Göteborgs universitet
Språkdata

Förslag till inrättande av ett organ för lagring och tillhandahållande av datamaskinellt läsbara texter, benämnt logotek

Göteborgs universitet
Språkdata
N. Allégatan 6
413 01 Göteborg
Professor Sture Allén, gb

1973-03-07

Till Konungen
Utbildningsdepartementet

Förslag till inrättande av ett organ för lagring och tillhandahållande av datamaskinellt läsbara texter, benämnt logotek

Med stöd av vad som anförs nedan får jag härmed anhålla, att ett fast serviceorgan, benämnt logotek, med uppgift att insamla och lagra datamaskinellt läsbara texter samt tillhandahålla dessa och bearbetningar av dem inrättas från och med 1974-07-01 och att detta organ placeras vid avdelningen för språklig databehandling vid Göteborgs universitet.

Underdånigst

Sture Allén

Innehåll

Sammanfattning

Miljontals ord löpande text framställs varje år i datamaskinellt läsbar form i landet. Rationell texthantering i ett stort och växande antal fall inom forskning, förvaltning och näringsliv kräver att det inrättas ett serviceorgan med uppgift att insamla, lagra och tillhandahålla maskinläsbara texter och bearbetningar av dem. Organet, benämnt logotek, föreslås placerat vid avdelningen för språklig databehandling vid Göteborgs universitet. Avdelningen har som en serviceåtgärd ställt textmaterial och bearbetningar till förfogande för avnämare inom många olika områden sedan 1966. Det programsystem för textbehandling som utvecklats är mycket omfattande. Inom kort installeras vid avdelningen en minidatoranläggning, som är specialutrustad för avancerad textbehandling. Som lagringsmedium föreslås magnetband. Den årliga kostnaden för verksamheten beräknas till 287 830 kronor enligt 1973 års löneplan.

Bakgrund och syfte

I en artikel i Dagens Nyheter 1970-09-29 med rubriken Åtta teser om texthantering föreslog jag att ett organ för arkivering och tillhandahållande av maskinläsbara texter skulle inrättas (se
./. bilaga 1). Som benämning på organet lanserade jag ordet logotek (egentligen 'ordmagasin'). Resonemanget var i korthet följande.

Det tryckta ordets användbarhet är alltför begränsad för att medge en rationell hantering av texter i ett stort och växande antal fall inom forskning, förvaltning och näringsliv. Tas datamaskinen till hjälp, kan uttestningen av beskrivningsmodeller, bearbetningen av material och utvärderingen av resultat göras till arbetets huvudpunkter, inte själva materialsamlandet. Vid landets sätterier framställs årligen mängder av maskinläsbara texter, som kasseras efter användningen. Dessa texter går förlorade därför att de inte faller inom något existerande organs verksamhetsområde. Det är därför nödvändigt att skapa ett sådant organ såsom komplement till biblioteken. Behovet ökar från år till år, och förlusterna av maskinläsbara texter blir större och större.

Det möter inga svårigheter att styrka behovet av det föreslagna logotekets service gentemot avnämarna. Inom avdelningen för språklig databehandling vid Göteborgs universitet respektive den forskningsgrupp som avdelningen utvecklats ur har vi kontinuerligt ställt material, metoder, program och resultat till förfogande för andra och biträtt vid planläggningen av olika uppgifter. Detta kan ses som en försöksverksamhet. På grund av att forskningsområdet dels är utpräglat tvärvetenskapligt, dels har ovanligt talrika tillämpningar har förfrågningarna varit många. Exempel på personer, institutioner och företag, som har kunnat dra nytta av
./. serviceverksamheten, ges i bilaga 2.

Ofta har det varit fråga om bearbetningar i olika avseenden av lagrad text: frekvenslistor över ord och andra språkliga enheter, finalalfabetiska ordlistor (baklängeslistor), beläggsamlingar, ord- och fraskonkordanser, listor över bokstavskombinationer m.m. Mottagarna spänner över ett vitt fält från språkvetenskap till teknologi, från praktisk pedagogik och informationssökning till grafisk industri. Se vidare avsnittet Tillämpningar i den bilagda informationsskriften från Göteborgs universitet rörande Språkdata
./. (bilaga 3).

