Foredrag ved Nordiske Datalingvistdage på IAML, Københavns Universitet, 9.-10. oktober 1979.

Henrik Holmboe:

Lemmatisering - hvilke af de ideelle krav til lemmatisering er opfyldelige eller opfyldte?

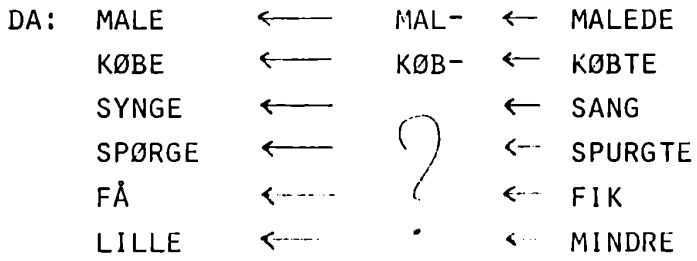
Lemmatisering er en term, der er kurant i snævre kredse, hvor termen anvendes på en måde, der vel ikke er entydig, men dog har et centralt betydningsområde, som alle er enige om hører med til termen. Hvis man konsulterer en række gængse lingvistiske terminologiske ordbøger eller oversigtsværker, konstaterer man imidlertid, at termen ikke er optaget og defineret i disse værker.

At lemmatisere betyder minimalt at henføre et ord fra en tekst til en bestemt type eller kategori, som det i teksten aktuelt forekommende ord kan påstås at være en bøjet form af. Dette forudsætter en analyse af ordet og eventuelt dets omgivelser i teksten, men behøver ikke at forudsætte informationer, der ligger uden for ordet og teksten selv.

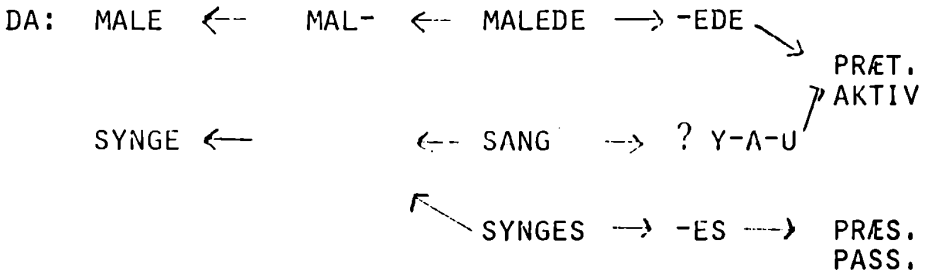
DA: MAL- ⟵ MALEDE
KØB- ⟵ KØBTE

Ud over dette minimale krav vil man eventuelt også forlange, at lemmatiseringen skal resultere i en henvisning til den leksikalske indgang, som ordet skal søges under i gængse ordbøger, hvad enten dette er en abstrakt form eller en bestemt forekommende bøjningsform af ordet. Dette vil implicere viden, der ikke nødvendigvis er til stede i ordet eller teksten på stedet.

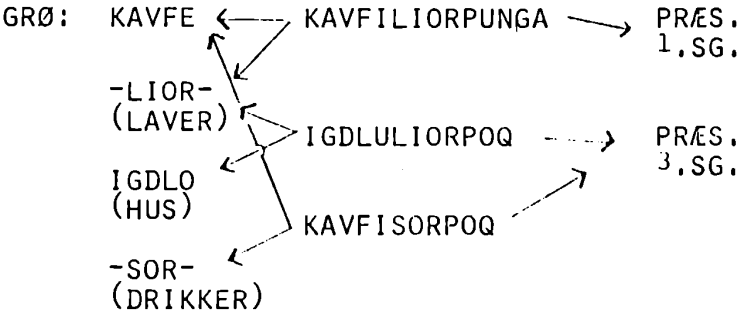
Lemmatisering - hvilke af de ideelle krav til lemmatisering er opfyldelige eller opfyldte?
Henrik Holmboe
Proceedings of NODALIDA 1979, pages 145-152



Endvidere vil man eventuelt forlange, at lemmatiseringen skal henføre ikke blot den del af ordet, der rummer dets centrale betydning, til en leksikalsk type eller kategori, med alle ordets dele til leksikalske eller morfologiske typer eller kategorier. Dette vil implicere en endnu større viden.

