NAVF'S EDB-SENTER FOR HUMANISTISK FORSKNING
VILLAVEI 10
POSTBOKS 53
5014 BERGEN - UNIVERSITET

Knut Hofland:

PRODUKSJON AV EN LEMMATISERT KONKORDANS TIL IBSENS SAMLEDE VERKER.

Ved NAVF's EDB-senter for humanistisk forskning er det i gang et prosjekt der målet er en lemmatisert konkordans til Ibsens skuespill og dikt. Prosjektet startet 1. april 1978 og skal gå over 3 år. Professor Harald Noreng er ansatt som faglig leder i hel stilling og konsulent Knut Hofland har ansvar for EDB-delen av prosjektet. Norges almenvitenskapelige forskningsråd finansierer tiltaket. Prosjektet disponerer ½ stilling til terminalarbeid og midler til timelønnet personale for vitenskapelig assistanse og til korrekturlesing og kontrollarbeid. All systemutvikling og programmering utføres av den faste staben ved NAVF's EDB-senter.

Tekstgrunnlaget er hundreårsutgaven av Ibsens verk og består av 28 skuespill derav 4 i to utgaver og et bind med dikt. Dette utgjør ca 3/4 million løpende ord.

Publiseringsmåten for konkordansen vil bli på mikrokort og det vil også bli aktuelt med en trykket utgave av en del av materialet, et sitatleksikon.

Ved tilrettelegging av teksten er det ført inn opplysninger om hvem en replikk er rettet til, type av scenehenvisning og markering av enderim etter type, der skuespill er på verseform.

Ibsen skifter rettskriving i 1870. Ved lemmatiseringen bruker en skrivemåten etter 1870 som grunnlag for oppslagsformen. Som en hjelp til brukeren av konkordansen (særlig i utlandet) innføres det ett sett med henvisninger. Dette gjelder fra moderne norsk rettskriving til Ibsens, fra Ibsen før 1870 til etter 1870 og ved sterk bøying fra ordform til oppslagsform.

Det er lagt opp til at hvert belegg kan få en nøye tilpasset kontekst. Deler av konteksten kan utelates og det kan også tilføyes opplysninger som f.eks. henvisning til pronomene. I tillegg til referansen vil det for hvert belegg være gitt opplysninger om hvem som har replikken og til hvem den er rettet.

Lemmatiseringen.

Tekstene lemmatiseres verk for verk. Den første teksten ble lemmatisert manuelt ut i fra en konkordans med 3 linjers kontekst.Det var her forslag til kontekst avgrensing basert på skilletegn. Ved de to neste verkene ble opplysningene fra det/de foregående verk stilt opp som forslag (se eksempel).

