The NeGra Export Format for Annotated Corpora (Version 3)

Thorsten Brants
`thorsten@coli.uni-sb.de`

July 1997

Sonderforschungsbereich 378
Ressourcenadaptive
Kognitive Prozesse

Universität des Saarlandes
Computerlinguistik

Projekt C3 Nebenläufige Grammatische Verarbeitung

Abstract

This paper describes the export format version 3 of corpora used in the NeGra project. We use a line-oriented and ASCII-based format that is both easy to read by humans and easy to parse by machines. It is intended for data exchange and for efficient processing with standard Unix tools and C programs.

1 File Level

Before the first sentence, the file format version is indicated by a line

`#FORMAT` ⟨*num*⟩

The keyword `#FORMAT` starts at the beginning of a line with no leading white space. The version described here is 3, so a file complying this format contains the line

```
#FORMAT 3
```

Each line in the file starting with two percentage signs (`%%`, no leading white space) are pure comments. These as well as empty lines (lines containing only white space) should be ignored by programs.

The text is encoded using codes 9 (tab), 10 (newline), 32 – 127 (ASCII characters) and 160 – 255 (ISO Latin1).

2 Tables

A file specifies descriptions of the sentences' origins and editors, and optionally tables of tags used in a corpus and the descriptions of these tags. The id's of origins and editors are stored with each sentence in the corpus. The optional description of tags is useful when moving a complete corpus from one site to another.

The beginning of a table is marked by the keyword `#BOT` (*beginning of table*) together with its name. The end is marked by `#EOT` (*end of table*):

```
#BOT ⟨table name⟩ \n
    ...⟨table data⟩...
#EOT ⟨table name⟩ \n
```

⟨*table name*⟩ is one of the following: `ORIGIN`, `EDITOR`, `WORDTAG`, `MORPHTAG`, `NODETAG`, `EDGETAG`, `SECEDGETAG`. These specify the sentence origins, editors, part-of-speech tags for words, morphological tags, tags for nodes, tags for edges, and tags for secondary edges, respectively.

2.1 Sentence Origins

The table `ORIGIN` is obligatory and describes the origins of sentences. Origins consist of three parts: a unique (positive integer) id, the origin name (max. 128 chars), and an optional comment (max. 160 chars), separated by two percentage signs from the name. Both origin name and comment may contain white space.

```
#BOT ORIGIN \n
⟨id1⟩ ⟨name1⟩ [%% ⟨comment1⟩] \n
⟨id2⟩ ⟨name2⟩ [%% ⟨comment2⟩] \n
...
#EOT ORIGIN \n
```

2.2 Editors

The table `EDITOR` is obligatory and describes editors that have edited sentences in the corpus. Editor entries consist of three parts: a unique id (integer greater or equal -1), the editor's login username (max. 16 chars), and the full name (max. 40 chars). The fields are separated by white space. Only the full name which occupies the rest of the line may contain spaces. Two id's have a fixed meaning: -1 marks 'automatic' and 0 'not named'.

```
#BOT EDITOR \n
-1   --     <automatic> \n
0    --     <not named> \n
⟨id1⟩ ⟨login1⟩ ⟨name1⟩ \n
⟨id2⟩ ⟨login2⟩ ⟨name2⟩ \n
...
#EOT EDITOR \n
```

2.3 Part-of-Speech tags

The table `WORDTAG` is optional. It describes part-of-speech tags for words in the corpus. Entries consist of four parts: a unique id (starting from 1, no intermediate missing number), the tag (max. 10 chars), a flag that identifies if this tag should be bound in a structure (`Y` or `N`; the NEGRA annotations use this flag to indicate that punctuation marks are not required to be bound to the structure, i.e. punctuation marks get a `N`, all others `Y`), and a description of the tag (max. 80 chars).

