CORPUS LINGUISTICS BASICS

Crash course for SFB-991 members

05.02.2013

Overview

- Monolingual corpora in SFB 991 (cont.)
  - German corpora: DeReKo, LCC Corpus, German Political Speeches
  - German (tree)banks: NEGRA, TIGER, TIGER Dependency Bank
  - English corpora: BNC
  - English (tree)banks: Penn Treebank, Penn Discourse Treebank, OntoNotes English Corpus, PARK 700 Dependency Bank
- BootCaT - simple utilities to bootstrap corpora and terms from the web

OVERVIEW

- MONOLINGUAL CORPORA IN SFB 991 (CONT.)
- BOOTCAT - SIMPLE UTILITIES TO BOOTSTRAP CORPORA AND TERMS FROM THE WEB

MONOLINGUAL CORPORA IN SFB 991

- DEREKO (DAS DEUTSCHE REFERENZKORPUS)
- HTTP://WWW.IDS-MANNHEIM.DE/KL/PROJEKTE/KORPORA/EU/INDEX.PHP?ID=198

DEREKO

The corpora of contemporary written German at the Institute for German Language in Mannheim:

- have some 5,4 billion words (by 29.02.2012) and constitute the biggest collection of machine-readable written German texts from the present period and the recent past;
- consist of fiction, science and popular-science texts, a large number of newspaper texts and other written texts;
- they are constantly extended;
- they are not explicitly designed to be ballanced;
- however, it is possible through **COSMAS II** to build virtual corpora, that are potentially ballanced or are suited for special research goals.

DEREKO CORPORA IN SFB 991

- Mannheimer Korpus 1 (mk1), Mannheimer Korpus 2 (mk2):
  - http://www.ids-mannheim.de/kl/projekte/korpora/archiv/mk.html
  - mk1 has 293 texts from the period 1950-1967, and ca. 2,2 million running-text tokens
  - mk2 has 52 texts from 1949, 1952, 1960 - 1974, and ca. 0,3 million running-text tokens
- Bonner Zeitungskorpus (bzk):
  - http://www.ids-mannheim.de/kl/projekte/korpora/archiv/bzk.html
  - bzk has 10840 texts from 1949, 1954, 1959, 1964, 1969 and 1974, and ca. 3,1 million running-text tokens

DEREKO CORPORA IN SFB 991

Example from mk1:

```
man kann nicht sagen, daß sie an dem Sohn, einem lang
aufgeschossenen Rotkopf, der dem verstorbenen Vater ähnlich
sah und übrigens für die humanistischen Studien wenig
veranlagt war, sondern vielmehr vom <orig
reg="Brückenbau">Brücken-</orig> und Wegebau träumte und
Ingenieur werden wollte, besonderen Anteil genommen hätte.

der Tochter dagegen, der hing sie an, ihrer einzigen
wirklichen Freundin.
```

OTHER GERMAN 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.

MONOLINGUAL CORPORA IN SFB 991

- GERMAN POLITICAL SPEECHES
- HTTP://PERSO.ENS-LYON.FR/ADRIEN.BARBARESI/CORPORA/INDEX.HTML

GERMAN POLITICAL SPEECHES

- The corpus contains the Presidency subcorpus and the Chancellery subcorpus.
- The Presidency subcorpus has a total of 1442 texts (2392074 tokens), from the period 01.07.1984-17.02.2012. It contains speeches of the following presidents: Richard von Weizsäcker (1984-1994), Roman Herzog (1994-1999), Johannes Rau (1999-2004), Horst Köhler (2004-2010) and Christian Wulff (2010-2012). The speeches were crawled from the online archive of the German Presidency (bundespraesident.de).
- The Cancellary subcorpus has a total of 1831 texts (3891588 tokens), from the period 11.12.1998-06.12.2011. It contains not only speeches by the chancellors Gerhard Schröder and Angela Merkel), but also a number of other state ministers that are linked to the head of the government and a few unrelated speeches from other politicians. The speeches were crawled from the online archive of the German Chancellary (bundesregierung.de).

GERMAN POLITICAL SPEECHES

Example from the Chancellary subcorpus:

```
<w xml:id="BRv2_t560w4" type="NN" lemma="Herr">
Herr
</w>
<w xml:id="BRv2_t560w5" type="NE" lemma="Lammert">
Lammert
</w>
<w xml:id="BRv2_t560w6" type="$," lemma=",">
,
</w>
<w xml:id="BRv2_t560w7" type="PPER" lemma="ich">
ich
</w>
<w xml:id="BRv2_t560w8" type="VVFIN" lemma="freuen">
freue
</w>
<w xml:id="BRv2_t560w9" type="PRF" lemma="ich">
mich
</w>
```

