Bente Maegaard og Hanne Ruus
Københavns Universitet, Amager
Njalsgade 80
DK 2300 København S.
DANMARK

# STRUKTURERING AF LINGVISTISKE DATA TIL BRUG VED MASKINOVERSÆTTELSE.

## EUROTRAs overførselsstruktur.

Siden feb. 1978 har forskere fra videnscentre inden for datamatisk lingvistik og maskinoversættelse i EF-landene arbejdet på at formulere et projekt til et europæisk maskinoversættelsessystem, EUROTRA.

Det er efterhånden blevet et meget stort problem for EF, at mængder af dokumenter skal foreligge på 6 sprog. Oversættelsesafdelingerne vokser og alligevel kan de knap klare efterspørgslen.

Til afhjælpning af oversættelsesbehovet har Kommissionen anskaffet et amerikansk system, SYSTRAN, der kan oversætte mellem visse par af EF-sprog (engelsk-fransk, fransk-engelsk). Dette systems oversættelser er imidlertid ikke tilfredsstillende og Kommissionen har derfor taget initiativ til at få udarbejdet en plan for et bedre system i Europa. Det europæiske system skal være flersproget, skal kunne oversætte mellem de 6 EF-sprog og skal kunne udvides til at omfatte nye sprog. I arbejdet med projektformuleringen har vi deltaget fra Danmark (siden sept. 1978).

I arbejdsgruppen deltager i øvrigt forskere fra følgende universiteter:

Leuwen

Manchester/Essex

Grenoble

Pisa

Saarbrücken

Arbejdsgruppens formand er englænderen Margaret King.

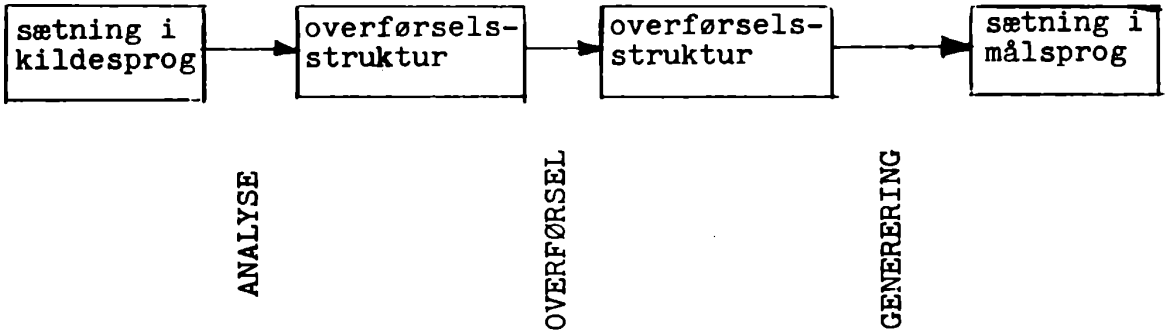
Det overordnede princip ved udformningen af projektet er, at det skal udvikles decentralt: Så meget som muligt skal udvikles separat og lokalt for det enkelte sprog af en arbejdsgruppe, der har det pågældende sprog som modersmål. Oversættelsesprocessen er derfor delt i tre dele: analyse, overførsel og generering, af hvilke de to, analyse og generering, er ensprogede og udvikles helt af

Strukturering af lingvistiske data til brug ved maskinoversættelse
Bente Maegaard, Hanne Ruus
Proceedings of NODALIDA 1979, pages 153-160

den enkelte arbejdsgruppe. Overførsel mellem sprogene skal derimod udarbejdes af grupperne i fællesskab, sprogpar for sprogpar.

Det er nok klart, at deltagerne i arbejdsgrupperne, der skal udvikle analyse og generering for et sprog og samarbejde med andre grupper om overførsel, må have det pågældende sprog som modersmål; men dette kunne også opnås ved et centralt system. Når man lægger så stor vægt på, at systemet skal udvikles decentralt, er det især fordi man herved styrker datamatisk lingvistiske miljøer i alle EF-landene.

