Wikipedia som indikator og motor for språkleg revitalisering

Trond Trosterud

UiT

Wikipedia finst for 287 språk, og på same måte som språka sjølv, varierer storleiken på dei ulike Wikipediaversjonane. Storleiken på språksamfunnet er likevel berre ein av fleire faktorar som har innverknad på kor vellukka Wikipedia kan bli.

Tema her er 6 mellomstore Wikipedia-versjonar (70000-400000 artiklar), alle for uralske språk i Russland, dei for moksja, erzja, vestmarisk, austmarisk, udmurtisk og komi. Vi vil sjå på i kor stor grad dei ulike wp-versjonane kan fungere som indikatorar for språksituasjonen i dei ulike språksamfunna.

Vi vil også sjå på i kor stor grad wp-versjonane er brukbare som språkteknologisk ressurs.

Presentasjon av dei 6 wikipedia-versjonane

Alle språka blir snakka i den austlege delen av europeisk Russland, der dei har status som offisielle språk i sine respektive republikkar. Rekna opp frå sør mot nord blir moksja og erzja snakka i Mordovia, dei mariske språka i Mari El, udmurtisk i Udmurtia og komi i Komi. Tabell 1 viser kor mange talarar språka har, og kor mange artiklar og ord det er i dei respektive Wikipedia-variantane (færøysk er med som samanlikningsgrunnlag).

Språk	WP	Talarar	WP-artiklar	Ord i WP	ord/talar	ord/artikkel	Aktive skrib
Moksja	mdf	200000	1221	69888	0,35	57,2	3
Erzja	myv	400000	1863	70693	0,18	37,9	7
Vestmarisk	mrj	36822	7190	258188	7,01	35,9	9
Austmarisk	mhr	414000	6661	378925	0,92	56,9	13
Udmurtisk	udm	324000	3582	144685	0,45	40,4	15
Komi	kv	156000	4373	202852	1,30	46,4	16
Færøysk	fo	66000	10929	1093696	16,57	100,1	29

Tabell 1: Talarar og Wikipediaversjonar for 6 uralske språk og færøysk, jf. Moosley (red.) 2010, https://meta.wikimedia.org/wiki/List_of_Wikipedias og https://en.wikipedia.org/wiki/Wikipedia:Database_download (31.1.2015).

For samtlege av dei uralske wp-versjonane er det ein handfull skribentar som står for brorparten av redigeringane, jf. tabell 2. Omlag 10 skribentar har stått for mellom 80 og 90% av redigeringane.

Sterkast er tendensen for vestmarisk, det minste språksamfunnet. Her står 7 skribentar for 92% av redigeringane. Den største av wp-versjonane, den for austmarisk, er også den best balanserte. Også for denne står likevel 9 skribentar for 82% av innhaldet.

mrj		myv		mdf		udm		kv		mhr	
skrb	wp	skrb	wp	skrb	wp	skrb	wp	skrb	wp	skrb	wp
7	0,92	14	0,95	19	0,92	13	0,86	14	0,89	18	0,91
6	0,91	9	0,91	10	0,83	5	0,77	5	0,80	11	0,87
2	0,79	8	0,89	3	0,63	3	0,67	4	0,77	9	0,82
1	0,66	4	0,64	3	0,63	1	0,45	2	0,50	1	0,22

Tabell 2: Skribentar, og deira kumulative del av dei respektive språkversjonane

Eit titals skribentar står for 70% av innhaldet på færøysk Wikipedia, for nynorsk er talet noko høgare. Derfrå er det svært grovt sett ei tidobling til bokmål, tysk og engelsk, jf. tabell 3.

fo		nn		nb		de		en	
skrb	wp	skrb	wp	skrb	wp	skrb	wp	skrb	wp
20	0,80	31	0,78	224	0,75	14845	0,87	149620	0,87
11	0,70	13	0,60	98	0,61	1161	0,51	3948	0,46
2	0,42	3	0,33	38	0,41	266	0,26	924	0,26
1	0,35	1	0,18	2	0,07	23	0,06	13	0,03

Tabell 3: Skribentar, og deira kumulative del av språkversjonane for eit par germanske språkversjonar

Skribentane

Fleire små wikipediaversjonar er dominert av det som vi kan kalle Wikipedia-samlarar, dvs. skribentar som finn seg små Wikipediaversjonar på språk dei ikkje kan, og gjer dei til sine. Det er større sjanse for å finne personar med kunnskap og interesse for å bygge Wikipedia i store enn i små språksamfunn, for minoritetsspråksamfunn kan det også vere vanskeleg å finne skribentar som kjenner leksikonsjangeren, eller skrivekyndige friviljuge i det heile. I beste fall kan slike utanforståande gjere ein nyttig jobb (hjelpe lokale skribentar, løyse tekniske problem, lage malar, til og med lage artiklar basert på sjablongar), men i verste fall kan dei fylle Wikipediaen med artiklar med større eller mindre språklege feil. Blir dei dominerande nok kan dei gjere det umogleg å bruke nettsøk som referanse til kva som er autentisk språkbruk, dersom ein synleg del av denne språkbruken er Wikipediaartiklar skrive av folk som ikkje kan språket dei skriv på.