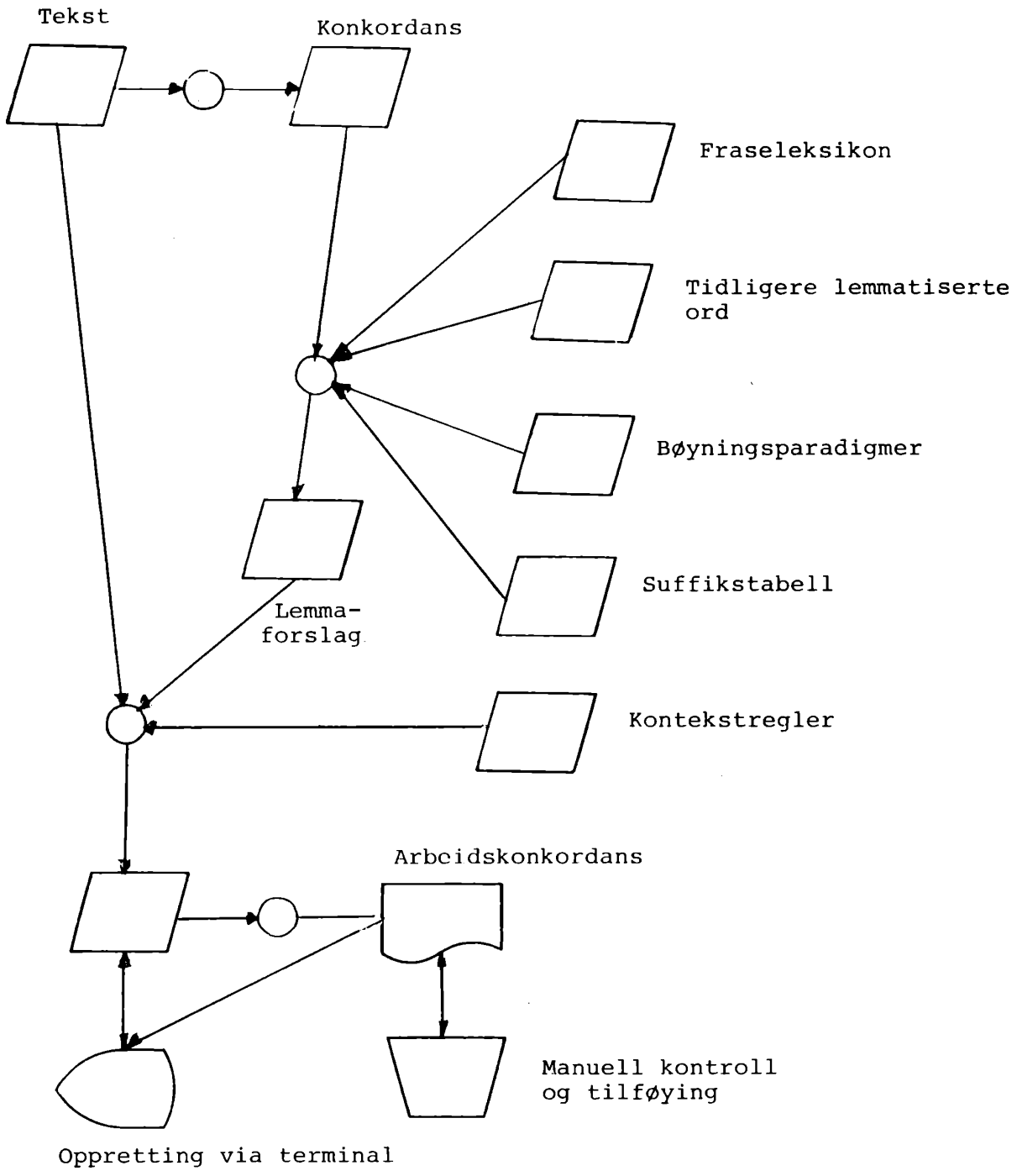
Produksjon av en lemmatisert konkordans til Ibsens samlede verker
Knut Hofland
Proceedings of NODALIDA 1979, pages 133-136

```
399 ********************************************************************************** 0 umeldt **********************
400 L= 02685, O= 01, A= 001, S= 0061, FRA= 15, TIL= 1, SH-NR=                          GF umeld
401    bitter Klage fra Din Læbe - / |EN GAMMEL SOLDAT (træder ind og nærmer sig Catilina) <1>   OK adj.
402  = Tilgiv mig, Herre! at jeg træder ind /  _umeldt_  ved Aftenstid her i din        BØ n. sg., ubest.   a)
403 Bolig; - / =  Tilgiv - - - |CATILINA. <15>  - - Sig frem! hvad vil Du her? /       BR adv.
404 ********************************************************************************** 0 umild **********************
405 L= 01286, O= 04, A= 002, S= 0077, FRA= 6, TIL= 7,8,9 & 10, SH-NR=                  GF umild
406       hører først, hvad jeg / vil Eder sige;  = - vi har os besluttet / med væbnet Haand   OK adj.   a)
407      at ville vinde, hvad / en haard, en  _umild_  Skjæbne har os nægtet; / =  nu   BØ c. sg., ubest.
408 undertrykkes vi, - vi ville herske; / vi lide Mangel, - Rigdom er vort Maal! / |PLEBE.
409 ********************************************************************************** 0 umuligt *********************
409 #1  O: umuligt  GF: umulig  OK: adj.  BØ: n. sg., ubest.  BR:  MO:
410 #2  O: umuligt  GF: umulig  OK: adj.  BØ: n. sg., ubest.  BR: adv.  MO:
412 L= 02034, O= 10, A= 003, S= 0097, FRA= 1, TIL= 4, SH-NR=
413    - / |CURIUS (fortvivlet) <1> Nei, nei, Du er forraadt, din Plan er røbet. / |CATILINA.
414   <4> Ha, raser Du?  = Nei, nei, det er  _umuligt!_  / =  |CURIUS. <1> Jo det er
415 saa, men fly i denne Stund, / end kan maaskee ved hurtig Flugt Du reddes. / |CATILINA.
416 ---------------------------------------------------------------------------------- 0 umuligt
417 L= 02037, O= 13, A= 003, S= 0097, FRA= 1, TIL= 4, SH-NR=
418      men fly i denne Stund, / end kan maaskee ved hurtig Flugt Du reddes. / |CATILINA. <4>
419    = O, skulde det - - - men nei, det er  _umuligt!_  - / =  |CURIUS (river sin Dolk
420 ud og rækker den hen til Catilina.) <1> Der, Catilina! tag, igjennemboer mit Bryst - / -
421 ********************************************************************************** 0 under **********************
421 #1  O: under  GF: under  OK: prep.  BØ:  BR:  MO:
422 #2  O: under  GF: under  OK: adv.  BØ:  BR:  MO:
423 #3  O: under  GF: unde  OK: verb  BØ: pr.  BP:  MO: unne
425 L= 00267, O= 02, A= 001, S= 0049, FRA= 1, TIL= 4, SH-NR=
426  det er vel Spøg? / |CATILINA. <4> En Spøg? ja vist! som al min Kjerlighed, / men Alvor er
427   dog, hvad jeg nys Dig sagde; - /  = da  _under_  sidste Fest igjennem Romas Gader /
428 i Procession Præstinderne jeg saae, / tilfældigviis paa een af dem jeg fæsted' / et  =
```

