

Evaluating Collocation Measures: Which Best Support Lexicographic Choice

Aleksandra Ignateva (Universitetet i Bergen)

This study investigates the effectiveness of statistical association measures in identifying strong Norwegian collocations and multiword expressions (MWEs), with a focus on their relevance to lexicographic practice. A set of 20 high-frequency Norwegian words -- covering adjectives, nouns, adverbs, and verbs -- was selected as collocation nodes. Collocations containing these target words were extracted from two annotated corpora: the Norwegian Newspaper Corpus (NNC) and the Lexicographic Corpus of Bokmål (LBK), both available through the Corpuscle interface.

After sorting the collocations and MWEs according to various association measures, they were compared to their occurrences in the online dictionary ordbokene.no for Bokmål. The dictionary collocations and MWEs were assigned scores based on how highly each association measure prioritized them. This ranking process aims to determine which association measures align the most closely with lexicographic judgment, thereby providing insights into the most suitable tools for dictionary work.

The following statistical measures were tested: collocation frequency, Log-likelihood, Mutual Information (MI), t-score, MI3 (cubic mutual information), $MI \cdot \log(\text{frequency})$, z-score, corrected z-score, logDice, and Pearson's chi-squared.

Among these, the collocation frequency consistently produced the highest scores across all experiments, suggesting it is the most reliable indicator of lexicographically relevant collocations. T-score, LogDice, MI3, and Pearson also showed strong alignment with dictionary choices. In contrast, the corrected z-score and Log-likelihood yielded the lowest scores. However, it is worth noting that Log-likelihood exhibited a distinct bias toward MWEs, more so than the other measures.

This research introduces a novel method that analyzes collocations and MWEs selected directly from dictionary entries and compares them with results obtained from two corpora — the Norwegian Newspaper Corpus and the Lexicographic Bokmål Corpus. The dictionary examples represent standardized usage in contemporary Norwegian. This study provides useful insights for lexicographers into which statistical association measures are most effective for identifying dictionary-relevant collocations and MWEs, while also revealing how well the measures align with their choices.

Evaluering av kollokasjonsmål: Hvilke støtter leksikografiske valg best

Denne studien undersøker hvor effektive statistiske assosiasjonsmål er for å identifisere sterke norske kollokasjoner og flerordsuttrykk (MWEs), med særlig fokus på deres relevans for leksikografisk arbeid. Et utvalg på 20 høyfrekvente norske ord – inkludert adjektiver, substantiver, adverb og verb – ble valgt ut som kollokasjonsnoder. Kollokasjoner som inneholder disse ordene ble hentet ut fra to annoterte korpus: Norsk Aviskorpus (NNC) og Leksikografisk Bokmålkorpus (LBK), begge tilgjengelige gjennom Corpuscle.

Etter å ha sortert kollokasjonene og uttrykkene etter ulike assosiasjonsmål, ble de sammenlignet med forekomster i nettordboken ordbokene.no for Bokmål. Kollokasjonene og uttrykkene fra ordboken ble gitt poeng ut fra hvor høyt hvert assosiasjonsmål rangerte dem. Denne rangeringsprosessen har som

mål å fastslå hvilke assosiasjonsmål som best samsvarer med leksikografisk vurdering, og dermed gi innsikt i hvilke verktøy som er mest hensiktsmessige for ordboksarbeid.

Følgende statistiske mål ble testet: kollokasjonsfrekvens, Log-likelihood, mutual information (MI), t-score, MI3 (kubisk mutual information), $MI \cdot \log(\text{frekvens})$, z-score, korrigert z-score, logDice og Pearsons kjikvadrattest.

Blant disse ble kollokasjonsfrekvens rangert jevnt over høyest i alle eksperimentene, noe som antyder at dette målet er den mest pålitelige indikatoren for kollokasjoner av leksikografisk relevans. T-score, LogDice, MI3 og Pearson viste også god overensstemmelse med leksikografiske valg. Derimot ga korrigert z-score og Log-likelihood de laveste poengene. Det er likevel verdt å merke seg at Log-likelihood hadde en tydelig tendens mot flerordsuttrykk, mer enn de andre målene.

Denne masteroppgaven introduserer en ny metode for å analysere kollokasjoner og flerordsuttrykk hentet direkte fra ordbokartikler og sammenligne dem med resultater fra to korpus – Norsk Aviskorpus og Leksikografisk Bokmålkorpus. Eksemplene fra ordboken representerer standardisert bruk av moderne norsk. Studien gir nyttige innsikter for leksikografer hvilke statistiske assosiasjonsmål som er mest effektive for å identifisere ordboksrelevante kollokasjoner og flerordsuttrykk, og viser samtidig hvor godt målene samsvarer med leksikografenes valg.