GÖTEBORGS UNIVERSITET
SPRÅKDATA

Lexikalisk databas

September 1977

PROJEKTET LEXIKALISK DATABAS

Vid Språkdata pågår förarbete för ett projekt kallat Lexikalisk databas. Projektet har som syfte att etablera en databas med svenskt språkligt material av huvudsakligen lexikalisk karaktär. En enspråkig svensk ordbok är tänkt som den primära avkastningen.

Det är naturligt att databasen av praktiska skäl till en början starkt präglas av ordbokens behov. Uppbyggnaden av databasen får emellertid betraktas som en egen uppgift, principiellt skild från framtagningen av ordboken. Det senare momentet blir främst en manifestation av databasens tillämpningsmöjligheter.

Den exakta lingvistiska informationen i databasen är ännu inte slutgiltigt fastställd. Bland de diskuterade uppgiftstyperna återfinns – utan strikt inbördes viktning – uttal, ortografi, avstavning, morfemindelning, flexion, ordklasstillhörighet, syntaktiskt konstruktionssätt, fraseologi, idiomatik, definition, specialbetydelse(r), synonymer, antonymer, stilnivå, användningsområde, frekvens/bruklighet, implicita värderingar samt etymologi. Som tunga informationstyper betraktas i synnerhet innehållslig definition och syntaktiskt konstruktionssätt.

Arbetet är fast datoranknutet. Orienteringen mot datamaskinen tar sig emellertid något olika uttryck under olika faser av projektet.

(1) Vid materialurvalet utnyttjas Logoteket. Det stoff Logoteket tillhandahåller har delvis genomgått maskinell bearbetning.

(2) Det kompletterande material som får tillföras databasen inkodas via textskärm. Likaså utförs i princip all analys och bearbetning interaktivt via textskärmen.

Projektet Lexikalisk databas
Bo Ralph
Proceedings of NODALIDA 1977, pages 79-81

Materialet presenteras i fast format på skärmen. Vi diskuterar i vilken utsträckning inkodningsprogrammen också skall kunna ge förslag till analys på olika punkter. Det kan nämnas att tidningsspråksprojektet har rymt sådana moment (med olika grader av förfining) som partiellt automatisk lemmatisering och tentativ automatisk uppdelning av ord i ordsegment, och liknande operationer har vidareutvecklats inom andra projekt vid institutionen (främst Algoritmisk textanalys).

(3) Databasen lagras som ett länkat nätverk. Detta kan ge praktiska fördelar, men framför allt är det modellbygget som lockar. Nätverksstrukturen förefaller oss ha något slags psykologisk realitet när det gäller lexikalisk lagring.

En nätverksstruktur kan naturligtvis få olika grader av komplexitet. Ambitionsnivån får till en början inte bli alltför hög, utan det får endast bli vissa typer av information som lagras på detta sätt.

(4) Under planeringsstadiet har en del experimentellt arbete utförts. Härvid har de maskinella faciliteterna utnyttjats.

Det teoretiska förarbetet har främst gällt stickordens konstruktionssätt och innehållsliga definitioner. Som allmän teoretisk ram har en form av kasusgrammatik valts. En viktig del av förundersökningarna har tagit sikte på just verbens kasusramar.

I definitionerna skall en minimalordlista (definitionsvokabulär) användas för undvikande av cirkularitet. Alla ord som används i definitionerna måste vara nedbrytbara till definitionsvokabulärens innehåll inom ramen för databasens eget material. Denna infallsvinkel förutsätter att orden ordnas i ett konvertibilitetssystem. Det krävs att alla ord väljs noggrant, inte bara de som skall ingå i definitionsvokabulären utan även de icke elementära men nedbrytbara orden i definitionerna. Förundersökningar har också ägnats åt detta problem.

(5) Maskinella kontroller blir efter hand aktuella. Till dels är dessa lingvistiskt triviala, men viktiga undantag finns. Exempelvis framstår kontrollen av definitionsvokabulärens användning i definitionerna som ett språkligt intressant företag.

(6) En uteslutande praktisk – men betydelsefull – poäng med den tänkta uppläggningen utgör slutligen möjligheten att utnyttja datorstyrd fotosättning vid produktionen av ordboken. Ett magnetband kan framställas på ett enkelt sätt eftersom databasen är välstrukturerad. Kontinuerlig uppdatering av databasen, och principiellt sett även ordboken, är genomförbar med enkla medel.

- - -

Denna rapport speglar arbete som utförts av flera personer gemensamt, främst vid Språkdata. Rapporten har formulerats av Bo Ralph.