CORPUS LINGUISTICS BASICS

Crash course for SFB-991 members

29.01.2013

OVERVIEW

- MULTILINGUAL CORPORA IN SFB 991
- MONOLINGUAL CORPORA IN SFB 991
- ANTCONC - A FREEWARE CONCORDANCE PROGRAM FOR WINDOWS, MACINTOSH OS X, AND LINUX

OVERVIEW

- MULTILINGUAL CORPORA IN SFB 991
- MONOLINGUAL CORPORA IN SFB 991
- ANTCONC - A FREEWARE CONCORDANCE PROGRAM FOR WINDOWS, MACINTOSH OS X, AND LINUX

MULTILINGUAL CORPORA IN SFB 991

- JRC-ACQUIS (EU LEGISLATIVE TEXTS IN ALL MEMBER-STATES LANGUAGES - BULGARIAN, CZECH, DANISH, GERMAN, GREEK, ENGLISH, SPANISH, ESTONIAN, FINNISH, FRENCH, HUNGARIAN, ITALIAN, LITHUANIAN, LATVIAN, MALTESE, DUTCH, POLISH, PORTUGUESE, ROMANIAN, SLOVAK, SLOVENE AND SWEDISH)
- HTTP://IPSC.JRC.EC.EUROPA.EU/INDEX.PHP?ID=198

JRC-Acquis corpus

Example from the English subcorpus:

```
<text>
<body>
<head n="1">
Agreement in the form of an Exchange of Letters extending the Protocol setting out, for the period from 1 June
2005 to 31 May 2006 , the fishing opportunities and the financial contribution provided for in the Agreement
between the European Economic Community and the Government of the Democratic Republic of S%atilde%o Tom%eacute
% and Pr%iacute%ncipe on fishing off the coast of S%atilde%o Tom%eacute% and Pr%iacute%ncipe
</head>
<div type="body">
<p n="2">Agreement in the form of an Exchange of Letters</p>
<p n="3">
extending the Protocol setting out, for the period from 1 June 2005 to 31 May 2006, the fishing opportunities
and the financial contribution provided for in the Agreement between the European Economic Community and the
Government of the Democratic Republic of São Tomé and Príncipe on fishing off the coast of São Tomé and
Príncipe
</p>
<p n="4">A. Letter from the Community</p>
<p n="5">Sirs,</p>
<p n="6">
I have the honour to confirm that pending negotiations on amendments to be made to the Protocol currently in
force ( 1 June 2002 to 31 May 2005) setting out the fishing opportunities and financial contribution provided
for in the Fisheries Agreement between the European Economic Community and the Government of the Democratic
Republic of São Tomé and Príncipe, we agree to the following interim arrangements:
</p>
```

MULTILINGUAL CORPORA IN SFB 991

- LCC (MOSTLY NEWSPAPER TEXTS IN CATALAN, DANISH, DUTCH, ENGLISH, ESTONIAN, FINNISH, FRENCH, GERMAN, ICELANDIC, ITALIAN, JAPANESE, KOREAN, NORWEGIAN, SERBIAN, SORBIAN, SPANISH, SWEDISH, TURKISH, ETC.)
- HTTP://CORPORA.INFORMATIK.UNI-LEIPZIG.DE/DOWNLOAD.HTML

LCC CORPUS

Example from the German subcorpus:

1 Bundeskanzlerin Merkel soll den Produktpiraten die Leviten lesen.
2 Werden in einem Haushalt mehrere getrennt Netze aufgebaut, so kann der Datentransfer per Passwort verschlüsselt werden und bleibt dann für die andren Teilnehmer unsichtbar.
3 WELT.de: Täuschen kann man auch anders.
4 Der Stand von Öl, Bremsflüssigkeit, Kühlwasser sollte geprüft werden - passt die Kilometerangabe auf dem Ölwechsel-Anhänger zum Tachostand?
5 "Es geht um den Ruf des Hauses.
6 Sie sprach sich für eine bessere Prävention aus.
7 Es gibt Beispiele.
8 Eine Sprechstundenhilfe kam ausgerechnet an diesem Morgen zu spät zur Arbeit, was ihr möglicherweise das Leben rettete.
9 "Die Zahlen gehen in allen Bereichen nach unten", sagte der Vizepräsident stolz.
10 Am Montag sollen in Moskau auch die Verhandlungen über ein Angebot Russlands beginnen, Uran für das iranische Atomprogramm in Russland anzureichern.
11 Training für die grauen Zellen: Lösen sie die kniffeligen Sudoku-Rätsel jetzt auch online.

