Henrik Holmboe

Grammatisk beredskab

Hvor stærkt et grammatisk beredskab skal man være i besiddelse af, når man skal lave en morfologisk eller syntaktisk analyse af en tekst? Ideelt set er kravet, at det grammatiske beredskab forstået som viden om grammatiske fænomener og kategorier skal være så stærkt, at man er i stand til korrekt at analysere en hvilken som helst korrekt sætning eller ytring af sproget. Dette ideelle krav gælder, hvad enten man konstruerer en morfologisk eller syntaktisk parser til sproget, eller man lærer et menneske, f.eks. en gymnasieelev eller en studerende, at analysere et givet sprog. Når man vil være i stand til at analysere en hvilken som helst sætning i sproget, betyder det altså, at for et hvilket som helst morfologisk eller syntaktisk fænomen, skal den for analysen nødvendige viden indgå i det grammatiske beredskab.

En anden formulering af spørgsmålet om grammatisk beredskab kunne være: Hvor lille et beredskab kan man klare sig med? Denne formulering forudsætter, at man er villig til at acceptere, at analysen ikke altid lykkes. Man skal altså overveje, hvor ofte man kan acceptere en utilstrækkelig analyse. Har man en rimelig komplet grammatik, vil det i regelen kræve en ganske stor indsats at gøre den helt komplet, og udbyttet af denne indsats vil være marginal for analyseresultatet; hvis man omvendt har en komplet grammatik, der af én eller anden grund er uhåndterlig, vil den sandsynligvis kunne reduceres betydeligt uden nævneværdig skade for det endelige analyseresultat.

Grammatisk beredskab
Henrik Holmboe
Proceedings of NODALIDA 1985, pages 83-93

Kravet til det reducerede grammatiske beredskab må blot være, at det er umiddelbart ekstenderbart i retning af den komplette grammatik, således at en ekstension ikke bliver en modsigelse af den reducerede grammatik.

Det ideelle grammatiske beredskab er i god overensstemmelse med Hjelmslevs empiriprincip, som han formulerede det i "Omkring Sprogteoriens Grundlæggelse", Kbh. 1943. Ifølge dette princip skal sprogbeskrivelsen være modsigelsesfri, udtømmende og den simplest mulige, dog således, at kravet om modsigelsesfrihed er overordnet kravet om udtømmende beskrivelse og kravet om udtømmende beskrivelse overordnet kravet om simpelhed. Det lille eller reducerede grammatiske beredskab er udtryk for, at simpelhed prioriteres op, således at simpelheden får lov at dominere, når blot beskrivelsen er tilnærmelsesvis udtømmende.

Som udgangspunkt for opstillingen af en reduceret grammatik har jeg valgt 3. bog af Caesars Gallerkrig. Antagelsen er nu, at det for den grammatiske analyse af denne tekst ikke er nødvendigt at have et morfologisk beredskab, der ville være tilstrækkeligt stærkt til at parse en hvilken som helst latinsk tekst fra den klassiske periode og til at konstruere alle mulige former af et hvilket som helst latinsk ord fra samme periode. Målet er altså at foretage reduktioner, der er tilnærmelsesvis omkostningsfrie for analysens slutresultat.

Ikke overraskende viser det sig, at nominalmorfologien ikke kan reduceres, men at verbalmorfologien kan. På grundlag af en

retrogradliste med frekvensangivelse af tekstens ord er formerne optalt og opstillet i skema 1 og skema 2. Skema 1 er de absolutte frekvenser, skema 2 er formernes relative frekvens i % i forhold til samtlige verbalformer. De 788 verbalformer udgør 21.5% af tekstens ord. Foruden de i skemaet medtagne former optræder der 2 eksempler på 1. pl.præs.indik.akt., hvilket er 0.25%. Kolonnen Akt.+ Dep. angiver, hvor mange former i alt der har aktiv betydning, kolonnen Dep.+Pass., hvor mange der har passiv form. Tallene viser herefter, at de finitte former er 3. person (sg. el. plur.), at aktive former er 8 gange så hyppige som passive former, at infinitiv og perf.partc. deler de fleste af de infinitte former, samt at der kun er ca. 30 forskellige former i corpus. Ved finitte verbalformer skal her forstås usammensatte verbalformer. Sammensatte verbalformer skal opfattes som en kombination af perf.partc. og finitte former. Da der i alt er 69 former af verbet SUM, betyder det, at man ved den maximale justering får sammensatte former for 63.3% finitte former og 36.7% infinitte former.

Skema 3 viser formerne ordnet efter rang. Den stiplede linie mellem rang 11 og 12 viser, at de 11 former, hvis frekvens er over gennemsnitsfrekvensen 3.3, dækker 79.3% af verbalformerne. Linien over rang 20 viser, at de former, hvis frekvens er mindre end 1%, dækker 6.8% af verbalformerne.