Two additional id's have a fixed meaning: -1 marks 'unknwon tags' (they can occur during corpus conversion) and 0 marks 'not bound'.

```
#BOT WORDTAG \n
-1    UNKNOWN N Unbekanntes Tag \n
0     --      N Nicht zugeordnet \n
⟨id1⟩ ⟨tag1⟩  ⟨f1⟩ ⟨description1⟩ \n
⟨id2⟩ ⟨tag2⟩  ⟨f2⟩ ⟨description2⟩ \n
...
#EOT WORDTAG \n
```

2.4 Morphological tags

The table `MORPHTAG` is optional. It describes morphological tags for words and phrase nodes in the corpus. Entries consist of three parts: a unique id (starting from 1, no intermediate missing number), the tag (max. 10 chars), and a description of the tag (max. 80 chars).

Two additional id's (-1 and 0) have a fixed meaning: 'unknwon tag' (can occur during corpus conversion) and 'not bound', resp.

```
#BOT MORPHTAG \n
-1    UNKNOWN Unbekanntes Tag \n
0     --      Nicht zugeordnet \n
⟨id1⟩ ⟨tag1⟩  ⟨description1⟩ \n
⟨id2⟩ ⟨tag2⟩  ⟨description2⟩ \n
...
#EOT MORPHTAG \n
```

2.5 Phrasal Categories

The table `NODETAG` is optional. It describes tags for phrases in the corpus. Entries consist of three parts: a unique id (starting from 1, no intermediate missing number), the tag (max. 10 chars), and a description of the tag (max. 80 chars).

Two additional id's (-1 and 0) have a fixed meaning: 'unknwon tag' (can occur during corpus conversion) and 'not bound', resp.

```
#BOT NODETAG \n
-1    UNKNOWN Unbekanntes Tag \n
0     --      Nicht zugeordnet \n
⟨id1⟩ ⟨tag1⟩  ⟨description1⟩ \n
⟨id2⟩ ⟨tag2⟩  ⟨description2⟩ \n
...
#EOT NODETAG \n
```

2.6 Edge Labels

The table EDGETAG is optional. It describes tags of edges in annotated structures. They denote grammatical functions in the NEGRA annotation scheme. Entries consist of three parts: a unique id (starting from 1, no intermediate missing number), the tag (max. 10 chars), and a description of the tag (max. 80 chars).

Two additional id's (-1 and 0) have a fixed meaning: 'unknwon tag' (can occur during corpus conversion) and 'not bound', resp.

```
#BOT EDGETAG \n
-1    UNKNOWN Unbekanntes Tag \n
0     --      Nicht zugeordnet \n
⟨id1⟩ ⟨tag1⟩  ⟨description1⟩ \n
⟨id2⟩ ⟨tag2⟩  ⟨description2⟩ \n
...
#EOT EDGETAG \n
```

2.7 Secondary Edge Labels

The table SECEDGETAG is optional. It describes secondary edge labelsused in the corpus. Entries consist of three parts: a unique id (starting from 1, no intermediate missing number), the tag (max. 10 chars), and a description of the tag (max. 80 chars).

Two id's (-1 and 0) have a fixed meaning: 'unknwon tag' (can occur during corpus conversion) and 'not bound', resp.

```
#BOT SECEDGETAG \n
-1    UNKNOWN Unbekanntes Tag \n
0     --      Nicht zugeordnet \n
⟨id1⟩ ⟨tag1⟩  ⟨description1⟩ \n
⟨id2⟩ ⟨tag2⟩  ⟨description2⟩ \n
...
#EOT SECEDGETAG \n
```

3 Sentence Level

The stored corpus is divided into sentences. The beginning of a sentence is marked by the keyword #BOS together with some additional information about the sentence. The end of a sentence is marked by #EOS and the sentence number:

```
#BOS ⟨num⟩ ⟨editor id⟩ ⟨date⟩ ⟨origin id⟩ [%% ⟨comment⟩] \n
    ...⟨sentence data⟩...
#EOS ⟨num⟩ \n
```

`#BOS` is the keyword that marks the beginning of the sentence. It starts at beginning of the line, with no leading white space.

⟨*num*⟩ is the unique sentence id greater or equal 1; the order in which the sentences are given in a corpus is not significant.

⟨*editor id*⟩ is a reference to the `EDITOR` table and indicates the last editor of the sentence.

⟨*date*⟩ is the date of annotation in Unix format (i.e., seconds since 1/1/1970)

⟨*origin id*⟩ is a reference to the `ORIGIN` table and indicates the origin of the sentence.