Eit godt døme på dette er brukaren Kmoksy. Kmoksy har i følgje brukarsida si (User:Kmolsy) tyrkisk som morsmål, og er aktiv på tyrkiske og uralske wp-versjonar. Han er den mest aktive

brukaren på mrj, udm og den komipermjakiske koi (som ikkje er med her), han er den nest mest aktive på mdf, og er blant dei 10 mest aktive brukarane på alle dei uralske wp-versjonane. Dette har t.d. ført til at mrj har ei svært god dekning på tyrkiske kommunar. Det går fram av brukarsida hans at den drivande motivasjonen er omtanke for minoritetsspråk og dei språklege rettane til språkbrukarane deira.

Eit døme på Wikipedia-samlarar med litt mindre omtanke er iñupiaq (ip). Her har skribentane tatt eit polysyntetisk språk, brukt ei ordbok, og rett og slett sett saman setningar etter engelsk syntaks. Fleire av bokstavane i iñupiaq finst ikkje i det engelske alfabetet. På 1980-talet vart det utvikla eit typesnitt der fleire engelske bokstavar (q, w, ..) rett og slett vart teikna om til ḅ, ḍ, osb. I og med at skribentane ikkje hadde kunnskap om iñupiaq i det heile publiserte dei rett og slett tekst med dei arbitrære engelske bokstavane i staden. Resultatet vart uforståeleg kaudervelsk.

Typologisk sett er dei søraustlege uralske språka svært lik tyrkisk, og det er ingen tvil om at Kmoksy har forstått syntaksen i dei setningane han har skrive. Artiklane han skriv er skjematiske, og inneheld enkle setningar. Resultatet ar eit språk som er ganske nært målspråket. Metodologisk er det likevel ikkje det same som ein tekst skrive av morsmålstalarar eller dyktige framandspråkstalarar, og dei er ikkje feilfrie.

Wikipedia som indikator på språksamfunnet

Wikipedia finst på 287 språk, ein liten del av dei 7000 språka som finst i verda. Det er likevel ikkje slik at det er dei 287 største språksamfunna som har wp-versjonar. Grovt sett er det mogleg å klassifisere fungerande wp-versjonar inn under ein av desse kategoriane: Dei er offisielle språk i sjølvstendige statar (estisk, fransk), dei er minoritetsspråk i vestlege land (nordsamisk, baskisk), eller dei er Ausbau-språk, dvs. nærståande dialektar av offisielle språk i vestlege land (lombardisk, plattysk, scots). Underrepresentert er særleg språka i tidlegare koloniar. Minoritetsspråka i Russland kjem i ei særstilling. Til skilnad frå minoritetar i andre delar av verda har dei lært å lese og skrive på skolen gjennom heile 1900-talet, og dei er dermed i stand til å lese, skrive og korrigere tekst på ein måte medlemmar av minoritetspråksamfunn i andre land ikkje er.

Der wp-versjonar som dei for iñupiaq, nordsamisk og til og med for grønlandsk er dominert av skribentar som ikkje har målspråket som morsmål, har omtrent halvparten av dei 10 mest aktive skribentane på dei uralske wp-versjonane i Russland målspråket som morsmål. Med eit aukande medvit om at morsmåla deira er i ein utsett posisjon, ei aukande vilje til å gjere noko med det, og betre tilgang til internett, fører dette til aukande interesse for desse wp-versjonane.

Wikipedia som ressurs for språksamfunnet

Dei uralske wikipediaversjonane i Russland inneheld mellom 70000 og 400000 ord, og dei utgjer frie og lett tilgjengelege tekstressursar. For små wp-versjonar er det å bruke Wikipedia som korpus ikkje uproblematisk. Viss store delar av teksten er skrive av skribentar som ikkje har språket som morsmål, og som i verste fall ikkje kan det i det heile, er det store sjansar for at språket i Wikipedia

rett og slett ikkje er representativt for språket slik det blir brukt av morsmålsskribentar.

Dei uralske wp-versjonane som er skildra her har vorte lasta ned og blir no brukt som korpus (samling av eksempelsetningar) for morfologisk intelligente elektroniske ordbøker for dei respektive språka. Bruken av ordbøkene blir logga, og resultata vil bli presentert på Wikipedia-akademiet.

Konklusjon

Viss vi ser bort i frå Ausbau-språk i Vest-Europa, er dei 5 wp-versjonane som blir handsama her større enn dei aller fleste wp-versjonane for språksamfunn av tilsvarande storleik. Ingen av dei har gått fri for Wikipedia-samlarar, men til skilnad frå wp-versjonar for minoritetsspråk i Vesten, har desse skribentane meir greie på målspråket sin grammatikk, og dei møter også meir skrivekyndige kolleger blant wikipedianarane med målspråket som morsmål. Om resultatet er tekst med kvalitet god nok til å bli brukt som ressurs for språksamfunnet, det vere seg som leksikon eller som språkteknologisk ressurs, står att å sjå.