UNIVERSITÄT LEIPZIG

Institut für Informatik

ASV

Automatische Sprachverarbeitung

Canonical Text Service und Text Re-Use am Beispiel des DTA-Korpus

Jochen Tiepmar

Abteilung für Automatische Sprachverarbeitung

Institut für Informatik

Universität Leipzig

jtiepmar@informatik.uni-leipzig.de

Europa fördert Sachsen.

ESF

Europäischer Sozialfonds

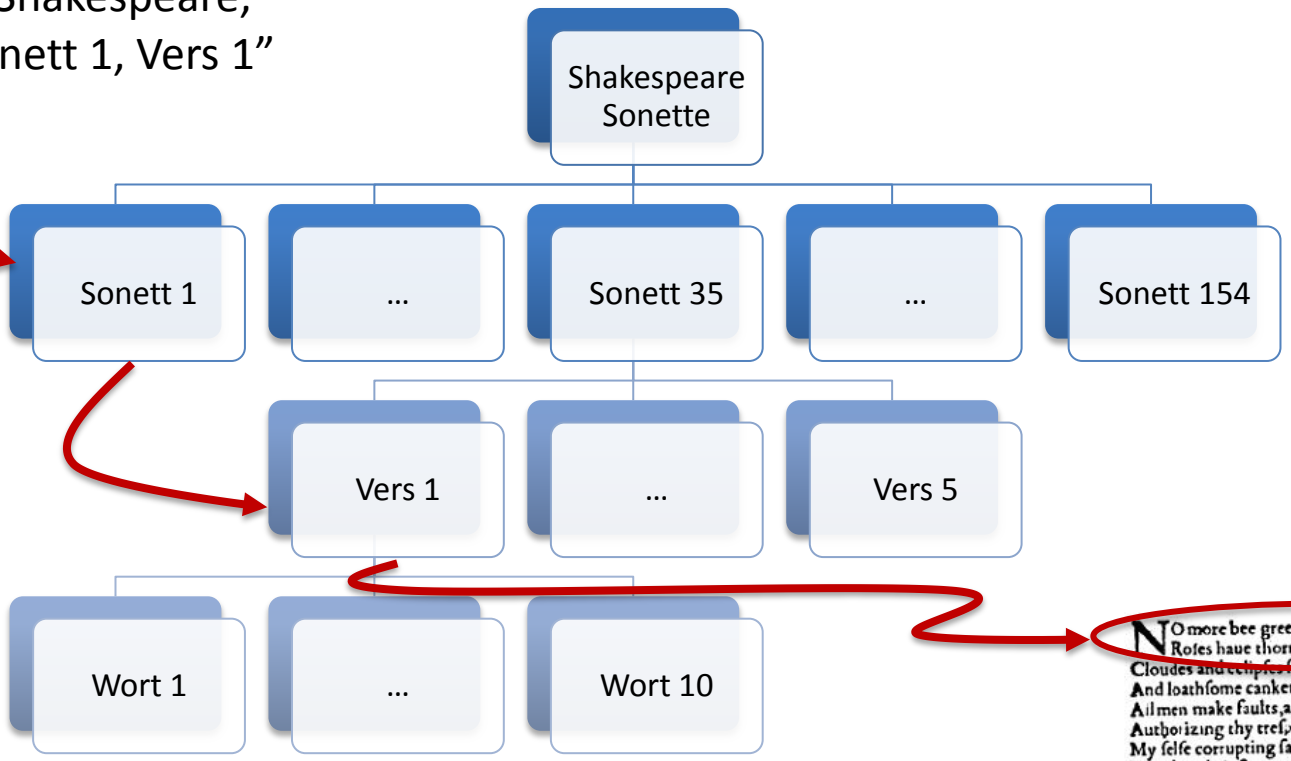
Grundlegendes

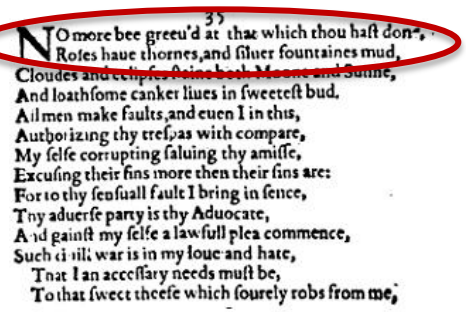
Canonical Text Services (CTS)

- Protokoll für webbasierten Textservice für Zitation
- Entwickelt im Homermultitext Projekt (www.homermultitext.org), Smith et.al.2009
  http://www.homermultitext.org/hmt-docs/specifications/ctsurn/
  http://www.homermultitext.org/hmt-docs/specifications/cts/
- Eindeutige Identifier (**U**nique **R**esource **N**ame, **URN**) spezifizieren Textabschnitte (passages)
- Implementierungen auf Basis von Tripelstore und XML-Datenbank vorhanden, waren aber nicht für unsere Zwecke nutzbar
- Diese (MySQL-basierte) Implementierung ist Teil des ESF-Projektes „Bibliothek der Milliarden Wörter“
- Der Inhalt dieser Präsentation kann auf www.urncts.de live nachvollzogen werden

Kanonische Zitation

"Shakespeare, Sonett 1, Vers 1"




