Eirik Lien.

DEMONSTRASJON AV PP*TT - EN PROGRAMPAKKE FOR KVANTITATIV TEKSTANALYSE

Nordiske datalingvistikdage,
København 9. - 10. oktober 1979

En god del av de kvantitative analyser og beregninger som en kan få en datamaskin til å hjelpe seg med, er svært like fra prosjekt til prosjekt. Det vil i de fleste tilfeller dreie seg om å lage listeprodukter, slik som alfabetisk ordliste (forlengs og baklengs sortert), frekvensordliste og konkordans. Sett fra en datamaskins synspunkt ligger forskjellen mellom to tekster i at de har forskjellig lengde.

Denne oppdagelsen har mange gjort, og det har igjen ført til at det rundt omkring ved humanistiske forskningsmiljøer er blitt laget en del program som produserer slike lister. Disse er gjerne i utgangspunktet laget for en spesiell tekst, men nettopp fordi forskjellen til en annen tekst ligger bare i lengden av dem, har en gjort programmene mer fleksible ved å bygge inn denne forskjellen ved hjelp av parametre. Etter hvert som behovet for nye analyser har økt, har en med utgangspunkt i det en har, lagt til nye program som fortsetter videre der det gamle programmet stanset.

Universitetet i Trondheim er ingen unntakelse. Ganske snart etter at jeg startet edb-tjenesten, så jeg behovet for slike program. I stedet for å starte med et bestemt prosjekt, startet jeg med å tenke etter hvilke analyser det kunne være behov for - og kunne derfor planlegge hele strukturen under ett.

Nærværende programpakke er derfor ment som et eksempel på hvordan systemet kan struktureres - den gjør alldeles ikke krav på å være genial og enestående. Jeg er fullstendig klar over at tilsvarende programpakker fins det mange andre steder, programpakker som gjør minst like god jobb.

Programpakken slik den er nå, gir listeprodukter på tre nivåer; ordnivå, stavingsnivå og grafemnivå. På ordnivå kan en få disse listeproduktene:
- initialalfabetisk ordliste med frekvens
- finalalfabetisk ordliste med frekvens
- frekvensordliste

Demonstration av PP*TT - en programpakke for kvantitativ tekstanalyse
Eirik Lien
Proceedings of NODALIDA 1979, pages 73-76

- konkordans
- lister basert på ordlengde

På stavingsnivå:
- alfabetisert stavingsliste med frekvens
- frekvenssortert stavingsliste
  Initiale og finale stavinger kan, med parameter, i begge disse programmene få spesielle tilleggstegn som markerer denne funksjonen og dermed skiller dem fra andre, ellers like stavinger.
- liste over stavingstyper, representert ved sifre, nemlig antall bokstaver i første konsonantgruppe, antall vokaler i stavingskjernen og antall konsonanter i andre konsonantgruppe.

På grafemnivå:
- "konkordans" - eller klic-liste ("key-letter-in-context")
- opptelling av hvilke tegnkombinasjoner som fins i teksten.
- frekvenssortering av disse kombinasjonene

Ved et referansesystem som går både på side og linjenummer og kapittel, avsnitt, periode og ordnr. i perioden, har en direkte referanse tilbake til teksten og kan f.eks. kontrastere utvalgte kapitler mot hverandre.

Systemet er hovedsaklig laget for satsvise kjøringer ("batch"). Den delen av programpakken som deler ordene opp i stavinger, er imidlertid laget for interaktiv bruk slik at en kan korrigere de forslagene til oppdeling programmet gjør.

Systemet er i laget slik at en med parametre kan velge hvilke tegn en vil oppfatte som skilletegn. En har også mulighet for å justere utskriften ved noen av listene og gi ekstrainformasjoner i enkelte tilfeller.

Programpakken er skrevet i NU ALGOL og går foreløpig bare på Universitetet i Trondheims UNIVAC-anlegg. Dette programmeringsspråket gjør at det ikke er særlig distribusjonsvennlig. Det har vært i drift siden høsten 1975, men har i år fått en del nye tillegg.

