Winnie Collin, Inger Schoonderbeek Hansen, Tina Thode Hougaard og Karen Schriver (udg.):
20. Møde om Udforskningen af Dansk Sprog, Århus 2024

CorpusEye: Et brugervenligt korpusværktøj til forskning og undervisning

Eckhard Bick
Syddansk Universitet

Store tekstkorpora (fx Biber 1998, Asmussen 2007), dvs. samlinger af fx litterære værker, internethøstninger eller posts fra sociale medier, har vist sig at kunne levere værdifulde, empirisk funderede bidrag til både forskning (fx Scheuer 2017) og undervisning (fx Christiansen & Jacobsen 2019) inden for en lang række humanistiske fag. Datamængden gør det imidlertid umuligt at benytte sig af rent kvalitative metoder, og kvantitativt baseret inspektion og visualisering, samt statistisk evaluering, har vist sig at være effektive redskaber til at opnå objektive, data-drevne resultater og supplere den kvalitative tilgang i en frugtbar hermeneutisk cirkel (Kalwa 2019).

Dette bidrag til MUDS præsenterer *CorpusEye* (https://corp.visl.dk/), et komplekst grafisk brugergrænsesnit til grammatisk-lingvistisk opmærkede korpora, og viser, hvordan korpuslingvistiske metoder (Bick, Gorbahn & Kalwa 2023) kan bruges til konkrete problemstillinger med et fokus på dels litterær, sproglig og kompa rativ tekstanalyse, dels på sociolingvistiske emner. Systemet understøtter bl.a. grammatisk-semantisk-baserede konkordanser og statistikker, relative og standardiserede frekvenstal, metadata-basErede søjlediagrammer til sammenligning af forfattere, værker eller perioder, samt fleksible todimensionelle scatter plots baseret på word embedding, for at sammenligne og "opmåle" beslægtede begreber, som fx ordene *flygtning, indvandrer, udlænding*. Sidstnævnte kan

også bruges til at kvantificere kønsstereotyper i forhold til fx. brands eller madvarer.

Opmærkede korpora

Opmærkede korpora på *CorpusEye* dækker i alt 10 germanske og romanske sprog med ca. 9 milliarder ord, men hovedfokusset vil her være på danske litterære tekster og danske sociale medier. Herudover bruges enkelte eksempler fra tysk og portugisisk for at understrege det tværsproglige, komparative aspekt. Det anvendte danske litteraturkorpus består af frit tilgængelige klassiske værker fra kilder som Det Kongelige Bibliotek, Gutenberg-projektet o.l., og indeholder i sagens natur (pga. copyright-reglerne) hovedsagelig klassiske værker fra første halvdel af 1900-tallet, 1800-tallet og tidligere, i alt 15,4 millioner ord. Det sociolingvistiske korpus (XPEROHS, Bick & Geyer 2023) dækker både tysk og dansk og består af internet-høstede data fra Twitter (2017-2023, i alt 514 millioner ord for dansk, 3 milliarder for tysk) og Facebook (2017-2018, i alt 34 mio. ord for dansk, 148 mio. for tysk). Samtlige tekster er morfologisk og syntaktisk analyseret vha. Constraint Grammar-parsere (Bick 2023c) mht. både lemma, ordklasse, bøjning, komposita, syntaktisk funktion og dependensstrukturer. Desuden findes der en gennemgående opmærkning på det semantiske niveau – noget der normalt ikke eksisterer i danske korpora. For substantiver og adjektiver bruges en hierarkisk ontologi af såkaldte semantiske prototyper[1], fx <Hprof> (= +human, erhverv), <tool> (værktøj), <food> (mad), <jsize> (størrelsesadjektiv). Verbernes semantiske klasser samt semantiske roller for afhængige sætningsled er baseret på det danske framenet[2] (Bick 2011). Fig. 1 viser et opmærket sætningseksempel, med lemma i []-parenteser, ordklasse og bøjning med store bogstaver, syntaktisk funktion (@), semantisk rolle (§) og dependensstruktur (#dependent->hoved). <...>-parenteser bruges til sekundære kategorier (fx <poss> = possessiv, <mv> = hovedverbum), kompositumsanalyse (fx

[1] Der bruges ca. 200 semantiske prototyper for substantiver og ca. 120 for adjektiver, jf.: https://edu.visl.dk/semantic_prototypes_overview.pdf og https://edu.visl.dk/semantic_prototypes_adj.pdf

[2] Kategoriinventaret rummer ca. 50 semantiske roller og ca. 500 frame-klasser, implementeret for alle danske verber og en del valens-bærende substantiver.

