Roald Skarsten
EDB-seksjonen v/HF
Univ. i Bergen

PRESENTASJON OG KOMMENTARER TIL NORSK UTGAVE AV C. MULLERS BOK "LINGVISTISK STATISTIKK"

Det er en glede å kunne presentere for deltakerne på de nordiske datalingvistikkdagene den norske utgaven av C. Mullers bok: Initiation aux méthodes de LA STATISTIQUE LINGUISTIQUE, Paris 1973. Den norske utgaven er et resultat av samarbeid mellom flere parter. Den er tilrettelagt ved NAVF's EDB-senter for humanistisk forskning i samarbeid med EDB-tjenestene for humanistiske fag ved universitetene i Oslo, Bergen og Trondheim. NAVF's EDB-senter for humanistisk forskning har i hovedsak dekket utgiftene. Undertegnede har organisert og ledet arbeidet med tilrettelegging og utgivelse av boken.

Boken er oversatt fra fransk av cand. philol. Kari Fonnes med assistanse fra amanuensis Ivar Fonnes som har hatt ansvaret for statistisk terminologi. Dessuten har universitetslektor Elsa Quale vært statistisk konsulent. Med utgangspunkt i E. Quales konsulentuttalelse har Ivar Fonnes og jeg bearbeidet oversettelsen på enkelte punkter, og noen merknader er satt inn i noter som "oversetters anmerkning" (o.a.). Konsulent Eirik Lien har lest igjennom det endelige manuskript og gitt språklige kommentarer.

Vi kan nesten si at den norske boken faktisk er bedre enn den franske originalen, fordi den er bearbeidet og utstyrt med noen oppklarende noter og fordi en del småfeil er rettet opp. Den fagstatistiske konsulenten hadde lyst til å rette opp mere enn det som er skjedd, men vi har måttet balansere mellom de ideale krav og den beskyttelse som åndsverkloven gir bokforfatteren.

Jeg antar at de av deltakerne som ikke har noen matematisk bakgrunn vil glede seg med oss over boken fordi den vil dekke et behov som tidligere har vært vanskelig å få tilfredsstilt for oss. Forfatterens intensjoner, slik han formulerer de i sitt forord, viser dette: Framstillingen er begrenset til å presentere og forklare både prinsipielt og praktisk hvordan statistiske metoder kan brukes på språklige og stilistiske fenomener. Hvilke spørsmål vi kan stille til lingvistisk statistikk og svar som statistikken kan gi er hele tiden en del av framstillingen. Slik gir boken elementære statistiske kunnskaper som kan brukes i praksis, og venner leseren til å bruke statistiske resonnementer og et algebraisk språk når det er snakk om å underbygge slike resonnementer.

Presentasjon og kommentarer til norsk utgave av C. Mullers bok "Lingvistisk Statistikk"
Roald Skarsten
Proceedings of NODALIDA 1981, pages 7-12

At et slikt udekket behov for en slik elementær innføring i bruk av statistiske metoder innen språk og litteratur ikke er noe spesielt norsk fenomen ser vi ved at boken også er oversatt til tysk. I 1971 oversatte Lothar Hoffmann en tidligere versjon (fra 1968): Initiation á la statistique linquistique. I sin anmeldelse av denne tidligere versjonen skrev Hoffmann bl.a.: "Mullers Darstellung des "Begriffs - und Methodenarsenals der Statistik entspricht dem neuesten Stand. Sie ist vorläufig der einzige systematische Versuch dieser Art auf dem Gebiet der Linguistik. Sie besticht durch Einfachheit und Klarheit.""

Da vi så søkte etter en bok som kunne dekke behovet var der i tillegg til Muller kommet en bok med et lignende formål fra Barron Brainerd, Weighing Evidence in Language and Literature (Toronto 1974), men den hadde ikke de pedagogiske fortrinn som Muller's bok har. Konferer Bruce A. Beatie i artikkelen, "Measurement and the Study of Literature", Computers and the Humanities, july/sept. 1979, "The Communication problem is even more obvious in an effort that purports to bridge the gap: Barron Brainerd's Weighing Evidence in Language and Literature.[8] Although the reviewer in Computers and the Humanities assured us that it was "clearly written", the literary colleagues to whom I showed it found Brainerd's book as incomprehensible as a textbook on tensor calculus or thin-layer chromatography." (s. 191)

Der er etterhvert en ganske omfattende forskningslitteratur på feltet, men denne har vært vanskelig tilgjengelig uten de nødvendige forkunnskaper, og nettopp dette angir Hoffmann som et argument for sin oversettelse. "Besonders mangelt es an einer verständlichen und systematischen Einführung in die Sprachstatistik, die der Zugang zu der im Ausland Zahlreich publizierten Forschungsergebnissen erschliesst. Die übersetzung des Buches von Ch. Muller soll diese Lücke schliessen helfen." (s. 5 i forordet)

Det kan forøvrig nevnes at vi tok kontakt med både Sture Allén og H. Spang-Hanssen som begge støttet tanken om en oversettelse av Mullers bok. Når vi allikevel ikke alltid har vært like sikker på riktigheten av det vi gjorde, så var det fordi den statistiske konsulent vi brukte ikke umiddelbart var begeistret for boken, ut fra et fagstatistisk synspunkt. Jeg skal komme tilbake til hennes synspunkter i kommentaravsnittet. Som et praktisk problem, bokutgivelse eller ikke bokutgivelse, sto vi imidlertid i den situasjon som karakteriseres fortrinnlig ved det engelske ordtaket: "Beggers can't be choosers."