Oversættelsesprocessen i EUROTRA kan fremstilles således:




Hvis man forestiller sig, at en engelsk tekst skal oversættes til dansk, vil det analysemodul, der anvendes, være udarbejdet af den engelske gruppe, overførselsmodulet i samarbejde mellem den engelske og den danske gruppe, og generering vil være udarbejdet af den danske gruppe. Analysemodul og genereringsmodul er uafhængige af henholdsvis målsprog og kildesprog; det samme engelske analysemodul benyttes altså ligegyldigt, hvilket sprog der skal oversættes til, hvorimod der udarbejdes to overførselsmoduler for hvert sprogpar. Det er derfor vigtigt, at overførslen begrænses til det allernødvendigste: det, som kræver, at man har adgang til informationer om begge sprog på én gang.

Inden for rammerne af det tilbudte, fælles programmel kan den enkelte arbejdsgruppe selv bestemme sin strategi, når blot resultater, der skal være input til næste modul, opfylder helt bestemte, veldefinerede krav, som fastlægges i projektbeskrivelsen. Resultatet af analysen, såvel som resultatet af overførslen, udmøntes i en overførselsstruktur: tekststykker og sætninger beskrives ved træstrukturer.

En overførselsstruktur består af et eller flere dependenstræer, hvor hver knude er forsynet med etiketter på forskellige sproglige analyseniveauer: morfosyntaktisk, syntaktisk og logisk-semantisk. Det er herved muligt at bevare ordstillingen fra kildeteksten i bunden af træstrukturen samtidig med, at man har oplysninger om sætningens dybdestruktur i træets etiketter.