Det är att märka att serviceverksamheten utövats som en extrainsats av ett arbetslag med eget forskningsprogram. Någon generell service har inte utlovats, utan biståndet har lämnats efter förfrågningar från sådana som fått kännedom om vårt arbete. Det verkliga servicebehovet är med säkerhet långt större. Aktiv upplysning om de möjligheter till rationalisering och kvalitetshöjning som datamaskinell textbehandling erbjuder skulle bekräfta detta.

Det utåtriktade arbete som bedrivs vid avdelningen hänger samman med vår uppfattning om en vetenskaplig institutions funktioner. De grundläggande funktionerna är forskning och i anslutning därtill undervisning. Det framstår som ett rättmätigt krav från samhällets sida, att institutionen dessutom informerar om sin verksamhet och sina resultat. Härigenom skapas ökad insikt i samhället och öppnas möjligheter för tillämpningar av vunna resultat. I åtskilliga fall kan detta leda till en mer eller mindre omfattande kontakt med avnämare inom olika samhällsområden. Denna tredje funktion hos den vetenskapliga institutionen kan sammanfattas under beteckningen service.

Förarbete

Logoteket är en synnerligen betydelsefull tillämpning av forskningen på ämnesområdet. Genom återkoppling kan logoteksarbetet också väntas ge fruktbara impulser till grundforskningen i och med att en kontinuerlig systemutveckling är nödvändig för att möta de uppkommande behoven. På grund av den vikt logoteket tillmätes har förarbeten utförts i flera olika avseenden.

På programvarusidan har vi utarbetat ett generellt textinläsningssystem, som är i drift vid Datasaabanläggningen vid Göteborgs datacentral för forskning och högre utbildning. Det accepterar hålremsor av olika slag inklusive sättremsor samt vidare hålkort och magnetband. Inläsningsrutinerna leder in i det övriga programsystemet, som nu omfattar ett par hundra program för olika slags kvalitativa och kvantitativa analyser och listningar.

Lagringsformen på det informationsbärande mediet är också av stor betydelse i detta sammanhang. Lovande förstudier över komplexa länkade lagringsstrukturer har utförts. Arbete pågår rörande optimal packning av text med hänsyn till de olika skrivtecknens frekvenser.

Den maskinella utrustningen vid avdelningen har hittills omfattat hålrems- och hålkortsstansar. Under 1973 installeras emellertid en minidatoranläggning. Denna är speciellt konfigurerad för textbehandling (inläsning, lagring, bearbetning) på basis av de gångna årens erfarenheter. Särskild vikt har bland annat lagts vid typuppsättningen (stora och små bokstäver, specialtecken, olika alfabet) och vid faciliteter för effektiv interaktiv bearbetning av text. I kravspecifikationen rörande kringutrustningen ingår specialkonstruerade remsläsare, magnetbandstation, skivminne, bildskärmar och radskrivare. Maskinens operativsystem skall tillåta interaktiv bearbetning vid flera terminaler, samtidigt med att maskintiden utnyttjas för satsvis bearbetning. Vid denna textorienterade anläggning kommer inläsningen och lagringen av text att ske i fortsättningen. Större satsvisa bearbetningar skall som hittills utföras vid datacentralen.

Ett tredje område av stor vikt för logoteksarbetet är den typografiska produktionstekniken. Genom vårt samarbete med olika förlag och tryckerier när det gällt att dels få tillgång till maskinläsbara texter, dels producera våra egna skrifter (i synnerhet ordböckerna) har vi fått detaljerad kännedom om de olika produktionsmetoderna. Det kan nämnas att projektet Datamaskinell undersökning av tidningsprosa bygger på maskinläsbara texter som insamlats från tidningssätterier. I detta fall rör det sig om hålremsor som styr blysättningsmaskiner. Vidare kan nämnas att ordböckerna har satts med den mest avancerade metod som föreligger, datorstyrd fotosättning. Manuskriptet levereras i detta fall i form av magnetband till tryckeriet. Nusvensk frekvensordbok 1 var den första svenska bok som producerades med denna teknik.