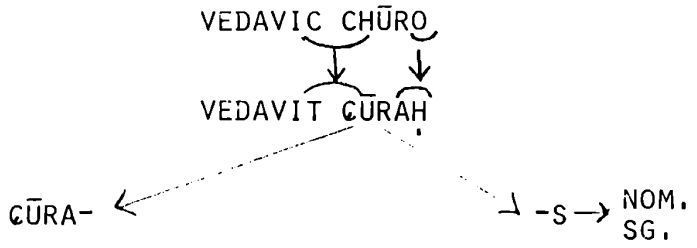


Men dette krav er vigtigt, hvis man vil opnå en mere generel definition af ordet lemmatisering, således at termen kan anvendes også i forbindelse med sprog, der f.eks. tillader mere end én orddel med det, vi vil kalde central betydning, inden for ét og samme ord:

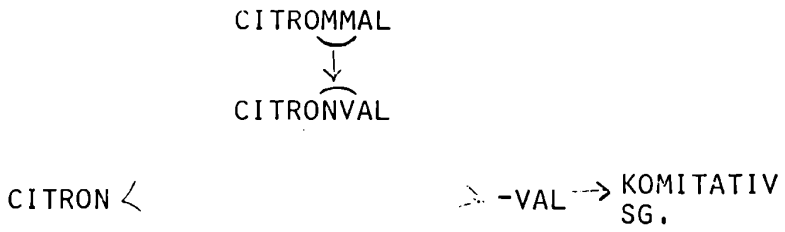


Lemmatiseringen skal også kunne opløse sandhi-fænomener internt og eksternt af forskellig kompleksitet (assimilation, fusion):

ASSIMILATION:
SANSKRIT:

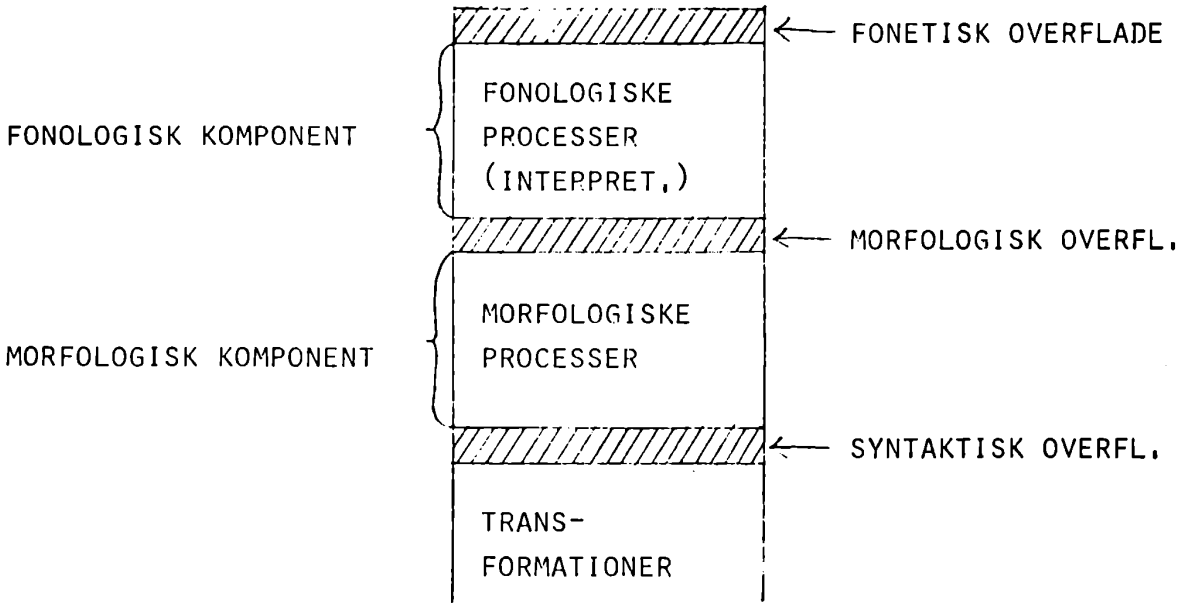


FUSION:
UNG.:

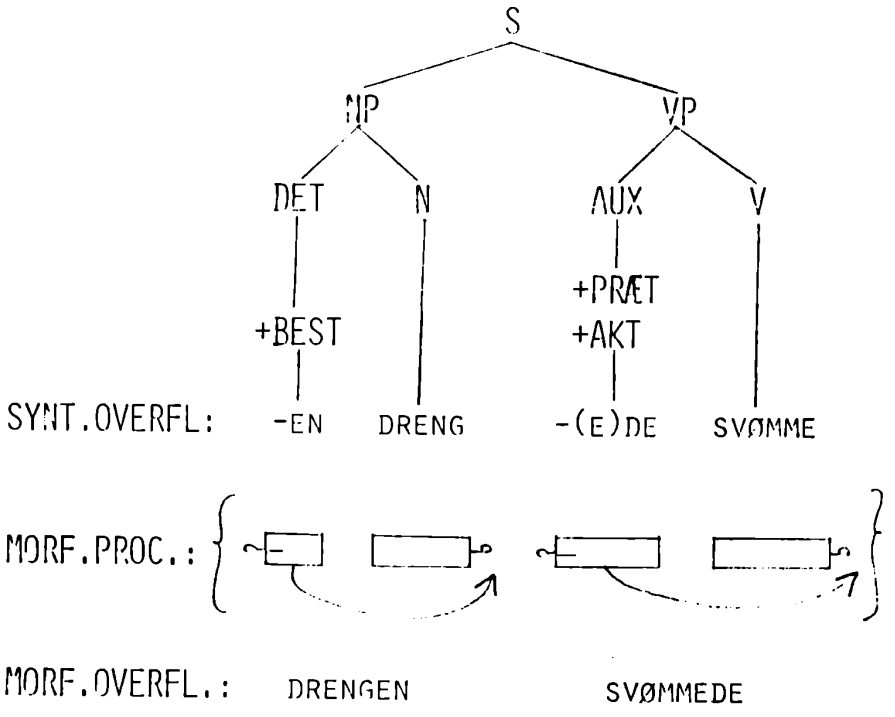


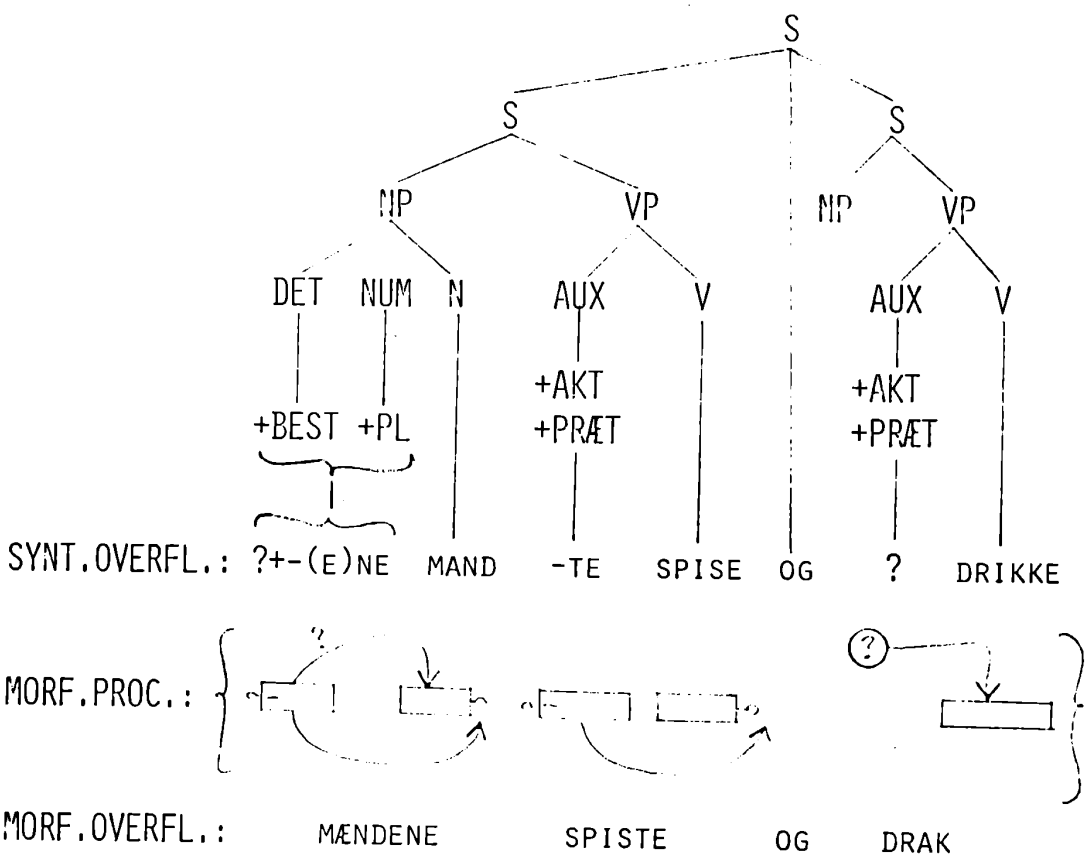
Når termen lemmatisering ikke kan siges at være en gængs lingvistisk term i almindelighed uden for leksikografers og datalingvisters kreds, skyldes det ikke, at lingvistikken ikke har beskæftiget sig med det problemkompleks, som lemmatisering omfatter, men at beskrivelsen af disse problemer skal søges under disciplinerne morfologi og fonologi og evt. morfofonologi. Morfologisk analyse og lemmatisering skal altså kunne opvise en række fælles resultater. Ser vi på en transformationsgrammatisk model, møder vi straks den vanskelighed i f.eks. Aspects-modellen, at TG beskæftiger sig meget lidt med morfologi. Den går næsten direkte fra transformationer af hovedsagelig syntaktisk natur til en fonologisk komponent, der producerer den fonetiske

repræsentation. Ind imellem skyder man de såkaldte readjustment-rules, hvis status ikke er ganske klar. En lidt ændret transformationsgrammatisk model kunne være følgende, idet kun de hér relevante dele er medtaget:



f.eks.:





Anskuet på denne måde er det lemmatiseringens opgave at bevæge sig fra den morfologiske overflade til den syntaktiske overflade.