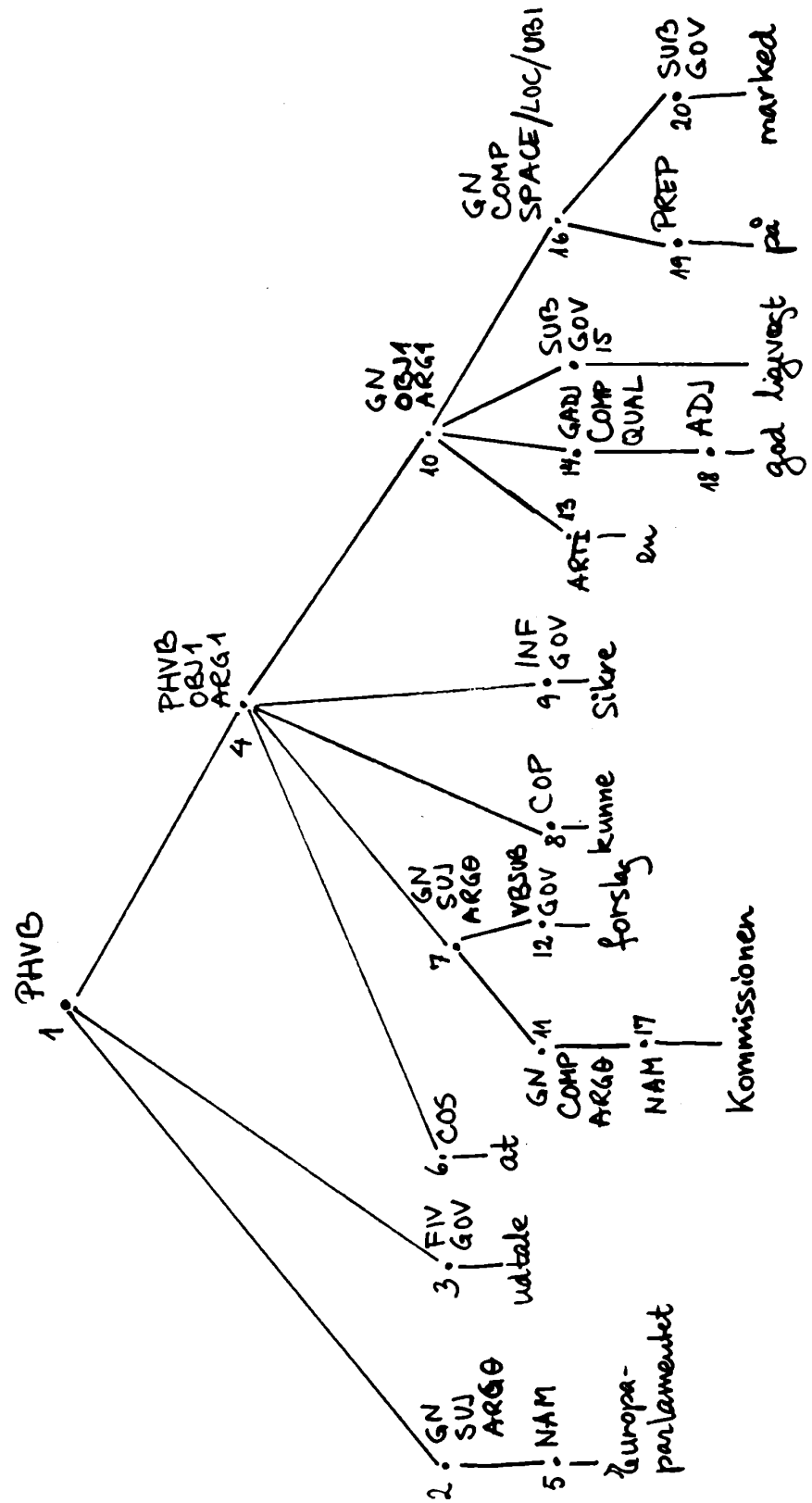



fig. 1.

Europaparlamentet udtalte at Kommissionens forslag kunne sikre en bedre ligevægt på markedet.

I fig. 1 er vist en overførselsstruktur for den danske sætning Europaparlamentet udtalte, at Kommissionens forslag kunne sikre en bedre ligevægt på markedet.

For hver moderknude (knuderne 1,4,7,10,16) gælder det, at en af dens døtre er kerne i forhold til sine søstre. Kernen mærkes med etiketten GOV på det syntaktiske niveau. Søstrene til kernen står i relation til denne gennem moderknuden. F.eks. står knude 2 og 4 i relation til kerneknuden 3. En kerneknude dominerer altid direkte et blad på træet, der indeholder henvisning til den leksikalske enhed, som indeholder det ord eller udtryk i inputsætningen, som indgår på den pågældende plads i træet. Etiketter på knuderne anføres i rækkefølgen morfosyntaktisk funktion, syntaktisk funktion, logisk-semantisk funktion: f. eks. er knude 2 en nominalgruppe (GN), der syntaktisk er subjekt i sætningen (SUJ). Logisk-semantisk er den dybdesubjekt (ARG0) i forhold til kerneknuden 3, som har den morfosyntaktiske etikette finit verbal (FIV) og syntaktisk er kerne (GOV). Træet under knude 4 er morfosyntaktisk en sætning (PHVB), syntaktisk direkte objekt (OBJ1) og logisk-semantisk dybdeobjekt (ARG1). Knude 2 dominerer knude 5, der indeholder oplysning om den morfologiske klasse, proprium (NAM) for Europaparlamentet.

Træets blade er leksikalske enheder og oplysninger om tekstens bøjningsformer findes i den knude, der dominerer den leksikalske enhed, f. eks. indeholder knude 18 oplysninger om, at god i teksten optræder i komparativ.

Knude 11 viser et tilfælde, hvor det logisk-semantiske subjekt ikke falder sammen med det syntaktiske subjekt. Man ser, at Kommissionen er ARG0 for forslag, idet det er Kommissionen, der foreslår noget. I denne model kan verbalsubstantiver altså have subjekter og objekter ligesom de tilsvarende verber, og det logisk-semantiske niveau benyttes til at beskrive dette forhold. Andre anvendelser af logisk-semantiske etiketter ser man i knude 16, hvor etiketten SPACE/LOC/UBI er brugt på et adverbial, der angiver stedet,hvor noget sker. Der findes en række etiketter, der benyttes ved stedsangivelser, og en tilsvarende række for tidsangivelser. Endelig kan man bemærke etiketten QUAL (knude 14), der anvendes på adjektiver i attributiv stilling, på relativsætninger mv. Disse logisk-semantiske etiketter svarer til semantiske roller (deep cases) i Fillmores forstand.

