

DWDS-DDC-Mini-Howto
Version 1.74
Kai Zimmer <zimmer@bbaw.de>

Inhalt

- 1) notwendige Dateien
- 2) Indexierung
- 3) Einbinden des Corpus
- 4) Neustart des ConcordDaemon
- 5) Logfile überprüfen

Kommandozeilenbefehle sind *kursiv* geschrieben.

1) Notwendige Dateien

Um ein Corpus zu indexieren benötigt man

- 1) eine Liste der zu indexierenden Dateien (z.B. Corpus.con) – die ist einfach zu erstellen, z.B. auf der Kommandozeile:
find . -name „.txt“ > Corpus.con*
Der Punkt gibt an ab welchem Verzeichnis find suchen soll, in diesem Fall in dem aktuellen Verzeichnis
- 2) eine Options-Datei (z.B. Corpus.opt), die DDC über den Typ der Daten (Text, HTML, XML) und andere Parameter informiert.
Ein paar Beispiele:
 - a. „klassische“ DWDS-XML-TEI Dokumente:

```
German
IndexType DWDS_Index
UseParagraphTagToDivide
EmptyLineIsNotSentenceDelim
DontUseIndentation
UserMaxTokenCountInOnePeriod 10000000
IndexMorphPatterns
OutputBibliographyOfHits
IndexPunctuation
Bibl 1 textClass /TEI.2/teiHeader/profileDesc/textClass/keywords/term
Bibl 0 author
/TEI.2/teiHeader/fileDesc/sourceDesc[@id="orig"]/biblFull/titleStmt/author
Bibl 0 date
/TEI.2/teiHeader/fileDesc/sourceDesc[@id="orig"]/biblFull/publicationStmt/date[
@id="first"]
Bibl 0 orig /TEI.2/teiHeader/fileDesc/sourceDesc[@id="orig"]/bibl
Bibl 0 scan /TEI.2/teiHeader/fileDesc/sourceDesc[@id="scan"]/bibl
Bibl 0 page
/TEI.2/teiHeader/fileDesc/sourceDesc[@id="orig"]/seriesStmt/idno[@type="page
"]
Bibl 0 body /TEI.2/text/body
```

b. "frei" indexierte XML Dokumente:

German
IndexType Free_Index
Indices [Token w normal]; [Lemma l normal]; [Pos p normal]
HitBorders [s:sentence:default]; [c:clause]
OutputBibliographyOfHits
IndexChunks
Bibl string 1 textClass /TEI.2/teiHeader/profileDesc/textClass/keywords/term
Bibl string 0 author
/TEI.2/teiHeader/fileDesc/sourceDesc[@id="orig"]/biblFull/titleStmt/author
Bibl integer 0 date
/TEI.2/teiHeader/fileDesc/sourceDesc[@id="orig"]/biblFull/publicationStmt/date[
@id="first"]
Bibl string 0 orig /TEI.2/teiHeader/fileDesc/sourceDesc[@id="orig"]/bibl
Bibl string 0 scan /TEI.2/teiHeader/fileDesc/sourceDesc[@id="scan"]/bibl
Bibl integer 0 page
/TEI.2/teiHeader/fileDesc/sourceDesc[@id="orig"]/seriesStmt/idno[@type="page
"]
Bibl string 0 body /TEI.2/text/body

c. XML Dokumente mit morphologischer Annotation:

German
IndexType MorphXML_Index
UserMaxTokenCountInOnePeriod 10000000
OutputBibliographyOfHits
Bibl string 0 body /ddc_document/text
Bibl string 0 author /ddc_document/header/author
Bibl integer 0 date /ddc_document/header/date
Bibl string 0 title /ddc_document/header/title
Bibl string 0 scan /ddc_document/header/bibl
Bibl string 1 textClass /ddc_document/header/textClass
Bibl integer 0 volume /ddc_document/header/idno[@type="volume"]

d. TEI-XML Dokumente mit Annotierung und Relevanz

German
IndexType Free_Index
Indices [Token w normal]; [Lemma l normal]; [Pos p normal]; [Thes t normal]
HitBorders [s:sentence:default]; [c:clause]
OutputBibliographyOfHits
TextHighlighting _&; &_; _&; &_
ResumeOnIndexErrors
CaseInsensitive
TfIdfRank 1.0
NearRank 1.0
PositionRank 1.0
Bibl string 1 textClass /TEI.2/teiHeader/profileDesc/textClass/keywords/term
Bibl string 1 title /TEI.2/teiHeader/fileDesc/titleStmt/title[@type="main"]
Bibl string 1 biblfilename /TEI.2/teiHeader/fileDesc/filename