Videre planer

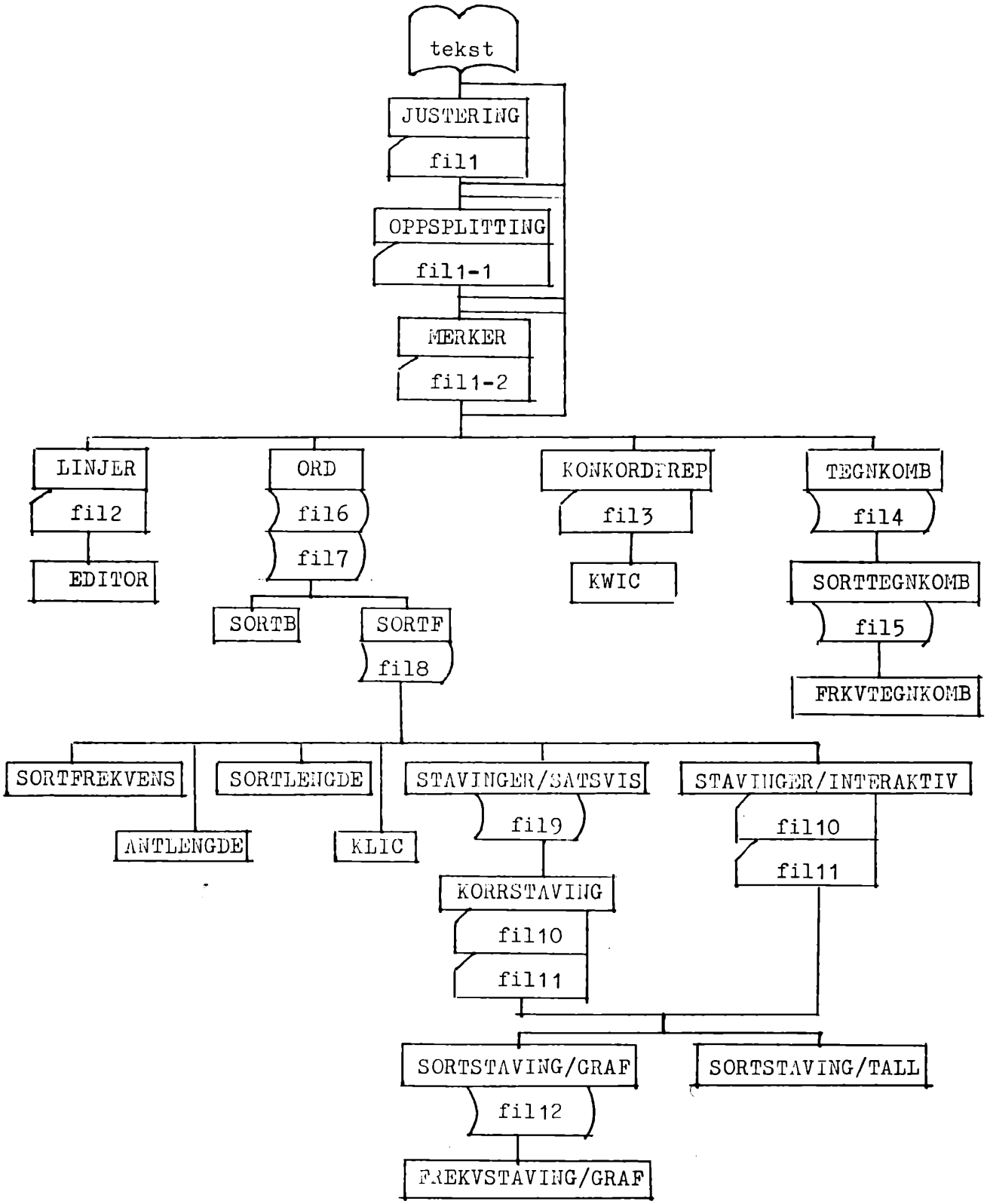
Det kan bli aktuelt å skrive om programmene til PASCAL, som ser ut til å bli det programmeringsspråket som nå slår igjennom.

Konkordansprogrammet bør bli mer fleksibelt ved at konteksten kan varieres, f.eks. til å omfatte den perioden ordet står i. Jeg har også planer om å arbeide inn program for analyse på frasenivå. Dessuten bør programmene for lemmatisering komme med i en slik programpakke. Likeså bør muligheten for å behandle merket ("tagget") tekst være med. Derimot vil nok analyse på setningsnivå ligge et godt stykke fram i tida.

Den metoden som er brukt, nemlig å ta vare på resultatene på filer, gjør det enkelt å innarbeide nye program fordi disse kan arbeide videre fra de filene som allerede er generert.

Trondheim 1979-09-17

Appendiks 1
Strukturen i PP*TT



EL/1979-10-03