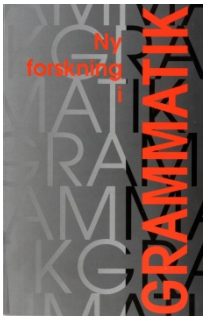
Ny Forskning i Grammatik

Titel: Søgestrenge på tværs af ordklasser

Forfatter: Hanne Jansen

Kilde: Ny Forskning i Grammatik 14, 2007, s. 125-144

URL: http://ojs.statsbiblioteket.dk/index.php/nfg/issue/archive



Søgbarhed
Artiklerne i de ældre numre af Ny Forskning i Grammatik (1993-2012) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Søgestrenge på tværs af ordklasser

Hanne Jansen

1. Indledning

Jeg vil i denne artikel igen tale om "spatialpartikler" – den fælles betegnelse jeg har valgt at bruge for en række præpositioner og retnings- og stedsadverbier, jf. bl.a. Jansen 2002 og 2004 og under udgivelse c – men denne gang fra en ny vinkel der inddrager brugen af elektroniske tekstkorpora og korpusværktøjer. Mine overvejelser udspringer af arbejdet inden for Mulinco-projektet, et samarbejde mellem forskere fra Institut for Engelsk, Germansk og Romansk (Engerom) og Center for Sprogteknologi (CST). Formålet med projektet er at udvikle en elektronisk flersproget korpusplatform (med originaltekster og oversættelser på dansk, engelsk, fransk, tysk, spansk og italiensk), der kan anvendes til både monolingvale og kontrastive studier af lingvistisk, sprogtypologisk, oversættelsesrelateret og stilistisk/litterær art (jf. Farø et al. 2005 og www.cst.dk/mulinco/index.html).

Mit overordnede mål er at undersøge hvordan elektroniske tekstkorpora og korpusværktøjer kan anvendes til at teste mine hidtidige iagttagelser og hypoteser om forskelle i brugen af spatialpartikler på dansk og italiensk. Jeg vil fokusere på mine overvejelser omkring konstruktionen af en søgestreng der kan finde de fænomener jeg er interesseret i, dvs. spatialpartikler på hhv. dansk og italiensk. Denne søgestreng – en kombination af formelle kriterier som søgeprogrammet søger på i korpus – kan siden bruges i mere komplekse søgestrenge svarende til de forskellige konstruktioner som spatialpartiklerne indgår i. Arbejdet med at definere søgestrenge belyser dels overordnede problemer med kategorisering, ikke mindst når kategorien forsøges opstillet på tværs af traditionelle ordklassedefinitioner, dels problemerne ved at gå fra manuel til computerstøttet analyse.

Dette vender jeg tilbage til efter en kort beskrivelse af hvad jeg mener med betegnelsen spatialpartikler, og hvorfor spatialpartikler

er interessante at beskæftige sig med i et dansk/italiensk perspektiv.

2. Spatialpartikler

Spatialpartikler er som sagt den fælles betegnelse jeg bruger for en række præpositioner og retnings- og stedsadverbier der alle har som deres primære og prototypiske funktion at kode rumlige relationer mellem entiteter – som på dansk *på, i, over, under* og *ind, ud, op, ned* og på italiensk *su, in, sopra, sotto* og *dentro, fuori, sù, giu.* Med rumlig relation mener jeg entiteternes placering i forhold til hinanden som koordinater i et tredimensionelt univers[1]. Udover dette grundlæggende semantiske fællestræk fremviser spatialpartiklerne også paralleller på det syntaktiske område, hvor det traditionelle kriterium der anvendes til at skelne mellem præpositioner og adverbier, nemlig + / ÷ styrelse, ikke altid forekommer særligt anvendeligt. En række leksemer kan anvendes på begge måder – *hun sprang* ***i vandet*** overfor *hun sprang* ***i***; *han løb* ***over vejen*** overfor *han løb* ***over*** eller *han løb* ***over til skolen*** – og burde derfor ifølge den traditionelle skelnen kategoriseres som skiftevis præposition og adverbium ("multiple class membership" jf. Croft 2001: 36-40), selvom der ofte ikke er forskel i semantik eller i syntaktisk adfærd i øvrigt. I kraft af deres funktion som "relatorer", udfylder såvel præpositioner som stedsadverbier endvidere ofte en central prædikativ rolle i sætningen, og kan i visse tilfælde overflødiggøre verbet: ***ud!***; ***i*** *seng med dig!*; ***på*** *med hatten,* ***op*** *på cyklen* og ***ned*** *i byen i en vis fart,* eller – dog kun på dansk – konstrueres med modalverber alene: *han ville* ***i*** *byen; hun skal* ***op***[2]. Endelig er der for også for de mest prototypiske medlemmer af gruppen paralleller hvad angår størrelsen: de er netop *particulae,* dvs. "små dele", hvilket gør dem særligt fleksible, ikke mindst i forbindelse med orddannelse, både som præfikser for andre ordklasser og i indbyrdes sammensætninger[3].

---

1. Det er klart at den sproglige kodning af spatiale relationer ikke blot er afbildning af en objektiv og for én gang givet "virkelig" verden, men at den altid er betinget af subjektets fysiske og "intentionelle" placering i forhold til denne verden; se bl.a. om den funktionelle vinkel versus den geometriske vinkel i beskrivelsen af kodning af rumlige relationer i Herskovits (1988).
2. Ikke alle modalverber kan dog i samme udstrækning konstrueres med spatialpartikel alene: *skulle, ville, måtte* kan (med visse begrænsninger); *burde, kunne, turde, gide* kan ikke (dog bruges *gide* + spatialpartikel hyppigt af børn).
3. For uddybende diskussion af fællestræk, se Jansen 2002, 2004 og under udgivelse b.