I knude 8 ser man, at kunne er markeret med den morfosyntaktiske etikette COP, dvs. som tilhørende ordklassen hjælpeverber. Dette skal ikke tages som udtryk for en bestemt mening om den mest hensigtsmæssige beskrivelse af modalverber, idet modalitet og hjælpeverber stadig er under overvejelse i arbejdsgruppen.

En vigtig generel egenskab ved træet er, at det er fladt: antallet af niveauer er begrænset, og herved spares både lagerplads og søgetider. Man kan f. eks. sammenligne dette træs grundstruktur med den tilsvarende i sædvanlig IC-analyse (Immediate Constituents) eller en tilsvarende kontekstfri grammatik:

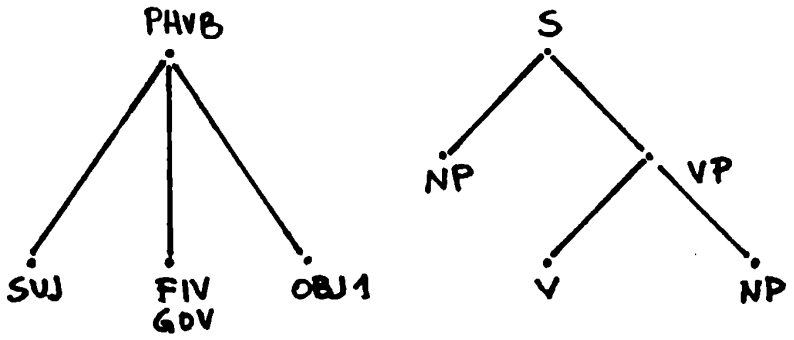



fig. 2.

Dependenstræet har blot tre grene gående ud fra roden, af hvilke en er markeret som GOV, og det giver den samme information som det andet træ (eller endda mere). Også ved beskrivelse af nominalgrupper er det flade træ enklere: se f. eks. en bedre ligevægt (det deltræ, der er domineret af knude 10).

En overførselsstruktur skal indeholde de oplysninger, der er nødvendige for at overføre tekstens indhold fra et sprog til et andet, men heller ikke flere. At de lingvistiske oplysninger på knuderne i overførselstræerne er nødvendige, ser man bedst gennem en beskrivelse af den vej, en overførselsstruktur tilbagelægger fra et sprog til et andet. Vi skal derfor vise i hovedtrækkene, hvordan en fremmedsproget overførselsstruktur ændres til dansk under udnyttelse af de informationer, der findes på træets knuder.