MONOLINGUAL CORPORA IN SFB 991

- THE NEGRA CORPUS
- HTTP://WWW.COLI.UNI-SAARLAND.DE/PROJECTS/SFB378/NEGRA-CORPUS/NEGRA-CORPUS.HTML

NEGRA

The NEGRA corpus version 2 consists of 355,096 tokens (20,602 sentences) of German newspaper text.
The texts are taken from the Frankfurter Rundschau.
The corpus is tagged for part-of-speech and completely annotated with syntactic structures.
The corpus is stored in an SQL database. Alternatively, the annotations can be represented in a line-oriented export format or in PennTreebank format

The following different types of information are coded in the corpus:

- Part-of-Speech tags (Stuttgart-Tübingen-Tagset (STTS))
- Morphological analysis (only for the first 60,000 tokens, the expanded STTS)
- The grammatical function in the directly dominating phrase.
- The category of nonterminal nodes (phrases).

NEGRA: EXAMPLE IN EXPORT FORMAT

```
%% word          tag       morph           edge parent    secedge comment
#BOS 1 1 985275570 1
Mögen            VMFIN     3.Pl.Pres.Konj HD   508
Puristen         NN        Masc.Nom.Pl.*  NK   505
aller            PIDAT     *.Gen.Pl  NK   500
Musikbereiche    NN        Masc.Gen.Pl.*  NK   500
auch             ADV       --        MO   508
die              ART       Def.Fem.Akk.Sg NK   501
Nase             NN        Fem.Akk.Sg.*   NK   501
rümpfen          VVINF     --        HD   506
,                $,        --        --   0
die              ART       Def.Fem.Nom.Sg NK   507
Zukunft          NN        Fem.Nom.Sg.*   NK   507
der              ART       Def.Fem.Gen.Sg NK   502
Musik            NN        Fem.Gen.Sg.*   NK   502
liegt            VVFIN     3.Sg.Pres.Ind  HD   509
für              APPR      Akk       AC   503
viele            PIDAT     *.Akk.Pl  NK   503
junge            ADJA      Pos.*.Akk.Pl.St     NK   503
Komponisten      NN        Masc.Akk.Pl.*  NK   503
im               APPRART   Dat.Masc  AC   504
Crossover-Stil   NN        Masc.Dat.Sg.*  NK   504
.                $.        --        --   0
#500             NP        --        GR   505
#501             NP        --        OA   506
#502             NP        --        GR   507
#503             PP        --        MO   509
#504             PP        --        MO   509
#505             NP        --        SB   508
#506             VP        --        OC   508
#507             NP        --        SB   509
#508             S         --        MO   509
#509             S         --        --   0
#EOS 1
```

NEGRA: EXAMPLE IN PENNTREEBANK FORMAT

```
      %% Sent 1
(
      (S
           (S-MO
                (VMFIN-HD Mögen)
                (NP-SB
                     (NN-NK Puristen)
                     (NP-GR
                          (PIDAT-NK aller)
                          (NN-NK Musikbereiche)
                     )
                )
                (ADV-MO auch)
                (VP-OC
                     (NP-OA
                          (ART-NK die)
                          (NN-NK Nase)
                     )
                     (VVINF-HD rümpfen)
...
                )
           )
)
```

MONOLINGUAL CORPORA IN SFB 991

- THE TIGER CORPUS, TREEBANK AND DEPENDENCY BANK
- HTTP://WWW.IMS.UNI-STUTTGART.DE/PROJEKTE/TIGER/TIGERCORPUS/
- HTTP://WWW.IMS.UNI-STUTTGART.DE/PROJEKTE/TIGER/TIGERCORPUS/LICENSE/
- (TO ACCESS THE DOWNLOADS FIRST OPEN AND ACCEPT THE LICENSE AGREEMENT

TIGER TREEBANK

The TIGER Treebank (Version 2.1) consists of app. 900,000 tokens (50,000 sentences) of German newspaper text.
The texts are taken from the Frankfurter Rundschau.
The corpus is tagged for part-of-speech and completely annotated with syntactic structures.
The annotations can be represented in Negra export format (cf. above) or in TIGER-XML format.

The following different types of information are coded in the corpus:

- part-of-speech tags and morphological analysis (based on STTS, but modified);
- the grammatical function in the directly dominating phrase;
- the category of nonterminal nodes (phrases).

