Automatisk konvertering til TEI-format

Introduktion

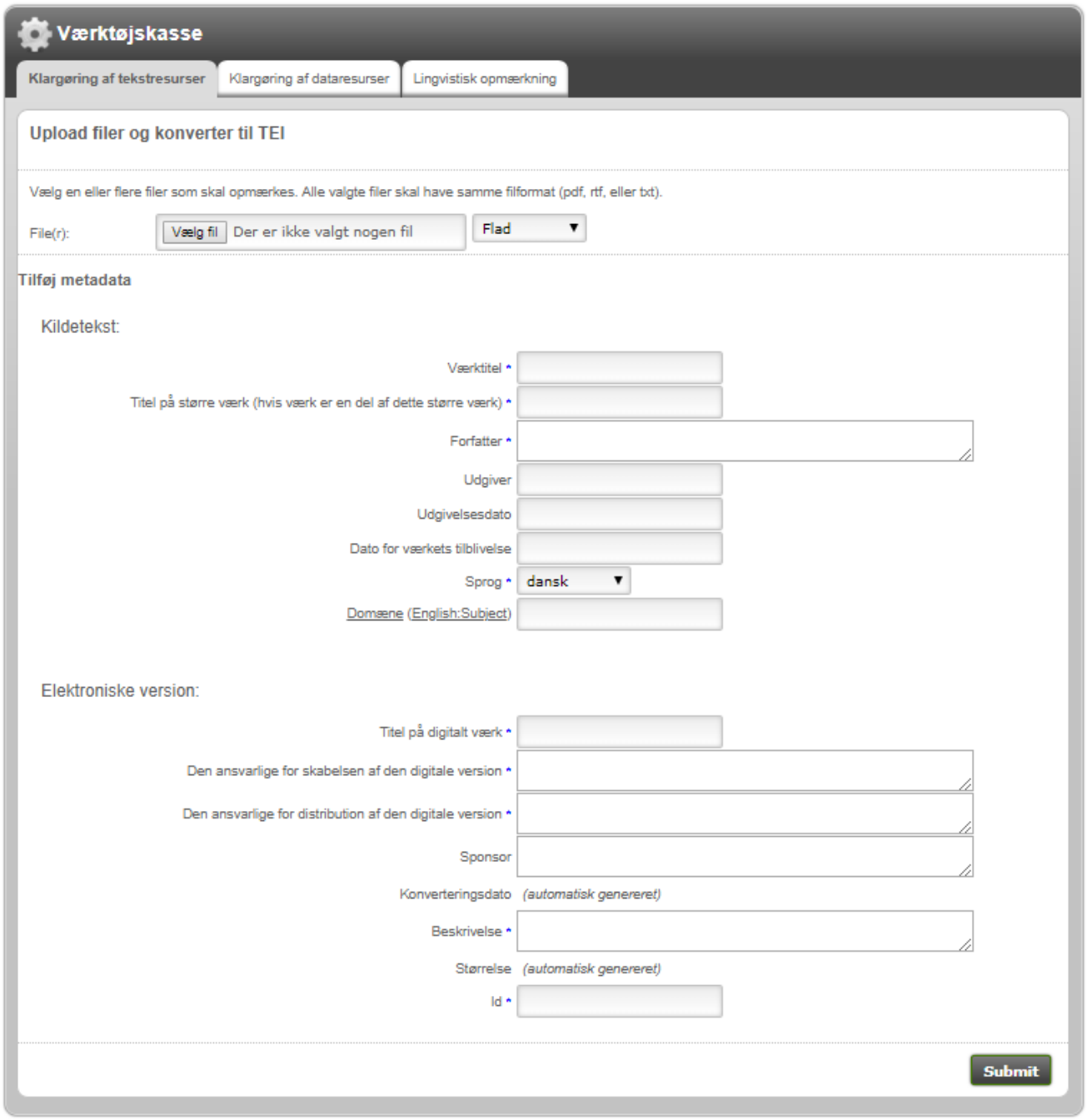
I denne ”tutorial” gennemgår vi hvordan du automatisk kan konvertere dine tekstfiler i txt-format eller RTF-format til xml-formatet TEIP5. Konverteringen følger det TEI-skema der blev defineret i DK-CLARIN-projektet 2008-2011. (Se dokumentation her: https://info.clarin.dk/clarin-dk-infrastrukturen/tutorials/text-format.pdf og https://info.clarin.dk/clarin-dk-infrastrukturen/tutorials/text-header.pdf)

I xml-filerne er al layout fjernet og teksten gemt med et ord på hver linje, linjerne (dvs. ordene) nummererede og afsnittene markerede. Pga. denne strukturering er det muligt at lave automatisk analyse af teksten; men ikke længere muligt at ”læse” teksten som før. Tekstfilen er blevet et forskningsobjekt.

Efter klargøringen skal du selv verificere kvaliteten af den automatiske proces, fx om det umiddelbart ser ud til at konverteringen ikke har ødelagt ordene.

Trin 1

Gå til NLP Tools (Værktøjskassen) – Klargøring af tekstresurser https://clarin.dk/clarindk/toolchains-upload.jsp, vælg den fil du vil have konverteret til TEI-formatet, og vælg filformat txt (flad) eller RTF.