Jeg vil ikke hævde, at en lemmatisering skal forløbe som et baglæns gennemløb af den morfologiske komponent, men resultaterne af de to skulle gerne være sammenfaldende: inputtet til den morfologiske komponent (eller den del af grammatikken, der rummer de såkaldte readjustment-rules) skal være det samme som outputtet fra en lemmatiseringsproces. Dvs. en lemmatisering "skal kunne" det samme som en grammatiks morfologiske komponent (hvad enten denne anskues som analyserende eller genererende), men ikke nødvendigvis på samme måde. Dette må være det ideelle krav til lemmatisering.

I lingvistikken har man beskrevet indholdet af den morfologiske komponent på lidt - men ikke meget- forskellig måde. Bl.a. J.H. Greenberg og P.H. Matthews har elaboreret

og forfinet den beskrivelse, vi finder hos Sapir i Language fra 1921, men for ikke at fortabe mig i detaljer vil jeg i store træk holde mig til Sapir, der omtaler følgende morfologiske eller grammatiske processer:

A: AFFIXERING:

	FUSING	JUXTAPOSING
PRÆ- IN- SUB-	+ SANDHI	- SANDHI

B: INTERN MODIFIKATION:

VOKALHARMONI
OMLYD } FJERNASSIMILATION { PROGRESSIV
REGRESSIV
AFLYD
STADIEVEKSLING

A: REDUPLIKATION:

PRÆFIGERING AF EN "DUBLETDEL".

