DET DANSKE
SPROG- OG
LITTERATURSELSKAB

1000 danske talemåder og faste udtryk

Delaflevering 2

Det Danske Sprog- og Litteraturselskab har udviklet et datasæt som indeholder 1000 danske talemåder og faste udtryk med overførte betydninger (herefter omtalt som talemåder)[1]. Datasættet er baseret på Den Danske Ordbogs indhold af faste udtryk og angiver for hvert udtryk den korrekte definition fra ordbogen. Derudover er der udarbejdet tre falske definitioner til hver talemåde. Modsat den korrekte definition stemmer disse tre ikke overens med hvordan den givne talemåde forstås og bruges i det almene danske sprog.

De falske definitioner er udarbejdet efter et system bestående af (1) en konkret fejlfortolkning, (2) en abstrakt fejlfortolkning og (3) en tilfældigt udtrukket definition fra en anden talemåde i datasættet og derfor i denne sammenhæng med sikkerhed forkert. Formålet med tre typer falske definitioner der er forkerte på hver sin måde, er for det første at gøre testsættet mere udfordrende fordi nogle definitioner lyder plausible, men alligevel ikke svarer til talemådens brug, for det andet at facilitere mere dybdegående og nuancerede analyser af sprogmodellers svar på de opgaver de sættes til at løse baseret på indholdet i datasættet.

I datasættet er de 4 svartyper randomiseret således at den korrekte definition kan forekomme i forskellige positioner.

Mere om datasættet

Filen *talemaader_leverance_2_uden_labels* indeholder følgende seks tab-separeret kolonner:

udtryk_id	Id-nummer for talemåden. Vha. dette Id-nummer kan datasættet linkes til udtrykket i Den Danske Ordbog
talemåde_udtryk	Selve talemåden. Ved varierende former (fx *der er forår (øretæver, amoriner, ..) i luften*) har vi valgt én form
A - D	En af definitionerne (enten den korrekte definition, en konkret fejlfortolkning, en abstrakt fejlfortolkning eller en tilfældig definition

[1] Nogle udtryk er markeret 'talemåde' i Den Danske Ordbog, men ikke alle hvor det kunne være relevant. Datasættet indeholder derfor ikke kun disse, men inkluderer også faste udtryk med overførte (metaforiske) betydninger, særligt udtryk der indeholder et centralt ord i dansk. Fælles for alle udtrykkene er at man ikke kan udlede betydningen ud fra enkeltordenes betydning.

CHRISTIANS BRYGGE 1 ♥ 1219 KØBENHAVN K ♥ TLF 33 13 06 60 ♥ FAX 33 14 06 08 ♥ POSTGIRO 7 03 24 39
♥ E-MAIL SEKRETARIAT@DSL.DK ♥

Dertil hører også filen *talemaader_leverance_2_kun_labels* som indeholder labels. Filen indeholder følgende seks tab-separeret kolonner:

udtryk_id	Id-nummer for talemåden. Vha. dette Id-nummer kan datasættet linkes til udtrykket i Den Danske Ordbog
talemåde_udtryk	Selve talemåden. Ved varierende former (fx *der er forår (øretæver, amoriner, ..) i luften*) har vi valgt én form
korrekt_def	Kolonne indeks fra 0-3 (0 = A, 1= B, osv.) hvor den **korrekte definition** er placeret.
falsk1	Kolonne indeks fra 0-3 (0 = A, 1= B, osv.) hvor den **konkrete fejlfortolkning** er placeret.
falsk2	Kolonne indeks fra 0-3 (0 = A, 1= B, osv.) hvor den **abstrakte fejlfortolkning** er placeret.
falsk3	Kolonne indeks fra 0-3 (0 = A, 1= B, osv.) hvor den **tilfældigt udvalgte definition** er placeret.

Datasættet udgives med en CC-by licens, hvilket betyder at Det Danske Sprog- og Litteraturselskab skal krediteres ved brug.

Kontaktperson til datasættet

Nathalie Hau Sørensen	nats (snabel-a) dsl.dk

Delafleveringen er anden aflevering ud af to til Digitaliseringsstyrelsen. Delaflevering et indeholder talemåderne med deres definition fra Den Danske Ordbog.