NO more bee greeu'd at that which thou haſt don:
Roſes haue thornes,and ſiluer fountaines mud,
Cloudes and eclipſes ſtaine both Moone and Sunne,
And loathſome canker liues in ſweeteſt bud.
All men make faults,and euen I in this,
Authorizing thy treſpas with compare,
My ſelfe corrupting ſaluing thy amiſſe,
Excuſing their ſins more then their ſins are:
For to thy ſenſuall fault I bring in ſence,
Thy aduerſe party is thy Aduocate,
And gainſt my ſelfe a lawfull plea commence,
Such ciuill war is in my loue and hate,
That I an acceſſary needs muſt be,
To that ſweet theefe which ſourely robs from me,

Kanonische Zitation

Dokumenteneinordung „von draußen“

Shakespeare → Sonnets → english → 1st edition

Textabschnitt innerhalb des Dokumentes

Sonnet 1 → Vers 1

Kombiniert

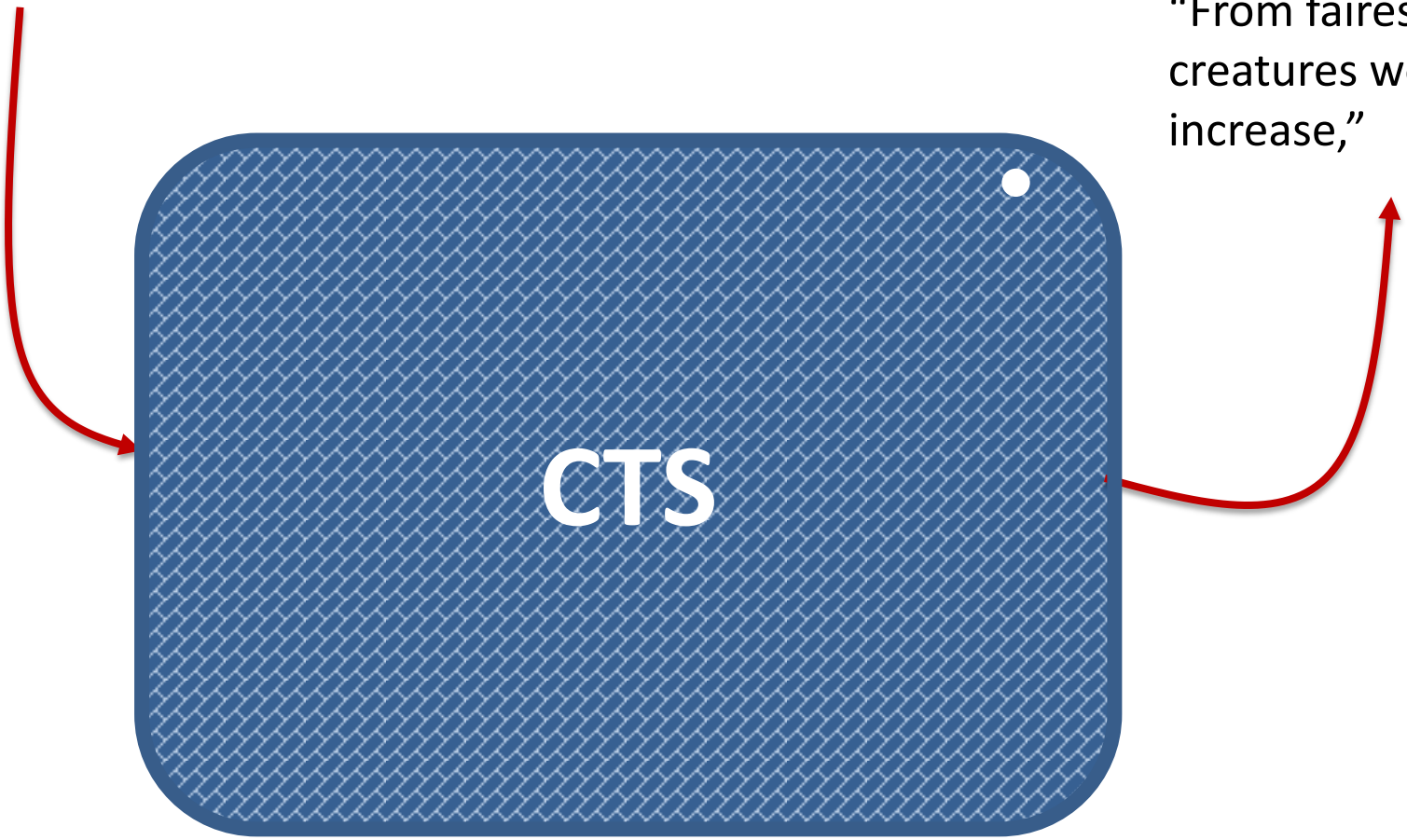
Shakespeare → Sonnets → english → 1st edition → Sonnet 1→ Vers 1

CTS-URN

urn:cts:demo:shakespeare.sonnets.en.1:1.1

Canonical Text Services (CTS)

urn:cts:demo:shakespeare.sonnets.en.1:1.1



"From fairest creatures we desire increase,"

Typen von Textabschnitten

Einfach (Kapitel 12, Vers 2, Lied 1985, Edition, Werk)

urn:cts:demo:shakespeare.sonnets.en.1:1.1

urn:cts:demo:shakespeare.sonnets.en.1:

Spanne

urn:cts:demo:shakespeare.sonnets.en.1:1.1-1.2

urn:cts:demo:shakespeare.sonnets.en.1:1-1.10.6

Teilabschnitt

urn:cts:demo:shakespeare.sonnets.en.1:1.1@creatures

Spanne über Teilabschnitte

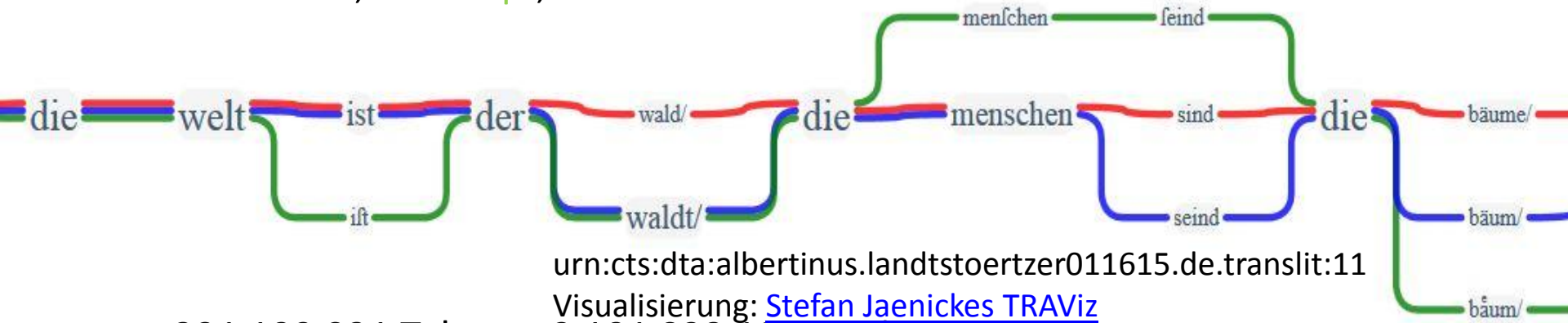
urn:cts:demo:shakespeare.sonnets.en.1:1.1@creatures-1.10@gaudy

urn:cts:demo:shakespeare.sonnets.en.1:1@creatures-1@gaudy[1]

Deutsches Text Archiv in CTS

1712 Werke in je 3 Editionen -> 5136 Editionen

translit, transcript, norm



urn:cts:dta:albertinus.landtstoertzer011615.de.translit:11

Visualisierung: Stefan Jaenickes TRAViz

321 192 031 Tokens , 2 191 023 188 Zeichen

Keine Einteilung in Kapitel oder Ähnliches, „nur“ Sätze

Tokens pro Edition

- Min 75
- Avg 62769,
- Max :588181

Weitere Datensätze

PBC

Parallel Bible Corpus

831 Editionen

247‘292‘629 Tokens

Perseus

3 „Versionen“ Plaintext, XML, Updated

greekLit, latinLit, (farsiLit, pdlrefwk)

407 bzw 1137 Editionen

6‘096‘120 bzw 27‘295‘030 Tokens

Statistiken

Testumgebung:

- Ubuntu-Server

```
ctsuser@ctstestbed:~$ free -h
             total       used       free     shared    buffers     cached
Mem:          1.0G       794M       207M         0B       109M       252M
-/+ buffers/cache:       432M       570M
Swap:         509M        21M       488M
```