Bibl integer 1 date
 /TEI.2/teiHeader/fileDesc/sourceDesc[@id="orig"]/biblFull/publicationStmt/date[
 @id="first"]
 Bibl string 1 author
 /TEI.2/teiHeader/fileDesc/sourceDesc[@id="orig"]/biblFull/titleStmt/author
 Bibl integer 1 PageRank
 /TEI.2/teiHeader/fileDesc/sourceDesc[@id="orig"]/relevance
 TextArea filename /TEI.2/teiHeader/fileDesc/sourceDesc[@id="orig"]/filename
 TextArea txttitle
 /TEI.2/teiHeader/fileDesc/sourceDesc[@id="orig"]/biblFull/titleStmt/title
 Textarea body /TEI.2/text/body

e. TEI-XML Dokumente mit Ausgabe von Annotationen und Relevanz:

German
 IndexType Free_Index
 Indices [Token w normal]; [Lemma l normal storage]; [Pos p normal storage];
 [Token2 v normal storage]; [Spk s normal storage]; [Mysql m normal storage]
 HitBorders [s:sentence]; [c:clause]; [u:utterance:default]
 OutputBibliographyOfHits
 ResumeOnIndexErrors
 TfIdfRank 1.0
 NearRank 1.0
 PositionRank 1.0
 Bibl string 1 textClass /TEI.2/teiHeader/profileDesc/textClass/keywords/term
 Bibl string 1 title /TEI.2/teiHeader/fileDesc/titleStmt/title[@id="main"]
 Bibl integer 1 date
 /TEI.2/teiHeader/fileDesc/sourceDesc[@id="orig"]/biblFull/publicationStmt/date[
 @id="first"]
 Bibl integer 1 page /TEI.2/text/body/pb/@n
 Bibl string 0 author
 /TEI.2/teiHeader/fileDesc/sourceDesc[@id="orig"]/biblFull/titleStmt/author
 TextArea body /TEI.2/text/body
 IndicesToShow 1 2 3 5 6
 InterpDelimiter ^

f. HTML-Dokumente (wie z.B. die DWDS-Homepage):

IndexType DWDS_Index
 German
 LocalPathPrefix pages/
 InternetPathPrefix www.dwds.de/pages/
 QueryOnlyFiles
 CaseInsensitive

2) Indexierung

Falls die unter 1) aufgeführten notwendigen Dateien vorliegen ist die eigentliche Indexierung denkbar einfach:

- die Umgebungsvariable RML sollte auf das RML-Verzeichnis zeigen, z.B.

```
export RML=/opt/ddc
```

- das Corpus indexieren mit
ConcordIndex Corpus.con

3) Einbinden des Corpus

Dazu sind folgende Schritte nötig:

- a) die Datei `$RML/ddc_local_corpora.cfg` anpassen:

```
// CorporaName IP PORT LocalName
dwds0 192.168.1.83 50005 /home/sokirko/M5/01/01.con
dwds1 192.168.1.83 50006 /home/sokirko/M5/02/02.con
dwds2 192.168.1.83 50007 /home/sokirko/M5/03/03.con
dwds3 192.168.1.83 50008 /home/sokirko/M5/04/04.con
dwds4 192.168.1.83 50009 /home/sokirko/M5/05/05.con
dwds5 192.168.1.83 50010 /home/sokirko/M5/06/06.con
dwds_hp 192.168.1.83 50100 /home/gneumann/texte/dwds_hp/pages/dwds_hp.con
```

In diesem Fall wurde am Ende der Datei das Corpus „dwds_hp“ hinzugefügt. Die IP-Adresse ist die des DDC-Servers, die Portnummer muss neu vergeben werden (traditionell ein TCP-Port > 50001) und muss einmalig sein (kurz: jedes neue Corpus braucht einen eigenen, neuen Port), dahinter steht der absolute Pfad zur Corpus.con Datei.