<N:lort~e+racist>) samt sentiment-markering (fx <Q->), semantiske prototyper (fx <Lh> = "funktionelt" sted) og framenetklasser (fx <fn:teach>).

I	[I] PERS 2P NOM @SUBJ> §COG #1->2
ved	[vide] <fn:know> <mv> V PR AKT @FS-STA #2->0
intet	[intet] <quant> INDP NEU S @<ACC §SOA #3->2
om	[om] PRP @<PIV #4->2
,	[,] PU @PU #5->0
hvordan	[hvordan] <interr> <amod> ADV @ADVL> #6->10
kvinder	[kvinde] <fem> <H> N UTR P IDF NOM @SUBJ> §AG #7->10
i	[i] PRP @N< #8->7
Mjølnerparken	[Mjølnerparken] <top> <Lh> PROP NOM @P< §LOC #9->8
opdrager	[opdrage] <fn:teach> <mv> V PR AKT @FS-P< §TP #10->4
deres	[de] <poss> PERS 3P GEN @>N #11->12
børn	[barn] <Hbio> N NEU P IDF NOM @<ACC §PAT #12->10
,	[,] PU @PU #13->0
lorteracister	[lorteracist] <N:lort~e+racist> <Q-> <Hideo> N UTR P IDF NOM @VOK #14->10

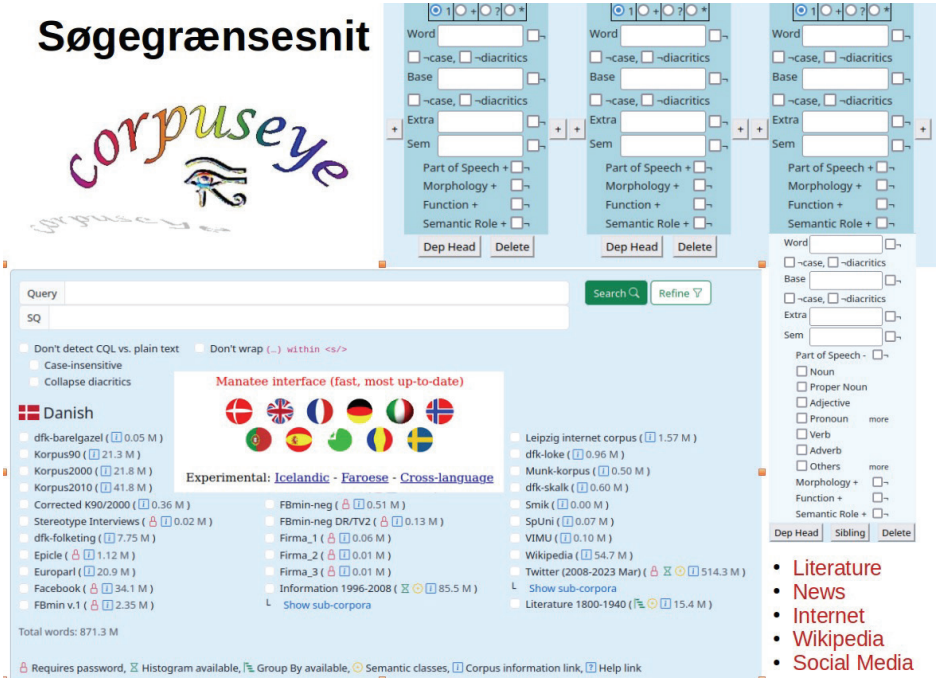
Figur 1: Opmærket eksempelsætning fra XPEROHS-korpusset

I det følgende vises ved hjælp af konkrete eksempler og forskningsresultater, hvordan disse korpora kan benyttes i *CorpusEye* til fx identifikation af kulturelle, etniske, religiøse og kønsbaserede stereotyper (Bick 2023b), brand-profilering og hate speech-forskning (Bick 2023a), i sociolingvistikken, eller til distant reading (Underwood 2017) i litterære studier. De ikke-danske eksempler bygger på den tyske del af XPEROHS samt et stort portugisisk blog-korpus (dos Santos et al. 2018) med 2 milliarder ord indsamlet mellem 2013 og 2018.

CorpusEye

CorpusEye (https://corp.visl.dk) er et grafisk søge-grænsesnit og en samling af grammatisk-semantisk opmærkede korpora, der dækker

de fleste germanske og romanske sprog og en lang række genrer. Fig. 2 viser, hvordan en søgning, efter valg af et eller flere korpora, opbygges af "ordmasker" med tekstfelter eller menuer for de enkelte kategorifelter i opmærkningen, fx lemma, syntaktisk funktion eller semantisk kategori. Felterne tillader regular expressions[3] og negering, og ordmaskerne selv kan markeres som optionelle eller som en-eller-flere.