a) ny klassifikasjon
b) forslag til klassifikasjon rangert etter hyppighet i tidligere verk
c) ny kontekstavgrensing

Utsnitt av arbeidskonkordans

Oversikt over lemmatiseringen.



Ved det videre arbeid vil det i økende grad bli tatt i bruk maskinell lemmatisering. Metoden som skal brukes er tradisjonell og bygger bl.a. på arbeid som har vært gjort andre steder (1), (2), (3).

En oversikt over metoden er gjort i figuren. Metoden er basert på at forskeren skal kunne arbeide med konkordanser utskrevet på papir. På bakgrunn av markeringer i konkordansen gjøres rettinger/ tilføyelser via terminal.

På grunnlag av de tekster som er lemmatisert blir det laget maskinelt et sett med tabeller som brukes ved den automatiske lemmatiseringen. Den nye teksten som skal lemmatiseres kjøres først ut i en konkordans med begrenset kontekst. Denne konkordansen kjøres mot et fraseleksikon som er en serie med ordkombinasjoner som opptrer samlet med en viss frekvens. Til hvert ord i denne kombinasjonen er det grammatiske opplysninger. Ordene i konkordansen flettes sammen med ordene i den lemmatiserte ordliste. Ved likhet brukes opplysninger fra ordlisten. I tillegg blir to og to ord sammenlignet tegn for tegn. Der hvor likheten opphører starter endelsen. Disse endelsene sjekkes mot ett sett med bøyningsparadigmer for Ibsens språk. Dersom de to endelsene kan høre til samme paradigme antas det at de to ordene er bøyningsformer av samme ord.

De ordene som ennå ikke har fått grammatisk klassifikasjon blir testet mot en suffikstabell, som er en liste med (grafiske) endelser og mulige grammatiske klassifikasjoner.

Ordene i konkordansen blir så sortert tilbake til sin opprinnelige rekkefølge. Det brukes da et sett med kontekstregler for å skille homografer. Disse kontekstreglene er en sekvens av grammatiske klassifikasjoner som kan opptre samlet, eller kombinasjoner som ikke kan forekomme.

Til slutt skrives det ut en arbeidskonkordans med 3 linjers kontekst.

(1) Staffan Hellberg: Automatisk lemmatisering, Språkdata, Gøtegorg 1971

(2) Green & Rubin: Automatic Grammatical Tagging of English, Brown University, Providence, 1971

(3) Gerd Willé: LEMMA - ein programsystem zur automatischen lemmatiserung deutscher wortformen, IKP, Bonn, 1978