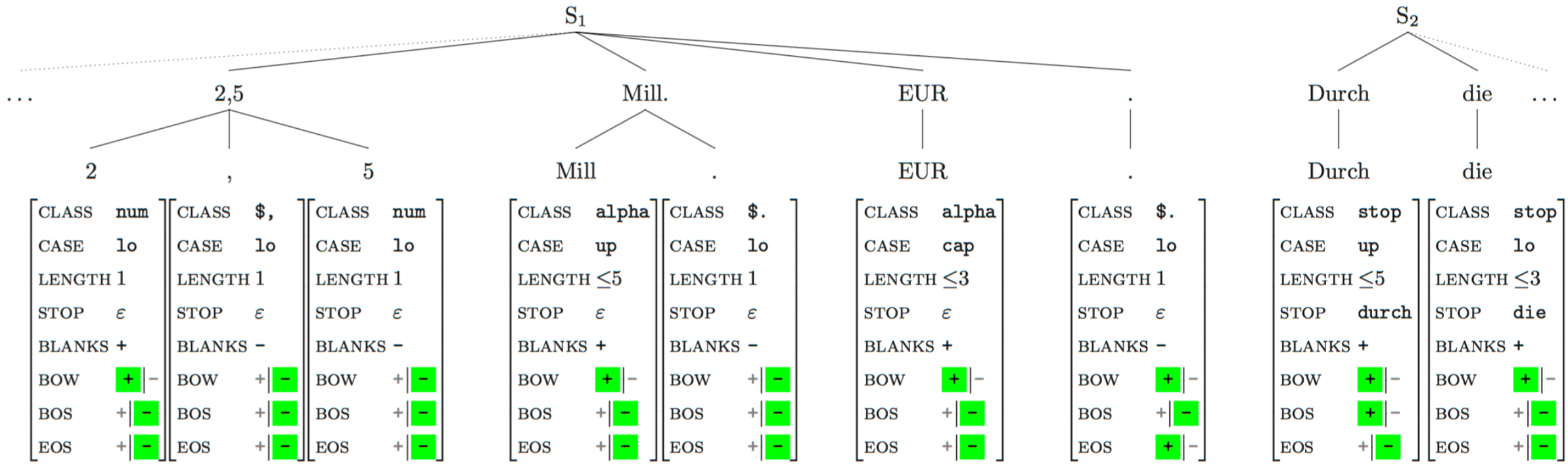
ARBEITSWEISE DES W.A.S.T.E. – SCANNERS

- Rohtext wird an Punkt und Satzzeichen getrennt.
  - auch Leerzeichen bilden nun Token
- Mindestens alle Wörter und Sätze sind getrennt
  - Abkürzungen, falsche Satzenden etc. stellen weiterhin Probleme dar

KLASSIFIZIERUNG ANSTELLE DER TOKENISIERUNG

- Durch Scanner ist Tokenisierung eigentlich abgeschlossen
- Für jeden Token gilt die Frage:
  - handelt es sich um einen Wortanfang?
  - handelt es sich ebenfalls um einen Satzanfang?

DIE KLASSIFIZIERUNG



f	rng(f)	**Description**
CLASS	{stop, alpha, num, ...}	typographical class
CASE	{lo, up, cap}	letter-case
LENGTH	$\{1, \leq 3, \leq 5, > 5\}$	segment length
STOP	finite set $\mathcal{S}_{\mathcal{L}}$	stopword text
BLANKS	$\{+, -\}$	leading whitespace?
ABBR	$\{+, -\}$	known abbreviation?
BOW	$\{+, -\}$	beginning-of-word?
BOS	$\{+, -\}$	beginning-of-sentence?
EOS	$\{+, -\}$	end-of-sentence?

Class	**Description**
stop	language-specific stopwords (see below)
roman	segments which may represent roman numerals
alpha	segments containing only alphabetic characters
num	segments containing only numeric characters
$.	period
$,	comma

WEITERGEGEBENES XML-FORMAT

```
<token ID="w78f">Raphael</token>
<token ID="w790">.</token>
<token ID="w791">Die</token>
<token ID="w792">Sonne</token>
<token ID="w793">to&#x0364;nt</token>
<token ID="w794">,</token>
<token ID="w795">nach</token>
<token ID="w796">alter</token>
<token ID="w797">Wei&#x017F;e</token>
<token ID="w798">,</token>
<token ID="w799">In</token>
<token ID="w79a">Bruder&#x017F;pha&#x0364;ren</token>
<token ID="w79b">Wettge&#x017F;ang</token>
<token ID="w79c">,</token>
<token ID="w79d">Und</token>
<token ID="w79e">ihre</token>
<token ID="w79f">vorge&#x017F;chriebne</token>
<token ID="w7a0">Rei&#x017F;e</token>
<token ID="w7a1">Vollendet</token>
<token ID="w7a2">&#x017F;ie</token>
<token ID="w7a3">mit</token>
<token ID="w7a4">Donnergang</token>
<token ID="w7a5">.</token>
<token ID="w7a6">Ihr</token>
<token ID="w7a7">Anblick</token>
<token ID="w7a8">giebt</token>
<token ID="w7a9">den</token>
<token ID="w7aa">Engeln</token>
<token ID="w7ab">Sta&#x0364;rke</token>
<token ID="w7ac">,</token>
<token ID="w7ad">Wenn</token>
<token ID="w7ae">keiner</token>
<token ID="w7af">&#x017F;ie</token>
<token ID="w7b0">ergru&#x0364;nden</token>
<token ID="w7b1">mag</token>
<token ID="w7b2">.</token>
<token ID="w7b3">Die</token>
<token ID="w7b4">unbegreiflich</token>
<token ID="w7b5">hohen</token>
<token ID="w7b6">Werke</token>
<token ID="w7b7">Sind</token>
<token ID="w7b8">herrlich</token>
<token ID="w7b9">wie</token>
<token ID="w7ba">am</token>
<token ID="w7bb">er&#x017F;ten</token>
<token ID="w7bc">Tag</token>
<token ID="w7bd">.</token>
```

Der Herr,
die himmlischen Heerscharen,
nachher Mephistopheles.

