Gustav Møller Grimberg
NEXT Vibenshus Gymnasium

Brug af komparative entropianalyser til datering og kvantificering af historiske divergenser mellem sprog

Matematikeren Claude E. Shannon formulerede i 1948 en teori om, hvordan informationen ved transmissioner af beskeder kunne kvantificeres. Helt centralt for denne teori er entropibegrebet. I det nærværende projekt vil jeg med afsæt i Shannons teori udarbejde en måde, hvorved entropien i sprog og den komparative lingvistik kan kombineres og derved skabe et redskab til kvantificering af forskelligheder mellem sprog. Dette gøres med et endeligt hovedsigte om at kunne formulere en metode, hvorigennem divergenser mellem sprog vil kunne dateres. Projektet afgrænses til at vedrøre dansk og svensk, så fokus kan bevares på det væsentligste. Dette leder til problemformuleringen:

- Er det muligt at kvantificere en historisk divergens mellem dansk og svensk gennem komparative entropianalyser, og såfremt dette er muligt, vil den kunne dateres?

Entropi i relation til sprog

Ved nærmere inspektion af et sprog tegner der sig unægteligt mønstre (Shannon, 1948). For et givent sprog vil en række bogstaver være tilbøjelige til at forekomme oftere end andre, og denne frekvensfordeling af bogstaver er ofte et særkende for et sprog. Man kan derfor ved sammenligning af frekvensfordelinger for forskellige sprog ofte identificere forskelle mellem forskellige sprogs skriftsystemer. Det er yderligere muligt at sammenligne frekvensfordelingen af bogstavsekvenser for et sprog, og herved opnås en dybere indsigt i sprogets struktur, da alle bogstaver har en ubestridelig relation til det foregående bogstav. Denne struktur lægger hjørnestenen i Shannons informationsteori, da entropien for en besked bestemmes på baggrund af frekvensfordelingen af bogstaver i beskeden.

En række begivenheder er givet: $x_1, \dots, x_n$ med de tilhørende sandsynligheder: $p_1, \dots, p_n$, således at p_i er sandsynligheden for at x_i optræder. Herfor er entropien H defineret ved:

$$H = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

hvor det gælder, at $p_i \in [0,1]\ \forall_i$ og $\sum p_i = 1$. Sandsynligheden for en begivenhed bestemmes ved at undersøge, hvor ofte den optræder ift. det samlede antal optrædener.[1] Det dog ikke tilstrækkeligt at bestemme entropien isoleret i de to skriftsprogsystemer, hvis man ønsker at detektere forskelligheder; det er nødvendigt at se på forholdet mellem frekvensfordelingen i de to systemer. Denne relative entropi, også omtalt som værende Kullback-Leibler entropien, er defineret således:

$$KL(p|q) = \sum_{i=1}^{n} p(i) \log_2 \left(\frac{p(i)}{q(i)} \right)$$

hvor $p(i)$ er sandsynligheden for, at den respektive begivenhed optræder i systemet, der tages udgangspunkt i, og $q(i)$ er sandsynligheden for, at den samme begivenhed optræder i systemet, der sammenlignes med. Anvendes dette mål for den relative entropi på danske og svenske tekster, vil det sandsynligvis være muligt at se en relativ entropi, hvilket skyldes, at frekvensfordelingen for ens

[1] Altså $p_i = \frac{n_i}{n_{total}}$. Optræder a eksempelvis 187 gange ud af 1000 bogstaver fås $p_a = \frac{187}{1000} = 0.187$

bogstaver (eller bogstavsekvenser) ikke er ens i de to sprog. Denne hypotese skal nu undersøges. Til dette benyttes et program programmeret til formålet, der bestemmer den relative entropi for to-bogstavssekvenser. I bestemmelserne tages der udgangspunkt i den danske og svenske bibel.

Den relative entropi i bibelen

Først bestemmes den relative entropi mellem Første Mosebog, Kapitel 1-10 på dansk og svensk. På en basis af 14637 optrådte to-bogstavsekvenser for dansk, n_{dansk} ,fordelt på 382 forskellige kombinationer, og 7981 optrådte to-bogstavssekvenser for svensk, n_{svensk}, fordelt på 350 forskellige kombinationer, fås en relativ entropi på $KL(p|q) = 0.263925$.[2] Dette understøtter fint hypotesen om at kunne identificere en relativ entropi mellem forskellige sprogs skriftsystemer.

Det er også væsentligt at undersøge den relative entropi mellem danske tekster. Bestemmes den relative entropi mellem Første Mosebog, Kapitel 1-10 og Første Mosebog, Kapitel 10-20 på dansk fås $KL(p|q) = 0.135597$.[3] Dette resultat kan bringe lidt tvivl om metodens gyldighed, da der ikke bør være nogen relativ entropi mellem tekster fra samme sprog. Sammenlignes denne entropi dog med $KL(p|q)$ mellem Første Mosebog, Kapitel 1-10 på dansk og svensk, ses en markant forskel, hvor den relative entropi mellem de ti kapitler skrevet på forskellige sprog er 94.7% højere end den relative entropi mellem ti forskellige kapitler af Første Mosebog skrevet på samme sprog. Bestemmes den relative entropi mellem Første Mosebog, Kapitel 1-10 på dansk og Første Mosebog, Kapitel 10-20 på svensk, ses det, at $KL(p|q) = 0.299576$.[4] Dette er 121% procent større end den relative entropi mellem de samme kapitler på samme sprog.

Konklusion

Det er blevet påvist, at det er muligt at identificere en relativ entropi mellem dansk og svensk, som sprogene er i dag, og eftersom det vides, at sprogene på et tidspunkt historisk var de samme (Braunmuller et al., 2002), vil det formodentlig være muligt at finde et tidspunkt, hvor den relative entropi mellem tekster skrevet i nuværende danske og svenske regioner var tæt på nul. De typologiske ændringer, som sprogene undergår med tiden, vil bevirke ændringer i den relative entropi, og en kronologisk gennemgang af den relative entropi sprogene imellem forventes derfor at afdække udviklingen fra en relativ entropi tæt på nul til det nuværende niveau. Det vil derfor med stor sandsynlighed være muligt at finde en periode, hvor den relative entropi begynder at ændre sig, og det vil i den forstand være muligt at lave en datering af sprogenes divergens.

Kilder

Braunmuller, K.B., Oscar Teleman, Ulf Naumann, Hans-Peter Karker, Allan Jahr, Ernst Hakon Widmark, Lennart Gun Elmevik, 2002. The Nordic Languages: An International Handbook of the History of the North Germanic Languages. Volume 1. De Gruyter.

Shannon, C.E., 1948. A Mathematical Theory of Communication. Reprinted with corrections from The Bell System Technical Journal.

[2] Baseret på $n_{dansk} = 14637$, fordelt på 382 forskellige to-bogstavs kombinationer, $n_{svensk} = 7981$, fordelt på 350 forskellige to-bogstavs kombinationer, og en n_{total} på 22618 (og 306 fælles to-bogstavs kombinationer for de to tekster).

[3] Baseret på $n_{dansk_1} = 14637$, fordelt på 382 forskellige to-bogstavs kombinationer, $n_{dansk_2} = 14100$, fordelt på 368 forskellige to-bogstavs kombinationer og en $n_{total} = 28737$ (og 334 fælles to-bogstavs kombinationer).

[4] Baseret på $n_{dansk} = 14637$, fordelt på 382 forskellige to-bogstavs kombinationer, $n_{svensk} = 16253$, fordelt på 317 forskellige to-bogstavs kombinationer og en $n_{total} = 30890$ (og 265 fælles to-bogstavs kombinationer)