INFORMATION STRUCTURE ANNOTATION IN DanPASS

Patrizia Paggio
patrizia@cst.dk
August 2006

The annotation of information structure in DanPASS consists in the assignment of topic (T) and focus (F) tags to the individual words in the corpus, following the guidelines below. Two annotators, Line Burholt Kristensen and Tina Ringkjær, were asked to tag the corpus. Two thirds of the corpus were tagged by both annotators independently, with an agreement of 0.7-0.8 (kappa score). Differences were evened out, and the guidelines revised accordingly. The last portion of the corpus was divided between the two coders. Patrizia Paggio supervised the whole annotation process.

1. TOPIC

The topic is a referent which the sentence can be said to be about, or to take its departure from. The first step in the annotation is to find the sentence topic according to the following test. If you can let the sentence under consideration be preceded by the question:

"Hvad (så) med X?" or "Hvad gør X?"

where X is replaced by a referring expression that corefers with one of the referring expressions in the sentence, then the referent of this expression is the topic. Example:

"Hvad med Peter?"
"Peter er Annas bror."
X = Peter = topic

Once the topic expression is found, tag all the words that the expression consists of with a capital "T". An example is provided at the end of this document.

2. FOCUS

Focus is meant to capture the new information in the sentence which is relevant for the building up of the discourse, that is either truly new referring expressions, or known information that is presented in a new relation, e.g.:

"Jeg mødte din søster i går".

Here, "din søster" is part of the focus because it participates in the new meeting relation. To find out where the focus starts and ends, look at the preceding context. For example, in the sequence:

"Hvem mødte du i går?"
"Jeg mødte din søster."

only "din søster" is focused, and the rest is new. Very often, however, things are not so clear-cut. The prosodic contour may also be helpful in determining the focus domain. For instance in the following sentence, the final adjunct has a weaker accent showing that it is outside the focus domain:

"Al'banien er nu på 'randen af "borgerkrig i'følge 'udenlandske observa'tører."

Once determined the focus domain, tag all the words in the focus domain with a capital "F".

3. BACKGROUND

In addition to topic and focus, a sentence will often also contain background information, that may be given or new, but is in any case peripheral, and serves the purpose of specifying the way in which eithr the topic or the focus are to be understood. Relative clauses, subordinate clauses and adjuncts (adverbials and PPs) are often part of the background.

The background does not have to be coded.

4. PRINCIPLES

4.1 GENERAL

Not all sentences have a topic.
All sentences have a focus.
Topic and focus are disjoint.

4.2 FOCUS CONSTITUENTS

The focus need not be coincidental with a sentence constituent.

4.3 FOCUS AND STRESS

There must be at least one main accent in a focus domain. There may be several.

4.4. NON-FOCUS AND STRESS

Words that are not part of the focus need not be deaccented.

4.5 DISCONTINUOUS FOCUS

We allow discontinous foci, that is foci that "contain" a topic or other background, e.g.:

"læg/F den/T der/F"

5. HEURISTICS

5.1 INITIAL SENTENCE

The first sentence in a text is always to be coded as entirely in focus, including conjunctions, that are usually not coded (see below):

"der/F var/F en/F gang/F en/F dronning/F som/F ønskede/F sig/F en/F datter/F".

5.2 ADVERBIAL SUBORDINATE CLAUSES

Two types are distinguished:

- If the clause is segmented as an independent clause, its focus and possibly topic must be annotated.

- If the clause is not segmented as an independent clause, it can either constitute background information, in which case it is not annotated, or it can be part of the focus, in which case it will all be annotated as focus.

5.3 COMPLEMENT CLAUSES IN EPISTEMIC CONSTRUCTIONS

They often express the main content of the sentence in which they occur. In this (default) case, they contain topic and focus of the overall sentence:

"det vil sige at den/T er/F gul/F"

However, the epistemic matrix predicate may be focused (emphatic stress):

"jeg ,,tror/F den er gul"

5.4 CLEFTS

In a cleft, the focus is distributed between the cleft head and the tail:

"det er den/F der er længst/F"

5.5 LEFT DISLOCATION

The dislocated referent is coded as focus, the resumptive pronoun as topic:

"vinduet/F det/T skal også have/F et/F rullegardin/F"
"efter/F banken/F der/T drejer/F du til/F højre/F ad/F Søndergade/F"

5.6 INITIAL ADVERBIALS

The corpus contains a large number of utterances where the first constituent is a locational PP containing a given referent that can be considered the topic of the sentence. Examples:

"Ovenover er der en grøn cirkel. Og oven over den grønne cirkel er der en lilla trekant."