Die drey Erzengel treten vor.

Raphael.

Die Sonne tönt, nach alter Weise,
In Brudersphären Wettgesang,
Und ihre vorgeschriebne Reise
Vollendet sie mit Donnergang.

Ihr Anblick giebt den Engeln Stärke,
Wenn keiner sie ergründen mag.

Die unbegreiflich hohen Werke
Sind herrlich wie am ersten Tag.

Gabriel.

Und schnell und unbegreiflich schnelle

```
<sentence ID="s6a" tokenIDs="w77f w780 w781 w782 w783 w784 w785 w786 w787 w788"/>
<sentence ID="s6b" tokenIDs="w789 w78a w78b w78c w78d w78e"/>
<sentence ID="s6c" tokenIDs="w78f w790"/>
<sentence ID="s6d" tokenIDs="w791 w792 w793 w794 w795 w796 w797 w798 w799 w79a w79b w79c w79d w79e w79f w7a0
    w7a1 w7a2 w7a3 w7a4 w7a5"/>
<sentence ID="s6e" tokenIDs="w7a6 w7a7 w7a8 w7a9 w7aa w7ab w7ac w7ad w7ae w7af w7b0 w7b1 w7b2"/>
<sentence ID="s6f" tokenIDs="w7b3 w7b4 w7b5 w7b6 w7b7 w7b8 w7b9 w7ba w7bb w7bc w7bd"/>
<sentence ID="s70" tokenIDs="w7be w7bf"/>
```

CAB („CASCADED ANALYSIS BROKER“)

- Prüft die Existenz einer modernen Schreibvariante historischer Wortformen

1. Unicode in Latin-1 Transformation	ohngefaͤhr	ohngefähr
2. Zurückführen auf phonetische Form	Theyl, Thayl, Teyl	[taɪl]
oder		
2.Ermitteln des ähnlichsten modernen Wortes	gläuben	glauben

DIE SUCHMASCHINE DDC

- Baut für jeden Text eine maschinenlesbare Indexdatei
- Zusatzinformationen für spätere Anfrage

```
<s>
 <l>Im  Im      Im      APPRART im      1392|1948|1459|1996     472     -       |text|</l>
 <l>Durchschnitt        Durchſchnitt    Durchschnitt    NN      Durchschnitt
1471|1948|1706|1999     472     -       |text|</l>
 <l>gilt        gilt    gilt    VVFIN   gelten  1721|1956|1780|2000     472     -
|text|</l>
 <l>ein ein     ein     ART     eine    1797|1951|1845|1985     472     -       |text|</l>
 <l>zahmer      zahmer  zahmer  ADJA    zahm    789|2001|921|2047       472     -
|text|</l>
 <l>Elephant    Elephant        Elefant NN      Elefant 931|2005|1100|2055      472     -
|text|</l>
 <l>ohngefähr   ohngefa&#x0364;hr       ungefähr        ADJD    ungefähr
1116|2001|1300|2060     472     -       |text|</l>
 <l>zweyhundert zweyhundert     zweihundert     CARD    zweihundert     1318|2001|1550|2060
472     -       |text|</l>
 <l>Thaler      Thaler  Taler   NN      Taler   1570|2001|1690|2060     472     -
|text|</l>
 <l>.   .       .       $.      .       1690|2001|1710|2060     472     -       |text|</l>
</s>
```

<l> latin-1 ; vorliegende Unicode-Form ; moderne Wortform (CAB) ; Syntaktische Kategorie ; Grundform ; ... </l>

FINALES XML-FORMAT

Zu jedem Wort:

- Token

```
<token ID="w7a6">Ihr</token>
<token ID="w7a7">Anblick</token>
<token ID="w7a8">giebt</token>
<token ID="w7a9">den</token>
```

- Grundform

```
<lemma tokenIDs="w7a6">ihr</lemma>
<lemma tokenIDs="w7a7">Anblick</lemma>
<lemma tokenIDs="w7a8">geben</lemma>
<lemma tokenIDs="w7a9">d</lemma>
```

- Wortart

```
<tag tokenIDs="w7a6">PPOSAT</tag>
<tag tokenIDs="w7a7">NN</tag>
<tag tokenIDs="w7a8">VVFIN</tag>
<tag tokenIDs="w7a9">ART</tag>
```

- Moderne Wortform

```
<correction tokenIDs="w7a0" operation="replace">Reise</correction>
<correction tokenIDs="w7a2" operation="replace">sie</correction>
<correction tokenIDs="w7a8" operation="replace">gibt</correction>
<correction tokenIDs="w7ab" operation="replace">Stärke</correction>
```

Im Beispieltext „Faust I“ entsteht ein XML Dokument mit ~126.000 Zeilen

(Bei 309 verarbeiteten Seiten und c.a. 30.000 Wörtern)