`%%` ⟨*comment*⟩ is an optional comment that is indicated by two percentage signs and occupies the rest of the line. The comment may contain spaces.

`#EOS` is the keyword that marks the end of the sentence. It starts at beginning of the line, with no leading white space.

4 Sentence Data

Each word in a sentence and each phrase node is stored in a separate line. The words are implicitly numbered starting with 0, the phrases are explicitly numbered starting with 500. A sentence contains at most 500 words (id's 0 – 499), and at most 500 phrase nodes (id's 500 – 999).

4.1 Words

The lines following a `#BOS` contain the words in the order as they appear in the sentence, one word per line. Words are implicitly numbered starting with 0. The line format is

⟨*word*⟩⟨*postag*⟩⟨*morphtag*⟩⟨*edge*⟩⟨*parent*⟩[⟨*secedge*⟩⟨*secprnt*⟩]* [`%%` ⟨*comment*⟩]`\n`

Columns are separated by any number of white spaces (space or tab). The first column (the word) starts at beginning of the line with no leading white space. Entries in columns *do not contain spaces* (except the comment that occupies the end of a line).

⟨*word*⟩ is the actual word. It may consist of a single #, but otherwise does not start with # (a leading # is reserved for keywords and node numbers)

⟨*postag*⟩ is the part of speech tag of the word.

⟨*morphtag*⟩ is the morphological tag of the word.

⟨*edge*⟩ is the edge label of the edge from the word to its primary parent (phrase)

⟨*parent*⟩ is the id of the primary parent.

Nodes without a primary parent are indicated by parent id 0, and an edge label consisting of two minus signs (`--`).

$\langle secedge \rangle$ is the label of a secondary edge from the word to a secondary parent

$\langle secprnt \rangle$ is the id of the secondary parent.

There may be any number of secondary parents, so $\langle sec\ edge \rangle$ and $\langle sec\ parent \rangle$ may be repeated any number of times.

`%%` separates the word data from an optional comment. It is separated by white space from the previous parent and the actual comment.

$\langle comment \rangle$ is an optional comment. It may contain white space.

`\n` is the newline character that terminates the data for each word in a sentence.

4.2 Phrases

The lines following the last word of a sentence contain the phrase nodes, one node per line. Nodes are explicitly numbered starting with 500. The line format is

`#`$\langle num \rangle \langle tag \rangle \langle morphtag \rangle \langle edge \rangle \langle parent \rangle [\langle secedge \rangle \langle secprnt \rangle]^*$[`%%` $\langle comment \rangle$]`\n`

`#`$\langle num \rangle$ is the character # immediately followed (*without* white space) by the node id (a number in the range 500 – 999).

The character `#` is first in the line, with no leading white space. The other columns are the same as for words.

Phrase nodes are numbered from 500 to $(499 + \langle num\ of\ phrases \rangle)$, no intermediate number is missing. The nodes are numbered bottom up, i.e., the id of some parent is always larger than the id's of all its children (this restriction is needed for the drawing algorithm employed in the NEGRA annotation tool). The order in which the nodes are given in the file is not significant.

5 Example

The sentence "*Schade, daß kein Arzt anwesend ist, der sich auskennt.*" is shown in figure 1 together with its structure. It is encoded as shown in figures 2 and 3. The sentence id is 12, it was annotated by Wojciech at Tue Nov 5 09:54:36 1996, and was collected from origin 1 (Stuttgarter Referenzkorpus).