Den endelige opstilling af den reducerede verbalmorfologi i skema 4 bygger på det princip, at det ikke er tilfældigt, at de hyppigste former er de hyppigste, men at det til gengæld er

ABS.		AKT.	+	DEP.	+	PASS.	I ALT
PRÆS.IND.	3.sg.	34	34		2	2	36
	3.pl.	41	42	1	6	5	47
- KONJ.		11	12	1	2	1	13
		7	9	2	4	2	11
IMPF.IND.		43	43		5	5	48
		38	41	3	10	7	48
- KONJ.		22	25	3	16	13	38
		46	48	2	6	4	52
PERF. IND.		49					49
		33					33
- KONJ./FUT:EX.		3					3
PLUSQ.PF.IND.		12					12
		17					17
- KONJ.		10					10
		13					13
FIN. I ALT		181	185	4	25	21	206
		198	206	8	26	18	430 224
INF. PRÆS.		88	105	17	44	27	132
- PERF.		7					7
PARTC.PRÆS.		4					4
- FUT.		6					6
- PERF.						167	167
GERUNDIV							42
INFIN. I ALT							358
VB. I ALT							788

%		AKT.	+	DEP.	+	PASS.	I ALT
PRÆS.IND.	3.sg.	4.3	4.3		0.3	0.3	4.6
	3.pl.	5.2	5.3	0.1	0.8	0.6	6.0
- KONJ.		1.4	1.5	0.1	0.3	0.1	1.6
		0.9	1.1	0.3	0.5	0.3	1.4
IMPF.IND.		5.5	5.5		0.6	0.6	6.1
		4.8	5.2	0.4	1.3	0.9	6.1
- KONJ.		2.8	3.1	0.4	2.0	1.6	4.8
		5.8	6.1	0.3	0.8	0.5	6.6
PERF. IND.		6.2					6.2
		4.2					4.2
- KONJ./FUT:EX.		0.4					0.4
PLUSQ.PF.IND.		1.5					1.5
		2.2					2.2
- KONJ.		1.3					1.3
		1.6					1.6
FIN. I ALT		23.0	23.5	0.5	3.2	2.7	26.1
		48.1			6.5		54.6
		25.1	26.1	1.0	3.3	2.3	28.4
INF. PRÆS.		11.2	13.3	2.2	5.6	3.4	16.8
- PERF.		0.9					0.9
PARTC.PRÆS.		0.5					0.5
- FUT.		0.8					0.8
- PERF.						21.2	21.2
GERUNDIV							5.3
INFIN. I ALT							45.4
VB. I ALT							100.0

RANG		%	ACCUMUL.
1:	perf.partc.pass.	21.2	
2:	inf.præs.akt.	11.2	
3:	perf.ind.akt.sg.	6.2	
4:	impf.konj.akt.pl.	5.8	
5:	inf.præs.pass.	5.6	50
6:	impf.ind.akt.sg.	5.5	
7:	gerundiv	5.3	
8:	præs.ind.akt.pl.	5.2	
9:	impf.ind.akt.pl.	4.8	
10:	præs.ind.akt.sg.	4.3	
11:	perf.ind.akt.pl.	4.2	79.3
			3.3
12:	impf.konj.akt.sg.	2.8	
13:	plusq.pf.ind.akt.pl.	2.2	
14:	impf.konj.pass.sg.	2.0	
15:	plusq.pf.konj.akt.pl.	1.6	
16:	plusq.pf.ind.akt.sg.	1.5	
17:	præs.konj.akt.sg.	1.4	
18:	impf.ind.pass.pl.	1.3	
19:	plusq.pf.konj.akt.sg.	1.3	
20:	inf.perf.akt.	0.9	6.8
21:	præs.konj.akt.pl.	0.9	
22:	præs.ind.pass.pl.	0.8	
23:	impf.konj.pass.pl.	0.8	
24:	partc.fut.akt.	0.8	
25:	impf.ind.pass.sg.	0.6	
26:	partc.præs.akt.	0.5	
27:	præs.konj.pass.pl.	0.5	
28:	perf.konj./fut.ex.akt.pl.	0.4	
29:	præs.ind.pass.sg.	0.3	
30:	præs.konj.pass.sg.	0.3	

	Ho. IND.		INF.		BI. KONJ.	
	AKT.	PASS.	A.	P.	A.	P.
P R Æ S.	-T -NT				-T	
I M P F.	-BAT -BANT	-BANTUR	-RE		-RET -RENT	-RETUR
P E R F.	-T -ERUNT		-ISSE			
P L U S Q. P F.	-ERAT -ERANT				-ISSET -ISSENT	
P.P.P.	-T- -(S)S-					
G E R U N D I V	-ND-					

tilfældigt, at de sjældneste former overhovedet er med i vores corpus. Eller formuleret lidt anderledes: Vi kan antage, at de hyppigste former i dette corpus også vil være de hyppigste i andre lignende, men at vi rimeligvis vil møde andre former end de sjældneste i stedet for dem, der -tilfældigvis - forekommer i dette corpus; men disse andre vil så nok være tilsvarende sjældne, - så sjældne, at en henvisning af dem til en punktkommentar ligger inden for rammerne af betegnelsen tilnærmelsesvis omkostningsfri for analysens slutresultat.