Både dansk og italiensk har præpositioner og stedsadverbier som en vigtig del af deres leksikon, og ovennævnte fællestræk mellem de to grupper findes på begge sprog, dog med færre tilfælde af ovennævnte "multiple class membership" på italiensk. Ud fra mine hidtidige undersøgelser af især oversættelsesdata[4] kan jeg dog konstatere en række forskelle i både inventar og anvendelse:

a) Frekvens – der anvendes i næsten alle slags tekster langt flere spatialpartikler på dansk end på italiensk; dog er det vigtigt at bemærke at frekvensen også i høj grad, på begge sprog, afhænger af teksttype, register og medium.

b) Variation – der er på dansk større variation og mere semantisk specificitet; variationen skyldes bl.a. at en række danske stedsadverbier har en hhv. dynamisk og statisk form (eller telisk og atelisk), som fx *ud/ude*, en distinktion som på italiensk (og mange andre sprog) ikke angives i leksemet selv, men må læses ud af konteksten.

c) Leksikaliseringsmønstre for bevægelsesverber – en af grundene til forskellen i frekvens er de typiske leksikaliseringsmønstre for bevægelsesverber på hhv. germanske og romanske sprog: hvor man på dansk almindeligvis koder bevægelsens retning separat med en spatialpartikel, vil retningen på italiensk typisk være udtrykt i verbalroden – jf. L. Talmys skelnen (1985) mellem *satellite-framed languages* og *verb-framed languages.*

d) Indbyrdes kombinatorik – som nævnt ovenfor er det karakteristisk for spatialpartiklerne på begge sprog at indgå i ordsammensætninger, ikke mindst indbyrdes; på dansk er frekvensen af sammenstillede spatialpartikler dog væsentligt højere end på italiensk[5].

e) +/÷ transitivitet – ved sammenligning af danske og italienske paralleltekster eller sammenlignelige tekster, vil man på dansk ofte finde en intransitiv konstruktion med spatialpartiklen som "mellemled", med fremhævelse af de spatiale relationer i den givne situation, hvor man på italiensk typisk anvender en transitiv konstruktion der

4. Jf. diskussion af brug af oversættelsesdata i kontrastiv lingvistik i (Jansen under udgivelse a).

5. En af grundene er den hyppige eksplicitering på dansk af subjektets placering i fht den spatiale relation, jf. følgende eksempel fra H.C. Andersens *Sneemanden* og den autoriserede italienske oversættelse: "*han saae stadig* ***ind*** *i Husholderskens KjelderEtage,* ***ned*** *i hendes Stue*" versus "*guardava fisso* [Ø] *nello scantinato della governante,* [Ø] *nella sua stanza*", hvor den italienske læser af konteksten må udlede at snemanden befinder sig ude og oppe i fht. til husholderskens stue.

i stedet fremhæver agens-patiens-relationen. De danske konstruktioner med spatial profilering er ikke så enkle at samle under én fælles beskrivelse, men der er efter min mening ikke desto mindre tale om en klar tendens som tydeligt ses i oversættelsesdata[6].

f) Topologi – på dansk ender mange sætninger i en spatialpartikel eller en spatialpartikel efterfulgt af styrelse, en tendens der ikke genfindes i nær samme grad på italiensk.

3. Fra manuel til computerstøttet analyse: Mulinco-platformen

De her kort opridsede forskelle – der kan samles i en overordnet hypotese om større fokus på den rumlige dimension på dansk – har jeg fundet belæg for i mine hidtidige undersøgelser, der er foretaget med manuel udvælgelse og analyse af især oversættelsesdata. Denne form for tilgang tillader dog kun at arbejde med en begrænset mængde data, hvilket gør det svært at foretage virkelig kvantitativt signifikante generaliseringer. Med et elektronisk tekstkorpus kan man arbejde med store mængder tekst, både ved indsamlingen af data og til en vis grad ved behandlingen (optælling og sortering) af data, og man har dermed et mere solidt empirisk grundlag[7]. De nævnte forskelle er for størstedelen ikke systembetingede, men især et spørgsmål om sprognorm og frekvens, og de er først rigtig iøjnefaldende når man ser samlet på dem, hvilket gør det særlig vigtigt at se på større tekstmængder.

Når man arbejder med et elektronisk tekstkorpus og elektroniske søgeværktøjer, må selektionskriterierne være entydigt og eksplicit formuleret, hvilket kan give problemer: de formaliserede og entydige selektionskriterier vil undertiden inkludere irrelevante data, eller om-

---

6. Jf. følgende eksempler:
"*sagde den unge Mand og pegede paa Sneemanden*" versus "*disse il giovane e indiò* [Ø] *il pupazzo di neve*".
"*»Det skrupknager i mig, saa deiligt koldt er det!«*" versus "*»Mi sento tutto scricchiolare* [mig jeg-føler helt knage] *con questo bel freddo!«*"
[om rimfrost] "*som strømmede der en hvid Glands ud fra hver Green.*" versus "*come se ogni ramo emettesse un fulgore bianco* [som om hver gren udsendte en glans hvid]."

7. Jf. den store interesse der indenfor de sidste årtier er blevet rettet mod brugen af elektroniske tekstkorpora i studier indenfor oversættelsesvidenskab, ofte kombineret med kontrastiv lingvistik, bl.a. Chesterman 2004, Malmkjær 1998, Munday 1998 og Olohan 2004.

vendt udelukke relevante data, dels fordi de kriterier man kan søge efter, afhænger af korpussets opmærkning, dels fordi computeren ikke kan håndtere flertydighed og glidende overgange.

Denne problematik er efter min mening central i kombinationen af manuel og computerstøttet analyse, og det er den jeg vil forsøge at illustrere med spatialpartiklerne som eksempel. Jeg har arbejdet med H. C. Andersens eventyr *Sneemanden* og den italienske oversættelse *Il pupazzo di neve*[8]. Teksten kan ikke bruges til en egentlig testning af ovennævnte hypoteser om forskelle mellem dansk og italiensk, da den kun er på i alt 1643 ord på dansk og 1922 ord på italiensk, og derfor ikke kan levere kvantitativt signifikante data. Til gengæld er det muligt manuelt at checke om de konstruerede søgestrenge rent faktisk fanger de fænomener der interesserer mig, og således løbende korrigere strengene.