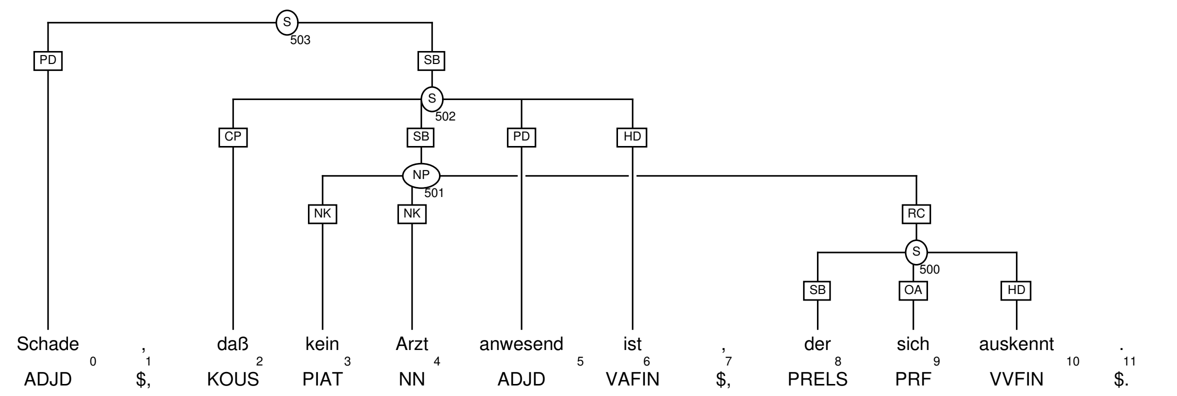

Figure 1: Example sentece

```
#FORMAT 3
#BOT ORIGIN
1       refcorpus %% Stuttgarter Referenzkorpus, Frankfurter Rundschau
#EOT ORIGIN
#BOT EDITOR
1       skut      Wojciech
#EOT EDITOR
#BOT WORDTAG
-1      UNKNOWN   N       Unbekanntes Tag, Fehler
0       --        N       nicht zugeordnet
1       ADJD      Y       Attributives Adjektiv
2       KOUS      Y       Unterordnende Konjunktion mit Satz
3       NN        Y       Normales Nomen
4       PIAT      Y       Attribuierendes Indefinitpronomen
5       PRELS     Y       Substituierendes Relativpronomen
6       PRF       Y       Reflexives Personalpronomen
7       VAFIN     Y       Finites Verb, aux
8       VVFIN     Y       Finites Verb, voll
9       $,        N       Komma
10      $.        N       Satzbeendende Interpunktion
#EOT WORDTAG
#BOT MORPHTAG
-1      UNKNOWN           unknown tag, error
0       --                not bound
1       3.Akk.Pl          3rd person, accusative, plural
2       3.Sg.Pres.Ind     3rd person singular, present, indicative
3       Masc.Nom.Sg       masculinum, nominative, singular
4       Masc.Nom.Sg.*     masculinum, nominative, singular, *
5       Pos               positive
6       *.*.*.*           underspecified
#EOT MORPHTAG
```

continued in next figure...

Figure 2: Encoding of the example sentence (part 1)

...continued from previous figure

```
#BOT NODETAG
-1      UNKNOWN unknown tag, error
1       NP      noun phrase
0       --      not bound
2       S       sentence
#EOT NODETAG
#BOT EDGETAG
-1      UNKNOWN unknown tag, error
1       NP      noun phrase
1       CP      complementizer
2       HD      head
3       NK      noun kernel modifier
4       OA      accusative object
5       PD      predicative
6       RC      relative clause
7       SB      subject
#EOT EDGETAG
#BOT SECEDGETAG
%% no secondary edges used
#EOT SECEDGETAG
#BOS 12 1 847184076 1
Schade              ADJD     Pos             PD      503
,                   $,       --              --      0
daß                 KOUS     --              CP      502
kein                PIAT     Masc.Nom.Sg.*   NK      501
Arzt                NN       Masc.Nom.Sg.*   NK      501
anwesend            ADJD     Pos             PD      502
ist                 VAFIN    3.Sg.Pres.Ind   HD      502
,                   $,       --              --      0
der                 PRELS    Masc.Nom.Sg     SB      500
sich                PRF      3.Akk.Pl        OA      500
auskennt            VVFIN    3.Sg.Pres.Ind   HD      500
.                   $.       --              --      0
#500                S        3.Sg.Pres.Ind   RC      501
#501                NP       Masc.Nom.Sg.*   SB      502
#502                S        3.Sg.Pres.Ind   SB      503
#503                S        *.*.*.*         --      0
#EOS 12
```

Figure 3: Encoding of the example sentence (part 2)