UPPSALA
UNIVERSITET

# Trädgårdens struktur

***Beáta B. Megyesi och Bengt Dahlqvist***
Institutionen för lingvistik och filologi
Uppsala universitet

UPPSALA UNIVERSITET

# Trädgårdens skapande

Mål:

Konvertera SUC2.0 till XCES med standoff annotering enligt ANC1.0.

Principen är att all information som finns i hela SUC2.0 skall bevaras.

UPPSALA UNIVERSITET

# SUC2.0 till XCES med standoff annotation

–filename.utf8

ursprunglig SUC-fil representerat som UTF-8

–filename-header.xml

korpusheader för beskrivning av innehållet

–filename.txt

en token per rad, ej meningssegmentering

–filename-logical.xml

info om text, body, byline, paragraf, list, label, item, line group

UPPSALA UNIVERSITET

# SUC2.0 till XCES med standoff annotation

–filename-s.xml

meningssegmentering

–filename-suctag.xml

token, typ av token, POS, morf och lemma

–filename-name.xml

named entities är uppmärkta

UPPSALA UNIVERSITET

# SUC2.0 till XCES med standoff annotation

–filename-abbr.xml

förkortningar är uppmärkta

–filename-foreign.xml

utländska ord är uppmärkta

–filename-num.xml

sifferuttryck är uppmärkta

UPPSALA UNIVERSITET

# SUC2.0 till XCES med standoff annotation

–filename-meta.xml

Uttryck som kan bestå av en eller flera token och

uttrycker metalingvistisk information

Dit hör: <distinct>, riktat tal/skrift <q>, citat <quote>,

referens <ref>, <mentioned>

UPPSALA
UNIVERSITET

# Exempel: filename.txt

Avspänningen

mellan

stormaktsblocken

och

nedrustningssträvanden

i

Europa

har

inte

mycket

motsvarighet

i

Mellanöstern

.

UPPSALA
UNIVERSITET

# Exempel: FILENAME_s.xml

```
<struct type="s" from="1" to="3">
   <feat name="id" value="aa01a-001" />
</struct>
<struct type="s" from="4" to="9">
   <feat name="id" value="aa01a-002" />
</struct>
...
```

UPPSALA UNIVERSITET

# Exempel: FILENAME_suctag.xml

```
<struct type="tok" from="1" to="1">
  <feat name="base" value="Smygrustning" />
  <feat name="toktype" value="w" />
  <feat name="pos" value="NN" />
  <feat name="morph" value="UTR SIN IND NOM" />
  <feat name="lemma" value="smygrustning" />
</struct>
<struct type="tok" from="2" to="2">
  <feat name="base" value="av" />
  <feat name="toktype" value="w" />
  <feat name="pos" value="PP" />
  <feat name="lemma" value="av" />
</struct>
```

UPPSALA
UNIVERSITET

# Exempel: FILENAME_name.xml

```
<struct type="name" from="5" to="6">
  <feat name="type" value="person" />
</struct>
<struct type="name" from="7" to="7">
  <feat name="type" value="inst" />
</struct>
<struct type="name" from="10" to="10">
  <feat name="type" value="place" />
</struct>
```

UPPSALA UNIVERSITET

# Exempel: FILENAME_foreign.xml

```
<struct type="foreign" from="164" to="164">
  <feat name="lang" value="en" />
</struct>
```

UPPSALA UNIVERSITET

# Exempel: FILENAME_abbr.xml

```
<struct type="abbr" from="7" to="7">
</struct>
<struct type="abbr" from="56" to="56">
</struct>
<struct type="abbr" from="1130" to="1130">
</struct>
<struct type="abbr" from="1859" to="1859">
</struct>
```

UPPSALA
UNIVERSITET

# Exempel: FILENAME_num.xml

```
<struct type="num" from="398" to="398">
</struct>
<struct type="num" from="425" to="425">
</struct>
<struct type="num" from="430" to="430">
</struct>
<struct type="num" from="455" to="455">
</struct>
```

UPPSALA
UNIVERSITET

# Exempel: FILENAME-meta.xml

```
<struct type="distinct" from="573" to="573">
   <feat name="type" value="formula" />
</struct>
<struct type="ref" from="697" to="699">
</struct>
<struct type="quote" from="1411" to="1448">
</struct>
```

UPPSALA UNIVERSITET

# Problem

Dokumentation och version av SUC2.0

– Taggar som finns i SUC2.0 men som inte finns i manualen

– Taggar som inte finns i SUC2.0 men finns i manualen

UPPSALA UNIVERSITET

# Att göra

- Diskutera lämpligt format (XCES med standoff ren eller med LAF à la ANC1 eller ANC2?)
- Corpus header
- Felsökning och felhantering
- DTD (inkl. character entities)
- Dokumentation