In some examples, we code the NP that expresses the topic, but not the whole prepositiomal phrase, as the topic:

"og oven over den/T grønne/T cirkel/T er der en/F lilla/F trekant/F"

In other cases, the PP can contain a focused element, as in:

"på/F din/F venstre/F hånd/F efter biblioteket/T ser du så slottet/F"

5.6 DETERMINERS

Determiners are coded in the same way as the rest of the NP even when followed by a pause.

5.7 REPETITIONS

They are normally not treated in isolation but grouped together with the rest of the constituent, e.g.:

"en/F en/F rød/F en/F rød/F trekant/F"

5.8 WORD FRAGMENTS

They are not tagged, even when they occur in repetitions:

"der er en s- der er en/F sti/F

5.9 PAUSES

Pauses are not tagged even when they occur in between tagged words.

5.10 CONJUNCTIONS

Conjunctions are normally not tagged. This concerns both coordinating and subordinating sentence conjunctions and coordinating conjunctions used withing phrases. There may be exceptions to this principle, e.g. if the conjunction is contrastively stressed:

"Jeg sagde OG, ikke ELLER".

The conjunction "om" in indirect interrogative sentences, however, is tagged with F, since it expresses the polarity of the question, which is (part of) the focus. Furthermore, conjunctions are coded with F when they occur in initial utterances (see earlier).

5.11 INTERROGATIVES

In wh-interrogatives, the focus is taken to be the interrogative pronoun itself. In polar interrogatives, we consider the whole sentence focused.

5.12 VERBS

Verbs that are presupposed in the context are not be included in the focus domain. However, in cases of doubt, a wide focus should be preferred over a narrow focus.

6. EXAMPLE

In the sentence "den lilla trekant er i midten", I want to tag "den lilla trekant" as the topic and "i midten" as the focus. The input textgrid fragment will look as follows:

```
intervals [33]:
    xmin = 15.750941788764642
    xmax = 15.87216227487791
    text = "den_"
intervals [34]:
    xmin = 15.87216227487791
    xmax = 16.04714220426846
    text = "l,illa_"
intervals [35]:
    xmin = 16.04714220426846
    xmax = 16.541452332956027
    text = "tr,ekant_"
intervals [36]:
```

```
        xmin = 16.541452332956027
        xmax = 16.71131240449736
        text = "+_"
    intervals [37]:
        xmin = 16.71131240449736
        xmax = 16.907097364916282
        text = ",er_"
    intervals [38]:
        xmin = 16.907097364916282
        xmax = 17.003183142430277
        text = "i_"
    intervals [39]:
        xmin = 17.003183142430277
        xmax = 17.39400734329822
        text = "m,idten_"
```

After the coding, the same fragment will look as follows:

```
    intervals [33]:
        xmin = 15.750941788764642
        xmax = 15.87216227487791
        text = "den_T"
    intervals [34]:
        xmin = 15.87216227487791
        xmax = 16.04714220426846
        text = "l,illa_T"
    intervals [35]:
        xmin = 16.04714220426846
        xmax = 16.541452332956027
        text = "tr,ekant_T"
    intervals [36]:
        xmin = 16.541452332956027
        xmax = 16.71131240449736
        text = "+_"
    intervals [37]:
        xmin = 16.71131240449736
        xmax = 16.907097364916282
        text = ",er_"
    intervals [38]:
        xmin = 16.907097364916282
        xmax = 17.003183142430277
        text = "i_F"
    intervals [39]:
        xmin = 17.003183142430277
        xmax = 17.39400734329822
```

text = "m,idten_F"

Sometimes a text field contains more than one word, as in:

intervals [291]:
xmin = 105.4296935919981
xmax = 105.66172311172834
text = "der_,er_"

In such a case, each word must be tagget, e.g.:

intervals [291]:
xmin = 105.4296935919981
xmax = 105.66172311172834
text = "der_F,er_F"

6. TEXTGRIDS AND TEXT FILES

Although topic and focus must be tagged in the text grids, text files are available to support comprehension of the monologues/dialogues. In the texts, accents have been deleted, but pauses kept. Sentence boundaries, that are indicated in the textgrids by the tag "boundary", are shown as line shifts.

Example:

nederst er der en blå firkant +

og= + ovenover + er der en + grøn cirkel =

og oven over den grønne cirkel =er der en + lilla trekant +

Here is an explanation of the notation for pauses and some other stuff (in Danish):

"+" er pause (op til tre plusser for en meget lang pause)

"=" er tøven (op til tre lig-med-tegn)

"{ord1 ord2}" ordene er realiserede som et sammentrukkent ord,
ofte "dar"

"{{ord1 ord2}}" ordene er ikke realiserede som et
sammentrukkent ord, men det kunne altså godt være sket.

Acknowledgements

This work was supported by the Carlsberg Foundation.