DK-CLARIN WP 5.1-arbejdspapir
Jørg Asmussen og Lene Offersgaard

Version 2.0 – 21. september 2008

Resumé

Nærværende papir giver en oversigt over centrale korpuslingvistiske begreber. Det skal dermed sikre en ensartet brug og forståelse af beskrivelser inden for DK-CLARIN WP 5.1, der omhandler korpusarbejde.

Definitioner

I det følgende gives en oversigt over centrale korpuslingvistiske begreber, som skal sikre en ensartet brug og forståelse af beskrivelser inden for DK-CLARIN WP 5.1, der omhandler korpusarbejde, indtil videre arbejdspapiret [Asmussen og Offersgaard, 2008].

De følgende definitioner er ordnet således, at der gås fra det overordnede, grundlæggende, til det mere specifikke, fra helheden til detaljerne. De afspejler i sagens natur især de organisatoriske principper for korpusarbejdet hos DSL. Helt overordnet skelnes mellem koncepterne *tekstsamling* og *korpus*, jf. [Kennedy, 1998, p. 4] og [Leech, 1991]:

> »A distinction is sometimes made between a corpus and a text archive [...] Whereas a corpus designed for linguistic analysis is normally a systematic, planned and structured compilation of text, an archive is a text repository, often huge and opportunistically collected, and normally not structured. It is generally the case, as Leech (1991: 11) suggested, that 'the difference between an archive and a corpus must be that the latter is designed or required for a particular "representative" function'.«

Tekstsamling/-arkiv (text collection/archive/repository): Samling af hele tekster eller tekstuddrag (samlebetegnelse *tekstenheder*), som et projekt eller en institution indsamler. Det er ikke på forhånd givet, at tekstenhederne i en sådan samling nogensinde faktisk behøver at komme til at indgå i et konkret korpus. Formålet med samlingen kan således også være dokumentation eller arkivering. En sådan samling er ofte resultatet af en rutinemæssig indsamling, hvor der som udgangspunkt ikke er defineret et konkret mål med hensyn til, hvilket korpus man ønsker at opbygge på baggrund af teksterne, endsige om man overhovedet vil opbygge et korpus. En sådan indsamling er ofte opportunistisk: Man tager det, der byder sig og overvejer så sidenhen, hvad det skal/kan bruges til. Tekstenhederne i en tekstsamling har et veldefineret minimum af annotationer på tekstniveau (metadata). Hvilke metadata der bruges,

afhænger af pågældende tekstsamlings art, dog kan metadata fastlægges på en måde, så tekstsamlingen principielt kan indeholde tekster af en hvilken som helst type, fx fra forskellige perioder eller på forskellige sprog. Tekstenhederne i en tekstsamling kan desuden være tokenopdelte, og de kan have flere niveauer af annotationer på tokenniveau. På baggrund af annotationerne på tekstniveau kan der udvælges en gruppe af tekstenheder, der kan eksporteres fra samlingen i forskellige formater, fx TEI eller CQP-importformatet. En sådan gruppe af tekstenheder, bevidst sammenstillet til et ganske bestemt formål, er et korpus.

Tekstbank (text bank): Mens tekstsamlingen er et teoretisk koncept, så er tekstbanken dets konkrete udformning: Den er med andre ord tekstsamlingen indlejret i en applikation til håndtering af den. En tekstbank vil således typisk være en database, der indeholder tekstsamlingens tekstenheder sammen med deres metadata. Tekstenhederne selv bør i tekstbanken være tokenopdelte, og hvert token bør her have et unikt reference-id, så det kan adresseres entydigt. En tekstbank stiller derudover en række »håndtag« til rådighed, dels til tilføjelse af metadata på tekstniveau eller annotationer på tokenniveau, dels til udvælgelse af en mængde tekster, der tilfredsstiller en bestemt metadataprofil (dvs. et korpus), og til eksport af dem i et ønsket format. Der kan ikke foretages egentlige sprogvidenskabelige undersøgelser vha. tekstbanken: Man kan kun søge på tekstannotationsniveau og kun i meget begrænset omfang på tokenniveau.

Korpus (corpus): En gruppe af tekstenheder fra tekstsamlingen, bevidst sammenstillet ud fra eksplicitte kriterier på baggrund af oplysningerne i tekstenhedernes metadata med det formål at kunne udføre bestemte sprogvidenskabelige undersøgelser, idet det antages, at tekstenhederne tilsammen, altså korpusset, udgør en repræsentativ stikprøve for den type sprog, der skal undersøges. Til de enkelte tekstenheder i et korpus knytter der sig i reglen de samme metadata, som også bruges i tekstsamlingen, eller en delmængde heraf, fx kun de relevante for det pgl. korpus. Til et korpus kan der knyttes metadata, der beskriver dets karakteristika, dvs. hensigten med det og udvælgelseskriterierne for teksterne i det. Rent praktisk vil et korpus typisk være udtrukket/eksporteret fra netop én tekstsamling, idet forskellige tekstsamlinger kan have forskellige typer af metadata, der stiller sig i vejen for en ensartet tilgang til tekstenhederne i dem. Et korpus med et eller flere annotationslag kan gøres søgbart i et konkordansværktøj, eller det kan processeres af andre applikationer, fx til statistiske formål.

