Dissertation Proposal: Diachronic NLP Analysis of "Therapy-Speak" (ADHD and Autism) in General Web Discourse

Jakob Lütkemeier

Deadline: 9 March 2026

Research questions. This dissertation asks: (1) How has the prevalence of ADHD- and autism-related terminology in general web discourse changed over approximately the last decade? (2) How has the *context* and framing around these terms shifted over time, consistent with the broader diffusion of "therapy-speak" into everyday language?

Significance. Public mental-health language has become more visible online, but it is unclear whether this reflects genuine changes in how people discuss neurodevelopmental conditions, shifting diagnostic awareness, or broader semantic broadening and "concept creep". A scalable, web-wide perspective complements platform-specific studies (e.g., Reddit or TikTok) and helps assess whether discourse changes are visible in the broader public web.

Proposed methodology. I will use Common Crawl as a large-scale web archive and implement a reproducible pipeline based on plaintext WET files. The dataset will be constructed diachronically by selecting one crawl per year (consistent temporal anchor) and sampling a fixed number of documents per year to estimate trends (Option A). Filtering will identify relevant pages containing "ADHD" and "autism" and key variants; ambiguous acronyms such as "ASD" will be counted only when disambiguated by nearby context (e.g., "autism" within ± 200 characters). To reduce dominance by large publishers, I will cap contributions per registered domain (e.g., max 50 documents per domain per year) and deduplicate content where feasible. In addition to trend estimation, I will build a smaller "deep corpus" of high-confidence hits for downstream NLP analysis (Option B), enabling analysis of context windows and framing changes over time. All figures and tables will be generated by code and included in LaTeX via file paths, avoiding manual copy–paste.