PROJEKTET ENGELSKT TALSPRÅK

Cecilia Thavenius

Projektet Engelskt talspråk i Lund, eller som det heter på engelska, Survey of Spoken English (SSE), är ett dotterprojekt till Survey of English Usage vid University College London. Projektet stöds sedan 1975 av Riksbankens Jubileumsfond. Projektledare är Professor Jan Svartvik.

MATERIAL

Projektets material är den s.k. 'London-Lund Corpus'. Den består av ca 1/2 miljon ord av talad engelska i ett flertal olika situationer. Materialet är inspelat, ortografiskt transkriberat, och prosodiskt och paralingvistiskt analyserat under ledning av Professor Randolph Quirk vid University College i London. De s.k. texterna av engelskt talspråk består av följande kategorier:

Material with origin in speech (100 "texts")

A Monologue (24)

- Prepared (but unscripted) oration 6
- Spontaneous
  - oration 10
  - commentary
    - sport 4
    - non-sport 4

B Dialogue (76)

- Conversation
  - surreptitious
    - intimate 24
    - distant 10
  - non-surreptitious
    - intimate 20
    - distant 6
  - telephone
    - intimate 10
    - distant 6

Projektet Engelskt talspråk
Cecilia Thavenius
Proceedings of NODALIDA 1979, pages 67-72

Talarnas identitet är skyddad, så vi har inte tillgång till deras namn. Däremot finns en del information om dem såsom yrke, ålder, kön och typ av relation mellan talarna. Men den djupare typ av socio-ekonomisk information, som t.ex. finns för Montrealkorpusen av talad franska saknas för detta material. När insamlandet av London-Lund korpusen påbörjades för ca 20 år sedan, hade man ännu inte börjat diskutera vikten av sådan kunskap för konversationsanalys.

Den prosodiska och paralingvistiska analysen, som helt har utförts vid Survey of English Usage i London, har varit oerhört tidskrävande. En timmes inspelat tal har tagit ca 80 timmar i anspråk för analysen. När analysen avslutats har materialet blivit liggande i London. Givetvis har det blivit utnyttjat för forskning av lingvister, som kommit till London och suttit där en tid och arbetat med materialet. Men tillgängligheten och användbarheten är inte särskilt stor när man tänker på det enorma arbete, som ligger bakom.

SYFTE

Projektet Engelskt talspråk har tre huvudsyften:

1 att med hjälp av dator göra materialet tillgängligt och sprida det till alla intresserade

2 att förse materialet med grammatisk 'tagging' för att möjliggöra en intressantare sortering, som går ovanför en ren lexikalisk ordnivå, och ge större kunskap om det engelska talspråkets grammatik

3 att producera forskning och forskningsresultat på grundval av materialet för att klargöra hur talspråket ser ut; flera doktorsavhandlingar är under arbete, och en del mindre studier har gjorts; dessutom finns planer på en lärobok i engelskt talspråk för skolor och universitet i Sverige, och en större talspråksgrammatik för internationellt bruk

ARBETSGÅNG

Efter fyra års intensivt arbete har vi nu kommit så långt att vi dels har ett färdigt magnetband med konversationstexterna, dels en bok, datasatt efter bandet, som beräknas komma ut i slutet av 1979. (A Corpus of English Conversation, ed. by Jan Svartvik and Randolph Quirk, Lund Studies in English, Lund: Gleerups/Liber, 1979). Bandet och boken omfattar alltså inte hela talspråksmaterialet, utan endast vad vi ansett vara den intressantaste delen, nämligen 'face-to-face conversation', inspelad med dold mikrofon. Detta material omfattar ca 170.000 ord, vilket utgör 34 prosodiskt markerade texter à 5000 ord. Resten av korpusen kommer att finnas tillgänglig på magnetband.

Vi fick materialet från London i form av maskinskrivna A6 kort. Vi gick först igenom det stora materialet flera gånger, redigerade bort en del paralingvistisk notation, som av tekniska skäl var svår att ta med, och påbörjade en ännu inte avslutad korrespondens med SEU i London för att kontrollera inkonsekvenser eller vad vi misstänkte var rena skrivfel. Själva har vi inte haft möjlighet att göra en sådan kontroll, eftersom vi inte har haft tillgång till ljudbandsmaterialet, och dessutom hållit på den principen, att eftersom analysen gjorts i London måste ändringar i denna komma därifrån. Men de allra flesta av våra frågor har visat sig relevanta, och vi har på detta sätt kunnat rätta ett par tusen fel i det ursprungliga materialet.