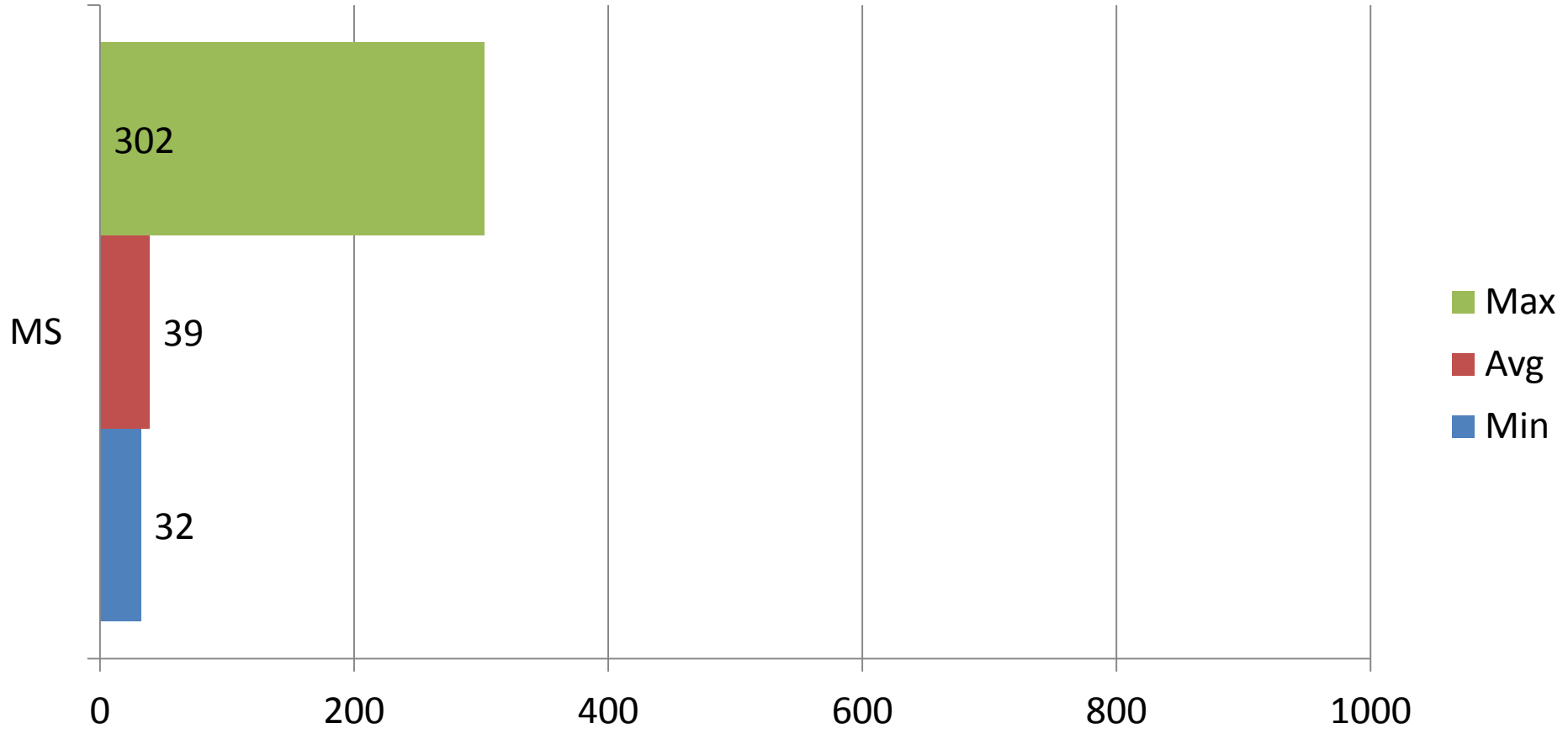
```
ctsuser@ctstestbed:~$ cat /proc/cpuinfo
processor       : 0
vendor_id       : AuthenticAMD
cpu family      : 15
model           : 6
model name      : Common KVM processor
stepping        : 1
microcode       : 0x1000065
cpu MHz         : 2399.998
cache size      : 512 KB
fdiv_bug        : no
hlt_bug         : no
f00f_bug        : no
coma_bug        : no
fpu             : yes
fpu_exception   : yes
cpuid level     : 5
wp              : yes
flags           : fpu de pse tsc msr pae mce cx8 api
bogomips        : 4799.99
clflush size    : 64
cache_alignment : 64
address sizes   : 40 bits physical, 48 bits virtual
```

Testsetup:

- Hole Liste aller Editionen
- Frage je den Abschnitt

  [URN_der_Edition]:1-2 ab

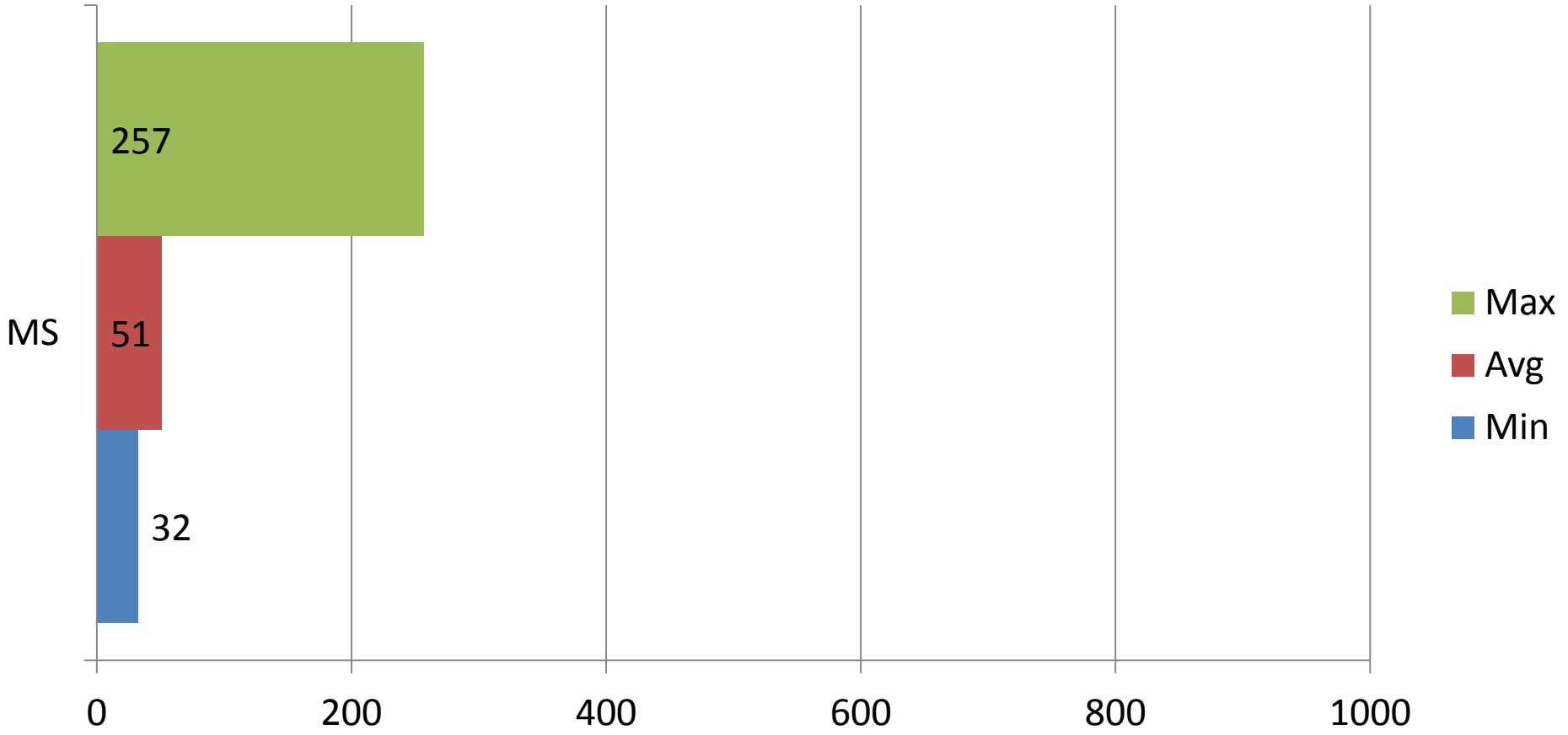
DTA Statistiken



5‘136 Editions . 334‘820‘482 Tokens. 2‘284‘090‘670 Characters.