Trin 2

Derefter skal du udfylde de mest basale metadata:

* betyder at metadataelementet er obligatorisk

Metadata	Betydning	Eksempel
*Værktitel	Titlen på kildeteksten. Evt. e-navn hvis resursen er skabt elektronisk	*Foraarets død*
Titel på større værk (hvis værk er en del af dette større værk)	Titlen på den del af værket som filen dækker, evt. samme som Værktitel hvis filen ikke hører under et samlet hele	*Kongens Fald*
*Forfatter	Forfatter af kildeteksten, kan evt. være anonym eller ukendt. Hvis der er flere forfattere, skrives en på hver linje	*Johannes V. Jensen*
Udgiver	Navnet på den organisation der er ansvarlig for udgivelsen eller distribution af værket. Evt. ens egen organisation.	*Gyldendal*
Udgivelsesdato	Format yyyy-mm-dd eller yyyy hvis den præcise dato ikke er kendt	*1900*
Dato for værkets tilblivelse	Bruges hvis man ønsker at angive at kildeteksten er skabt på et andet tidspunkt end udgivelsesdato. Format yyyy-mm-dd eller yyyy hvis den præcise dato ikke er kendt	*1898*
*Sprog	Sproget i den udgivelse der her uploades udtrykt i ISO639-1 Se: https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes	da
Domæne	Emne eller domæne, gerne med henvisning til et anerkendt klassifikationssystem som DK5 eller DDC der bruges af biblioteksvæsner. Emneangivelse er oftest relevant når det drejer sig om fagtekst	*Dansk skønlitteratur (DK5-86)*
*Titel på digitalt værk	Titlen på den elektroniske version. Kan være identisk med Værktitel hvis resursen er skabt elektronisk eller man ønsker at resursen skal være synlig i CLARIN under Værktitlen	*Kongens Fald - TEIP5-version*
*Personansvarlig for digital version	Person der er ansvarlig for *skabelse* af den elektroniske version. Hvis der er flere, skrives en på hver linje.	*Peter Petersen*
*Ansvarlig for distribution af digital version	Person eller organisation der er ansvarlig for *distribution* af den elektroniske version. Hvis der er flere, skrives en på hver linje.	*NorS, Københavns Universitet*
Sponsor	Her kan angives projekt, institution, bevilling hvis man skønner at det er relevant.	*CLARIN-DK-projektet*
Konverteringsdato	Dato for digital version af værket i formatet yyyy-mm-dd eller yyyy. Udregnes og tilskrives automatisk.	*2011*
*Beskrivelse	Kommentarer om hvilke forskningssammenhænge teksten kan bruges i, eller andre oplysninger det er relevant at huske eller at dele med andre	*Kongens fald i xml TEIP5 format er en del af JVJ online tilvejebragt gennem DK-CLARIN-projektet til brug i tekstanalyse*
Størrelse	Værkets størrelse. Udregnes og tilskrives automatisk.	*17.563 words*
Id	Tilskrives automatisk.	*2100106490*

Når du har klikket *Submit* genereres tekstfilen i TEIP5-formatet. Der genereres desuden en segmenteret fil som er et mellemtrin og en index-fil der beskriver den tekniske tilblivelsesproces.

Trin 3

Du kan se indholdet i en browser og du kan redigere filen med en tekst editor som f.eks. *Notepad* eller med en XML-editor som f.eks. 'Oxygen' eller Microsofts 'Visual Studio'. Hvis du åbner filen i fx Word bliver indholdet fortolket, dvs. at alle xml-koderne er usynlige.

Metadataene vil se således ud:

```
<teiHeader type="text">
   <fileDesc>
      <titleStmt>
         <title>Kongens Fald - TEIP5-version</title>
         <sponsor>CLARIN-DK-projektet</sponsor>
         <respStmt>
            <resp>a_annotation</resp>
               <name><name>Peter Petersen</name>
                     <note type="method">flat2cbf</note>
                     <date when="2016-06-28"/></name>
         </respStmt>
      </titleStmt>
   <extent><num type="words">1100</num></extent>
      <publicationStmt>
         <distributor>Nordisk Forskningsinstitut, Københavns Universitet</distributor>
         <idno type="ctb">20160609-830-step3</idno>
         <availability status="free">
            <ab type="public"/>
         </availability>
   </publicationStmt>
   <notesStmt>
      <note>Kongens fald i xml TEIP5 format er en del af JVJ online tilvejebragt gennem DK-CLARIN-projektet til brug i
tekstanalyse</note>
</notesStmt>
   <sourceDesc>
      <biblStruct>
         <analytic>
            <title xml:lang="da">Foraarets død</title>
            <author><name>Johannes V. Jensen</name></author>
         </analytic>
         <monogr>
            <title>Kongens Fald</title>
            <imprint>
            <publisher n="n/a">Gyldendal</publisher>
           <date when="1900"/>
          </imprint>
      </monogr>
   </biblStruct>
</sourceDesc>

osv. ….

</teiHeader>
```

Og tekstdelen således:

```
                  <body>
                  <p n="1">
```

```
                <w xml:id="i1.1">1</w>
</p>
<p n="2">
                <w xml:id="i2.1">Vejen</w>
                <c xml:id="i2.2" type="s"/>
                <w xml:id="i2.3">bøjede</w>
                <c xml:id="i2.4" type="s"/>
                <w xml:id="i2.5">tilvenstre</w>
                <c xml:id="i2.6" type="s"/>
                <w xml:id="i2.7">over</w>
                <c xml:id="i2.8" type="s"/>
                <w xml:id="i2.9">en</w>
                <c xml:id="i2.10" type="s"/>
                <w xml:id="i2.11">Bro</w>
                osv. ….
</body>
```

Hvis det er klart at processen ikke er lykkedes, kan man gå ind i en af filerne rtf-filen eller txt-filen, rette de ting der er gået galt og uploade filen igen startende ved Trin 1. Metadata skal desværre indtastes igen; men de gamle metadataværdier kommer frem som forslag når man taster det første bogstave.

Det er ikke tilrådeligt at rette i tekstdelen i xml-filen da hvert enkelt ord er unikt nummereret hvilket er påkrævet i den videre lingvistiske processering. Metadataene kan man dog rolig rette i.