MULTILINGUAL CORPORA IN SFB 991

- MULTEXTEAST (ORWELL'S "1984" IN BULGARIAN, CZECH, ENGLISH, ESTONIAN, HUNGARIAN, MACEDONIAN, PERSIAN, POLISH, ROMANIAN, SERBIAN, SLOVAK AND SLOVENE, ALIGNED AT SENTENCE LEVEL)
- HTTP://NL.IJS.SI/ME/

MULTEXT-EAST

The MULTEXT-East project developed an annotated multilingual corpus and lexical resources for six languages: **Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene**, as well as for **English**, as the 'hub' language of the project. The results were release in 1998.

Version 4 of the MULTEXT-East dataset extends the resources by several new languages: **Croatian, Lithuanian, Macedonian, Persian, Polish, Resian, Russian, Serbian, Slovak** and **Ukrainian**. All the resources are uniformly encoded in XML, in TEI P5.

MULTEXT-EAST, VERSION 4

It contains, among others, the following resources:

✦ Morphosyntactic specifications. Languages: Bulgarian, Croatian, Czech, English, Estonian, Hungarian, Macedonian, Persian, Polish, Resian, Romanian, Russian, Serbian, Slovak, Slovene, Ukrainian.

✦ Lexica. Languages: Bulgarian, Czech, English, Estonian, Hungarian, Macedonian, Persian, Polish, Resian, Romanian, Russian, Serbian, Slovak, Slovene, Ukrainian.

✦ Linguistically annotated '1984'. Languages: Bulgarian, Czech, English, Estonian, Hungarian, Macedonian, Persian, Polish, Romanian, Serbian, Slovak, Slovene.

MULTEXT-EAST

Example from the annotated English version of Orwell's "1984":

```
<text xml:id="Oen." xml:lang="en">
<body>
<div type="part" xml:id="Oen.1">
<div type="chapter" xml:id="Oen.1.1">
<p xml:id="Oen.1.1.1">
<s xml:id="Oen.1.1.1.1">
<w xml:id="Oen.1.1.1.1.1" lemma="it" ana="#Pp3ns">It</w>
<w xml:id="Oen.1.1.1.1.2" lemma="be" ana="#Vmis3s">was</w>
<w xml:id="Oen.1.1.1.1.3" lemma="a" ana="#Di">a</w>
<w xml:id="Oen.1.1.1.1.4" lemma="bright" ana="#Af">bright</w>
<w xml:id="Oen.1.1.1.1.5" lemma="cold" ana="#Afp">cold</w>
<w xml:id="Oen.1.1.1.1.6" lemma="day" ana="#Ncns">day</w>
<w xml:id="Oen.1.1.1.1.7" lemma="in" ana="#Sp">in</w>
<w xml:id="Oen.1.1.1.1.8" lemma="April" ana="#Ncns">April</w>
<c xml:id="Oen.1.1.1.1.9">,</c>
<w xml:id="Oen.1.1.1.1.10" lemma="and" ana="#Cc-n">and</w>
<w xml:id="Oen.1.1.1.1.11" lemma="the" ana="#Dd">the</w>
<w xml:id="Oen.1.1.1.1.12" lemma="clock" ana="#Ncnp">clocks</w>
<w xml:id="Oen.1.1.1.1.13" lemma="be" ana="#Vais-p">were</w>
<w xml:id="Oen.1.1.1.1.14" lemma="strike" ana="#Vmpp">striking</w>
<w xml:id="Oen.1.1.1.1.15" lemma="thirteen" ana="#Mc">thirteen</w>
<c xml:id="Oen.1.1.1.1.16">.</c>
</s>
```