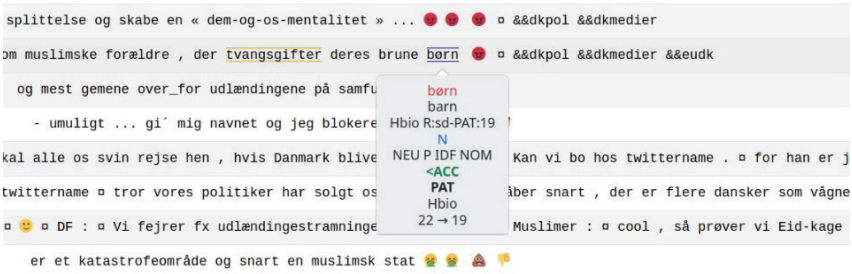
Figur 2: *CorpusEye*, søgegrænsesnit

Kongstanken med redskabets ergonomiske aspekter er dels en grafisk tilgang til både søgninger og resultater, dels en iterativ arbejdsproces, hvor kvantitativ-statistisk analyse og kvalitativ inspektion under-

[3] Dvs. udtryk som 'øko.+' eller 's[a-z]+(e[nt]|er?ne)', hvor førstnævnte matcher sammensætninger med 'øko-' og sidstnævnte matcher substantiver, der begynder med 's' og ender med en bøjningsendelse for bestemthed.

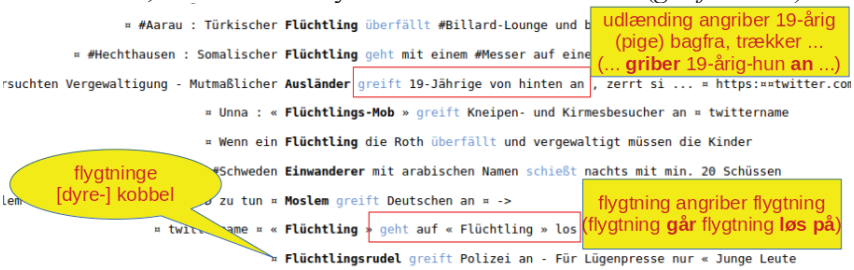
støtter hinanden iterativt. Hvis opgaven fx lyder på at kortlægge det danske leksikon (herunder at finde nye ord!) omkring begrebet *økologi,* kan man starte med to ordmasker, (1) lemma="økologisk" + (2) N (substantiv), hvad der vil levere en såkaldt *konkordans* af sætninger med *'økologisk landbrug', 'økologisk vare'* osv. Man kan herefter danne sig et overblik ved at frekvenssortere bigrammerne (ordparrene), finde mere sjældne, men interessante eksempler på listen og klikke sig tilbage til en – nu selektiv – konkordans. Man kan også bede *CorpusEye* om at vise relativ i stedet for absolut frekvens, for at identificere de mest typiske korrelater af *'økologisk'*. Her normaliseres frekvensen af korrelaterne ved at dividere med deres generelle frekvens i korpusset. Både *'landbrug'* og *'vare'* ses således ofte sammen med *'økologisk',* men i sproget som sådan er *'vare'* meget mere frekvent end *'landbrug',* så i en relativ frekvenssortering vil *'vare'* ryge langt ned i listen og gøre plads til sjældne, men karakteristiske kollokater som *'spisemærke', 'mejeriproducent'* og *'gårdkylling'*. For at finde nye ord kan man bruge kompositumsanalyse og søge på ord, hvor opmærkningen har markeret *'øko-'* som forstavelse ("F:øko\+.*" i ordmaskens ekstra-felt). Dette leverer, som konkordans eller frekvensliste, ord som *økoflipper, økomærke, økohøne, økokød, økotosse, økofascisme, økobonde, økokollaps, økosamfund.*

En særlig feature ved *CorpusEye* er, at der kan generes subkorpora *on the fly* (SQ=*subquery*). Her filtreres hovedkorpusset i en sekventiel søgning forud for hovedsøgningen – en metode, der er velegnet til at genere tættere, og dermed mere læselige, konkordanser, eller til at kombinere søgekriterier uanset hvor i sætningen de forekommer. Fx kan man skabe et immigrationssubkorpus ved at bruge *"(indva ndr|flygtning|udlænding|muslim|islam|arab).*"* i lemma-feltet i for-søgningen, og så *"emo-angry"* (alle former for vrede emojis) i hovedsøgningen. Fig. 3 viser en del af den resulterende konkordans, med en *mouse over*-rude, der viser opmærkningen for et af ordene.