Därefter har vi kört in hela materialet på magnetband från bildskärmsterminal, som är ansluten till Lunds Datacentral. Datacentralen har sedan producerat utskrifter åt oss. Dessa utskrifter har sedan kontrollästs mot original av minst tre och i de flesta fall fyra personer. Sedan har materialet rättats efter varje korrekturläsning via bildskärmsterminalen. Varje text har körts ut och kontrollästs sex gånger, och vi hoppas nu att vi kommit så nära fulländningens stadium som möjligt. Men det har alltså tagit fyra år. F.n. håller Liber

Tryck i Stockholm på med datasättning efter magnetbandet. Programmering och provkörningar är nu klara, och vi håller på med korrekturet. Boktexten kommer att få detta utseende:

> A 45 ΔOld and Middle ₐEnglish GRAPHÓLOGY■ 47 «or ‖something ☆like ΔTHÀT■ 48 ‖you SÉE■»☆

B 49 ☆‖WÉLL■ 50 ‖you give☆ them the ₐLÒT {‖you SÉE■} ■ ☆ · ☆ 51 ‖that's the ☆☆PÓINT■☆☆ 52 «and» ‖make sure that there's ₐSÓMETHING■ 53 [ə:] ‖fairly ₐCLÒSELY RELÁTED■

A 54 ☆‖[m̀]■☆ 55 ☆☆‖[m̀]■☆☆

> B 56 «to ‖what they've STÙDIED■»

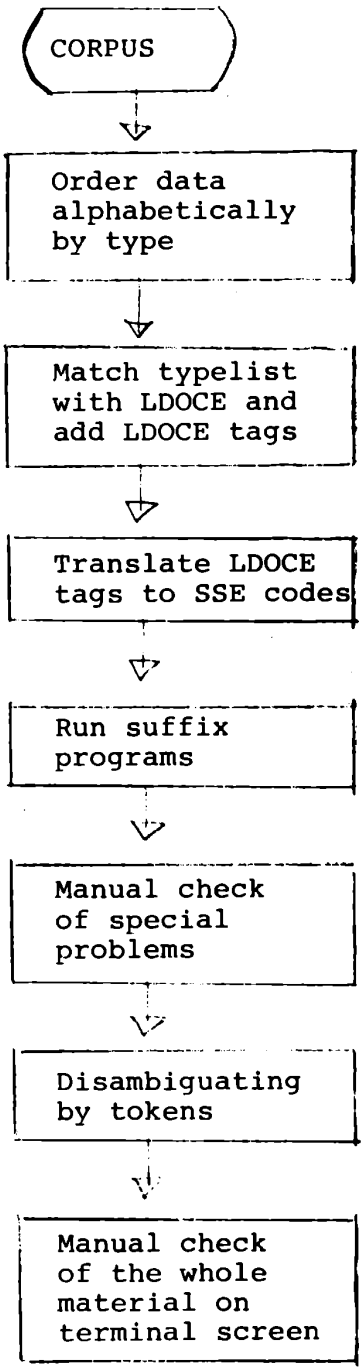
A 57 «it's ‖just 'one ΔQUÈSTION that they have to do ÍSN'T it■»

B 58 ‖well there were [ə] ΔÓNE■ 59 or ‖TWÓ we've ☆got on THÉRE■☆ 60 ‖you SÉE■

A 61 ☆‖yes ΔÍ SÈE■☆ · 62 ‖YÉS■ · 63 ‖YÉS■ - 64 [ə:m] · ‖one ÒTHER thing SÁM■ - 65 [ə:m] - ‖DEΔLÀNEY■ - 66 a CA‖NÁDIAN■ 67 ☆«who» ‖graduated☆

B 68 ☆«[ə] ‖WHÉRE did you put those THÌNGS■☆ 69 ‖just one» · ‖let me put this in my BÀG■ 70 «or» I'll «‖walk AₐWÀY with▷out it■» - - -

Vårt andra huvudsyfte var att förse det prosodiskt markerade materialet med grammatisk tagging. Där har vi efter ett långt förarbete och många och långa diskussioner kommit fram till följande procedur av ordklassanalysen. Longman Dictionary of Contemporary English från 1978, det s.k. LDOCE, finns på ett magnetband, som vi erhållit från Longmans. Vi kommer nu att utnyttja LDOCE-bandet för att få en så automatisk märkning som möjligt av vårt material. Det intressanta med detta tillvägagångssätt består i att man kan ta reda på hur långt man kan komma med automatiska metoder. Enkelt beskrivet blir gången följande:



Word class analysis

Materialet sorteras alltså först i en alfabetisk ordtypslista. Därefter matchas ordtypslistan med LDOCE och får LDOCE's grammatiska information, som sedan översätts till våra egna koder. Sedan körs suffixprogram på ord som slutar på -s, -er, -est, -ed, -ing. Vi kontrollerar sedan vissa problematiska kategorier manuellt i ordtypslistan. Därefter går maskinen över till att arbeta med den löpande texten och disambiguerar i de fall ord har fått mer än en 'tag'. Disambigueringsprogrammen grundar sig på sådant som position och statistik. Allra sist sker en manuell interaktiv kontroll och bearbetning vid terminal.

Efter ordklassanalys går vi över till 'phrase-tagging'. Tillvägagångssättet blir då att maskinen går igenom varje s.k. 'tone unit' i det löpande materialet, och söker från höger till vänster inom 'tone uniten' efter fraserna i följande ordning: verbfras, adverbfras, adjektivfras, nominalfras, prepositionsfras. Även 'phrase-tagging' fasen avslutas med en manuell kontroll vid terminalen. I mån av tid och ekonomiska resurser ska sedan en analys av satsfunktioner utföras.