Teksterne i Mulinco-korpusset er lemmatiseret (således at man ved at søge på lemma-formen får vist alle de forskellige bøjningsformer lemmaet forekommer i) og morfologisk opmærket med en Part of Speech- eller POS-tagger (et program der automatisk opmærker alle forekomster efter ordklasse, således at søgeprogrammet kan søge på ordklasser)[9]. Teksterne er kodet, så de kan håndteres af søgeværktøjet CQP (Corpus Query Processor)[10] der gør det muligt dels at søge på ordformer, lemmaer og POS-tags, dels at konstruere søgestrenge der kombinerer forskellige leksikalske og morfologiske kriterier vha. diverse symboler for inklusion, eksklusion, disjunktion, intervaller, søgegræn-

8. *Sneemanden* og respektive oversættelser på engelsk, fransk, tysk, spansk og italiensk indgår i Mulinco-korpusset og er udgangspunkt for en samling artikler om forskellige sprogteknologiske og kontrastive problemstillinger (jf. Maegaard & Schøsler, under udgivelse); de overvejelser jeg præsenterer i denne artikel, er fremlagt mere detaljeret dér.
9. For dansk bruges en POS-tagger udviklet af Center for Sprogteknologi bestående af 49 tags; for italiensk et tagsæt udviklet af Marco Baroni (pt. ved Università di Trento) bestående af 36 tags. En af grundene til at der er flere tags i den danske tagger, er bla. at både pronominer og substantiver er mere specifikt opmærket på dansk; jf. fx N_DEF_SING, N_DEF_SING_GEN, N_INDEF_SING, N_INDEF_SING_GEN, som ville være POS-taggene for hhv. KATTEN, KATTENS, KAT, KATS.
10. Se Query tool CQP i IMS Corpus Workbench © 1993-2003 opbygget af Institut für maschinelle Sprachverarbeitung.

ser etc. CQP kan også gruppere og vise søgeresultaterne samt udføre forskellige simple frekvensanalyser[11].

Fig. 1 viser søgeprogrammets skærmbillede (*interface*) med en søgning i *Sneemanden* på ord opmærket som tilhørende ordklassen præposition [pos="PRÆP"]. Søgningen giver i alt 138 forekomster der vises i deres kontekst [= *sentence*]. For at se hvor mange forskellige præpositioner (*unique strings*) det samlede antal forekomster (*matches*) fordeler sig på, kan man ved brug af funktionerne *Freq Case* og *Freq no Case* (der neutraliserer store og små bogstaver) få opstillet en frekvensliste over de fundne forekomster.

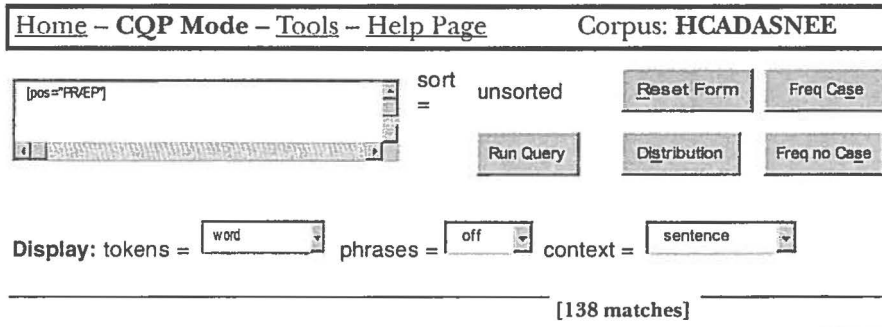


Fig. 1

Udfordringen ligger i, med disse redskaber, at definere en søgestreng der kan bruges til maskinel søgning af spatialpartikler, dvs. kan fortælle computeren hvilke forekomster i teksterne den skal finde og udpege, liste, tælle eller på anden måde behandle for mig.

11. For en mere detaljeret redegørelse for CQP's opbygning og den valgte POS-tagger, se Offersgaard & Olsen og Henriksen, Offersgaard & Povlsen i Maegaard & Schøsler (under udgivelse).

4. Spatiale adverbier

Da jeg med kategorien "spatialpartikler" (SPAT) ønsker at gå på tværs af de traditionelle ordklasser præpositioner og adverbier, kan jeg i mine søgestrenge ikke nøjes med at søge på POS-tags.

Det er særligt tydeligt i forbindelse med adverbier. Adverbier er en meget heterogen klasse der ud over stedsadverbier indeholder mådesadverbier, modalpartikler, diskursmarkører, tidsadverbier etc. For at afgrænse søgningen til kun de spatiale adverbier er det nødvendigt at operere med en leksikalsk definition, dvs. en liste af relevante leksemer som søgemaskinen kan søge på. Dette er muligt, da der er tale om en (rimeligt) lukket klasse, i modsætning til fx substantiver, verber og adjektiver. For at indkredse de relevante leksemer har jeg med søgestrengen [pos="ADV"] søgt alle ord tagget (dvs. opmærket) som adverbier i hhv. den danske og italienske tekst. De 215 danske og 162 italienske forekomster (*matches*) kan dernæst ordnes efter hyppighed vha. *Freq no Case*-funktionen, der reducerer *tokens* til *types*. Af de hhv. 75 danske og 58 italienske forskellige adverbier (*unique strings*) har jeg manuelt udvalgt hhv. 18 og 9 spatiale adverbier (SPAT_adv), jf. tabel 1:

Matches	SPAT_adv (DA)	Matches	SPAT_adv (IT)
8	ind	4	giù
8	op	3	fuori
8	ud	3	sotto
7	ned	2	accanto
3	hen	2	dentro
2	af	2	intorno
2	derind	2	laggiù
2	for	2	sopra
2	frem	1	lassù
2	om		
2	ovenpaa		
1	ad		
1	derinde		
1	deroppe		
1	forneden		
1	rundt		
1	til		
1	udenfor		

Tabel 1. Adverbier

For størstedelen af forekomsterne har den manuelle frasortering af ikke-spatiale adverbier været ret uproblematisk (fx *ikke, nok, meget, igen, aldrig, rødt, underligt* og tilsvarende på italiensk), men i nogle tilfælde vil min afgrænsning af kategoriens medlemmer givetvis kunne diskuteres. Deiktiske stedsadverbier som *her, der* og tilsvarende *ci, qui, lí* er fx ikke inkluderet (når *derinde, derind* og *lassù* er medtaget i listen i tabel 1, er det pga. den efterhængte spatialpartikel), ligesom adverbiet *væk* (eller *borte*) og tilsvarende *via* heller ikke medregnes. Når disse adverbier på trods af deres utvivlsomt (også) spatiale indhold ikke er inkluderet, er det fordi de ikke som deres primære eller centrale funktion har kodningen af spatiale relationer i en eller anden form for tredimensionelt rum, men derimod enten den deiktiske relation (afstand i forhold til udsigeren) eller det faktum at noget ikke er der/ikke kan ses. De kan muligvis betragtes som mere perifere medlemmer af klassen af spatialpartikler, og såvel paralleller som forskelle fortjener afgjort at blive undersøgt nærmere i netop dette lys; i nærværende sammenhæng vil jeg dog begrænse mig til at se på de former der efter min mening udgør de centrale medlemmer af klassen.

Et andet problem angår de forskellige sammenskrevne former. Det gælder sammenskrivninger med deiktiske udtryk på såvel dansk (*derind, derinde, deroppe*) som italiensk (*lassù, laggiù*). Det ville være ønskværdigt at kunne underordne disse forekomster de tilsvarende former uden deiktiske udtryk (altså *ind, inde, oppe, sù* og *giù*), dvs. opfatte dem som en slags varianter over samme lemma – ikke mindst fordi sammenskrivningen ikke altid finder sted. På især dansk gælder problemet også de mange kombinationer af adverbier og præpositioner i sammenskreven form, som *ovenpaa, forneden, udenfor.* Også her gælder det at kombinationen nogle gange sammenskrives, andre gange ikke (alt efter om den er med eller uden styrelse). Bortset fra at reglerne for sammenskrivning ikke følges konsekvent af sprogbrugerne, så giver de i denne sammenhæng problemer omkring optælling af spatialpartikler. De sammenskrevne former må inkluderes som leksemer i leksemlisten for at finde mulige forekomster, men de fundne forekomster tæller dermed som én spatialpartikel i den samlede søgning, selvom man kunne argumentere for at der er tale om to spatiale relationer (ligesom ved de tilsvarende ikke-sammenskrevne former). Som det fremgår, indregner jeg ikke grammatikaliserede italienske

former som *intorno* eller *accanto* som sammenskrevne former i nærværende sammenhæng, selvom det er indlysende at de sproghistorisk er sådanne; det afgørende er om de af sprogbrugeren opfattes som en toleddet størrelse, hvilket på dansk understreges af muligheden for ikke at sammenskrive.

Listen over SPAT_adv der skal indgå i den samlede søgestreng for spatialpartikler (SPAT), vil kunne suppleres med søgninger på [pos="ADV"] i andre tekster for på den måde at nå frem til en tilnærmelsesvis udtømmende liste over spatiale adverbier. En søgning på [pos="ADV"] i Mulinco-platformens samlede danske H.C. Andersen-korpus[12] giver i alt 3764 forekomster og (ordnet vha. *Freq no Case*) 265 unikke strenge, hvoraf 88 kan udskilles som spatiale. Af disse er 18 gengangere i forhold til forekomsterne i *Sneemanden*, mens andre tilføjer nogle af de basale spatialpartikler der mangler i *Sneemanden* (*midt, langs, bag*). Helt overvejende supplerer de dog med flere sammenskrevne former, både SPAT + SPAT (som *foroven, fremad, henimod*) og deiktiske udtryk + SPAT (*herhenne, herind, derud, derunder*), samt kompletterer mønstret af teliske/ateliske former for en lang række adverbier (*hen/henne, om/omme* etc.).

En søgestreng for de spatiale adverbier kan udformes som en kombination af en leksemliste og en søgning på [pos= "ADV"] for at sikre at fx *for, om* eller *af* kun medtages, når de er tagget som adverbier, jf. følgende søgestrenge for SPAT_adv. konstrueret på basis af *Sneemanden* på hhv. italiensk og dansk, samt den udvidede danske søgestreng konstrueret på basis af det samlede H.C. Andersen-korpus:

SPAT_adv (IT) på basis af *Il pupazzo di neve*:

[pos="ADV"&word="giù|fuori|sotto|accanto|dentro|intorno|laggiù|sopra|lassù"%c][13]

SPAT_adv (DA) på basis af *Sneemanden*:

[pos="ADV"&word="ud|ind|op|ned|hen|frem|rundt|derinde|udenfor|deroppe|for|om|af|til|derind|ovenpaa|ad|forneden"%c]

12. Som omfatter 7 eventyr i alt, inklusive *Sneemanden*.

13. Tilføjelsen %c neutraliserer store og små bogstaver.

Udvidet SPAT_adv (DA) på basis af det samlede H.C. Andersen-korpus:

[pos="ADV"&word="ud|ind|op|ned|hen|frem|rundt|tilbage|ude|inde|oppe|derinde|over|midt|bag|forbi|udenfor|deroppe|nede|derom|for|om|deri|forud|imellem|af|derpaa|herinde|til|oven|deraf|derind|dernede|derover|derude|hvori|indenfor|ovenpaa|ved|ad|derhen|derop|derovre|dertil|derved|fornede|henne|langs|ligoverfor|ovre|tvers|under|indeni|rundtomkring|derfra|derhenne|derigjennem|derud|derunder|forfra|foroven|fremad|henimod|herfra|herhenne|herind|hernede|herneden|herovre|herpaa|hertil|herved|hvorpaa|neden|omkring|omme|opad|ovenover|paa|rundtom"%c]

5. Spatiale præpositioner

I sammenligning med adverbierne udgør præpositioner en mere homogen klasse, hvor langt de fleste centrale medlemmer er spatiale i deres grundbetydning. Derfor får man også umiddelbart mere anvendelige resultater med samlede søgninger på hhv. dansk [pos="PRÆP"] og italiensk [pos= "PRE.*"], hvor tilføjelsen [.*] udvider søgningen til også at gælde for sammenskrivninger af præposition og artikel ("le preposizioni articolate", som fx *al, alla, allo, ai, alle, agli, all'*, der alle er "artikelbøjninger" af præpositionen *a*).

Tabel 2 viser de fundne forekomster ordnet vha. *Freq no case*, på dansk ialt 138 forekomster fordelt på 15 forskellige præpositioner, på italiensk i alt 204 fordelt på 10. Som det fremgår af den italienske del af tabellen, må sammentællingen af præpositioner med og uden artikel foretages manuelt.

Matches	PRÆP	Matches	PRE + PRE.:det
42	i	56+22=78	di
20	paa	16+36=52	a
17	med	13+10=23	in
14	af	10+10=20	da
9	til	2+8=10	su
7	for	10+1=11	con
6	om	6	per
6	fra	2	come
5	over	1	sotto
5	under	1	tra
3	ved		
1	efter		
1	hos		
1	op_til		
1	ud_fra		

Tabel 2. Præpositioner

Som sagt giver den samlede søgning på præpositioner i *Sneemands*-teksterne et umiddelbart ret fornuftigt resultat, hvor kun enkelte ikke-spatiale præpositioner skal frasorteres. Det er fx ret uproblematisk at frasortere følgende ikke-spatiale præpositioner: *efter* (primært temporal), *come* (i betydningen *ligesom*, dvs. sammenlignende), samt også *med* og tilsvarende italiensk *con* (og de negative pendanter *uden* og *senza*), der ganske vist kan siges at sammenføje entiteter, men i en mere generisk "ledsage"-relation, ikke primært spatialt. En mulighed for at gøre frasorteringen mere dækkende og præcis er at søge bredt på [pos="PRÆP"] i et større antal tekster og udvide mængden af ekskluderede elementer. For dog at undgå uforudsete ikke-spatiale præpositioner (dvs. forekomster som ikke er blandt de eksplicit ekskluderede elementer), når søgestrengen anvendes på andre ikke så narrative og deskriptive tekster, har jeg valgt også her at eksplicitere hvilke leksemer inden for de fundne præpositioner der skal MED i søgningsresultaterne, dvs. følgende søgestreng: [pos="PRÆP"&word="x|y|z..."%c].

På basis af søgningerne i *Sneemanden* og *Il pupazzo di neve* (jf. tabel 2) har vi nu følgende søgestrenge for spatiale præpositioner bestående af en leksemliste kombineret med en søgning på ordklasseopmærkning (pos="PRÆP" eller tilsvarende italiensk pos="PRE.*"), således at kun

de forekomster af leksemet der er tagget som præposition, bliver medtaget i søgeresultatet.

SPAT_præp (IT) på basis af *Il pupazzo di neve*:

[pos="PRE.*"&word="di|d'|del|della|dello|dell'|dei|delle|degli|a|al|alla|allo|all'|ai|alle|agli|in|nel|nella|nello|nell'|nei|nelle|negli|su|sul|sulla|sullo|sull'|sui|sulle|sugli|da|dal|dalla|dallo|dall'|dai|dalle|dagli|per|sotto|tra"%c]

SPAT_præp (DA) på basis af *Sneemanden*:

[pos="PRÆP"&word="i|paa|af|til|for|om|fra|over|under|ved|hos|op_til|ud_fra"%c]

Udvidet SPAT_præp (DA) på basis af det samlede HCA-korpus:

[pos="PRÆP"&word="i|paa|af|til|for|om|fra|over|under|ved|hos|op_til|ud_fra|mellem|ad|igjennem|gjennem|mod|imod|foran|omkring|bag|imellem|henover|giennem|indenfor|ned_til|udenfor|udi|over_for|overfor|ud_for"%c]

Som det fremgår, er jeg i den italienske søgestreng nødt til at medtage alle præpositionernes "artikelbøjninger", når jeg vil begrænse det samlede søgeresultat af [pos="PRE.*"] til de spatiale præpositioner der er anført i leksemlisten. Endvidere kan man bemærke at de danske lister indeholder flere tilfælde af flerordsforbindelser SPAT + SPAT, som den automatiske *tokeniser* (ordinddeler) har valgt at sammenholde som én ordform (én token) med indsættelse af underscore (*op_til*, *ud_fra*, *over_for*, *ud_for* og *ned_til*). Som i de ovennævnte tilfælde af sammenskrivning vil også de af maskinen sammenholdte former vanskeliggøre en ensartet optælling af spatialpartiklerne.

Hvis man sammenligner søgestrengene (og tabel 1 og 2), ser man at flere leksemer (fx *for*, *til* og *om* på dansk, og *sotto* på italiensk) dukker op ved søgning på såvel adverbium som præposition (jf. ovennævnte "multiple class membership") – et argument for at arbejde med en samlet leksemliste (dvs. uden "dubletter") kombineret med en søgning på enten adverbium eller præposition:

SPAT (DA) **[pos="PRÆP|ADV"&word="x|y|z|..."%c]**
SPAT (IT) **[pos="PRE.*|ADV"&word="x|y|z|..."%c]**

6. Problemer med afgrænsningen spatial/ikke-spatial

Der er dog stadig en del af de genererede forekomster som jeg helst vil undgå. Problemerne skyldes at spatialpartiklerne er så polyfunktionelle – ikke mindst på grund af deres funktion som forbindere ("relator"-funktionen) der "lånes" til andre domæner end det spatiale. Inden for især klassen af præpositioner gør den det vanskeligt at gennemføre en konsistent indkredsning af de relevante leksemer.

