

Grønne streker i Word bestemt av UiOs språkforskere

Mange har fått grønne streker når de skriver i Word, men få kjenner historien bak. – Arbeidet har vært både gøy og vanskelig, sier UiO-professor Janne Bondi Johannessen.



Janne Bondi Johannessen leder UiOs tekstlaboratorium.

Hver gang Microsoft Word gjør oss oppmerksomme på en sannsynlig grammatisk feil i det vi har skrevet, er det en frukt av mange års grunnforskning ved UiOs tekstlaboratorium, ledet av Janne Bondi Johannessen.

De grønne strekene i Word er historien om forskning de fleste har merket resultatene av, men nok ikke kjent historien bak.

Mellom oppdragsforskning og grunnforskning

Det var et finsk firma som var mellomledd mellom Microsoft og Tekstlaboratoriets professor Janne Bondi Johannessen og senioringeniør ved laboratoriet Kristin Hagen. Det var et oppdrag de fikk utenfra, men oppdragsforskning var det likevel slett ikke, understreker de to. Forskningen var nemlig allerede gjort.

– Det eneste vi trengte å gjøre, var å operasjonalisere den på en måte som kunne brukes i akkurat denne sammenhengen, sier de. Det tok noen måneder. Forskningen derimot har gått over mange år, sier Johannessen.

- ***Les også: Fanger måtte danse som Michael Jackson***

Historien om en grammatisk tagger

Det hele begynte med at Johannessen og andre for hele 20 år siden startet det såkalte taggerprosjektet - et prosjekt tilknyttet Norges forskningsråd. En tagger er et avansert digitalt analyseverktøy.

Målet med prosjektet var ubeskjeden: å utvikle et automatisk verktøy som både kunne identifisere ord, klassifisere dem og analysere dem i forhold til hverandre, i setninger. Resultatet er nå, etter flere oppdateringer, kjent som Oslo-Bergen-taggeren (OBT).

Det er ikke et lite stykke språkteknologisk innovasjon det er snakk om. Taggeren er den beste i sitt slag og har en treffsikkerhet på 96,5 %, helt uten menneskelig hjelp.

Norsk ordbank

Et biprodukt av taggerprosjektet er Norsk ordbank. Ordbanken er en såkalt fullformsordbok, et leksikon med alle tilgjengelige bøyingsformer av alle norske ord.

I dag er det en videreutviklet utgave av Ordbanken som ligger til grunn for blant annet bøyingsinformasjonen i Bokmåls- og Nynorskordboka på nett, og det elektroniske scrabblespillet Wordfeud. Det er samlingsenhetene på ILN som har arbeidet mest med Ordbanken.

- ***Les også: Masterstudent graver i Afrikas historie***

Den siste mila

Det er altså den grammatiske taggeren som utgjør grunnlaget for de grønne strekene i Word. Men et viktig stykke arbeid sto likevel igjen da de takket ja til bestillingen fra Microsoft.

Den grammatiske taggeren forutsetter i utgangspunktet korrekt språk, mens en grammatikkontroll må forstå feilaktig språkbruk, og så foreslå rettinger. Da måtte man utlede og konstruere tillegg til de reglene som lå til grunn for taggerens analyser, samt formulere de forklaringene den enkelte Word-bruker får når en feil er funnet.

- Selv om taggeren fra før gjorde det meste riktig, var dette et sannhetens øyeblikk, understreker Kristin Hagen, som sammen med Pia Lane (nå professor ved SFF-senteret MultiLing) jobbet mest intenst med akkurat denne fasen.
- Det er veldig vanskelig å lage en slik grammatikkontroll, men desto morsommere når man får det til. Når du får en regel til å virke er det det deiligste i verden, ler Hagen.

Og de fikk det til. På en brøkdels sekund sjekker det ferdige verktøyet både samsvarsbøyning, verbformer, ordstilling, plassering av adverb, rett kasus og enda mye mere til, selv om langt fra alle feil blir oppdaget, understreker hun.

Betydning for samfunnet

De grønne strekene i Microsoft Word er blant de tydeligste sporene som Tekstlaboratoriet har etterlatt seg utenfor academia. Samfunnsrelevans og praktisk betydning for folk utenfor akademiske kretser har hele tiden vært hovedmålet til språkforskerne. Det kommer tydelig til uttrykk gjennom et høyt antall vitenskapelige publikasjoner gjennom en årrekke og en tagger med åpen lisens.

Les mer om forskning som forandrer her!

Tekstlaboratoriets mange norske korpus er tagget med OBT-taggeren, og har i skrivende stund om lag 3000 registrerte forskere og studenter som brukere rundt om i verden. Et korpus er en søkbar digitalisert samling av skriftlig eller muntlig tekst med informasjon om ordklasser.

Og laboratoriet er kanskje bare så vidt i gang, tenker Johannessen og Hagen. De har videreutviklet taggeren for norsk talemål, og den er en nå en av bærebjelkene for de nyeste storsatsingene ved Tekstlaboratoriet, nemlig ulike talespråskorpus for oslodialekt, samt norske, nordiske og samiske dialekter før og nå.

Talespråskorpusene gjør det mulig for dem selv og andre forskere å bringe forskningen på norsk og andre språk et uant antall steg videre i jakten på målet om mer kunnskap. Det enorme materialet man nå har til disposisjon, har nemlig ingen hatt før.

Av Arve T. Thorsen & Fillip-André Baarøy

Publisert 13. juli 2017 14:01 - Sist endret 16. juni 2020 09:55

