

# Unsupervised POS-Tagging Employing Efficient Graph Clustering

Chris Biemann, University of Leipzig, Coling/ACL SRW 2006

biem@informatik.uni-leipzig.de

## Tagset for NORWEGIAN

This tagset was automatically induced by merely presenting large amounts of raw text to the algorithm. Numbers following hashmarks # correspond to tag labels, numbers in brackets give the number of words per tag in the lexicon. Maximally, four randomly chosen entries per tag are given. Special characters are transformed for technical reasons, e.g. `_ESENTS_` is the full stop, `_CLOSEBRAC_` is a closing bracket etc.

#1(63401):herligheter, treningsopplegget, polarlitteratur, nærtrafikken #2(24094):latinen, europeer, sædbank, prat #3(4185):sjikanerer, florent, avklarte, kopierte #4(7282):køfritt, ovnsbakt, vettug, trinnvis #5(3926):intensivere, avhøres, tilkalle, flørte #6(5477):vågale, skrekkslagent, urnorske, støyutsatte #7(12296):Brundalen, Dell, Birdstep, Hochfilzen #8(2212):utfordret, strammet, kanongode, anvendt #9(3191):Samleie, Frykten, Fuglene, Juridisk #10(16303):Sirseth, Miccoli, Inkster, Kristoffersen #11(2042):Yousef, Amund, Helene, Oddbjørn #12(2524):språkkonsulent, overdommer, borgermester, kommunestyrerepresentant #13(3079):energi-indeksen, inntektssvikten, forventningspresset, singelsalget #14(1467):Village, Intelligent, Loop, Golde #15(464):Tråkket, Forstår, Skjuler, Dropp #16(1029):øyrikets, Fosens, motebransjens, diktaturets #17(565):finanspolitisk, verbal, vitenskapsteoretisk, nettverksbasert #18(90):femti, seks-syv, fire, fjorten #19(98):GRIMSTAD, JERUSALEM, VENNESLA, HORTEN #20(809):tennisproffen, filmkunstneren, tøffingen, seriedrapsmannen #21(53):torsdagens, halvkjørt, lørdagskveldens, stopperens #22(1420):reporterens, lånets, kandidatenes, datamaskinenes #23(854):can, hear, immediately, teach #24(1846):varehusets, Svens, Esats, skogeiernes #25(300):kvassere, jevnere, løsere, finere #26(25):Nordre, Sierra, Indre, Galleri #27(60):skyttelettrafikk, avsted, skrånende, østover #28(236):FrP-politiker, Redaktør, Styrrer, Lensmannsførstebetjent #29(313):cubansk, belgisk, spansk, singaporsk #30(185):RAV4, Xantia, firesylindret, Fiat #31(25):derved, heretter, trolig, utvilsomt #32(3):fremfor, glem, framfor #33(28):tyvende, nyaste, femtende, syttende #34(11):REUTER, AVSTEMNING, Scanpix, REUTERS #35(83):svindeltaltale, straffedømt, utuktssiktede, sårede #36(15):ubeseiret, storspillt, sittet, vørt #37(89):Wireds, magasinet, Telemarksavisas, Varietys #38(15):adm, osv, hmm, hmmm #39(5):hotel, Hotell, Hotel, Panorama #40(358):oslolagets, Lyngdals, Färjestads, Eiks #41(436):Kostbar, Vellykket, Nyttig, Tynn #42(5):ca, Kl, nr, kl #43(78):Snille, Fargerikt, Tolv, Halvannet #44(794):91-åringen, småbarnsfaren, Nokas-ranerne, herredsretten #45(5):nordøst, sørvest, sørøst, nord-vest, #46(62):hjelpeorganisasjonen, kvinnebladet, musikkavisen, bransjebladet #47(18):herredsrett, tingsrett, Skifterett, Tingrett #48(5):trådte, trer, trådde, trådt #49(16):akselererende, uant, oppadgående, stigande #50(41):tilsammen, anslagsvis, norskeste, cirka #51(69):Frankrike-VM, ABC-teatret, 