At ekskludere den praktisk talt betydningstomme "universalforbinder" *di* fra søgestrengen – dvs. stryge den fra leksemlisten – er ret uproblematisk; det er kun i meget få kontekster den har bevaret sin spatiale betydning. For at skabe balance mellem den italienske og den danske søgning, må man også overveje dansk *af*, hvilket dog er mere diskutabelt, da *af* bruges spatialt i en del kontekster. Er det muligt at få maskinen til at foretage en kun delvis inklusion? Det er det, hvis man udnytter det faktum at *af* tagges som hhv. PRÆP og ADV alt efter om den står med eller uden styrelse. Hvis man udelader *af* i den samlede leksemliste og i stedet efter leksemlisten tilføjer **[pos="ADV"&word="af"%c]**, vil søgningen inkludere fx *af* tagget som ADV i: "*Sneemanden tog af*".

Desværre vil man ikke få inkluderet *af* i spatiale konstruktioner med styrelse, som fx *stige af hesten* eller *tage af bordet*. Ofte optræder den spatialt anvendte præposition *af* dog i forbindelse med andre spatialpartikler: *ud af, ind af, op af, ned af*, jf. fx "*Ilden staaer den ud af Munden*". Disse anvendelser kan dog findes ved at søge på *af* tagget som præposition, men kun hvis den følger umiddelbart efter *ud, ind, op* eller *ned*, jf. følgende streng der omfatter to ord: **[word="ud|ind|op|ned"%c] [pos="PRÆP"&word="af"%c]**. Denne delstreng finder faktisk en del af de forekomster, hvor præpositionen *af* er anvendt spatialt. Når delstrengen efterstilles leksemlisten, opstår der desværre et problem: ordene *ud, ind, op* og *ned* indgår allerede i denne, og når søgningen har fundet én forekomst i teksten der opfylder søgekriterierne, kan den ikke gå tilbage til samme forekomst i forbindelse med næste delsøgning. Søgningen på *ud af, ind af, op af, ned af* må derfor foregå separat, og de fundne forekomster må adderes til søgeresultatet af den samlede streng.

Andre præpositioner, som fx *for* og *om*, volder problemer fordi de også kan fungere som konjunktioner, hhv. sideordnende SKONJ og underord-

nende UKONJ. Den automatiske opmærkning af tekst disambiguerer i mange tilfælde *for* og *om*'s forskellige funktioner, og ingen *for* og *om* der er korrekt tagget som hhv. SKONJ og UKONJ kommer med i søgeresultatet. Men der er en del taggingfejl, hvor det ikke i den automatiske opmærkning af teksten er lykkedes at disambiguere de forskellige funktioner af *for* og *om*. Konjunktionalt anvendte *for* og *om* er således blevet fejlagtigt tagget som enten adverbier eller præpositioner, jf. følgende eksempler: "*men Du maa ikke rasle med Lænken ,/TEGN for/ADV saa knækker det i mig*" og "*har været Herskab ;/TEGN for/PRÆP det var jeg hos Huusholdersken*". Disse taggingfejl kan man prøve at undgå ved i søgestrengen at eksplicitere de kontekster hvor der (højst sandsynligt) er tale om konjunktioner. Det kan fx være kontekster hvor *for* eller *om* følger efter et komma, semikolon, udråbstegn, dvs. [pos="TEGN"], som alle de korrekt taggede konjunktioner *for* og *om* netop gør. I min søgestreng har jeg således valgt at "hjælpe" computeren til et mere præcist søgeresultat ved at udelade *for* og *om* i den samlede leksemliste, og i stedet lave en separat søgestreng der udelukker de forekomster af *for* og *om* der, selvom de er tagget som ADV eller PRÆP, følger efter et TEGN. Jeg vil altså kun medtage *for* og *om* i min søgning, hvis de følger efter en token forskellig fra et TEGN ("forskellig fra" skrives som "!="), altså:

[pos!="TEGN"] [pos="PRÆP|ADV"&word="for|om"%c][14].

En lignende næsten "konjunktional" anvendelse, hvor det spatiale indhold er væk eller i alle tilfælde meget svækket, finder vi i en række forekomster med *for, om, ved, til* efterfulgt af *at* + infinitiv, fx "*lagde sig saa ind i sit Huus for/PRÆP at sove*". For ikke at få disse kontekster med i søgeresultatet, må de pågældende ordformer, som ovenfor, udelades af den samlede leksemliste og i stedet må følgende streng tilføjes:

[pos="PRÆP|ADV"&word="for|om"%c] [word!="at"].

Samme overvejelser gælder den tilsvarende brug af italienske præpositioner som *per, da, a* efterfulgt af infinitiv, [pos="VER:infi.*"]. Her har jeg ligeledes valgt at ekskludere disse kontekster ved følgende separate søgestreng for *per, da, a*:

14. I det samlede HCA-korpus betyder det faktisk en frasortering af 34 forekomster af de i alt 358 forekomster af *for* og *om* tagget som PRÆP eller ADV.

[pos="PRE.*|ADV"&word="per|a|al|alla|allo|all'|ai|alle|agli|da|dal|dalla|dallo|dall'|dai|dalle|dagli"%c] [pos!="VER:infi.*"].