Figur 3: Konkordans *Invandrer* + *vred emoji*, mouse over-kategorirude

Hver ordmaske kan udvides med en tilknyttet kategorimaske for dependensrelationer, *head* (hoved) og *sibling* (søskende). Sidstnævnte gør det muligt at koble subjekt og objekt sammen, uanset hvor i sætningen de forekommer, og førstnævnte muliggør fx framesøgninger. Metoden kan generere meget "tætte" excerpter fra meget store korpora, som det ses i det tyske eksempel i fig. 4, hvor der er søgt på et subjekt med et "integrations"-lemma i "attack"-frames <fn:attack>. Bemærk, hvordan *CorpusEye* markerer både konkordansens kerneord (subjektet, sort) og frame-kernen (verbet, blå), og at framen identificeres, selv hvor det tyske verbum er brudt i to (*greift .... an*).



Figur 4: indvandrer + "attack"-frame, tysk Twitter

Word embeddings

Word embeddings er en af *CorpusEyes* mest effektive metoder til undersøgelse af stereotyper. Word embeddings anvendes i store statistiske sprogmodeller (LLMs) til at approksimere et ords betydning, distribution og slægtskab med andre ord ud fra ordkonteksten i sætningerne i store korpora. Konkret er der tale om maskinlærte ord-vektorer i et højdimensionelt koordinatsystem,

bygget ud fra samforekomster af ord og repræsenteret i neuronale netværkslag med flere hundrede knuder. I *CorpusEye* bruges Word2vec-metoden (Mikolov et al. 2013) med 300 koordinater og den rækkefølge-følsomme Skipgram-metode. Med udgangspunkt i den lingvistiske opmærkning anvendes der i *CorpusEye* i stedet for ordformer disambiguerede lemma-ordklassepar[4]. For hvert korpus fører *CorpusEye* en vektordatabase over alle ord, der gør det muligt at beregne den semantiske lighed mellem to ord som længden af differencevektoren i det 300-dimensionelle *embedding*-rum.

Til brug ved kvantitativ stereotyp-forskning kan *CorpusEye* generere såkaldte vectorplots (eller *scatterplots*) for fundne konkordans-ord, hvor ordene indsættes i et bruger-defineret 2-dimensionelt koordinatsystem, hvor både x- og y-aksen defineres igennem polare (normalt antonymiske) stereotyp-vektorer. Her kan både enkeltord og ordgrupper (med semikolon) benyttes som + og – poler for x- eller y-aksen, fx (a) *mand – kvinde,* (b) *ung – gammel,* (c) *forurening; fossil – vedvarende; grøn; miljøvenlig.*

Fordi den grammatiske opmærkning også gælder emoticons og emojis, der "lemmatiseres" som adverbier i ni semantiske klasser (fx happy, laughing, angry), er det muligt at bruge en af akserne til sentiment-analyse, fx:

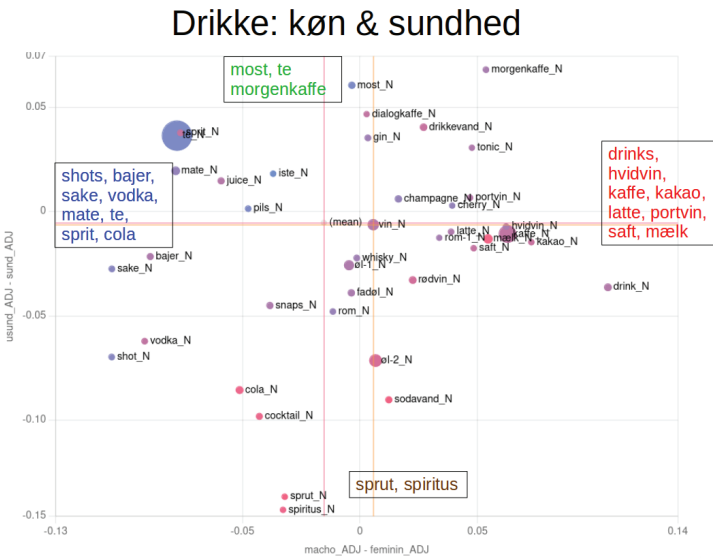
• *emo-sad; emo-angry; emo_skeptical (neg.) – emo-happy; emo-love (pos.)*

Et givent ords x- og y-koordinater beregnes herefter som differencen af vektorligheden med de respektive poler på den pågældende akse. Med x-aksen som *mand-kvinde* og y-aksen som *ung-gammel,* og SIM='similarity' (differencevektor) bliver det for et givent ord xxx:

x = SIM(mand,xxx) – SIM(kvinde,xxx), **y** = SIM(ung,xxx) – SIM(gammel,xxx)

Fordi associeringsstyrkerne (målt som vektor-lighed) mellem mål-ordene og de (polare) ord, der vælges til at definere de to koordinater, er et nedkog over hele korpussets sætningskontekster, kan man argumentere, at de af bilder stereotypiske associeringer i en gennemsnitlig bidragyders (tekst- eller post-forfatters) mentale sproglige landkort.