Kännedomen om den grafiska industrins arbetsmetoder har gjort det möjligt för oss att utforma maskin- och programvaran på ett sådant sätt, att maskinläsbara texter och ordsamlingar på olika slags hålremsor och magnetband kan omhändertas på ett rationellt

sätt. Bland svenskt material som för närvarande är tillgängligt vid avdelningen kan nämnas en miljon ord ur morgontidningar 1965, 200 000 ord ur Hufvudstadsbladet 1967, romaner av Sivar Arnér, Dagmar Edqvist, Per Anders Fogelström, Lars Gyllensten, Eyvind Johnson och Björn Runeborg, poesi av Hjalmar Gullberg och Esaias Tegnér samt psalmboken 1937. Till ordbanken har särskilt kunnat fogas bland annat ordmängden i indexvolymen till uppslagsverket Focus och det s.k. nummerbandet (90 magnetbandsrullar) med det svenska personnamnsbeståndet per 1973-01-01. För närvarande omfattar arkivmaterialet inklusive bearbetningar av olika slag drygt 300 magnetband.

Självfallet är det inte meningen att insamla och lagra allt som sätts med hjälp av maskinläsbart medium. Som en fjärde punkt skall här framhållas, att vi har gjort en fullständig undersökning av årstrycket på svenska i Sverige och Finland 1970 (exklusive accidenstrycket). Avsikten har varit att få grepp om omfånget och fördelningen på olika texttyper och därigenom ett fast underlag för det urval som skall göras. Till grund för undersökningen lades uppgifterna i Svensk bokförteckning, Svensk tidskriftsförteckning samt Inländsk och Utländsk tidningstaxa.

En koncentrerad bild av resultatet ger bilagan Årstrycket på svenska 1970 (bilaga 4). De signa materialet fördelats på är de som förekommer i Sveriges allmänna biblioteksförenings klassifikationssystem (fördelningen på detaljerade signa finns i utskrifter från körningen). Det framgår att de drygt 11 000 titlarna ger drygt 3 miljarder löpande ord, varav 2,5 miljarder (77 %) kommer på tidningarna. De största huvudsigna utanför tidningarna är O Samhälls- och rättsvetenskap, H Skönlitteratur och Q Ekonomi och näringsväsen.

Uppläggning

Som huvudkälla till textflödet in i logoteket framstår som nämnts sätterierna. Det har varit lätt att samarbeta med förlagen och tidningarna när det gällt insamling av hålremsor. Det kan påpekas att förlagen bör vara jämställda med andra avnämare i fråga om service från logoteket. Av betydelse för dem är som tidigare angetts tillämpningar inom den grafiska industrin. I vissa fall

kan de också vara intresserade av att återrekvirera en lagrad text. Dock är att märka att varje lagrad text är rensad från rent typografiska signaler, vilket i sin tur är en garanti mot missbruk av texterna.

Urvalet av texter föreslås ske efter två linjer. Grundläggande blir en strategi baserad på undersökningen av årstrycket. Som mål uppställs då ett sampel, som i görligaste mån avspeglar årstryckets fördelning på (huvud)signa. Urvalet blir härigenom representativt för svenskt skriftspråk i en klart angiven mening. Självfallet får signum Ä Tidningar särbehandlas med tanke på dess kolossala omfång.

Den andra urvalslinjen innebär ett slags viktning. Det är rimligt att införliva sådana texter som enligt logotekets bedömning framstår som särskilt betydelsefulla eller från någon synpunkt särskilt intressanta. Som exempel kan nämnas Svensk författningssamling, som för närvarande är aktuell. Detsamma bör principiellt gälla texter som föreslås eller efterfrågas från olika håll.

Texturvalets storlek kan inte helt preciseras på nuvarande stadium. Som ett minimum för det första arbetsåret kan tio miljoner löpande ord anges. I omfång motsvarar detta omkring 15 000 tidningsartiklar eller ett par tre hundra böcker. När verksamheten upparbetats, bör mängden kunna flerfaldigas.

Från logotekets synpunkt är remsor och band som innehåller slutkorrigerade texter av störst intresse. Också texter i tidigare produktionsled kan emellertid vara värdefulla. Logoteket kan självfallet inte åta sig detaljgranskning av alla inkommande texter. Däremot måste en allmän kontroll av textens karaktär ske i anslutning till registreringen. I många fall, särskilt när det gäller vetenskapliga undersökningar av enskilda texter, torde gången bli den, att logoteket ställer en utskrift till förfogande. Denna kontrolläses av mottagaren, som markerar önskade korrektioner. Dessa utförs sedan på bandet av denne eller logoteket efter överenskommelse från fall till fall.