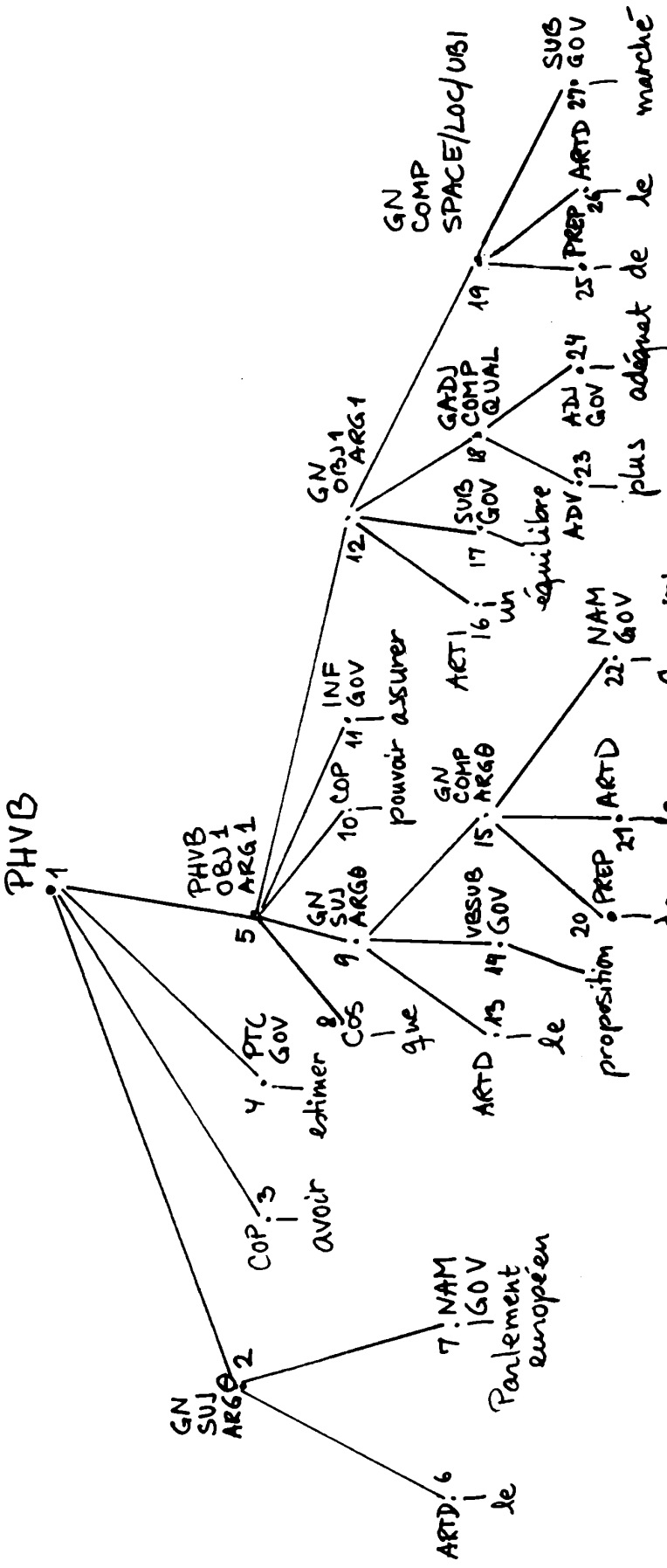
Fig. 3 viser en overførselsstruktur for den franske sætning Le Parlement européen a estimé que les propositions de la Commission pouvaient assurer un équilibre plus adéquat du marché.

Overførslen af det franske træ består af en leksikalsk overførsel, dvs. at de franske leksikalske enheder på træets blade skiftes ud med tilsvarende danske, mens strukturelle ændringer i træet henvises til genereringsfasen.

Det skaber ingen særlige problemer at udveksle Parlement européen med Europaparlamentet, her er den morfologiske etikette proprium (NAM) en ekstrasikring. Tilsvarende veksles équilibre til ligevægt.

Når det franske ord har flere oversættelser til dansk, bliver overførslen imidlertid mere kompliceret og man får brug for etiketterne på de forskellige niveauer. For eksempel kan estimer (knude 4) betyde agte, vurdere, mene og assurer (knude 11) forsikre og sikre. Ved overførsel af assurer skal i dette tilfælde vælges det danske sikre. Dette valg træffes ved at udnytte informationer i de knuder, som assurer er kerne for: når assurer optræder med et ARG0, hvis kerne er non-human, vælges sikre. Ved overførslen af estimer

fig. 3.




Le Parlement européen a estimé que les propositions de la Commission pouvaient assurer un équilibre plus adéquat du marché.

vælges mene, fordi det franske verbums objekt (OBJ1, knude 5) har form af en sætning (PHVB). I den ene valgsituation udnyttes altså semantiske informationer - dels den logisk-semantiske etikette i en knude (ARG0), dels et semantisk træk, der er anført ved en leksikalsk enhed i ordbogen (non-human) - i den anden har vi udnyttet syntaktisk (OBJ1) og morfosyntaktisk information (PHVB) fra etiketterne.

