

Murre24: Dialect Identification of Finnish Internet Forum Messages

Contributions

- Manually annotated dataset of around 4000 Finnish internet forum messages,
- Automatic annotation of all 94M messages in the forum corpus,
- Evaluation of five different language identification tools on the task, and
- Discussion on the variation and language change in the dataset

Data

- Suom24 (Finland24) is the largest internet forum in Finland
- Messages posted on the forum from 2001 to 2020 have been published for academic use in the Language Bank of Finland
- The objective of the current work is to
 - identify the languages and Finnish dialects used on the forum and
 - follow the change of their distributions over time

Manual annotation

- Around 4000 messages annotated manually by variety:
 - Standard Finnish, seven traditional dialects (see Map), Helsinki slang and a colloquial style
- Based on phonology and morphology
- Split to three folds of train and test splits (90-10)
- Labeled as S24 to not be confused with the forum name
- An additional random test set taken to see how the methods perform with the actual distribution of varieties in the corpus

Methodology

- Train five different dialect identification tools with the manually annotated data and external datasets from the Web and dialect interviews
- Plus an ensemble model
- Two traditional models (SVM and Naive Bayes), two off-the-shelf models re-trained (fastText and HeLI) and one neural model (FinBERT)
- Annotation in three stages: first languages, then standard vs. non-standard Finnish and finally the nine non-standard varieties
- The end goal is to collect a dataset of dialectal Finnish internet forum messages, entitled Murre24 ('Dialect24')



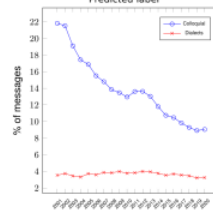
Model	S24 + Web + SKN		
	Balanced		
NB	0.45 \pm 0.01	0.25 \pm 0.01	0.59 \pm 0.01
SVM	0.89 \pm 0.02	0.90 \pm 0.01	0.91 \pm 0.02
fastText	0.87 \pm 0.01	0.86 \pm 0.02	0.87 \pm 0.01
HeLI	0.86 \pm 0.02	0.59 \pm 0.01	0.76 \pm 0.03
FinBERT	0.76 \pm 0.23	0.21 \pm 0.00	0.21 \pm 0.00
Ensemble	0.89 \pm 0.00	0.89 \pm 0.02	0.89 \pm 0.02
Random sample			
NB	0.06 \pm 0.00	0.73 \pm 0.00	0.75 \pm 0.01
SVM	0.72 \pm 0.02	0.86 \pm 0.01	0.86 \pm 0.00
fastText	0.77 \pm 0.02	0.85 \pm 0.00	0.85 \pm 0.01
HeLI	0.80 \pm 0.02	0.37 \pm 0.02	0.62 \pm 0.01
FinBERT	0.58 \pm 0.37	0.73 \pm 0.00	0.73 \pm 0.00
Ensemble	0.73 \pm 0.02	0.85 \pm 0.00	0.86 \pm 0.01

Model	S24 + Web + SKN		
	Balanced		
NB	0.56 \pm 0.01	0.55 \pm 0.01	0.18 \pm 0.00
SVM	0.80 \pm 0.01	0.81 \pm 0.01	0.78 \pm 0.01
FastText	0.74 \pm 0.02	0.75 \pm 0.02	0.70 \pm 0.00
HeLI	0.68 \pm 0.02	0.68 \pm 0.02	0.63 \pm 0.01
FinBERT	0.81 \pm 0.00	0.81 \pm 0.00	0.78 \pm 0.00
Ensemble	0.82 \pm 0.01	0.82 \pm 0.01	0.78 \pm 0.00
Random sample			
NB	0.86 \pm 0.01	0.79 \pm 0.02	0.09 \pm 0.00
SVM	0.84 \pm 0.01	0.83 \pm 0.01	0.74 \pm 0.00
FastText	0.76 \pm 0.01	0.76 \pm 0.01	0.71 \pm 0.01
HeLI	0.72 \pm 0.01	0.73 \pm 0.01	0.80 \pm 0.01
FinBERT	0.85 \pm 0.02	0.87 \pm 0.01	0.82 \pm 0.02
Ensemble	0.85 \pm 0.01	0.85 \pm 0.01	0.81 \pm 0.00

Results

- Language identification with HeLI-OTS
- Most messages (97% in Finnish)
- For standard vs. non-standard Finnish
- SVM offers best performance followed by the ensemble model (weighted F1)
- FinBERT and NB are unstable
- For final dialect identification
- FinBERT is stable and performs best followed by the ensemble model
- The FinBERT model often mislabels Häme dialect to the colloquial style and Savo dialect to Northern Ostrobothnia

True label \	CO	FN	HÄ	NO	SA	SE	SO	SW
CO	36	0	0	5	0	0	2	1
FN	0	29	0	0	2	0	0	0
HÄ	2	0	29	0	0	0	0	0
NO	6	1	0	16	0	1	0	2
SA	2	0	0	1	26	2	0	0
SE	0	0	0	1	8	25	0	0
SO	1	1	0	0	2	25	0	1
SW	2	0	0	4	3	0	1	31
SW	1	0	0	2	0	0	1	29



Variation and change

- Of 93.7M messages, 90.8M are in Finnish, 15.9M in non-standard Finnish and 3.5M in dialectal Finnish
- Most messages thus in standard Finnish and most non-standard messages in colloquial style
- Usage of the colloquial style is in steady decline
- Traditional dialects are not used much but constantly

