Udnyttelse af korpora ved leksikografisk arbejde

1. Tekstkorpora som redskab ved udarbejdelsen af traditionelle ordbøger
2. Brugen af "råt" tegnsprogskorpus i Ordbog over Dansk Tegnsprog

Noter fra præsentation ved seminaret

Teckenspråken i Norden – Korpusarbete

Stockholm, 7-8 november 2008

Thomas Troelsgård
–
Ordbog over Dansk Tegnsprog
Center for Tegnsprog og Tegnstøttet Kommunikation - KC
Professionshøjskolen UCC

Tekstkorpora som redskab ved udarbejdelsen af traditionelle ordbøger

Korpus kan bruges til at finde:

- betydninger
- faste udtryk
- eksempelsætninger

Eksempler på tekstkorpora:
Språkbanken (svensk) http://spraakbanken.gu.se/konk/

Eksempel: Enkel søgning: ***plätt***

Se ned over konkordansen og find:

- *betydning*: 'lille kage'
- *betydning*: 'lille område'
- *fast udtryk*: "lätt som en plätt"

Der er kun søgt på den nøjagtige streng "plätt". Hvis bøjningsformer skal medtages, kan man søge med trunkering: ***plätt*,*** dog med den "bivirkning" at også sammensætninger og andre ord der måtte begynde på 'p-l-ä-t-t-' kommer med.
I grundigt taggede korpora kan dette undgås, ved at man kun søger på et bestemt ord og dets former.

Klik i feltet "frekvens" og opdater søgningen.
Der vises da en frekvensliste over alle de former der er medtaget i søgningen. Derved kan man hurtigt få et overblik over konkordansens indhold og fordelingen mellem de forskellige former.

Søg på: ***"lätt + plätt" :3*** (tekster hvor 'lätt' og 'plätt' forekommer med max. 3 ords afstand).
Herved findes forekomster på udtrykket "lätt som en plätt":

I et større korpus ville man også kunne finde varianter af faste udtryk ved at søge på f.eks. "lätt som en". På internettet findes således også: "lätt som en tvätt" og "lätt som en omelett".

KorpusDK (dansk) http://ordnet.dk/korpusdk
Dette korpus har tilknyttet en lang række ekstrafunktioner. F.eks. kan man ´sortere konkordanser efter venstre eller højre kontekst, og man kan reducere konkordansens størrelse, hvilket er nyttigt ved undersøgelse af højfrekvente ord.

Man kan også få vist hyppige omgivelser/naboord:

Eksempel: Søg på *plet (= fläck)*
Klik på "Naboord" i venstre kolonne:

Venstre kontekst

	ordform	ordklasse	score	signifikans
1	blinde	adj	10.68	
2	brune	adj	10.21	
3	røde	adj	10.19	
4	lysende	adj	9.93	
5	rammer	v	9.93	
6	våd	adj	9.88	
7	våde	adj	9.64	
8	blind	adj	9.62	
9	bare	adj	9.62	
10	mørk	adj	9.59	
11	mørke	adj	9.37	
12	hvide	adj	9.33	
13	sorte	adj	9.31	
14	rød	adj	9.31	
15	ramt	v	9.22	
16	ramte	v	9.22	
17	ramme	v	9.20	
18	sort	adj	9.12	
19	gule	adj	8.99	
20	lyse	adj	8.62	
21	fjerne	v	8.15	

Højre kontekst

	ordform	ordklasse	score	signifikans
1	landkortet	n	13.19	
2	kinderne	n	11.28	
3	huden	n	9.32	
4	halsen	n	8.15	
5	jord	n	7.54	
6	midt	adv	6.07	
7	på	prp	4.98	
8	når	ks	4.39	
9	hvor	adv	3.97	
10	som	indp	3.90	
11	hans	pers	3.64	
12	mellem	prp	3.61	
13	eller	kc	3.53	
14	der	indp	3.50	
15	da	ks	3.38	
16	i	prp	3.16	
17	sin	det	2.85	
18	af	prp	2.81	
19	med	prp	2.79	
20	den	art	2.64	

På denne måde kan man finde udtryk som "blind plet", "gul plet" og "hvide pletter på landkortet". Ved at klikke på ordet i kolonnen "ordform", kan man se en konkordans over forbindelsen af "plet" og det pågældende ord.