OVERVIEW

- MULTILINGUAL CORPORA IN SFB 991
- MONOLINGUAL CORPORA IN SFB 991
- ANTCONC - A FREEWARE CONCORDANCE PROGRAM FOR WINDOWS, MACINTOSH OS X, AND LINUX

MONOLINGUAL CORPORA IN SFB 991

- LANCASTER CORPUS OF MANDARIN CHINESE
- HTTP://WWW.OTA.OX.AC.UK/HEADERS/2474.XML
- HTTP://WWW.LANCS.AC.UK/FASS/PROJECTS/CORPUS/LCMC/LCMC/LCMC_INFO.HTM

LANCASTER CORPUS OF MANDARIN CHINESE

Example from LCMC:

```
<text ID="A" TYPE="Press reportage">
<file ID="A01">
<p>
<s n="0001">
<w POS="a">大</w>
<w POS="n">墙</w>
<w POS="f">内外</w>
<c POS="w">——</c>
<w POS="ns">北京市</w>
<w POS="n">监狱</w>
<w POS="n">纪实</w>
<c POS="w">（</c>
<w POS="m">三</w>
<c POS="w">）</c>
</s>
</p>
<p>
```

MONOLINGUAL CORPORA IN SFB 991

- POLISH NATIONAL CORPUS
- HTTP://NKJP.PL/INDEX.PHP?PAGE=0&LANG=1

POLISH NATIONAL CORPUS

Example from the 1-million annotated subcorpus, the morphosyntax level (other levels are text, words, senses, segmentation, named (entities) and groups):

```
<seg corresp="ann_segmentation.xml#segm_1.11-seg" xml:id="morph_1.11-seg">
      <fs type="morph">
       <f name="orth">
        <string>Następnie</string>
       </f>
       <!-- Następnie [53,9] -->
       <f name="interps">
        <fs type="lex" xml:id="morph_1.11.1-lex">
         <f name="base">
          <string>następnie</string>
         </f>
         <f name="ctag">
          <symbol value="adv"/>
         </f>
         <f name="msd">
          <symbol value="" xml:id="morph_1.11.1.1-msd"/>
         </f>
        </fs>
       </f>
       <f name="disamb">
        <fs feats="#an8003" type="tool_report">
         <f fVal="#morph_1.11.1.1-msd" name="choice"/>
         <f name="interpretation">
          <string>następnie:adv</string>
         </f>
        </fs>
       </f>
      </fs>
     </seg>
```

MONOLINGUAL CORPORA IN SFB 991

- BULGARIAN MULTEXT-EAST
- BULGARIAN TREEBANK
- HTTP://WWW.BULTREEBANK.ORG/RESOURCES.HTML

BULTREEBANK-MORPH

Example from Bulgarian Treebank, morphologically annotated part:

```
<s>
<pt>"</pt>
<w aa="Adv" ta="Adv">Защо</w>
<w aa="Pron" ta="Pron">се</w>
<w aa="Verb" ta="Verb">нарича</w>
<w aa="Adv;Conj;Part" ta="Adv">така</w>
<w aa="Pron" ta="Pron">никой</w>
<w aa="Conj;Part" ta="Part">не</w>
<w aa="Verb" ta="Verb">знае</w>
<pt>"</pt>
<pt>,</pt>
<w aa="Verb" ta="Verb">казва</w>
<w aa="Noun" ta="Noun">полк</w>
<pt>.</pt>
</s>
```

MONOLINGUAL CORPORA IN SFB 991

- MACEDONIAN MULTEX-EAST
- MACEDONIAN (SERVICES AND TOOLS FROM THE TEXT LABORATORY, UNIVERSITY OF OSLO)
- HTTP://WWW.TEKSTLAB.UIO.NO/GLOSSA/HTML/INDEX_DEV.PHP?CORPUS=MAK