Konkordansværktøj (concordancer): Den rolle, som en tekstbank spiller i forhold til en tekstsamling, spiller et konkordansværktøj i forhold til et korpus. Vha. et konkordansværktøj er det muligt at udføre avancerede sprogvidenskabelige undersøgelser i et korpus og få resultatet præsenteret i form af såkaldte KWIC-konkordanser.[1] Et udbredt og hidtil uovertruffent konkordansværktøj er CQP, som er en del af *The IMS Open Corpus Workbench*, jf. http://cwb.sourceforge.net. CQP bruges af DSL, CST og DSN. INSS bruger åbenbart ikke noget konkordansværktøj endnu.

[1] KWIC står lidt upræcist for *keyword in context*, ofte vil der dog snarere være tale om en *gruppe* af ord, der matcher et søgeudtryk, som brugeren har formuleret.

Statistikværktøj (statistics tool): Bruges til at lave statistiske undersøgelser med i et korpus, fx frekvensundersøgelser (kan normalt også klares af et konkordansværktøj) eller mere avancerede kollokabilitetsundersøgelser efter forskellige statistiske beregningsmetoder som fx *mutual information*, *t-score*, *log-likelihood*,[2] der ofte søger at give svar på, hvilke (grupper af) ord (eller andre tekstfænomener) der optræder signifikant hyppigt sammen med andre (eller netop ikke gør det), eller hvorved et korpus adskiller sig fra et andet, eller en bestemt tekstenhed fra et bestemt korpus. Ofte kan statistikværktøjer bruges som en mere struktureret indgang til konkordanssøgning, især ved højfrekvente fænomener.

Annoteringsværktøj (annotation tool): En applikation, der hjælper en bruger med at annotere et korpus, enten manuelt, tekstord for tekstord, eller som en automatisk, algoritmisk proces, fx i form af en ordklassetagger.

Tekstenhed (text unit): Enten en (kortere) heltekst eller et uddrag af en heltekst. Tekstsamlinger, som opbygges til korpusformål, vil normalt underopdele lange tekster, fx romaner, i mindre enheder, så et korpus kan komponeres mere finkornet, altså balanceres bedre, så det ikke får »overvægt« af nogle få lange tekster. En tekstsamling, der skal kunne bruges til korpusformål, er derfor opdelt i tekstenheder af overskuelig længde.

Tekstord (running word): Sekvens af bogstaver og/eller cifre, som står mellem mellemrum og interpunktionstegn i en tekst.

Token (token): Tupel bestående af et tekstord og andre attributter. Efter tokenopdelingen består et token i DSL's tekstbank af tekstordet selv (attributnavn *ortho*), en lettere normaliseret version heraf (*word*),[3] interpunktionstegn, som står efter tekstordet (*punct*), mellemrum og tegn (fx anførselstegn), som står foran tekstordet (*space*), id'et på den tekstenhed, som det tilhører (*id*) og tokenindekset inden for pgl. tekstenhed (*tix*). Tilsammen sikrer tekst-id'et og tokenindekset, at hvert eneste token i tekstbanken entydigt kan identificeres. Hver annotation, som tilføjes på tokenniveau, udgør et yderligere attribut i tuplet. Teksten *Han siger, at de viste »KB-agenten« bort.* bliver således tokenopdelt i

```
('Han', 'han', '', '', 4800012345, 0)
('siger' 'siger', ',', ' ', 4800012345, 1)
('at', 'at', '', ' ', 4800012345, 2)
('de', 'de', '', ' ', 4800012345, 3)
('viste', 'viste', '', ' ', 4800012345, 4)
('KB-agenten', 'kbagenten', '«', ' »', 4800012345, 5)
('bort', 'bort', '.', ' ', 4800012345, 6)
```

Visse tokenizere giver også interpunktionstegn tokenstatus. Imidlertid kan det være problematisk ved søgeforespørgsler i et konkordansværktøj. Fx ville søgeforespørgslen (udtrykt i CQP's formalisme)

```
[word=''siger''][word=''at'']
```

[2] Jf. fx [Kilgarriff, 2001] for en sammenlignende evaluering.
[3] En almindelig konkordanssøgning i DSL's korpora udføres som default på dette attribut.

i så fald ikke kunne finde ovenstående teksteksempel, men skulle i stedet formuleres som

```
[word=''siger''][word='',''][word=''at'']
```

eller for at finde eksempler med med og uden komma efter *siger* som

```
[word=''siger''][word='',''][word=''at'']
```

– en unødig komplikation, der på mange brugere nok ville virke kontra-intuitiv.

Litteratur

[Asmussen og Offersgaard, 2008] Asmussen, J. og Offersgaard, L. (2008). Korpus-workflow. Rapport, DK-CLARIN.

[Kennedy, 1998] Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. Longman, London.

[Kilgarriff, 2001] Kilgarriff, A. (2001). Comparing Corpora. *IJCL*, 6(1):97-133.

[Leech, 1991] Leech, G. (1991). The state of the art in corpus linguistics. I Aijmer og Altenberg, redaktører, *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. Longman, London.