En annan viktig källa att avtappa utgör materialet från TT. Logoteket bör abonnera på detta stoff och utveckla särskilda rutiner för omhändertagande av ordströmmen på teleprinterremsorna. Mycket

av förnyelsen i ordförrådet kan avläsas i dessa texter. De inbjuder bland annat också till intressanta samhällsvetenskapliga undersökningar.

En tredje källa till texter är kopiering av magnetband innehållande material som inkodats för olika, oftast vetenskapliga ändamål inom och utom landet. Löpande förteckningar i tidskriften Computers and the Humanities ger en god bild av tillgången på sådana texter. I den mån transkriptioner av talat språk föreligger på maskinläsbart medium är de naturligtvis också av stort intresse.

Ytterligare en källa, som normalt dock endast kan komma i fråga under särskilda omständigheter, är nystansning av texter inom logoteket. En möjlighet är att sådan stansning utföres inom ramen för arkiv- eller beredskapsarbete.

Utvecklingen av tekniken för optisk klartextläsning måste noggrant följas. Än så länge är restriktionerna alltför kraftiga och kostnaderna alltför höga för att den optiska läsningen skall kunna hävda sig i det här aktuella avseendet.

Det framgår – och får anses självklart – att huvudföremålet för logotekets verksamhet skall vara texter på svenska. Någon principiell begränsning till svenska bör emellertid inte göras. Det finns all anledning att exempelvis stödja forskningen rörande främmande språk genom att lagra utländska texter när tillfälle bjuds. En betydelsefull uppgift blir också att förmedla kontakter.

Ett naturligt led i arbetet med texterna är att successivt kumulera ordförrådet i dem och på så sätt bygga ut den ordbank som föreligger. Härigenom erhålls ett kraftfullt referenslexikon över den levande svenskan med uppgifter om ordens bruklighet. Värdet av ett sådant är uppenbart.

Som fysiskt lagringsmedium erbjuder sig i första hand magnetband. Det medger kompakt och billig lagring av så vitt man vet tillfredsställande beständighet. En magnetbandsrulle rymmer betydligt över en miljon löpande ord även vid spatiös lagring. Priset är omkring 100 kronor. På sikt blir det aktuellt att pröva lag-

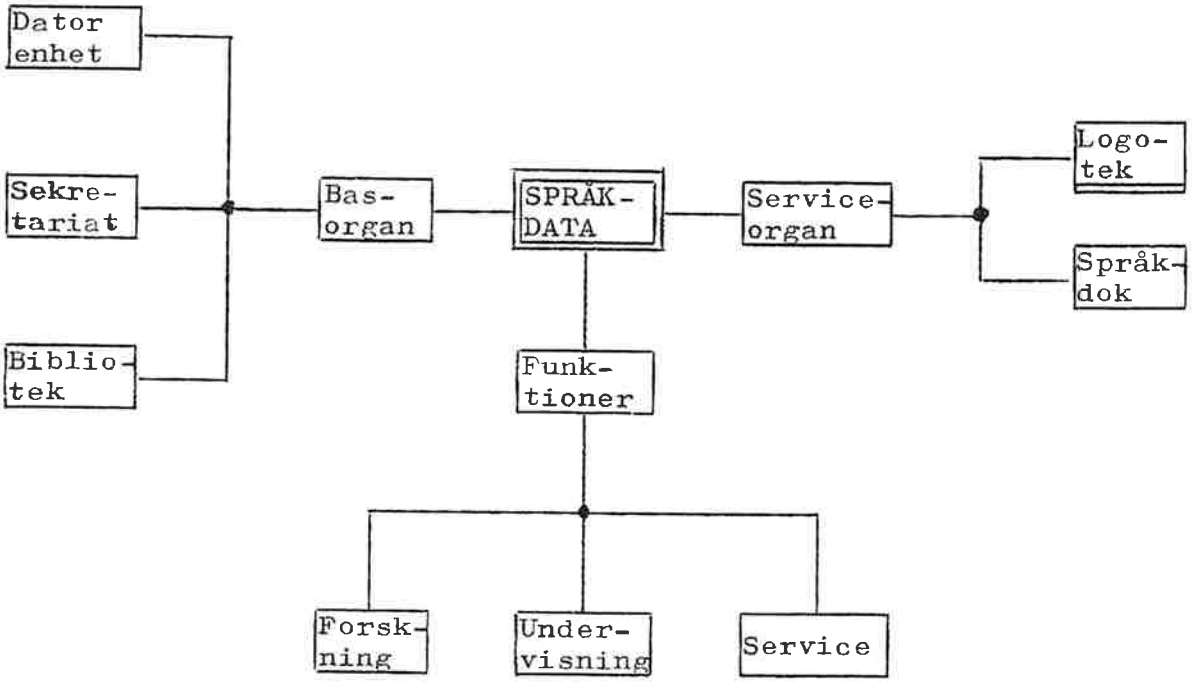
ring på mikrofilm som genererats från magnetband. Detta alternativ är intressant, eftersom mikrofilm är ögonläsbar. Särskilt intressant blir det, när mikrofilmen i sin tur kan läsas in i datorn på optisk väg och sålunda fungera som enda lagringsmedium. Kostnadsnivån måste naturligtvis också beaktas.