Noen lurer kanskje på hvorfor jeg kommer inn på disse forholdene. Boken foreligger der, og ferdig med det! En god grunn er at vi ønsker å ha en mest mulig positiv bakgrunn for de faglige kritiske synspunkter som vi skal komme tilbake til, og en annen god grunn for å presentere boken er at vi ønsker salg på den!

Jeg iler tilmed å forsikre om at de forannevnte personer ikke har noen økonomisk gevinst av salget. Et eventuelt tap på boken bæres av Universitetsforlaget. Hvis der imidlertid blir et visst salg på boken, så har vi en mulighet til å få utgitt fortsettelsesboken: Principes et méhodes de statistique lexicale. (Paris 1977) I den første versjonen fra 1968 besto boken av to deler som hver svarer til disse to nye bøkene, bortsett fra at de sistnevnte er blitt betraktelig utvidet og forbedret. I en anmeldelse av den siste boken skriver Michel Dubrocard om den første versjonen fra 1968, "The work was soon out of print, which was a measure of its success."

Om den boken som danner grunnlaget for den norske oversettelsen skriver han forøvrig at den er "considerably improved," og om fortsettelsesboken: "this new book is not merely a serviceable introductory manual indispensible for all newcomers in the field of statistical linguisties. It is also required reading for specialists and advanced scholars,",Computers and the Humanities,January - March 1979, s. 84.

Det er også verdt å nevne at det er utgitt et eget øvelseshefte til bøkene, som det også kunne være aktuelt å vurdere i oversettelsessammenheng, etter som statistikerne stadig vekk fremhever at praktiske statistiske øvelser er et nødvendig ledd i læringsprosessen.

Boken og forfatteren skulle hermed være presentert og anbefalt på det sterkeste. Det skal bare nevnes at den har vært benyttet som kursbok i et nasjonalt sommerkurs, og at boken i Norge allerede har vist sin fortreffelighet. (Hvis dere finner forbausende få trykkfeil i boken så skyldes det bl.a. kursdeltakernes innsats). I tiltro til at Norge er et representativt utvalg som gir grunnlag for en positiv slutning om bokens skjebne i den nordiske populasjon våger jeg å gå over fra presentasjonen til kommentarene.

Den alvorligste innvendingen mot Mullers bok fra den fagstatistiske konsulenten gjaldt presentasjonen av hypotesetesting. Hun skrev bl.a.: "Det jeg kritiserer hos Muller er hans slurv og nonchalanse i den manglende formulering og presisering av dette utgangspunktet for flere av signifikanstestene.

Først i avsnitt 13.8 kommer han inn på modellbegrepene språkytring. Da har han allerede gjennomgått flere eksempler på signifikanstesting og brukt sannsynlighets-teoretiske begreper som stokastisk variabel, sannsynlighetsfordeling, forventning etc. etc. - uten denne rammen som skulle gi begrepene mening.

I avsnitt 16.3 kommer han igjen inn på denne tankegangen som, så vidt jeg kan se, danner grunnlaget for og gir mening til det meste av analysene i de siste to tredjedelene av boka (kap. 9 og utover)."

Først må jeg her si at Muller allerede i kap. 3.8 med overskriften SPRÅK OG YTRING kommer inn på dette fundamentale forhold i forbindelse med det generelle spørsmål om POPULASJON OG UTVALG som er hovedoverskriften for kap. 3. "På den annen side kan vi introdusere den klassiske distinksjon mellom språk og ytring, altså mellom den potensielle og den realiserte. Da må vi betrakte enhver ytring som en realisering, altså som et utvalg av språket til den talende eller skrivende. For å observere språk må vi gå omveien om ytringene. Enhver statistikk baserer seg nødvendigvis på tekster, dvs. på utvalg av språket. Derav følger at hver gang vi mener å trekke en konklusjon om språk ut fra en statistisk analyse, så resonnerer vi ved induksjon; vi gjør oss opp en mening ut fra et utvalg. Vi må derfor benytte regneoperasjoner som er utviklet for slike analyser."