Den Danske Ordbog, en seksbindsordbog over moderne dansk, er korpusbaseret. Det vil bl.a. sige at betydningsrækkefølgen er påvirket af deres frekvens i korpus, og at de fleste eksempelsætninger er hentet fra korpus. En del af ordbogens korpus indgår i KorpusDK, og man kan således genfinde mange af ordbogens eksempler i KorpusDK:

Udsnit af højresorteret konkordans over "plet":

øjet rakte. Høje graner, der stod så tæt, at kun **pletter** af flimrende sollys formåede at kaste et gyldent skær mellem

præsident altid sørgede for at vælge folk, der var **plettet** af fortiden, så han kunne presse dem. "Folk i Vesten

ikke undgås, at især små børn, der leger ude, får **pletter** af græs på tøjet. Men hvis de fugtes med opvaskemiddel

sommerens rekordregn lyste græsset saftiggrønt, med **pletter** af gult: blade fra små grupper birke. Ingen mennesker var

at de var gået i stykker- en anden ville rense **pletter** af hans bukser. De pleier ham rigtigt, fordi mange af

Udsnit fra artiklen *plet*[1] (= 'fläck' i delbetydelsen "litet nedsmutsat ställe") i Den Danske Ordbog:

med.92

• mindre, snavset område på tøj, møbler e.l., fx med en rest af noget man har spildt □ *det kan ikke undgås, at især små børn, der leger ude, får pletter af græs på tøjet* Femina84

definition

exempel-mening

Brugen af "råt" tegnsprogskorpus i Ordbog over Dansk Tegnsprog

Ordbog over Dansk Tegnsprog er ikke korpusbaseret, men ved udarbejdelsen af ordbogen har vi anvendt videomateriale som hjælpemiddel ved forskellige redaktionelle opgaver.

Videooptagelser af informanter er brugt som ”råt”, ikke-transskriberet korpus til at finde:

- tegn
- varianter
- eksempelsætninger

Tegn

Eksempel: Til brug for undersøgelse af bynavnetegn har vi optaget informanter der sidder foran et Danmarkskort og diskuterer hvilke tegn de kender for de forskellige byer.

Varianter

Ved variantundersøgelser har vi bl.a. bedt informanterne om at læse en række titler (bøger, film m.v.) og derefter sige titlerne på tegnsprog. Titlerne var udvalgt så der i nogle af dem var ord der med stor sandsynlighed ville resultere i et bestemt tegn der varierer i formen. F.eks. indgik eventyret "Den lykkelige familie". Alle informanter brugte samme tegn for 'lykkelig', men med tre forskellige håndformer:

Eksempelsætninger

Ved hvert informantmøde blev informanterne bedt om at indtale en "dagbog" på ca. 5 min. Dagbøgerne blev siden gennemset dels med henblik på at finde tegn og betydninger, dels for at finde naturlige brugseksempler, der kunne bruges i ordbogen. Eksemplerne blev derefter tilpasset, så de kunne forstås ude af sammenhængen, og evt. anonymiseret.

Leksikografiske muligheder med et ”rigtigt”, søgbart tegnsprogskorpus

Ved den fremtidige udvikling og opdatering af Ordbog over Dansk Tegnsprog vil et "rigtig", annoteret tegnsprogskorpus kunne udnyttes til en lang række opgaver:

- Kontrol af tegnforrådet – Har vi alle de centrale og hyppige tegn med?
- Mere præcis og udtømmende beskrivelse af tegnenes betydning(er).
- Bedre overblik over tegnvariation.
- Nem adgang til autentiske eksempel-sætninger.
- og meget andet...

I den kommende udgave af ordbogen regner vi med at vise eksempelsætningerne opstillet som konkordans, således at man kan se alle forekomsterne af et bestemt tegn i sætningerne, og ikke kun de sætninger der er placeret i den pågældende artikel. Eksempel på konkordans over eksempelsætninger:

PEG VIDENSKAB FÆRDIG ANALYSERE FLY FÆRDIG LANDE~1 PROFORM MEN GRÆS
BOG TO BIND FØRST BIND FÆRDIG ANDEN ENDNU-IKKE
LÆSE BA(HA) FÆRDIG ANDRE FORTSÆTTE OVERBYGNING
PEG SIGE PEG PEG FÆRDIG BEDØVE PEG
NU JEG REN SAMVITTIGHED JEG FÆRDIG BESØGE FAR
SPISE-MIDDAG FÆRDIG BRUN DESSERT
SPARK KAMP FÆRDIG DANMARK VINDE EN NUL DANMARK MED VM
NY LEJLIGHED PEG TILFÆLDIG PEG NY PEG BYGGE FÆRDIG ENDNU-IKKE TO KØBE PEG
JEG HUS HAFT STJÆLE PRÆSENTATIONSGESTUS TYV FÆRDIG FANGE
KAMP JEG SPILLE-BRÆTSPIL KAMP REGN KAMP FÆRDIG EIN VEJR

Bilag: Fakta om Ordbog over Dansk Tegnsprog

The Danish Sign Language Dictionary

The dictionary is (as of May 2008) freely accessible at: www.tegnsprog.dk

The dictionary was developed and edited at the Centre for Sign Language and Sign Supported Communication – KC in close cooperation with the Danish Deaf Association (DDL).

The editorial staff can be contacted through: info@tegnsprog.dk

Main features

The entries

The sign lemmas are represented through:

- A video clip of the base form (and potentially of one or more variants). A photo.
- A unique gloss, which represents the sign throughout the dictionary, regardless of its actual meaning in a given context.
- Drawings of the first location and handshape occurring in the description of the sign's manual features.

Clickable entry-level cross-references are given to:

- homonyms
- examples of lexicalised classifier verbs (referred to from classifier entries)

Polysemous signs are subdivided into separate meanings. Meanings with a semantically opaque relation to the basic meaning of the sign, are considered homonyms, and established as separate entries.

Common sign compounds, which meaning cannot be deducted as the sum of the meanings of the signs comprised in the compound, are described as separate meanings.

Each meaning is described through one or more Danish equivalents. Equivalents that can be used as mouthing for the sign in the particular meaning are marked with a special mouth symbol.

If no appropriate equivalent can be given, or if the equivalents do not fully cover the meaning of the sign, a prose description of the sign's use or function is provided.

Non-Danish mouth movements that often co-occur with the sign in a particular meaning are listed below the equivalents.

Clickable meaning-level cross-references are given to:

- synonyms
- short forms / long forms
- related number incorporations
- signs with a similar Danish equivalent, but a different meaning (the equivalent is homonymous or strongly polysemous).

A meaning can be accompanied by information about special restrictions in its use.

For the majority of meanings one or more usage example is provided.
The examples are rendered as video clips accompanied by glosses and Danish translations.

Both glosses and translations are searchable through text search.
All glosses that match dictionary entries are clickable (link to the appropriate entries).
Most of the usage examples are elicited from video recordings of a group of Sign Language consultants (all native signers) affiliated to the dictionary project.

Search facilities
The signs can be looked up through:

- Handshape (with a possibility of specifying particular handshapes for the active and the passive hand).
- Location
- Text, including phrase search ("...") and wildcard search (* and ?).
- Topic
- Combinations of the above.

Handshapes to be used in the search criteria are selected in groups of related handshapes. However, selected handshapes can be deselected individually, which gives the user the opportunity to select single handshapes or custom groups of handshapes, independently of the default handshape grouping.

The search result is by default ordered by relevance: Handshape and location matches are ordered according to their appearance in the sign. Text search matches are weighted in the following order: glosses, equivalents, glosses in usage examples, words in translations of usage examples. Matches with equal relevance are ordered first by location, then by handshape. Furthermore, the user can choose to sort the entire search result either by location or handshape.

Approximate key figures

2.000 sign entries.
Due to limited resources and a strict project time frame, the first edition of the dictionary describes only the basic vocabulary of Danish Sign Language.
Among the entries are special entries for the following common "building blocks" of Danish Sign Language:

- Classifiers (49 entries)
- The manual alphabet (29 entries)
- The Mouth-Hand-System (14 entries)
- Affixes (6 entries)

500 variants.
The variants are shown as secondary video clips under the base form entries. All recorded variants are included in searches on manual features.

75 entry-level cross-references.

3.000 meanings.

6.000 Danish equivalents.

150 prose descriptions of special use or function.

2.500 meaning-level cross-references.

3.500 usage examples.
The examples contain about:

- 3.400 glosses (27.500 running glosses)
- 7.000 words in the translations (44.500 running words)