MULTEXT-EAST, MACEDONIAN

Example from the annotated Macedonian version of Orwell's "1984":

```
<p xml:id="Omk.1.1.1">
<s xml:id="Omk.1.1.1.1">
<tok xml:id="Omk.1.1.1.1.1">
<w>Беше</w>
<ana type="lexicon">
<fs>
<f name="lemma">
<string>сум</string>
</f>
<f name="msd">
<vColl>
<fs copyOf="#Vaii2s"/>
<fs copyOf="#Vaii3s"/>
</vColl>
</f>
</fs>
</ana>
</tok>
<S/>
```

MONOLINGUAL CORPORA IN SFB 991

- SYNTAGRUS
- HTTP://WWW.RUSCORPORA.RU/EN/INDEX.HTML
- HTTP://WWW.RUSCORPORA.RU/INSTRUCTION-SYNTAX.HTML
- RUSSIAN IN THE RUN PROJECT:
- HTTP://WWW.HF.UIO.NO/ILOS/ENGLISH/RESEARCH/PROJECTS/RUN/

SYNTAGRUS

Example from SynTagRus:

```
...
<source>Российская газета № , 14 февраля 2008 </source>
<title>"Ископаемое добро" </title>
</inf>
<body>
<S ID="1">
<W DOM="3" FEAT="S ЕД ЖЕН ИМ НЕОД" ID="1" LEMMA="ЭКОНОМИКА" LINK="предик">Экономика</W>
<W DOM="1" FEAT="S МН СРЕД РОД НЕОД" ID="2" LEMMA="ЗНАНИЕ" LINK="1-компл">знаний</W>
<W DOM="_root" FEAT="V СОВ ИЗЪЯВ НЕПРОШ ЕД 3-Л" ID="3" LEMMA="ТРЕБОВАТЬ">потребует</W>
<W DOM="5" FEAT="A ЕД СРЕД РОД" ID="4" LEMMA="КОМПЛЕКСНЫЙ" LINK="опред">комплексного</W>
<W DOM="3" FEAT="S ЕД СРЕД РОД НЕОД" ID="5" LEMMA="РАЗВИТИЕ" LINK="1-компл">развития</W>
<W DOM="7" FEAT="A ЕД ЖЕН РОД" ID="6" LEMMA="ВЕСЬ" LINK="опред">всей</W>
<W DOM="5" FEAT="S ЕД ЖЕН РОД НЕОД" ID="7" LEMMA="СТРАНА" LINK="1-компл">страны</W>
</S>
```

OVERVIEW

- MULTILINGUAL CORPORA IN SFB 991
- MONOLINGUAL CORPORA IN SFB 991
- ANTCONC - A FREEWARE CONCORDANCE PROGRAM FOR WINDOWS, MACINTOSH OS X, AND LINUX

AntConc

- Getting started

  http://www.youtube.com/watch?v=_z9wwX7eR-Y

- Concordance tool - basic features

  http://www.youtube.com/watch?v=9TsqFVrUYO0

- Concordance tool - advanced features

  http://www.youtube.com/watch?v=pUN-flv-Cmw

- Concordance plot tool

  http://www.youtube.com/watch?v=YNDvcxonmP4

ANTCONC

- FILE VIEW TOOL

  http://www.youtube.com/watch?v=Yalioyw_74M

- CLUSTERS TOOL

  http://www.youtube.com/watch?v=5kosY5xlsLE

- N-GRAMS TOOL

  http://www.youtube.com/watch?v=-vv663HjTso

- FILE COLLOCATES TOOL

  http://www.youtube.com/watch?v=An6gb1W3pFM

AntConc

- Word List Tool

  http://www.youtube.com/watch?v=Zb71yaBP_lI

- Keyword List Tool

  http://www.youtube.com/watch?v=JvasTvQY7kU

- Working with Lemmas

  http://www.youtube.com/watch?v=WxCjhbCRE4A

THANK YOU!