Integration of Finnish web services in WebLicht

Presentation in Freudenstadt 2010-10-16 by Jussi Piitulainen

Who we are

FIN-CLARIN

University of Helsinki

The Language Bank of Finland

CSC - The Center for Science

What we do

Corpora and tools

Finnish wordnet,
tree bank, parse bank

legal agreements,
access rights

What we have

Finnish and Finland-Swedish corpora (newspapers)

Swahili corpus

small Uralic languages

tools, notably HFST implemented morphological analysis

Access to the Language Bank requires authorization and authentication

Finnish WebLicht integration exercise so far

Installed

A version of omorfi, an analyzer of Finnish morphology, at University of Helsinki CS department

Installed as a chain of three tools in WebLicht 0.3: wrapper of plain text as text element, tokenizer, and then omorfi

Straightforward (modulo a protocol change?)

Not installed

Other HFST morphologies (French, German, Swedish)
Other HFST output formats

Shortcomings

WebLicht omorfi is not current
Tokenizer is too simple (but easy to improve)
No disambiguation of analyses in context

First service: wrapping of plain UTF-8 in D-Spin XML

Input: plain text (Wikipedia on "Orava" 2010-10-08)

```
Oravan tunnistaa helposti pitkästä tuuheasta hännästä.
Oravien silmät ovat melko suuret ja eteenpäin työntyvät,
```

Output: text element in XML (all XML rendered by Firefox)

```
<ns4:D-Spin>
  <MetaData>
    <source/>
  </MetaData>
  <ns2:TextCorpus lang="fi">
     <ns2:text>
        Oravan tunnistaa helposti pitkästä tuuheasta
        hännästä. Oravien silmät ovat melko suuret ja
```

Orava Düsseldorfissa (Wikipediasta)

Eichhörnchen im Düsseldorfer Hofgarten, fotografiert von Ray eye

Second service: tokenization of the text

Output: tokens element after text element

```
    säksätystä, undulaattimaista kujerrusta tai
    kurnutusta ja niiskuttelua.
  </ns2:text>
  <ns2:tokens>
    <ns2:token ID="t0">Oravan</ns2:token>
    <ns2:token ID="t1">tunnistaa</ns2:token>
    <ns2:token ID="t2">helposti</ns2:token>
    <ns2:token ID="t3">pitkästä</ns2:token>
    <ns2:token ID="t4">tuuheasta</ns2:token>
    <ns2:token ID="t5">hännästä</ns2:token>
    <ns2:token ID="t6">Oravien</ns2:token>
    <ns2:token ID="t7">silmät</ns2:token>
    <ns2:token ID="t8">ovat</ns2:token>
```

Third service: morphological analysis of the tokens

Output: morphology element after tokens element

```
 <ns2:token ID="t92">kurnutusta</ns2:token>
 <ns2:token ID="t93">ja</ns2:token>
 <ns2:token ID="t94">niiskuttelua</ns2:token>
</ns2:tokens>
<ns2:morphology tagset="UniHel Morph Convention">
 <ns2:analysis tokID="t0" ID="a0">##orava+noun+1+sg+ins#
 <ns2:analysis tokID="t0" ID="a1">##orava+noun+1+sg+gen#
 <ns2:analysis tokID="t0" ID="a2">##orava+noun+1+sg+acc#
 <ns2:analysis tokID="t1" ID="a3">##tunnistaa+verb+53+ac
 <ns2:analysis tokID="t1" ID="a4">##tunnistaa+verb+53+ac
 <ns2:analysis tokID="t2" ID="a5">##helppo+noun+1+b+Dsti
 <ns2:analysis tokID="t2" ID="a6">##helposti+99+99##</ns
 <ns2:analysis tokID="t3" ID="a7">##pitkä+noun+1+sg+ela#
```

No fourth service!

Chain for Finnish ends short

At least WebLicht suggests no further services

Even in absence of specifically Finnish services, no reason not to do e.g. statistics on token occurrences

Interface problems?

Something trivial missing or off?

Or all WebLicht interface specifications too rigid?
There seem to be many trivial wrapper services
Would further output formats cause similar multiplication?

Visions of a possible future

assuming we want to find and match data and tools on the web

Strengths

Interoperability with open standards, XML, UTF-8, WebLicht input and output specifications

Challenges

Large data volumes require computer resources

Copyright and privacy may require authorization and authentication

Finnish opportunities?

Current omorfi
Other HFST morphologies and output formats

Finnish wordnet: related words? Finnish tree bank, parse bank

Serve concordances from databases?

OAI-PMH Metadata: soon there for corpora; could try on individual documents?