- b) die Datei `$RML/ddc_server.cfg` anpassen:

```
// CorporaName IP PORT
server 192.168.1.83 50011
dwds0 192.168.1.83 50005
dwds1 192.168.1.83 50006
dwds2 192.168.1.83 50007
dwds3 192.168.1.83 50008
dwds4 192.168.1.83 50009
dwds5 192.168.1.83 50010
dwds_hp 192.168.1.83 50100
```

Anpassungen wie oben in der `ddc_local_corpora.cfg` – aber ohne Angabe des Pfades zur Corpus.con Datei. Der Sinn dieser Datei liegt in der Zusammenfassung verschiedener DDC-Server (z.B. einem Cluster) in einem weiteren Frontend-Server, muss aber in jedem Fall ausgefüllt werden. Das Schlüsselwort „server“ muss zwingend in der ersten Zeile stehen.

- c) die Datei `$RML/ddc_xml_server.cfg`:

```
server 192.168.1.83 50011
```

Diese Datei ist speziell für die Abfrage mit `ddc_xml` gedacht. Normalerweise muss hier nur einmal der Server-Eintrag vorgenommen werden, über den die einzelnen Corpora dann an `ddc_xml` herausgegeben werden.

4) Neustart des ConcordDaemon:

Abschliessend muss der ConcordDaemon neu gestartet werden :

```
ConcordDaemon stop  
ConcordDaemon start
```

Achtung, bei selbstkompilierten Installationen (nicht aus dem RPM) kann der Daemon auch „ConcordDaemontst“ heissen.

Der komplette Startvorgang kann (abhängig von der verwendeten Corpusgrösse) mehrere Minuten dauern. Dann ist es hilfreich das Logfile beim Wachstum zu beobachten (s.u.).

5) Logfile überprüfen:

Ob der Daemon richtig gestartet wurde kann man im Logfile überprüfen, unter \$RML/Logs/concord/Datum.log . Datum ist dabei durch das aktuelle Datum zu ersetzen:

```
tail -f 05January2005.log
```

für die obige Konfiguration sieht das etwa so aus:

```
16:11:30 > Entering ConcordDaemon  
16:11:30 > SocketInitialize  
16:11:30 > InitLemmatizers  
16:11:34 > LoadCorpora  
16:11:34 > Start loading corpora from /home/sokirko/RML/Bin/ddc_local_corpora.cfg  
16:11:34 > Found 7 Hosts  
16:11:34 > Loading corpora index for /home/sokirko/M5/01/01.con  
16:11:58 > init graphan  
16:11:58 > ok  
16:11:58 > Loading corpora index for /home/sokirko/M5/02/02.con  
16:12:24 > init graphan  
16:12:24 > ok  
16:12:24 > Loading corpora index for /home/sokirko/M5/03/03.con  
16:12:53 > init graphan  
16:12:53 > ok  
16:12:53 > Loading corpora index for /home/sokirko/M5/04/04.con  
16:13:18 > init graphan  
16:13:18 > ok  
16:13:18 > Loading corpora index for /home/sokirko/M5/05/05.con  
16:13:38 > init graphan  
16:13:38 > ok  
16:13:38 > Loading corpora index for /home/sokirko/M5/06/06.con  
16:14:00 > init graphan  
16:14:00 > ok  
16:14:00 > Loading corpora index for /home/gneumann/texte/dwds_hp/pages/dwds_hp.con  
16:14:01 > init graphan  
16:14:01 > ok
```

```
16:14:01 > Create Listener dwds0 (192.168.1.83:50005)
16:14:01 > Create Listener dwds1 (192.168.1.83:50006)
16:14:01 > waiting for accept
16:14:01 > Create Listener dwds2 (192.168.1.83:50007)
16:14:01 > waiting for accept
16:14:01 > Create Listener dwds3 (192.168.1.83:50008)
16:14:01 > waiting for accept
16:14:01 > Create Listener dwds4 (192.168.1.83:50009)
16:14:01 > waiting for accept
16:14:01 > waiting for accept
16:14:01 > Create Listener dwds5 (192.168.1.83:50010)
16:14:01 > Create Listener dwds_hp (192.168.1.83:50100)
16:14:01 > waiting for accept
16:14:01 > LoadServer
16:14:01 > waiting for accept
16:14:01 > Create Listener server (192.168.1.83:50011)
16:14:01 > waiting for accept
```

Wichtig sind die “waiting for accept”-Einträge. Fehlen diese ist der Daemon entweder noch nicht vollständig gestartet (z.B. bei grossen Corpora) oder es liegt noch irgendein Konfigurationsfehler vor.