Durchschnittliche Abschnittslänge: 203 Characters

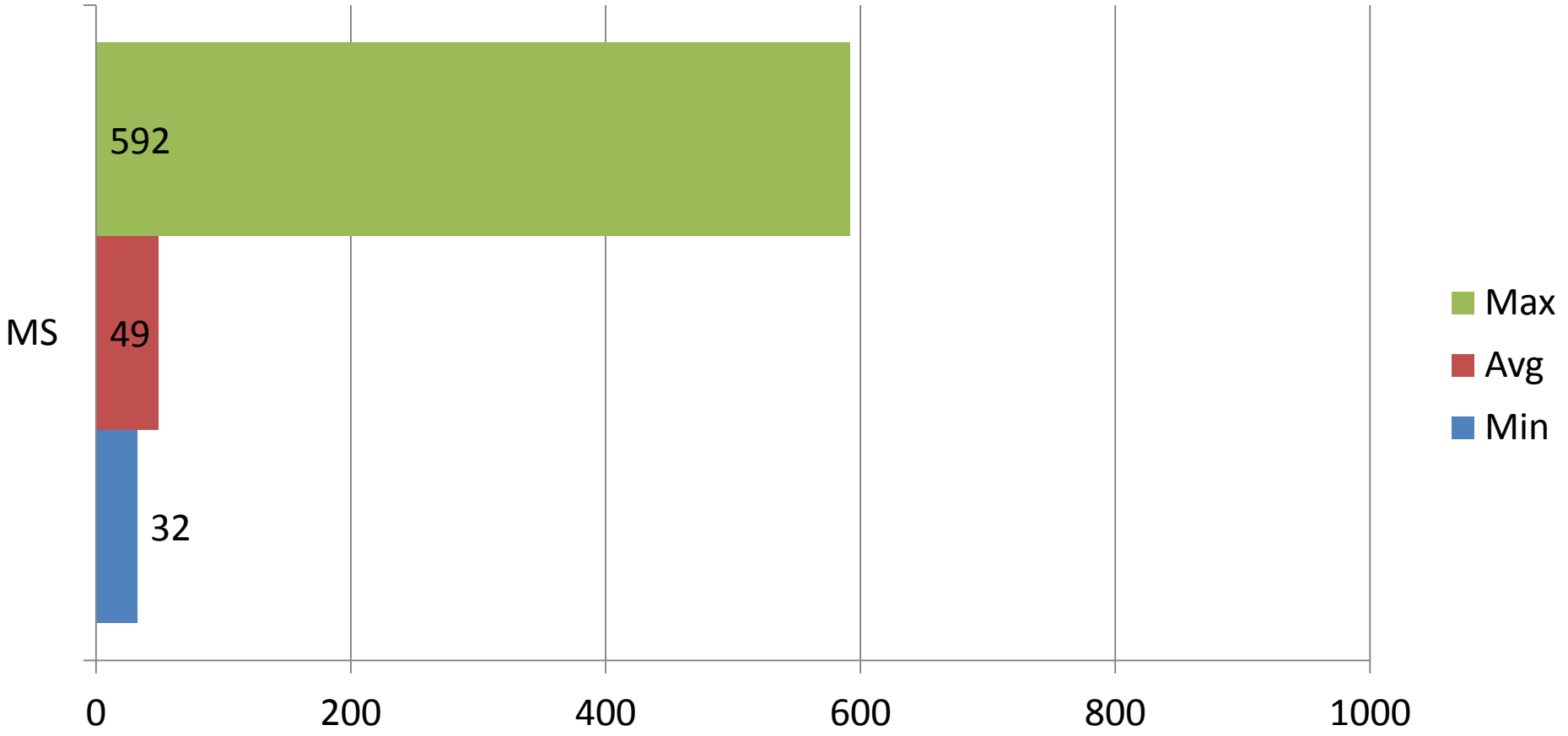
Perseus_xml Statistiken



407 Editions . 6‘096‘120 Tokens. 44‘217‘523 Characters.

Durchschnittliche Abschnittslänge: 26‘838 Characters

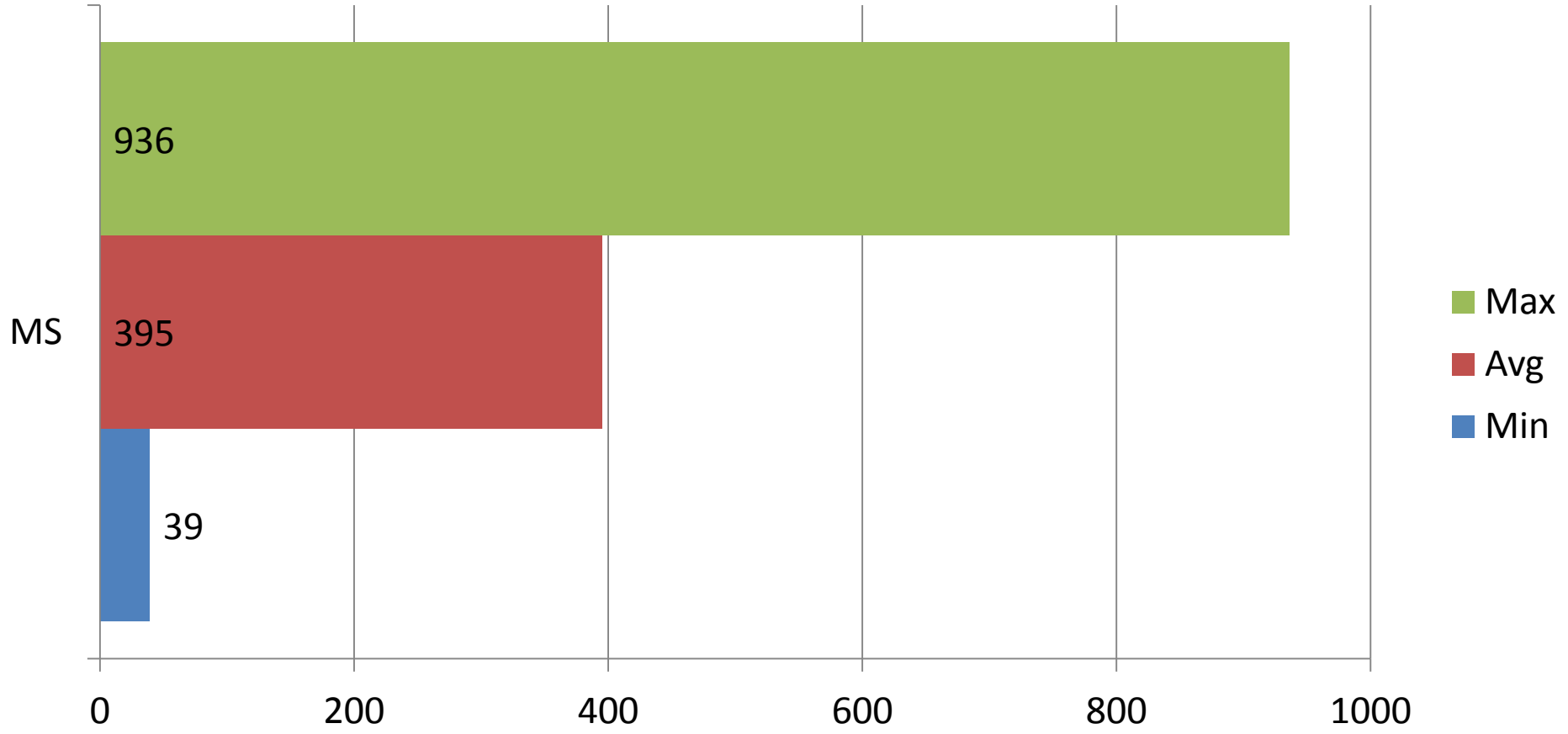
Perseus_new Statistiken



1137 Editions . 27‘295‘030 Tokens. 222‘456‘065 Characters.

Durchschnittliche Abschnittslänge: 28‘930 Characters

PBC Statistiken



831 Editions . 247‘292‘629 Tokens. 1‘357‘136‘926 Characters

Durchschnittliche Abschnittslänge: 352‘634 Characters

1 Milliarde Wörter?

CTS-Instanz	Anzahl Tokens
DTA	334‘820‘482
PBC	247‘292‘629
Perseus_new	27‘295‘030
Perseus_xml	6‘096‘120
Perseus_plain	5‘525‘132
insg	621‘029‘393

Hat zufällig jemand ein paar Wörter übrig???