1970-tallet, 1997-sesongen #52(695):Hærlands, lærlingenes, trondheimsmannens, reineierens #53(4):kunnet, villt, skullet, måttet #54(10):forberedelsene, adgang, Fram, Frem #55(713):finsk-svenske, pro-vestlige, indianske, uruguyanske #56(33):urorammede, sørvestlige, sydøstlige, nordlige #57(14):to-timers, times, kvarters, måneds #58(781):LO-sjefens, Djupedals, Nixons, Håkons #59(3):tilhørte, tilkommer, tilhører #60(192):Jordanske, Interne, Finnen, Nakne #61(2):hvis, dersom #62(4):reiser, dro, drar, reiste #63(13):president, statsminister, utenriksminister, forsvarsminister #64(2):leilighet, turnering #65(2):anklaget, beskyldt #66(2):gjelder, gjaldt #67(2):ordre, pålegg #68(3):artIngress>, artBrodtekst>, bildetekst> #69(2):selv, sjølv #70(2):forklart, uttalt #71(2):mener, meiner #72(2):Prix, Center #73(3):vegne, tvers, bekostning #74(2):prosjektet, arrangementet #75(2):planer, ambisjoner #76(2):Men, Og #77(3):Landsforening, Hovedorganisasjon, Sentralforbund #78(2):synes, syns #79(3):endret, forandret, bedret #80(36):kommunen, Høyre, Statoil, Israel #81(2):video, sykkel #82(2):person, politimann #83(2):jobber, arbeider #84(4):men, mens, ettersom, hvorav #85(7):vis, Lanka, Trafford, Brygge #86(2):avisen, avisa #87(2):ham, henne #88(2):AS, A\_S #89(2):Børs, Lufthavn #90(2):talsmann, talskvinne #91(5):NTB-AFP, NTB-Reuters, NTB-DPA, NTB-Reuters-AFP #92(2):fra, frå #93(2):drept, knivstukket #94(2):hva, hvem #95(11):Høyres, Arbeiderpartiets, SVs, Aps #96(2):Han, Hun #97(2):Institutt, Senter #98(2):inneholder, inneholdt #99(3):Hvilke, Hvilken, Hvilket #100(3):pressemelding, årrekke, børsmelding #101(2):stede, salgs #102(2):fram, frem #103(3):høsten, sommeren, våren #104(2):sammenlignet, sammenliknet #105(2):musikken, maten #106(2):problemer, vanskeligheter #107(2):arbeidskraft, matvarer #108(2):selskapet, partiet #109(2):mann, kvinne #110(2):million, milliard #111(2):rundt, omkring #112(3):NTB, Dagbladet, Dagbladet\_no #113(2):kan, vil #114(2):samme, same #115(2):stoppet, stanset #116(2):trekker, trakk #117(2):bidrar, bidro #118(2):kamerat, venninne #119(3):i, på, mot #120(2):høy, lav #121(3):lenger, engang, nødvendigvis #122(2):Du, Man #123(2):representerer, representerte #124(2):Gode, Dårlig #125(2):advarer, advarte #126(2):pågrepet, arrestert #127(239):få, bli, ta, gå #128(13):selskaper, grupper, verdier, oppgaver #129(2):loven, grunnloven #130(15):selskap, parti, prosjekt, program #131(2):se, høre #132(2):peker, pekte #133(2):minste, lengste #134(2):flere, fleire #135(4):filmen, boken, boka, romanen #136(2):fortsetter, fortsatte #137(3):stammer, Bortsett, Tall #138(2):tiden, tida #139(2):Målet, Planen #140(2):heter, het #141(2):taket, nesen #142(2):være, vera #143(2):spesielt, særlig #144(3):vår, vårt, mitt #145(2):områdene, miljøene #146(2):Deres, Selve #147(2):dagene, ukene #148(2):Fjord, Color #149(18):spilleme, jentene, kvinnene, guttene #150(2):lar, lot #151(2):kontakt, samtaler #152(5):Antall, Stort, Relaterte, Gult, Rødt #153(3):stått, ligget, opptråd #154(2):sønn, datter #155(2):gripe, stramme #156(2):dommen, kjennelsen