Ligesom det er omstændeligt at lære et menneske en udtømmende morfologi, kræver det også mange regler og mange regelniveauer at lære en datamat en morfologi, således som det er sket i f.eks. lemmatiseringsprogrammel, der er opbygget over den almindelige morfologis regler. Et lemmatiseringsprogram til analyseformål skal bl.a. kunne afgøre, om en given kontextform kan henføres til et bestemt lemma, om kontextformen "a" er en bøjningsform af lemmaet "b". Formålet kunne være, at man ønsker udskrevet alle belæg af et ord i dets forskellige bøjningsformer fra en tekst. Igen kan vi stille spørgsmålet: hvor mange fejl er vi villige til at acceptere i forbindelse med en metode, der er simplere end den klassiske lemmatisering? Mit datamateriale har igen været latin, dvs. et sprog af en overvejende eksternt flekterende, suffigerende type, eller "simpelt" udtrykt, hvor bøjning er noget, der foregår i slutningen af ordet. Udgangspunktet var endvidere, at metoden hellere måtte medtage for meget i første approximation end udelade ord, der burde have været medtaget.

En almindelig morfologisk analyse skal kunne identificere en stamme og det sæt af endelser, denne stamme kan være forsynet med. Endelserne kan være simple eller komplekse; man kan evt. diskutere, om man vil tale om en udvidet stamme med en simpel endelse eller en konstant stamme med komplekse endelser. Både stamme og endelse er ret konstante, dog kan specielt vokalalternationer forekomme, og ved overgangen mellem stamme og endelse kan sandhifænomener sløre stammens udlyd eller endelsens forlyd. Den udtømmende viden om alt dette er ganske kompleks.

Den simplere fremgangsmåde er at definere stammen som en konstant søgestreng, som altid skal forekomme som den initiale delstreng, hvis et ord skal betragtes som en mulig bøjningsform af et lemma. Hvis man søger med stammen som udgangspunkt, bliver resultatet af søgningen for upræcis; alle ønskede former kommer med, men resultatet indeholder for meget "støj". For at eliminere denne støj må man gennemføre en mere eller mindre detaljeret undersøgelse af det som følger efter stammen, hvilket ikke nødvendigvis er det samme som det, som morfologien forstår ved endelser. Jo mere detaljeret, desto nærmere vil analysen komme den egentlige morfologi i kompleksitet og viden.

Det har til min egen overraskelse vist sig, at følgende meget simple metode giver et resultat med meget lidt støj:

> for at et ord skal være en mulig bøjningsform, skal det 1) indeholde en bestemt stamme 2) have en total længde på et antal bogstaver, som ligger mellem stammen + den korteste mulige endelse og stammen + den længste mulige endelse

- dette forsynet med en -que og -ve-test.

Man kan sagtens forestille sig argumenterne for, at denne metode skulle give et upræcist resultat; derfor min overraskelse over, at resultatet blev meget "støjfrit" - på ca. 11.000 løbende ord (Plinius naturhistorie bog 36).

Efterskrift:

Spørgsmålet om, hvorvidt den reducerede grammatiks inventar svarer til det morfologiske inventar, man møder tidligt i modersmålstilegnelsen, - om, hvorvidt dette evt. genspejler en universel rækkefølge i erhvervelsen af morfologiske kategorier, er vanskeligt at besvare, fordi undersøgelser af primært materiale, dvs. børns spontane sprogproduktion, er så sjældne. Man har heller ikke mig bekendt undersøgelser af den mothertalk eller caretaker-talk, som børn i vid udstrækning er henvist til at lære ders modersmål af.

Jeg har derimod foretaget en lille undersøgelse af et par børnebøger for helt små børn og kunnet konstatere, at dette sprog, som man læser op for små børn, har følgende karakteristika:

genetiv forekommer ikke

indefinitte former er meget hyppigere end definitte.

definitte nominalsyntagmer får deres definithed ikke ved morforlogisk men ved syntaktisk markering, dvs. ikke ved efterhængt artikel, men ved rækkefølgen Det Adj N.

Jeg har så undersøgt nogle børnestile fra 3. - 7. klasse, og det ser ud til, at den skriftlige sprogproduktion i 3 kl. har meget færre genetiver og også færre definitte former end i de højere klasser.

Materialet er som sagt spinkelt og kan ikke bære vidtgående konklusioner. Min egen hypothese vil gå ud på, at man ikke skal vente at finde en universel rækkefølge i tilegnelsen af morfologiske kategorier analogt med den, Roman Jakobsen introducerer for fonologien. Spørgsmålet om, hvad der er morfologiens /a/ og /m/ og hvad der er dens /f/ og /s/, er forkert stillet; den rækkefølge, der nok kan konstateres, vil efter min mening skulle forklares conceptuelt eller semantisk, og ikke morfologisk.