1 Milliarde künstlicher Wörter

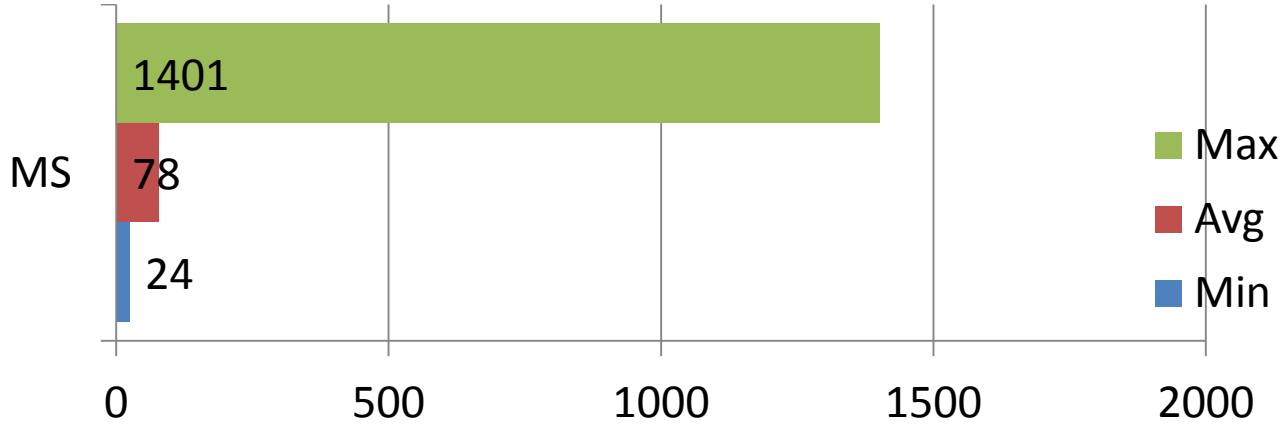
1 CTS mit 100 000 zufällig generierten Editionen

1’281’272’600 Tokens (min. 3/edition, max. 69’118/edition)

Test

1) Alle Editionen auflisten

2) Konsturiere vollen Textabschnitt für jede Edition (kein XML)



Language Support – UTF8

```
<div1 n="1" type="multilang">
  <div2 n="1" xml:lang="de" type="singlelang"> Das Haus auf dem Berg. äöüÄÖÜß. </div2>
  <div2 n="2" xml:lang="ar" type="singlelang"> داس هاوس عوف ماركا بيرغ </div2>
  <div2 n="3" xml:lang="zh(tradition)" type="singlelang"> 達斯豪斯奧夫DEM伯格 </div2>
  <div2 n="4" xml:lang="zh(simple)" type="singlelang"> 达斯豪斯奥夫DEM伯格 </div2>
  <div2 n="5" xml:lang="gu" type="singlelang"> દાસ હાઉસ ઔફ ડેમને બર્ગ </div2>
  <div2 n="6" xml:lang="gr" type="singlelang"> Το σπίτι στο λόφο </div2>
  <div2 n="7" xml:lang="am" type="singlelang"> Տունը բլրի վրա </div2>
  <div2 n="8" xml:lang="bn" type="singlelang"> পাহাড় ঘর </div2>
  <div2 n="9" xml:lang="bg" type="singlelang"> Къщата на хълма </div2>
  <div2 n="10" xml:lang="he" type="singlelang"> הבית על הגבעה </div2>
  <div2 n="11" xml:lang="in" type="singlelang"> पहाड़ी पर घर </div2>
  <div2 n="12" xml:lang="jp" type="singlelang"> 丘の上の家 </div2>
  <div2 n="13" xml:lang="ji" type="singlelang"> די הויז אויף די בערגל </div2>
  <div2 n="14" xml:lang="Khmer" type="singlelang"> ផ្ទះនៅលើភ្នំនេះ </div2>
  <div2 n="15" xml:lang="kr" type="singlelang"> 언덕에 집 </div2>
  <div2 n="16" xml:lang="lo" type="singlelang"> ເຮືອນກ່ຽວກັບຫນ່ວຍ </div2>
  <div2 n="17" xml:lang="mr" type="singlelang"> टेकडी वर घर </div2>
  <div2 n="18" xml:lang="ne" type="singlelang"> पहाडी मा घर </div2>
  <div2 n="19" xml:lang="Persisch" type="singlelang"> خانه در تپه </div2>
  <div2 n="20" xml:lang="Punjabi" type="singlelang"> ਪਹਾੜੀ 'ਤੇ ਘਰ ਨੂੰ </div2>
  <div2 n="21" xml:lang="ru" type="singlelang"> Дом на холме </div2>
  <div2 n="22" xml:lang="tl" type="singlelang"> மலை மீது வீட்டில் </div2>
  <div2 n="23" xml:lang="te" type="singlelang"> కోండ మీద హౌస్ </div2>
  <div2 n="24" xml:lang="th" type="singlelang"> บ้านอยู่บนเนินเขา </div2>
  <div2 n="25" xml:lang="ur" type="singlelang"> پہاڑی پر گھر </div2>
  <div2 n="26" xml:lang="vn" type="singlelang"> Ngôi nhà trên đồi </div2>
</div1>
```