Dette er nå allikevel bare en detalj, hovedpoenget for E. Quale er at Muller ikke alltid klarer å fastholde sine i utgangspunktet riktige formuleringer når det kommer til konkrete eksempler i boken, og hun ville ha gitt dette forhold en bredere omtale i boken. Til nye lesere vil jeg derfor understreke nødvendigheten av å være særdeles oppmerksom på kap. 3.8, kap. 13.8 og kap. 16. Jeg synes Mullers synspunkter her er så viktige for forståelsen av boken og for å sette kritikken i sitt rette perspektiv at jeg vil sitere litt grundig fra kap. 16. "Vi går ut fra at moderpopulasjonen ikke består av selve teksten, men av språket, i Saussures betydning av ordet; et "språk" der egenskaper estimeres ut fra det store utvalg som stykket utgjør. Det er altså ikke det franske språk, heller ikke århundrets språk eller det språk som ble brukt i komedier på vers på den tiden, heller ikke det språk som ble brukt i Pierre Corneilles komedier på vers. Det er en bestemt språklig tilstand, nemlig Corneilles mens han skrev L'Illusjon; en latent mulighet som ikke bare er bestemt av språkets egne lover, men også av forfatterens person, av den genre han brukte (stilistiske årsker) og av det emne han hadde valgt (tematiske årsaker); et "språk" hvorav stykket og utdraget bare er utvalg av forskjellig størrelse.

Og språket, eller den språklige tilstand, kan pr. definisjon ikke observeres umiddelbart."

Det som så skjer i en del eksempler er at Muller, når han formulerer nullhypotesen, knytter denne til utvalget (teksten) istedetfor til den bakenforliggende populasjon (modell). Hypotesetestingens hensikt eller funksjon er jo å undersøke om forskjellen mellom f.eks. to gjennomsnitt i to utvalg er stor nok til at vi kan konkludere med at de er utvalg fra forskjellige populasjoner, og at nullhypotesen om at de er fra samme populasjon kan forkastes. Et eksempel fra Mullers bok kan illustrere Quales påstand. "Le cid har 4,01% adjektiver, Phédre har 5,99%. L'Illusion comique har 18% substantiver mens Matamore's rolle i stykket har 20%. I begge tilfeller må man gå ut fra nullhypotesen: "de to tragediene er utvalg av samme populasjon hvor proporsjonen av adjektiver er stabil. Forskjellen kommer av tilfeldige variasjoner" - "proporsjonen av substantiver i stykket er stabil. Det avvik som observeres i rollen bygger på tilfeldige variasjoner i de utvalg som er trukket fra populasjonen". (s. 108) Dette eksemplet viser hvordan det i det første tilfellet er brukt en riktig formulering av nullhypotesen. "De to tragediene er utvalg av samme populasjon...", mens det i det andre tilfellet "proporsjonen av substantiver i stykket er stabil...". Her er altså nullhypotesen knyttet til utvalget (til stykket) og da blir det meningsløst med hypotesetesting. Det er nærliggende å tale om slurv i dette tilfellet, men der er flere slike tilfeller, f.eks. i kap. 14.6 hvor nullhypotesen igjen formuleres i tilknytning til det aktuelle utvalg av et gitt språklig fenomen."(Den er holdbar hvis vi kan forkaste nullhypotesen:"tilfeldig fordeling av (a) i hele teksten.")", men det er ikke bare formuleringen av nullhypotesen dette gjelder, det kan også gjelde konklusjonene som trekkes, de knyttes til utvalget istedetfor til populasjonen, dvs. "språket". Et annet eksempel kan man bl.a. finne i kap. 18, med følgende konklusjon: "...det stilistiske fenomen (spesielt mange substantiver i dette utdraget) er reelt." (s. 142) Et annet eksempel finnes i kap. 21.4. Generelt savnes ofte modellformuleringen, leseren må selv ha den i tankene.

E. Quale finner at den sannsynlighetsteoretiske grunntanken blir uklar fordi forbindelsen mellom modell og virkelighet fremstilles for mangelfullt og til dels uriktig, som vi har sett. Det eneste vi kan gjøre med dette er å henstille til leserne å sette de uheldige formulerte eksempler inn i sin rette sammenheng, slik den er presentert i de nevnte Viktige avsnitt om språk og ytring. Når disse kritiske kommentarer fra fagkonsulenten er nevnt vil jeg også ta med noen av hennes positive kommentarer. De første kapitlene karakteriserer hun som preget av pedagogisk nennsomhet og omhu, og anbefales på det varmeste, og den deskriptive statistikken får god omtale . Ved fornyet lesning av den induktive statistikken (i lys av språkytring-modellen) hadde hun fått et mere positivt syn på denne delen. Hennes kommentar tildet oppsummerende avslutningskapittel består av bare ett ord: Nydelig.

"Det er respektabelt og oppløftende at en lingvist har påtatt seg det krevende arbeide å forsøke å formulere og introdusere det statistiske begrepsapparat blant språkforskere, hvis forhold til tall og formler man må formode ikke er av det mest fortrolige slaget." Hun synes "Det er viktig at boka er skrevet, at den er oversatt til norsk, og jeg synes den bør utgis." Hun ville dog ha foretrukket at de problemer som er påpekt i den induktive delen kunne vært kommentert i et tillegg, slik at studentene venner seg til presise oppstillinger av nullhypotese og alternativ hypotese i tilknytning til den aktuelle modell (populasjon).

Mullers bok skulle hermed være presentert og kommentert med sine pro et contra på behørig måte, ut fra dens egne forutsetninger.