Vokalharmoni og i visse tilfælde omlyd kan anskues som henholdsvis progressiv og regressiv fjernassimilation og er altså forudsigelige eller redundante. Dette gælder ikke aflyd som vi kender den fra f.eks. germanske stærke verber eller intern flexion i arabisk.

Dette er altså fundamentalt to typer af processer:

A. Den ene vil jeg kalde <u>forøgelse</u>, dvs. noget forøges med noget andet. Herunder hører affixering og reduplikation. Lemmatiseringsopgaven er her at identificere det, der er blevet forøget, og forøgelsen.

B. Den anden vil jeg kalde <u>mønsterændring</u>, dvs. ét mønster erstattes af et andet; f.eks.

B: ARAB.: KITĀB ⟵ KUTUB (PL)
RASŪL ⟵ RUSUL (PL)
BAJT ⟵ BUJUT (PL)
JAUM ⟵ AJJĀM (PL)
AJŪĀM (*PL)

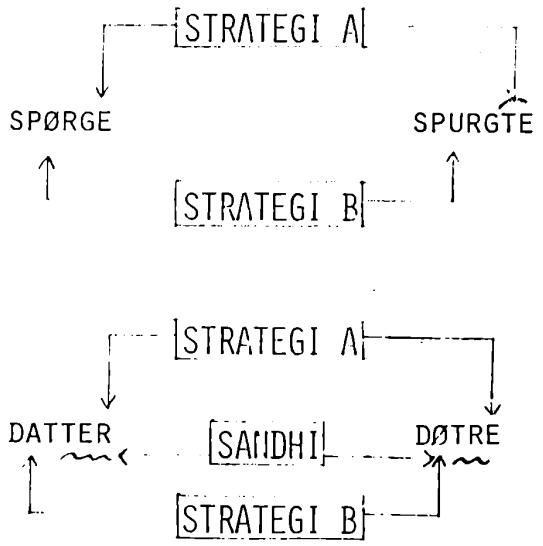
DA.: SPRINGE ⟵ SPRANG (PRÆT)
GÅS ⟵ GÆS (PL)

Herunder hører intern modifikation.

Lemmatiseringsopgaven er hér at skelne mønsteret fra baggrunden og derpå identificere disse to. I begge tilfælde kan intern og extern sandhi sløre billedet. Der vil så vidt jeg kan se være tale om to helt forskellige lemmatiseringsstrategier alt efter, om man skal identificere en forøgelse eller et mønster.

Resultaterne og erfaringerne viser, at man lettest kan automatisere lemmatisering af agglutinerende strukturer uden sandhi. Sandhi-fænomener er brydsomme, men ikke uovervindelige. I systemer, der mestrer problemer af denne type, vil strategien over for mønsterproblemerne være at henvise disse til undtagelseslister. Denne strategi er langt fra ideel, men praktisk og anvendelig, så længe man beskæftiger sig med sprog, hvori de agglutinerende fænomener er de hyppigste og de flekterende befinder sig i relativt små, lukkede klasser, men strategien ville være uanvendelig, hvis forholdet var det omvendte.

Af de ideelle krav til lemmatisering mangler man at opfylde dem, der vedrører mønstergenkendelse og formodentlig også metoder til at styre, hvornår den ene og hvornår den anden strategi skal bringes i anvendelse.



Den almindelige opfattelse af, hvad der er svært og let, svarer nøje til, hvad vore maskinelle metoder i dag kan klare: tyrkiskens morfologi, der er agglutinerende næsten uden sandhi, er lettere end f.eks. ungarskens, der er agglutinerende med sandhi. Noget sværere er f.eks. sanskrit med sin blanding af agglutinerende og flekterende morfologi med udstrakt sandhi og sværest er klassisk arabisk med sin internt flekterende morfologi med en del sandhi.

Henrik Holmboe
Institut for Lingvistik
Aarhus Universitet