Mot automatiska metoder för att upptäcka språkförändring

I takt med att vår värld och vår livsstil förändras, förändras även vårt språk. Vi lär oss nya ord, skaffar nya betydelser på existerande ord eller förändrar betydelser så att de passar in för att beskriva vår värld och vår tid. Vi glömmer fort och när vi tittar tillbaka, tex i gamla tidningsmaterial så är det inte alltid lätt att förstå vad som menats. Vem kommer tex ihåg *yuppienallen* eller vad ordet *guzz* betyder?

Generellt sett kan vi dela upp språkliga förändringar i två kategorier, den första rör ord vars betydelser ändras över tid medan den andra rör ord som ersätter varandra för samma betydelse. Ordet rock är ett exempel på ett ord som fått en tillagd betydelse, utöver att vara ett ytterplagg är det även en musikstil och faller därför i den första kategorin. I den andra kategorin faller en betydelse som "slug" där ordet *fin* en gång använts för att utrycka denna betydelse och nu används just *slug* eller *listig*. I denna senare kategori faller även namnförändringar, tex personer, städer och länder som byter namn.

När det gäller informationssökning i gammalt material, tex tidningar eller böcker, så orsakar ordbyten problem för att finna relevant material. Detta gäller oavsett om den som söker är en person, forskare eller annan, eller ett datorprogram. Anta att vi vill hitta material om första världskriget från perioden då kriget pågick, vid den tiden kallades kriget inte för första världskriget och en sökning med denna sträng skulle inte ge oss alla relevanta dokument.

När vi väl har hittat relevant material måste vi kunna tolka innehållet korrekt och där ställer betydelseändringar till det för oss. *Han var en grym person.* Hur detta skall tolkas beror naturligtvis på när meningen skrevs och att atomatiskt finna dessa ord vars betydelser har ändrats över tid, samt att veta hur förändringarna skett är av högsta vikt för att hjälpa människor och datorprogram som behöver tolka äldre (och inte alltid så gamla) texter.

Vi kommer att bygga verktyg att studera vårt språk och dess förändringar i större skala. När får ett ord en ny betydelse och hur länge är betydelsen aktiv? Vilka andra ord förändras när ett ord får en ny mening? Problemet är av högsta vikt då allt mer historiskt material blir öppet och enkelt tillgängligt men är även intressant i sociala medier där språket ändras fort. Det lockar forskare från alla domäner, framförallt digial humaniora, att forska i historiskt material och leta svar på ett automatiskt och storskaligt vis. Dessa forskare är inte, och bör inte vara, experter på historisk lingvistik för att kunna få tillgång till denna information. Allt ifrån attityden till retorik genom historien, till abstraktion av marknaden och olika politiska partiers användning av ord studeras och gynnas av att hantera språkliga förändringar automatiskt.

Problemen med att finna dessa förändringar är många och stora. Ordböcker och andra resurser kan användas till viss grad, men finns sällan i digitalt format, täcker inte alla epoker eller domäner och är tänkta som referenser. För att modellera den faktiska användningen av språket bör vi istället använda oss av automatiska metoder och börja med att finna betydelsen av ord ur en text. Detta kallas för *betydelse induktion* och är i sig ett mycket svårt problem. Vi kommer att studera ordens betydelser genom dess grannar enligt devisen " *You shall know a word by the company it keeps* (Firth, J. R. 1957:11)".

När vi väl har funnit vad orden betyder i varje tidsperiod så jämför vi betydelser över tid för att finna förändringar. Tidigare försök som gjorts har fokuserat på engelska och oftast delar av eller olika aspekter av problemet och ännu saknas t.ex. både automatiska utvärderingsmetoder och data att utvärdera på. Mycket fokus har legat på att hitta olika typer av förändringar utan att mäta eller filtrera brus. Nya, effektiva metoder använder distributionell semantik för att projicerar orden till vektorer och analyserar förändringar i dessa och kan då svara på *att* med inte *vad* som ändrats. Vi kommer att använda oss av en kombination av distributionell semantik och betydelseinduktion för kunna svara både på *vad* som ändrats och *när*. Vi kommer att använda de mycket stora samlingar av svensk text på Språkbanken, i ett världsunikt samarbete mellan semantiker och språkteknologer som med sina respektive expertiser har mycket goda förutsättningar att finna nya, automatiska metoder samt att i större skala svara på existerande hypoteser och deras generalisering till andra datamängder och tidsepoker.