TIGER: EXAMPLE IN TIGER-XML FORMAT

```
    <s id="s4231">
<graph root="s4231_VROOT" discontinuous="true">
<terminals>
<t id="s4231_1" word="In" lemma="in" pos="APPR" morph="--"/>
<t id="s4231_2" word="Japan" lemma="Japan" pos="NE" morph="Dat.Sg.Neut"/>
<t id="s4231_3" word="wird" lemma="werden" pos="VAFIN" morph="3.Sg.Pres.Ind"/>
<t id="s4231_4" word="offenbar" lemma="offenbar" pos="ADJD" morph="Pos"/>
<t id="s4231_5" word="die" lemma="der" pos="ART" morph="Nom.Sg.Fem"/>
<t id="s4231_6" word="Fusion" lemma="Fusion" pos="NN" morph="Nom.Sg.Fem"/>
<t id="s4231_7" word="der" lemma="der" pos="ART" morph="Gen.Pl.Masc"/>
<t id="s4231_8" word="Geldkonzerne" lemma="Geldkonzern" pos="NN" morph="Gen.Pl.Masc"/>
<t id="s4231_9" word="Daiwa" lemma="Daiwa" pos="NE" morph="Nom.Sg.*"/>
<t id="s4231_10" word="und" lemma="und" pos="KON" morph="--"/>
<t id="s4231_11" word="Sumitomo" lemma="Sumitomo" pos="NE" morph="Nom.Sg.*"/>
<t id="s4231_12" word="zur" lemma="zu" pos="APPRART" morph="Dat.Sg.Fem"/>
<t id="s4231_13" word="größten" lemma="groß" pos="ADJA" morph="Sup.Dat.Sg.Fem"/>
<t id="s4231_14" word="Bank" lemma="Bank" pos="NN" morph="Dat.Sg.Fem"/>
<t id="s4231_15" word="der" lemma="der" pos="ART" morph="Gen.Sg.Fem"/>
<t id="s4231_16" word="Welt" lemma="Welt" pos="NN" morph="Gen.Sg.Fem"/>
<t id="s4231_17" word="vorbereitet" lemma="vorbereiten" pos="VVPP" morph="Psp"/>
<t id="s4231_18" word="." lemma="--" pos="$." morph="--"/>
</terminals>
<nonterminals>
<nt id="s4231_500" cat="PP">
<edge label="AC" idref="s4231_1"/>
<edge label="NK" idref="s4231_2"/>
</nt>
<nt id="s4231_501" cat="CNP">
<edge label="CJ" idref="s4231_9"/>
<edge label="CD" idref="s4231_10"/>
<edge label="CJ" idref="s4231_11"/>
</nt>
...
</nonterminals>
</graph>
```

TIGER DEPENDENCY BANK

The TiGer Dependency Bank (TiGer DB) covers sentences 8,001 - 10,000 of the TIGER Corpus and is created as a dependency-based gold standard for German parsers. Its annotation is close to the annotation of PARC 700 dependency bank.

MONOLINGUAL CORPORA IN SFB 991

- THE BRITISH NATIONAL CORPUS (BNS)
- HTTP://WWW.NATCORP.OX.AC.UK

THE BNC

http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html

The XML Edition of the BNC contains 4049 texts and occupies (including all markup) 5,228,040 Kb, or about 5.2 Gb.

In total, it comprises just under 100 million orthographic words (specifically, 96986707), but the number of w-units (POS-tagged items) is slightly higher at 98363783.

The tagging distinguishes a further 13614425 punctuation strings, giving a total content count of 110691482 strings.

The total number of s-units tagged is about 6 million (6026284). Counts for these and all the other elements tagged in the corpus are provided in the corpus header.

In the following tables both an absolute count and a percentage are given for all the counts. The percentage is calculated with reference to the relevant portion of the corpus, for example, in the table for "written text domain", with reference to the total number of w-units in written texts. Note that punctuation strings are not included in these totals. The reference totals used are given in the first table below.