Organisation

Det föreslås att logoteket placeras vid avdelningen för språklig databehandling vid Göteborgs universitet. Skälen härtill är flera. Avdelningen har mångårig erfarenhet av forskning på området och har bedrivit praktiskt servicearbete sedan 1966. Erforderligt förarbete har utförts på flera olika plan. Till förfogande står en dataanläggning som är specialutrustad för avancerad textbehandling. Arkivutrymme finns i avdelningens lokaler. Härtill kan fogas att en förläggning till Göteborg ligger i linje med den inriktning mot decentralisering av databanker, för vilken bland annat beredskapsskäl talar (jämför dataindustriutredningens lägesrapport Data och näringspolitik, SOU 1973:6, s. 125 f.). Ett organ för databehandling är för övrigt på ett helt annat sätt oberoende av geografiskt läge än andra organ, eftersom det kan nås via terminaler från hela riket. Som en parallell kan nämnas att Norges allmänvetenskapliga forskningsråd har placerat sin nationella datacentral för humanistisk forskning i Bergen.

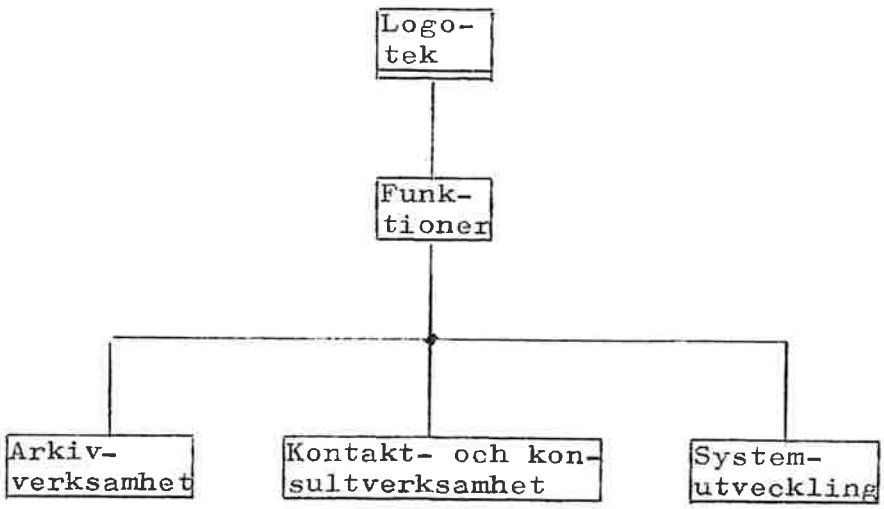
Inledningsvis angavs tre funktioner som väsentliga för en vetenskaplig institution: forskning, undervisning och service. Utövandet av funktionerna kräver vissa basorgan och vissa serviceorgan. Den enkla skissen överst på nästa sida visar strukturen inom avdelningen för språklig databehandling.

Av basorganen har datorenheten, som ställs till förfogande av Utrustningsnämnden, tidigare kortfattat beskrivits. Sekretariatet omfattar en heltidsanställd kontorsskrivare och en halvtidsanställd amanuens, som avlönas av Statens humanistiska forskningsråd, samt en arkivarbetare. I utrustningen ingår skriv- och räknemaskiner samt kopierings-, reproduktions- och bindningsapparatur. Biblioteket består för närvarande av omkring 1000 bokvolymer och ungefär lika många broschyrer.



Den externa verksamheten kanaliseras i två serviceorgan, logoteket och Språkdok. Det sistnämnda är ett organ för dokumentation av språkvetenskaplig forskning.

I logotekets inre struktur kan tre huvudfunktioner urskiljas.