[4] Fx for ordformen *køber* skelnes der således mellem *køber_N* (substantiv) og *købe_V* (verbum).

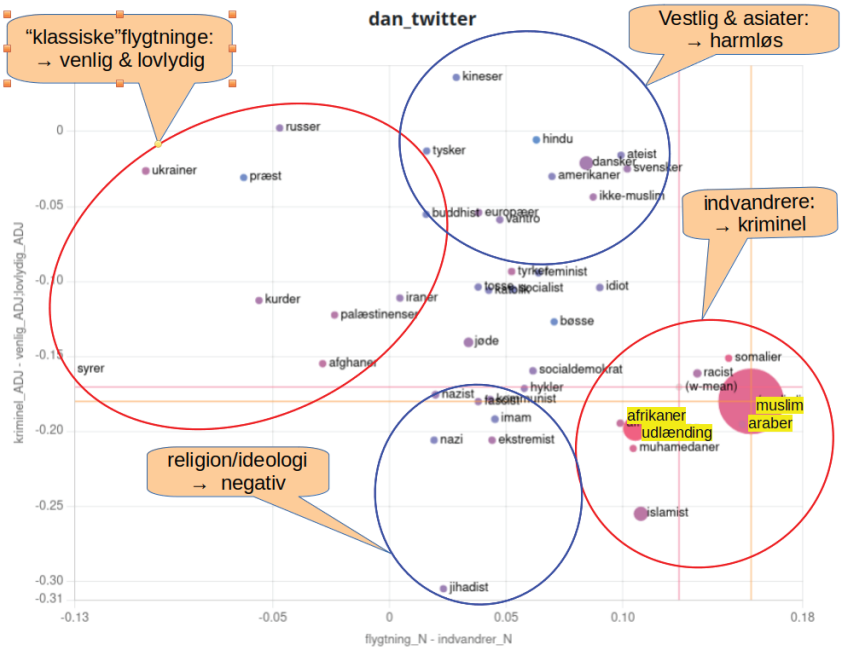


Figur 5: vektorplot ”drikke”, stereotyper: køn og sundhed

Eksemplet i fig. 5 viser et vektorplot[5] for en søgning på den semantiske klasse <drink> med stereotyp-akserne ’køn’ (her: *macho – feminin)* og ’sundhed’ *(usund – sund)*. Det kan umiddelbart aflæses, at spiritus betragtes som meget usund, og most og te som meget sund, og at manddom forbindes med shots og bajere, og femininitet med latte og portvin. Selvfølgelig kan en sådan kvantitativ-vektorbaseret stereotypanalyse ikke stå alene, men den kan hjælpe brugeren med at fokusere den kvalitative inspektion[6] af et stort korpus.

[5] Ordpunkternes størrelse viser ordets relative hyppighed i søgningsresultatet, og rød- og blå-komponenterne i prikkernes farve viser den **absolutte** korrelationsstyrke med hhv. x- og y-aksens pluspoler. Denne sidste information er nemlig ikke umiddelbart synlig i x|y-koordinaterne selv, idet der er tale om **differencen** mellem korrelationerne med hhv. plus- og minuspolerne.

[6] Næste skridt i den kvantitativ-kvalitative metode kunne således være et klik på et interessant ord i den frekvensliste over <drink>-ord, der blev brugt som udgangspunkt for scatterplottet. Klikket får CorpusEye til at generere en – noget kortere og derfor mere inspicerbar – subkonkordans.



Figur 6: vektorplot "udlænding", stereotyper: flygtning vs. indvandrer, kriminalitet