Ved udgangen fra den leksikalske overførsel er den franske overførselsstrukturs leksikalske enheder altså erstattet med de tilsvarende danske. De fleste funktionsord som artikler, konjunktioner og præpositioner udskiftes dog ikke, da deres udfyldning i målsproget først kan bestemmes ud fra syntaktiske og logisk-semantiske etiketter på de knuder, der dominerer dem.

Den overførselsstruktur, der er output fra overførselsmodulet, fungerer som input til genereringsmodulet. Ved generering skal konstituenterne i sætningen og i de enkelte led anbringes i korrekt orden. F. eks. skal knuderne 16,17,18 i det franske træ ændre rækkefølge, således at man fra un équilibre plus adéquat når til en mere passende ligevægt.

Ved genereringen skal også vælges verbaltider og til brug for dette valg må de nødvendige oplysninger findes i overførselsstrukturens etiketter. Det franske træ skal således indeholde information, der gør det muligt at vælge dansk præteritum, mente, som oversættelse for den sammensatte franske verbaltid i knude 3 og 4, a estimé.

Til bestemmelse af funktionsord, der ikke er oversat ved overførslen, benyttes den lingvistiske analyse af konteksten, som er udmøntet i etiketterne, og oplysninger fra den danske ordbog, - f. eks. om hvilke præpositioner et verbum konstrueres med. I den her betragtede overførselsstruktur for den franske sætning skal que i knude 8 udskiftes med at, fordi det indleder en sætning, der er objekt (OBJ1) for mene.

Præpositionen de i nominalgruppen domineret af knude 19, du marché, skal oversættes ved på. Denne afgørelse træffes ved hjælp af den logisk-semantiske etikette SPACE/LOC/UBI sammenholdt med ordbogsoplysninger for den leksikalske enhed marked.

Som et sidste eksempel på, hvad genereringsfasen må omfatte, skal nævnes anvendelse af morfologiske regler: substantiver, adjektiver mm. skal bøjes, der skal være kongruens visse steder osv. Som et lidt mere indviklet tilfælde kan man betragte de knuder, der i det franske træ er domineret af knude 15, de la Commission. De skal på dansk blive til Kommissionens. Den regel, der sørger for det, vil gå ud på, at en nominalgruppe, der er ARG0 for et VBSUB, på dansk sættes i genitiv. De logisk-semantiske etiketter bruges altså ikke blot ved overførsel, men også i genereringsfasen.

Med denne gennemgang af en overførselsstruktur for en forholdsvis enkel, dansk sætning og den skitsemæssige overførsel af en tilsvarende fransk overførselsstruktur fra fransk til dansk har vi vist, hvilken slags informationer

der skal bruges i det planlagte oversættelsessystem; vi har derimod kun antydet, hvor disse informationer skal komme fra.

Det vil afhænge af den enkelte gruppes valg af analysestrategi, hvornår de vil fremtage og bruge oplysningerne på de forskellige lingvistiske niveauer i deres arbejde hen mod overførselsstrukturer: én gruppe kan støtte sig mest på semantiske oplysninger, en anden på syntaktiske oplysninger i analysefasen. Det er dog fastlagt, at alle lingvistiske oplysninger skal hentes fra grammatikker og ordbøger, der er adskilt fra og uafhængige af det anvendte programmel.

Vi vil således kunne udbygge og anvende den danske morfologiske analyse, som vi bruger i DANwORD (se f. eks. SAML III, 4, 5), ligesom en del af de frekvensoplysninger, vi fremtager i DANwORD, vil kunne indgå i arbejdet med at formulere regler for valg af den rette oversættelse ved flertydige ord og udtryk. Selv med den forholdsvis udførlige lingvistiske analyse, som man sigter mod i EUROTRA, kan man nemlig ikke vente, at man altid har informationer nok til at vælge mellem to oversættelser ud fra træk i de lingvistiske omgivelser, ligesom man heller ikke kan forvente, at analysen af alle perioder vil resultere i én overførselsstruktur. I sådanne tilfælde vil oplysninger om hyppige og sjældne ord og konstruktioner være en god støtte til at træffe det rigtige valg.