På italiensk ville jeg også gerne frasortere de forekomster af præpositionen *da* (med grundbetydning *fra*), hvor denne anvendes som agensmarkør i passivkonstruktion, jf. "*fu preso **da** qualcosa che non conosceva* [*han blev grebet **af** noget som han ikke kendte*]. Problemet er at denne anvendelse ikke kan defineres ud fra ordklasseopmærkning (*da* tagges her som præposition) og heller ikke ved en særlig leksikalsk/morfologisk kontekst. En eksklusion af denne anvendelse af *da* ville kræve en syntaktisk opmærkning, der fx angav aktive og passive konstruktioner, hvad Mulinco-platformen pt. ikke råder over. Hvad angår den tilsvarende brug af *af* som agensmarkør, er denne allerede stort set elimineret med eksklusionen af *af* som præposition (hvilket selvfølgelig medfører en vis asymmetri mellem de to søgestrenge).

Man kan indvende at der er mange andre tilfælde, hvor de nu indkredsede spatialpartikler anvendes ikke-spatialt, som fx ved domæneskift til temporal anvendelse, jf. "*og **fra** den Tid har jeg staaet i Lænke*", eller i et utal af faste udtryk hvor spatialpartiklen på anden måde bruges i overført forstand, jf. "*det blev Enden **paa** det*". I hvor høj grad er der i disse tilfælde stadig tale om en spatial relation mellem to entiteter? Det drejer sig om glidende overgange, som computeren i sagens natur har svært ved at håndtere, men hvor det også for en menneskelig analysator, der i modsætning til computeren kan arbejde med flere betydningslag på én gang, kan være svært at gennemføre en konsistent skelnen. Jeg er derfor endt med at tage den principbeslutning at medregne alle forekomster hvor spatialpartiklerne ikke fungerer som syntaktisk forbinder (som det italienske *di* eller de forskellige tilfælde af konjunktional anvendelse), idet jeg antager at spatialpartiklen selv ved domæneskift til temporale eller endnu mere abstrakte kontekster bliver ved at signalere rumlig relation. Denne beslutning kan diskuteres, og det er indlysende at en nærmere undersøgelse af de glidende overgange fra klar spatial anvendelse til klar ikke-spatial anvendelse vil være påkrævet i en senere fase (omend en sådan detaljeret semantisk kategorisering synes svær at forene med et computer-operationelt synspunkt).

De samlede søgestrenge ser nu ud som følger:

SPAT (IT) på basis af *Il pupazzo di neve*:
([pos="PRE.*|ADV"&word="in|nel|nella|nello|nell'|nei|nelle|negli|sul|sulla|sullo|sull'|sui|sulle|sugli|sotto|tra|giù|fuori|accanto|dentro|intorno|laggiù|sopra|lassù"%c])|([pos="PRE.*|ADV"&word="per|da|dal|dalla|dallo|dall'|dai|dalle|dagli|a|al|alla|allo|all'|ai|alle|agli"%c] [pos!="VER:infi.*"])

SPAT (DA) på basis af *Sneemanden*:
([pos="PRÆP|ADV"&word="i|paa|fra|over|under|hos|op_til|ud_fra|udi|ndop|ned|hen|frem|rundt|derinde|udenfor|deroppe|derind|ovenpaa|ad|forneden"%c])|([pos="ADV"&word="af"%c])|([pos!="TEGN|SENT"] [pos="PRÆP|ADV"&word="for|om"%c] [word!="at"])|([pos="PRÆP|ADV"&word="ved|til"%c] [word!="at"])

samt den specifikke søgestreng for *af* i konteksten *ud|ind|op|ned* + *af*, som køres separat og hvis søgeresultat adderes manuelt til det samlede resultat:

[word="ud|ind|op|ned"%c] [pos="PRÆP"&word="af"%c])

Resultatet af søgningerne er for *Il pupazzo di neve* 116 forekomster, fordelt på 13 forskellige spatialpartikler, og for *Sneemanden* 154 forekomster fordelt på 24 forskellige spatialpartikler.

7. Konklusion

Spørgsmålet er om disse overvejelser kan bruges til andet end at konstruere en specifik søgestreng der kan bruges i en specifik tekst. Hvad er, med andre ord, fordelen fra et mere overordnet synspunkt? Jeg mener at der er flere argumenter der taler for det utvivlsomt meget tidskrævende arbejde:

1) I en mindre tekst eller tekstsamling, som *Sneemanden* eller det ovennævnte H.C. Andersen-korpus på syv eventyr, er en manuel søgning givetvis både hurtigere og mere præcis. Men hvis man vil have et mere komplet, detaljeret og statistisk signifikant billede af spatialpartiklernes

adfærd og frekvens, er det nødvendigt med større korpora (parallelle og sammenlignelige) samt elektroniske søgemuligheder. Hvis det lykkes at definere en søgestreng der i store mængder tekst (på tværs af sprog, genrer, medium etc.) kan finde tilnærmelsesvis alle forekomster af det fænomen man er interesseret i, så vil den tid man lægger i konstruktionen af søgestrengen, blive tjent ind igen. Næste fase i mit arbejde med søgestrengen for spatialpartikler vil være dels at undersøge dens effektivitet i andre tekster, dels at anvende den i udvidede søgestrenge der kan fange de forskellige konstruktioner som spatialpartiklerne indgår i (for både at præcisere og teste mine hypoteser om forskelle mellem dansk og italiensk[15]).

2) De forskellige elektroniske søgeresultater der er kommet frem undervejs i konstruktionen af den endelige søgestreng, har i sig selv peget på nogle fænomener og korrelationer (såvel monolingvalt som kontrastivt) der ved en manuel søgning muligvis ikke ville være kommet frem, eller ikke ville have stået så klart. Det drejer sig fx om de slående mange kombinationer af spatialpartikler der dukker op i sammenskreven form ved søgning i H.C. Andersens tekster – hvilket peger på kombinationen af spatialpartikler som et endnu mere markant træk på dansk end forventet, og et område der afgjort fortjener mere opmærksomhed. I et kontrastivt perspektiv kan nævnes forholdet mellem *af/fra* på dansk og *di/da* på italiensk, hvor overvejelserne omkring hvornår og i hvor høj grad det spatiale aspekt er bevaret, lægger op til en nærmere undersøgelse og sammenligning af disse leksemers semantiske og syntaktiske muligheder.