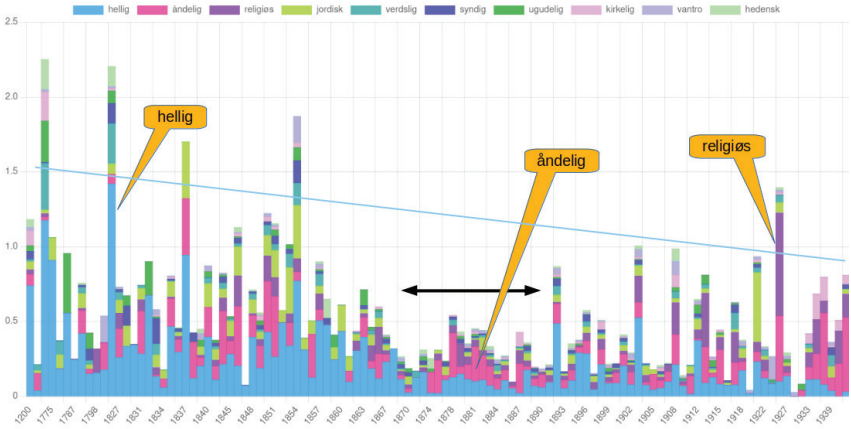
Et andet eksempel på stereotyp-visualisering er immigrations-vektorplottet i fig. 6. Her er den bagvedliggende søgning, foretaget i et *on-the-fly*-genereret immigrations-subkorpus en semantisk søgning på *(Hattr|Hideo|Hnat),* dvs. substantiver, der betegner personer mht. nationalitet, ideologi eller adjektiviske attributter. I plottet, med stereotyp-akserne *flygtning-indvandrer* (x-akse) og *kriminel – venlig;lovlydig* (y-akse) fremstår ukrainere og syrere som de mest "flygtningeagtige", mens somaliere, muslimer og arabere er de mest "indvandreragtige", hvor sidstnævnte betragtes som mere kriminelle end førstnævnte. Vesterlændinge og asiater anses for at være de mest venlige og lovlydige, med jihadister og islamister i den anden ende.

Distant reading

Distant reading (fx Moretti 2000), der måske kunne oversættes med "afstandslæsning" (modsat "nærlæsning") er en digital

”overbliksmetodologi” for især litterære tekster. Metoden er velegnet til at forberede en mere målrettet og kvalitativ inspektion og kan bruges til en statistisk sammenligning af forfattere, værker eller genrer mht. lingvistiske mønstre, ordforråd eller leksikalsk spredning. Den bruges desuden til automatisk ekstraktion af nøglebegreber (topic modelling) eller til at generere lister og relationsnetværk over fx figurerne i en roman.

Et eksempel kunne være et forskningsprojekt ang. rollen af dyr og planter i dansk litteratur (spoiler alert: flere blomster i romantikken, færre under det moderne gennembrud, Oehlenschläger, Claussen og Winther går op i roser [> 50%], der i øvrigt ved kvalitativt eftersyn viser sig ofte at blive brugt til erotisk metaforik), eller spørgsmålet om hvor fremtrædende religion har været i dansk litteratur, på tværs af forfattere, værker eller tid. CorpusEye understøtter her en 3-trins analyse. Først gennemføres en konkordanssøgning på relevante ord, evt. med indgrænsende kontekst, fx en søgning på adjektiver markeret som religiøs (<jrel> og domænemarkøren <D:rel>). Som trin 2 frekvenssorteres de fundne ord *(hellig, åndelig, religiøs, jordisk, verdslig, syndig, ugudelig, kirkelig, vantro, hedensk ...)*, hvor man så kan klikke videre til delkonkordanser for de enkelte ord eller sortere på korre-lerende substantiver *(hellig skrift, fader, ånd, kirke, jomfru eller hedensk tid, kvinde, gud)*

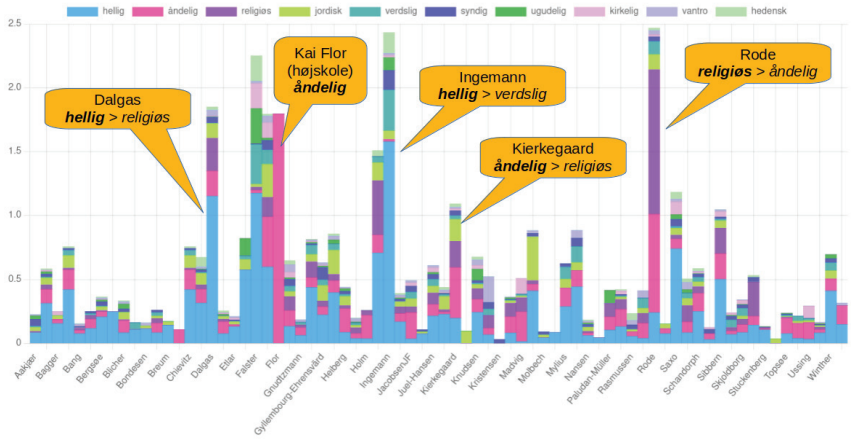


Figur 7: søjlediagram ”religiøse adjektiver”, med tidsakse

Som trin 3 benytter *CorpusEye* teksternes meta-information til at gruppere de fundne ord i søjlediagrammer, hvor hver søjle står for fx en forfatter, forfatterens køn, et værk eller et år (eller en kombination af disse), og hvor den enkelte søjle enten angiver den samlede forekomst af alle de fundne ord (fx som % hits per sætning) for den pågældende forfatter osv. (simpelt søjlediagram), eller den relative andel af de mest frekvente (eller manuelt valgte) ord ved at "summere" ordene som farvede afsnit på søjlen ("stacked" søjlediagram).