Future Work

URNs als IDs

ähnlich CITE-Architektur http://www.homermultitext.org/hmt-docs/cite/

Nutzen der URNs als IDs für andere Projekte

Standardisierung

CTS - Ausgabe als Standardformat für GUI

Unabhängigkeit von eigentlicher Textstruktur

Eigenschaften der URNs

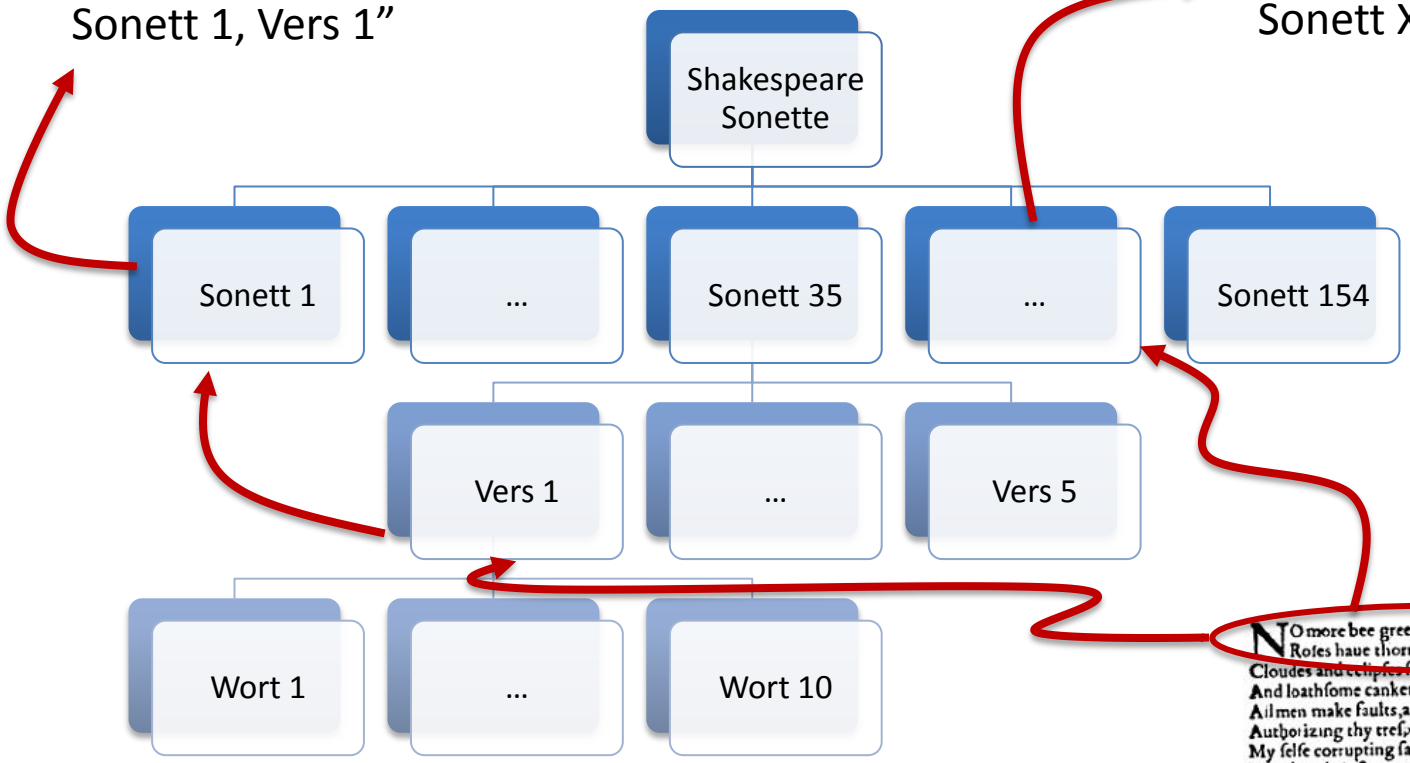
Sprachkürzel in URN für Übersetzung nutzen

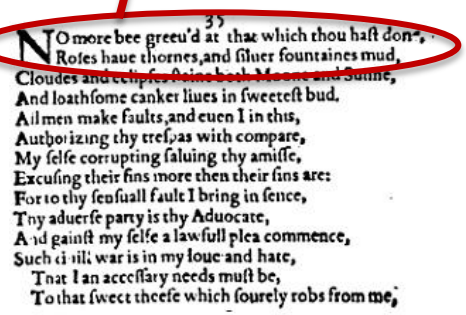
urn:cts:demo:shakespeare.sonnets.en.1:1.1-1.2

Future Work – Volltextsuche

"Shakespeare, Sonett 1, Vers 1"

"Shakespeare, Sonett X, …"





35

NO more bee greeu'd at that which thou haſt don:
Roſes haue thornes,and ſiluer fountaines mud,
Cloudes and eclipſes ſtaine both Moone and Sunne,
And loathſome canker liues in ſweeteſt bud.
All men make faults,and euen I in this,
Authorizing thy treſpas with compare,
My ſelfe corrupting ſaluing thy amiſſe,
Excuſing their ſins more then their ſins are:
For to thy ſenſuall fault I bring in ſence,
Thy aduerſe party is thy Aduocate,
And gainſt my ſelfe a lawfull plea commence,
Such ciuill war is in my loue and hate,
That I an acceſſary needs muſt be,
To that ſweet theefe which ſourely robs from me.

Future Work – Volltextsuche

- „Rückrichtung“
- Text ReUse, Zitatsuche, Duplikatsuche, Plagiatsuche
- Analyse beliebig großer Textabschnitte (Bigramm, Trigramm,…) ohne zusätzliche Datenaufbereitung
- Textspannen über mehrere Texteinheiten (Sätze, Kapitel) schwierig

Vielen Dank

```
<cts:GetPassage>
 - <cts:request>
      <cts:requestName>GetPassage</cts:requestName>
      <cts:urn>urn:cts:presentation:last.slide.en:</cts:urn>
   </cts:request>
 - <cts:reply>
    - <cts:passage>
         <div1 n="1" type="line">Thank you for your attention.</div1>
         <div1 n="2" type="line">Do you have any questions?</div1>
      </cts:passage>
   </cts:reply>
</cts:GetPassage>
```