3) Endelig tvinger arbejdet med korpusværktøjet én til at definere det man leder efter på en både eksplicit og systematisk måde. Det medfører

15. Det er indlysende at et søgeresultat der kobler større søgninger på bestemte konstruktioner med spatialpartikel til de konstruktioner som er valgt i de respektive paralleltekster, vil være meget nyttigt. Dette kræver at teksterne er aligneret, dvs. at enheder (ord, sætninger, perioder) i kildeteksten er elektronisk forbundet med de enheder de er oversat til i måltekst, således at man ved søgning på et ord eller en konstruktion også automatisk får vist de sekvenser i den oversatte tekst, hvor gengivelsen af pågældende ord eller konstruktion forekommer. Manuel alignering er en meget tidskrævende opgave, mens automatisk alignering til gengæld ofte er ret fejlbehæftet, ikke mindst i forbindelse med litterære tekster, hvor oversætteren af stilistiske hensyn ofte foretager mange omrokeringer og opbrydninger på både sætnings- og periodeniveau.

at de kriterier der ligger til grund for ens dataselektion, er åbne for kritisk efterprøvning. Og det betyder endvidere at man i arbejdet med at eksplicitere søgekriterierne, er tvunget til selv minutiøst at overveje hvilke kriterier man rent faktisk definerer en given kategori eller konstruktion ud fra. Derved får man mulighed for at afdække eventuelle inkonsistenser i ens opstilling af kategorier og ens dataselektion, samtidig med at man bliver opmærksom på hvor svært det er at give en systematisk beskrivelse af ens ofte umiddelbare forståelse af sproget og dets nuancer. Mennesket kan i sin kategorisering arbejde analogt, dvs. finde lighedstræk i det der ikke er identisk, og dermed håndtere undtagelser, grænsetilfælde og glidende overgange. Computeren behandler derimod sproglige data ud fra kriterier der skal være så entydige at de kan omsættes til binære modsætninger, dvs. digitaliseres. Det har været interessant og lærerigt at prøve at forene den analoge og digitale tilgang i konstruktionen af en søgestreng.

Henvisninger

Andersen, H. C. (1861). Sneemanden, i E. Dal & E. Nielsen (red.). (1966) *Nye Eventyr og Historier 2. række 1861-66*, København: Danske Sprog- og Litteraturselskab. (Mulincokorpus)

Andersen, H. C. (2001). Il pupazzo di neve (oversættelse ved Bruno Berni), i B. Berni (red.) *Fiabe e storie*, Roma: Donzelli. (Mulincokorpus)

Chesterman, A. (2004). Beyond the particular, i A. Mauranen & P. Kujamäki (red.) *Translation Universals: Do they exist?*, Amsterdam: Benjamins, 33-49.

Farø, K. et al. (2005). *MULINCO – MUltiLINgual COrpus of the University of Copenhagen. Behovsanalyse. Rapport 1*, KUA : CST & Engerom (www.cst.dk/mulinco/index.html)

Henriksen, L., L. Offersgaard & C. Povlsen (under udgivelse). Ordklasse-tagging, i B. Maegaard & L. Schøsler (red.).

Herskovits, A. (1988). Spatial Expressions and the Plasticity of Meaning, i B. Rudzka-Ostyn (red.) *Topics in Cognitive Linguistics*, Amsterdam/Philadelphia: John Benjamins, 271-297.

Jansen, H. (under udgivelse a). Oversættelsesstudier, kontrastiv lingvistik og elektroniske tekstkorpora, i B. Maegaard & L. Schøsler (red.).

Jansen, H. (under udgivelse b). Spatialpartikler og søgestrenge. Hvordan

fanger man spatialpartikler i et elektronisk oversættelseskorpus?, i B. Maegaard & L. Schøsler (red.).

Jansen, H. (under udgivelse c). Construals in literary translation: spatial particles and spatial imagery, i Y. Gambier, M. Shlesinger & R. Stolze (red.) *Translation Studies: Doubts and Directions. Selected Contributions from the EST Congress, Lisbon 2004.* Amsterdam: John Benjamins.

Jansen, H. (2002). Spatialpartikler. Forstudier om brugen af præpositioner og lokative adverbier på hhv. italiensk og dansk, i H. Leth Andersen et al. (red.) *Ny forskning i grammatik. Fællespublikation 9. 2001.* Odense: Syddansk Universitetsforlag, 121- 140.

Jansen, H. (2004). Spatialpartiklen mellem periferi og centrum. Om spatialpartikler og spatial kodning på dansk og italiensk, i I. Korzen et al. (red.) *Ny forskning i grammatik. Fællespublikation 11. 2003.* Odense: Syddansk Universitetsforlag, 121-139.

Maegaard, B. & L.Schøsler (red.). (under udgivelse). *En Snemand på syv måder. En indføring i sprogteknologiske og kontrastive problemstillinger og metoder.* København: Museum Tusculanum.

Malmkjær, K. (1998). Love thy neighbour: will parallel corpora endear linguists to translators?, *Meta, XLIII, 4,* 534-541.

Munday, J. (1998). A computer-assisted approach to the analysis of translation shifts, *Meta, XLIII, 4,* 1-16.

Offergaard, L. & S. Olsen (under udgivelse). Brug af en korpus-platform, i B. Maegaard & L. Schøsler (red.).

Olohan, M. (2004). *Introducing Corpora in Translation Studies.* London: Routledge.

Talmy, L. (1985). Lexicalisation patterns: semantic structure in lexical forms, i T. Shopen (red.) *Language Typology and Syntactic Description. Vol III.* Cambridge University Press. 57-149.