Et umiddelbart resultat for religionsadjektiverne på en tidsakse (fig. 7) er, at forekomsten samlet set er faldende over perioden, med en særlig flad kurve under det moderne gennembrud (1870-1890), og at *hellig* udfases til fordel for *åndelig* og *religiøs*.

Et tilsvarende søjlediagram for forfattere (fig. 8) lægger op til en typologi ud fra den relative forekomst af de dominerende "religiøse" adjektiver, med *hellig* som det mest "bogstavelig"[7] religiøse adjektiv, *åndelig* som en mere oplyst og mere filosofisk term, og *religiøs* som et neutralt-beskrivende meta-ord:



Figur 8: søjlediagram "religiøse adjektiver", forfattersammenligning

[7] Sammen med *hedensk, ugudelig, vantro* og *syndig* er ordet overrepræsenteret i den tidlige del af tidaksen.

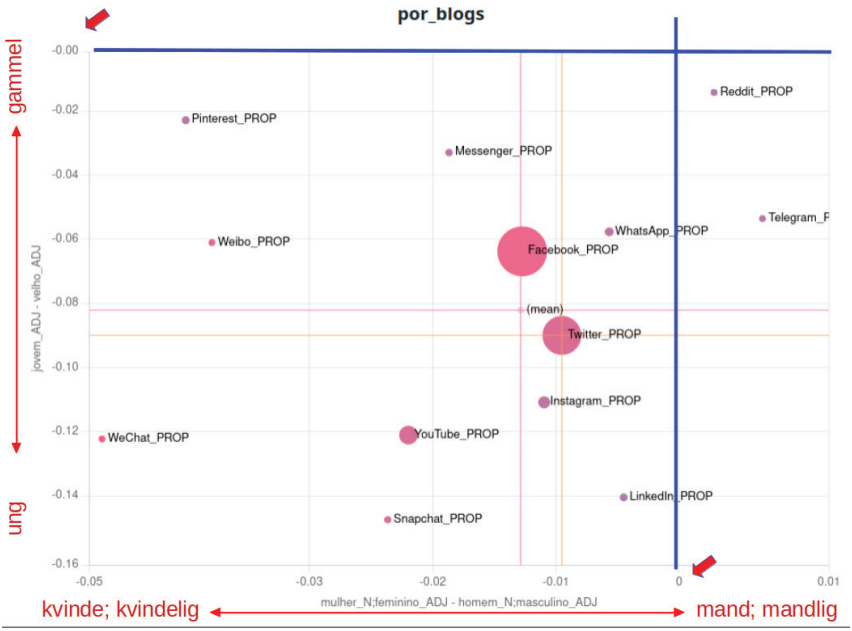
- Latin & salmer: Christian Falster 1690-1752 & B.S. Ingemann 1789-1862*hellig > verdslig > ugudelig (Ingemann: syndig)*
- *Filosoffen: Søren Kierkegaard 1813-1855: åndelig > religiøs > jordisk*
- *Højskolen: Kaj Flor 1886-1965: 100% åndelig, 0% hellig*
- *Lyrisk refleksion: Helge Rode 1870-1937: religiøs > åndelig (> hellig)*

Brand-profilering

Brand-profilering, afslutningsvis, er en korpus-applikation fra hinsides humanioras hellige haller. De fleste firmaer er interesserede i image-pleje og brand-profilering, og (aktuelle) korpora kan i hvert fald hjælpe med det andet. *CorpusEye* kan gøre status over, hvordan omverdenen opfatter et givent brand – har det en grøn eller "fossil" profil, moderne eller gammeldags[8?] Her kan der bruges metoder, der minder om stereotypafdækning, ikke mindst vektoranalyse. Fig. 9 retter fokus på sociale mediers køns- og aldersprofil i det portugisiske blog-korpus, med vektorpolerne *mulher;feminino (kvinde;kvindelig)* og *homem;masculino (mand;mandlig)* for x-aksen og tilsvarende *jovem (ung)* vs. *velho (gammel)* for y-aksen.

Plottet viser, at de billed- og video-priorite$rende medier Snapchat, YouTube og Instagram har den yngste profil, med mere tekst-orienterede medier som Messenger, WhatsApp og Reddit i den anden ende af y-aksen. På x-aksen har WeChat og Weibo tydelig mere kvinde-tække end Telegram, Reddit og LinkedIn. Både køns- og aldersprofileringen skal dog ses i det lys, at nul-linjerne ligger asymmetrisk, således at sociale medier som sådan opfattes som mere "kvindelige" og især mere "unge". Hvad angår sværvægterne, indtager Twitter rollen som medianpunkt, mens Facebook som brand placerer sig lidt til den ældre side, og en anelse til den kvindelige.