MONOLINGUAL CORPORA IN SFB 991

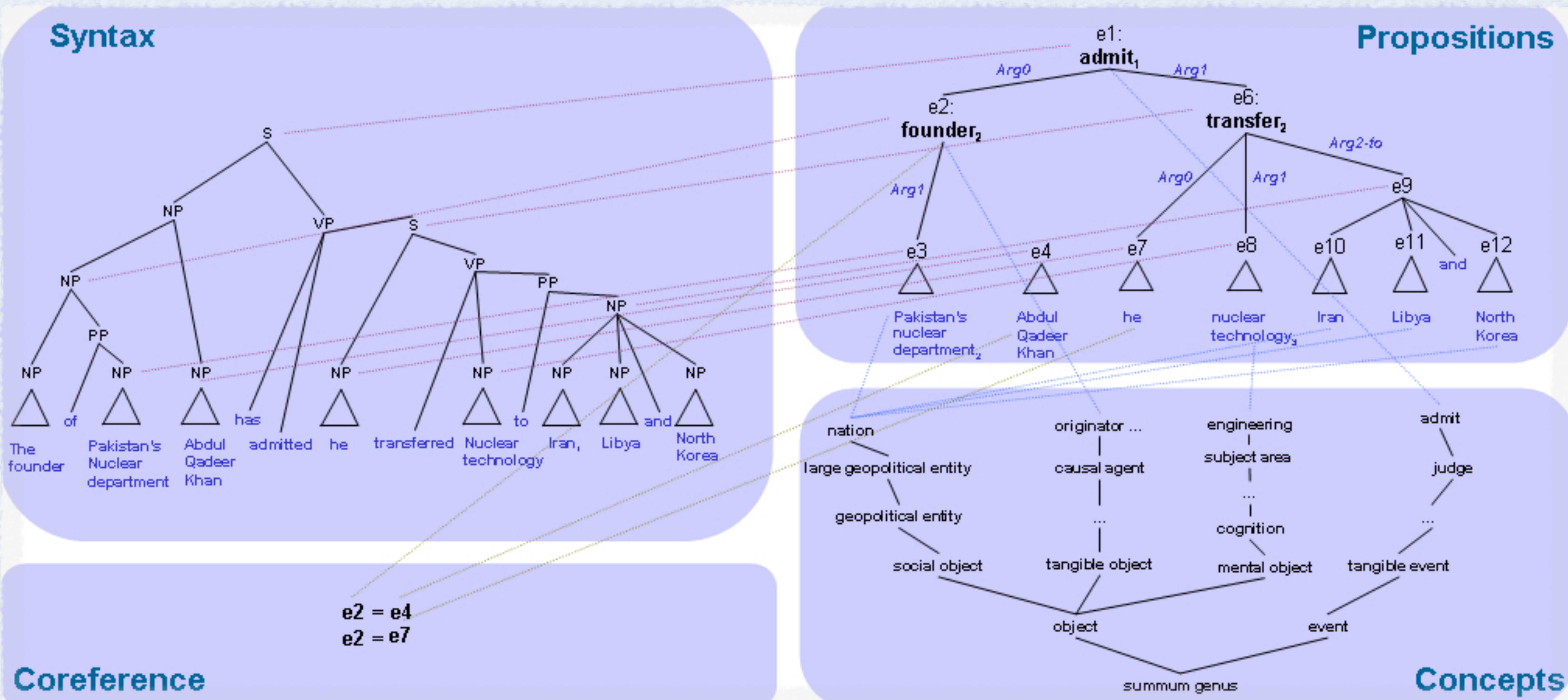
- THE PENN TREEBANK
- HTTP://WWW.CIS.UPENN.EDU/~TREEBANK/
- THE PENN DISCOURSE TREEBANK
- HTTP://WWW.SEAS.UPENN.EDU/~PDTB/
- ONTONOTES
- HTTP://WWW.BBN.COM/ONTONOTES/

ONTONOTES

Ontonotes aims to annotate a large corpus comprising various genres of text:
news, conversational telephone speech, weblogs, use net, broadcast, talk shows,
in three languages: English, Chinese, and Arabic
with structural information (syntax and predicate argument structure)
and shallow semantics (word sense linked to an ontology and coreference).
OntoNotes builds on two time-tested resources,
following the Penn Treebank for syntax
and the Penn PropBank for predicate-argument structure.
Its semantic representation will include word sense disambiguation for nouns and verbs,
with each word sense connected to an ontology, and coreference.
Over the course of the five-year program, the current goals call for annotation of over a million words each of English and Chinese, and half a million words of Arabic.

http://www.bbn.com/ontonotes/

ONTONOTES


Syntax
Propositions
Coreference
Concepts
S
NP
VP
PP
The founder
of
Pakistan's Nuclear department
Abdul Qadeer Khan
has
admitted
he
transferred
Nuclear technology
to
Iran,
Libya
and
North Korea
e1: admit₁
Arg0
Arg1
e2: founder₂
e6: transfer₂
Arg2-to
e3
e4
e7
e8
e9
e10
e11
e12
Pakistan's nuclear department₂
Abdul Qadeer Khan
he
nuclear technology₃
Iran
Libya
North Korea
e2 = e4
e2 = e7
nation
large geopolitical entity
geopolitical entity
social object
originator ...
causal agent
...
tangible object
engineering
subject area
...
cognition
mental object
admit
judge
...
tangible event
object
event
summum genus

OVERVIEW

- MONOLINGUAL CORPORA IN SFB 991 (CONT.)
- BOOTCAT - SIMPLE UTILITIES TO BOOTSTRAP CORPORA AND TERMS FROM THE WEB

BOOTCAT

The BootCaT front-end is a graphical interface for the BootCaT toolkit.

(http://bootcat.sslmit.unibo.it)

It guides you through the process of creating a simple web corpus.

It only needs a list of "seeds" (terms that are expected to be typical of the domain of interest) as input.

Download link (do scroll down to get to the downloads for Windows, Mac and Linux/Unix):

http://bootcat.sslmit.unibo.it/?section=download

THANK YOU!