Arkivverksamheten omfattar urval, införskaffande, lagring och katalogisering av maskinläsbara texter jämte vissa bearbetningar av dem. Kontakt- och konsultverksamheten omfattar information om

logotekets möjligheter, rådgivning och handledning i de enskilda fallen samt samplanering av aktiviteterna inom området på olika håll. Systemutvecklingen omfattar utarbetande och programmering av procedurer och algoritmer för logotekets uppgifter och i samband därmed bevakning av den tekniska utvecklingen. Aktivt arbete på förbättring och nyutveckling av metoderna för inläsning, lagring, bearbetning och utskrift av språkligt material är en grundförutsättning för en effektiv serviceverksamhet.

Personal

För att leda det dagliga arbetet med arkivverksamheten och svara för huvudparten av den viktiga kontakt- och konsultverksamheten behövs en "logotekarie". Innehavaren bör med tanke på de tidigare skisserade uppgifterna vara en docentkompetent lingvist med god erfarenhet av språklig databehandling.

Systemutvecklingen kräver dels en lingvist med doktorsexamen (motsvarande), som förutom gedigen allmänlingvistisk utbildning har dokumenterad förmåga till algoritmiskt tänkande, dels en systemman/systemprogrammerare, som har informationsvetenskaplig kombinerad med lingvistisk utbildning och god erfarenhet av språklig databehandling. I den senares uppgifter ingår att omsätta algoritmerna i program, att svara för de modifikationer av befintliga program som det löpande arbetet kräver och att ansvara för logotekets körningar i datamaskin. Tillsammans har dessa båda befattningshavare avgörande betydelse för logotekets funktion.

För skrivarbetet i samband med insamling, registrering, information osv. samt stansning krävs en sekreterare/stansoperatör.

Möjligheterna att besätta de fyra aktuella befattningarna med lämpliga personer är för närvarande mycket goda.

Som föreståndare för logoteket inträder på ett naturligt sätt ledaren för avdelningen för språklig databehandling. På honom ankommer att dra upp de stora linjerna för arbetet, att ytterst bära det vetenskapliga ansvaret och att ansvara för ekonomin.

Insamlingen av hålremsor och magnetband sker lämpligen med hjälp av särskilda kontaktpersoner vid sätterierna. Modellen med sätteriombud visade sig fungera väl vid insamlingen av materialet till projektet Datamaskinell undersökning av tidningsprosa.

Budget (1973 års löneplan)

Löner			
(1)	"Logotekarie" (docentkompetent lingvist), U 20 + 138	59 952	
(2)	Forskningsassistent (lingvist med doktorsexamen), U 19	55 416	
(3)	Systemman/systemprogrammerare (i paritet med byrådirektör), A 28	58 296	
(4)	Sekreterare/stansoperatör (i paritet med kontorsskrivare), A 13	29 448	
(5)	Uppdragstillägg för föreståndare 12x410	4 920	
(6)	Arvode till sätteriombud 700 timmar à 15	10 500	
(7)	Lönekostnadspålägg 24 %	52 448	270 980
Material			
(8)	Kontorsmaterial, telefon, expenser	4 000	
(9)	Facklitteratur	3 000	
(10)	Magnetband	2 500	9 500
Resor			
(11)	Resor, konferenser	3 000	3 000
Utrustning			
(12)	Skrivmaskin	1 500	1 500
Förvaltningskostnad			
(13)	Förvaltningskostnad 1 %	2 850	2 850
		S:a	287 830

Budgetmotivering

(1) - (4) Lönesättningen är densamma som för befattningar med motsvarande kompetenskrav vid universitetens institutioner respektive (i fråga om punkt 3) de akademiska datacentralerna.

(5) Uppdragstillägget är detsamma som det universitetslektor vid universitetsfilial uppbär för prefektgöromål m.m.

(6) Sätteriombuden får sköta leveransen av hålremsor och magnetband till logoteket utanför sin ordinarie arbetstid.

(8) - (9) Verksamheten drar vissa löpande utgifter och kräver möjlighet att anskaffa viss speciell litteratur, till stor del utländsk.

(10) Posten avser 25 magnetband för lagring av text och bearbetningar.

(11) Samarbetet med sätterierna och bevakningen av den tekniska utvecklingen fordrar vissa resemöjligheter.

(12) Posten avser en elektrisk skrivmaskin.