[8] For flere eksempler, se https://corp.visl.dk/corpuseye_manual.pdf og https://corp.visl.dk/Corpuseye_IKS.pdf



Figur 9: vektorplot "sociale medier", brand-profil køn & alder

Litteratur

Asmussen, Jørg (2007). *Introduktion til korpuslingvistik*. https://ja.dsl.dk/papers/cbs_2007/terminlogi-og-korpuslingvistik-outline.pdf [tilgået 23.8.2024].

Biber, Douglas et al. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: CUP.

Bick, Eckhard (2011). A FrameNet for Danish. I *Proceedings of NODALIDA 2011*. NEALT Proceedings Series, Vol 11, 34-41. Tartu: Tartu University Library. (ISSN 1736-6305).

Bick, Eckhard (2023a). Derogatory Linguistic Mechanisms in Danish Online Hate Speech. I Isabel Ermida (red.), *Hate Speech in Social Media: A Linguistic Approach* (s. 165-202). London: Palgrave Macmillan.

Bick, Eckhard (2023b). Concepts of Europe in Danish and German Social Media: A Corpus-Linguistic Study. I Katja Gorbahn, Erla Hallsteinsdóttir & Jan Engberg (red.), *Exploring Interconnectedness - Constructions of European and National Identities in Educational Media (*s. 187-212). Palgrave Macmillan & Springer.

Bick, Eckhard (2023c). VISL & CG-3: Constraint Grammar on the Move: An application-driven paradigm. I Arvi Hurskainen, Kimmo Koskenniemi & Tommi Pirinen (red.), *Rule-Based Language Technology*. NEALT Monograph Series vol. 2, 112-140. University of Tartu.

Bick, Eckhard & Geyer, Klaus (2023). Das deutsch-dänische XPEROHS-Korpus: Hassrede in sozialen Medien. I Marc Kupietz og Thomas Schmidt (red.), *Neuere Entwicklungen in der Korpuslandschaft der Germanistik: Beiträge zur IDS-Methodenmesse 2022*, 115-115. Narr Francke Attempto Verlag (CLIP - Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache).

Bick, Eckhard; Gorbahn, Katja; Kalwa, Nina (2023). Methodological Approaches to the Digital Analysis of Educational Media: Exploring Concepts of Europe and the Nation. I *Katja Gorbahn, Erla Hallsteinsdóttir & Jan Engberg (red.), Exploring Interconnectedness - Constructions of European and National Identities in Educational Media* (s. 143-186). Palgrave Macmillan & Springer.

Christiansen, Mads & Jacobsen, Malou (2019). Korpuslingvistik i fremmedsprogsundervisningen. Med tysk i gymnasiet som eksempel. *Globe: A Journal of Language, Culture and Communication*, 8, 57-67.

Kalwa, Nina (2019). Die Konstitution von Konzepten in Diskursen: Zoom als Methode der diskurslinguistischen Bedeutungsanalyse. I Sandro Moraldo et al. (red.), Sprach(kritik)-kompetenz als Mittel demokratischer Willensbildung. *Greifswalder Beiträge zur Linguistik*, bd. 12, 11-26.

Li, Yang & Yang, Tao (2018). Word Embedding for Understanding Natural Language: A Survey. I Siddarth Srinivasan (red.), *Guide to Big Data Applications. Studies in Big Data*, bd. 26 (s. 83-104). Springer, Cham.

Mikolov, Tomas; Ilya Sutskever; Kai Chen; G. S. Corrado; Jeffrey Dean (2013). Distributed representations of words and phrases and their compositionality. I Christopher J.C. Burges et al. (red.), *Advances in neural information processing systems - Proceedings of NIPS 2013*. 3111–3119.

Moretti, Franco (2000). *Conjectures on World Literature*. New Left Review 1 jan/feb 2000, 54-68. NLR Ltd.

dos Santos, Henrique D.P.; Woloszyn, Vinicius; Vieira, Renata (2018). BlogSet-BR: A Brazilian Portuguese Blog Corpus. I *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 661-664. ACL.

Scheuer, Jann (2017). Hans mening om hendes krop: Portrætter af to køn i danske tekstkorpusser. I Inger Schoonderbeek Hansen, Tina Thode Hougaard og Kathrine Thisted Petersen (red.), *16. møde om Udforskningen af Dansk Sprog* (s. 343-366). Aarhus: Aarhus Universitetsforlag.

Underwood, Ted (2017). A Genealogy of Distant Reading. *Digital Humanities Quarterly* 2017, bd. 11, nr. 2