see next page

#157(2):You, Love #158(2):Ja, Nei #159(2):kommer, kjem #160(2):mer, meir #161(2):bok, roman #162(2):landslaget, kontinentet #163(3):Hva, Hvordan, Hvorfor #164(2):minner, minnet #165(4):Manchester, Real, Aston, FC #166(2):gjør, gjorde #167(2):Kristelig, Sosialistisk #168(3):står, sto, stod #169(3):hovedstad, Lake, di #170(2):sønneren, datteren #171(2):slappe, skryte #172(2):grensen, grensa #173(2):stadion, Stadion #174(2):Administrerende, Daglig #175(2):t\_v, t\_h #176(2):ulike, forskjellige #177(28):prosent, minutter, meter, poeng #178(2):verdt, verd #179(5):fått, hatt, gitt, opplevd, måttet #180(4):Over, Rundt, Minst, Nærmere #181(2):dreier, dreide #182(3):spillet, systemet, nettverket #183(3):sam, inkludert, deriblant #184(2):ligger, lå #185(2):fokus, jakten #186(2):stund, halvtid #187(2):sørget, sørger #188(2):han, hun #189(2):mye, mykje #190(2):fungerer, fungerte #191(3):årene, månedene, åra #192(2):May, Jun #193(4):skjedde, skjer, foregår, foregikk #194(3):Jeg, Vi, Dere #195(2):gamle, gammel #196(2):økning, nedgang #197(2):fortsatt, fremdeles #198(2):ødelagt, knust #199(24):Bush, Røkke, Clinton, Blair #200(2):Hvis, Dersom #201(2):understreker, presiserer #202(2):millioner, milliarder #203(2):ble, blei #204(2):prinsesse, prins #205(2):ber, ba #206(2):lignende, liknende #207(2):jente, gutt #208(2):båten, maskinen #209(12):mai, januar, september, mars #210(2):van, von #211(2):heller, overhodet #212(2):sitte, bo #213(2):ulykken, uhellet #214(2):hjelper, hjalp #215(2):positiv, negativ #216(12):Selskapet, Kommunen, Laget, Klubben #217(3):eksempel, øvrig, lengst #218(2):plassert, lagret #219(3):kvinnen, jenta, gutten #220(2):brukes, benyttes #221(2):hele, heile #222(3):statsministeren, forsvarsministeren, midtbanespilleren #223(2):består, besto #224(2):Quart, MTV #225(2):driver, drev #226(2):ikke, ikkje #227(2):overkant, underkant #228(2):snakket, pratet #229(3):minutt, divisjon, juledag #230(2):gjennomført, iverksatt #231(2):vilje, myndighet #232(4):Universitetet, Høgskolen, Sentralsjukehuset, Regionsykehuset #233(2):ga, gav #234(2):Verdens, Kommunenes #235(3):vanskelig, umulig, fristende #236(2):makt, muskler #237(2):konkluderer, konkluderte #238(2):Klokken, Klokka #239(2):skadet, skadd #240(2):regel, oftest #241(2):senere, seinere #242(4):sammen, Sammen, parallelt, befatning #243(2):bedre, dårligere #244(2):natten, natta #245(2):kreve, foreslå #246(3):politikkerne, myndighetene, bøndene #247(2):tilknyttet, underlagt #248(2):sin, sitt #249(8):lørdag, fredag, mandag, søndag, onsdag, torsdag, tirsdag, igår #250(3):øvrig, ansattes, berørte #251(2):Grenland, Ham #252(3):sikret, skaffet, pådratt #253(2):involvert, innblandet #254(3):uken, uka, måneden #255(2):utviklet, spredd #256(2):Årsaken, Grunnen #257(3):hans, deres, hennes #258(2):de, dei #259(4):siktet, tiltalt, ansvarlig, mistenkt #260(2):føler, følte #261(2):fyller, fylte #262(20):Nå, Her, Derfor, Dermed #263(11):USAs, Israels, Sveriges, Russlands #264(2):finnes, fins #265(2):populære, profilerte #266(2):sent, seint #267(2):jordskjelv, ektepar #268(2):Kverneland, Ringnes #269(2):dermed, følgelig #270(5):bor, deltok, deltar, bodde, inngår #271(2):førte, fører #272(2):likevel, allikevel #273(2):bøker, romaner #274(3):meste, motsatte, dobbelte #275(2):ferdig, ferdige #276(3):er, var, blir #277(2):foreløpig, dessverre #278(4):interesse, respekt, risiko, sans #279(2):flyet, toget #280(16):operasjonsleder, informasjonssjef, rektor, forsker #281(2):70-tallet, 60-tallet #282(12):produksjon, oppfølging, forvaltning, innkjøp #283(2):listen, lista #284(2):fremover, framover #285(7):Fredag, Mandag, Lørdag, Onsdag #286(2):Sri, Soria #287(3):Mål, Pris, Publisert #288(19):sakene, bilene, spørsmålene, filmene #289(4):jo, selvsagt, selvfølgelig, naturligvis #290(3):dager, måneder, timer #291(2):stasjon, ungdomsskole #292(2):sykdommen, streiken #293(4):the, and, with, my #294(2):danser, maler #295(2):x, | #296(2):skjedd, oppstått #297(2):altså, iallfall #298(2):utstyr, materiell #299(2):El, Le #300(2):havnet, havner #301(3):glad, glade, takknemlig #302(2):-, — #303(2):opp, ned #304(3):kveld, ettermiddag, formiddag #305(20):professor, ordfører, konsernsjef, generalsekretær #306(4):forhold, natt, motsetning, henhold #307(2):fleste, færreste #308(3):høyere, lavere, svakere #309(2):hindre, forhindre #310(5):forsøkt, begynt, prøvd, greid, rukket #311(2):land, EU-land #312(2):interessant, realistisk #313(2):red, dvs #314(2):vi, jeg #315(2):til, for #316(2):venstre, høyre #317(3):Ingen, Mange, Noen #318(2):både, Både #319(5):en, et, ei, ein, eit #320(2):trener, manager #321(2):tyder, tydet #322(2):verken, hverken #323(1):som #324(1):da #325(1):De #326(1):\_OPENBRAC\_ #327(1):ingen #328(1):bare #329(1):ville #330(1):det #331(1):alle #332(1):når #333(1):måtte #334(1):over #335(1):av #336(1):\_ESENTE\_ #337(1):Den #338(1):nok #339(1):\_NUMBER\_ #340(1):ut #341(1):- #342(1):fordi #343(1):andre #344(1):gjennom #345(1):viser #346(1):man #347(1):har #348(1):fikk #349(1):For #350(1):Av #351(1):og #352(1):like #353(1):saken #354(1):helt #355(1):seg #356(1):si #357(1):om #358(1):tilbake #359(1):meg #360(1):dem #361(1):litt #362(1):VG #363(1):tar #364(1):tror #365(1):ha #366(1):å #367(1):skal #368(1):grunn #369(1):nå #370(1):alt #371(1):under #372(1):tatt #373(1):oss #374(1):dette #375(1):mest #376(1):du #377(1):mellom #378(1):ved #379(1):må #380(1):\_KOM\_ #381(1):\_ESENTS\_ #382(1):folk #383(1):kunne #384(1):med #385(1):\_CLOSEBRAC\_ #386(1):gang #387(1):politiet #388(1):gikk #389(1):før #390(1):Etter #391(1):siden #392(1):annet #393(1):tid #394(1):noe #395(1):her #396(1):ser #397(1):godt #398(1):kom #399(1):også #400(1):\_ #401(1):En #402(1):etter #403(1):den #404(1):slik #405(1):der #406(1):\_ESENTQ\_ #407(1):enn #408(1):får #409(1):hadde #410(1):første #411(1):så #412(1):hvor #413(1):tidligere #414(1):blitt #415(1):disse #416(1):mange #417(1):I #418(1):år #419(1):uten #420(1):noen #421(1):sine #422(1):\_QUOT\_ #423(1):skulle #424(1):Da #425(1):at #426(1):\_COL\_ #427(1):blant #428(1):eller #429(1):inn #430(1):På #431(1):langt