Contributions

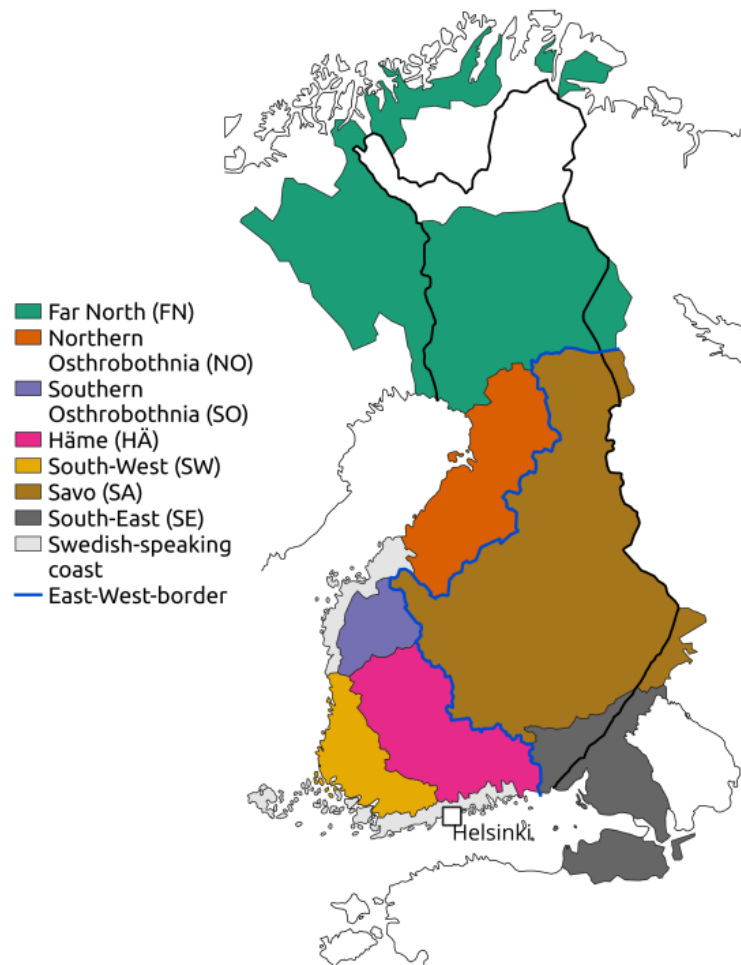
- Manually annotated dataset of around 4000 Finnish Internet forum messages,
- Automatic annotation of all 94M messages in the forum corpus,
- Evaluation of five different language identification tools on the task, and
- Discussion on the variation and language change in the dataset

Data

- Suomi24 ('Finland24') is the largest Internet forum in Finland
- Messages posted on the forum from 2001 to 2020 have been published for academic use in the Language Bank of Finland
- The objective of the current work is to
 - identify the languages and Finnish dialects used on the forum and
 - follow the change of their distributions over time

Manual annotation

- Around 4000 messages annotated manually by variety:
 - Standard Finnish, seven traditional dialects (see Map), Helsinki slang and a colloquial style
 - Based on phonology and morphology
- Split to three folds of train and test splits (90-10)
- Labeled as S24 to not be confused with the forum name
- An additional random test set taken to see how the methods perform with the actual distribution of varieties in the corpus



Methodology

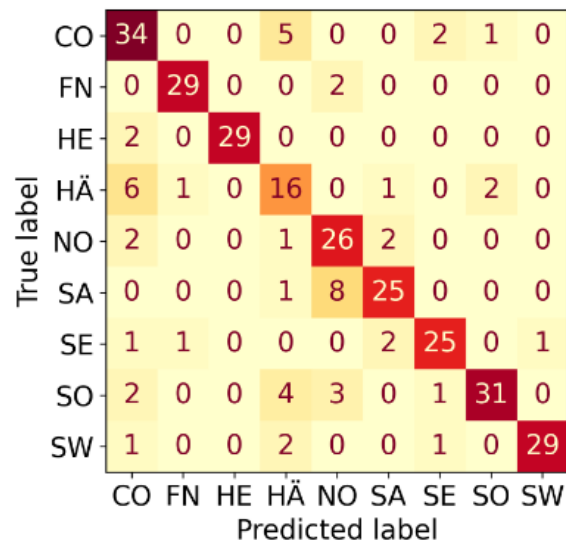
- Train five different dialect identification tools with the manually annotated data and external datasets from the Web and dialect interviews
 - Plus an ensemble model
- Two traditional models (SVM and Naïve Bayes), two off-the-shelf models re-trained (fastText and HeLI) and one neural model (FinBERT)
- Annotation in three stages: first languages, then standard vs. non-standard Finnish and finally the nine non-standard varieties
- The end goal is to collect a dataset of dialectal Finnish Internet forum messages, entitled Murre24 ('Dialect24')

Results

- Language identification with HeLI-OTS
 - Most messages (97% in Finnish)
- For standard vs. non-standard Finnish
 - SVM offers best performance followed by the ensemble model (weighted F1)
 - FinBERT and NB are unstable
- For final dialect identification
 - FinBERT is stable and performs best followed by the ensemble model
- The FinBERT model often mislabels Häme dialect to the colloquial style and Savo dialect to Northern Ostrobothnia

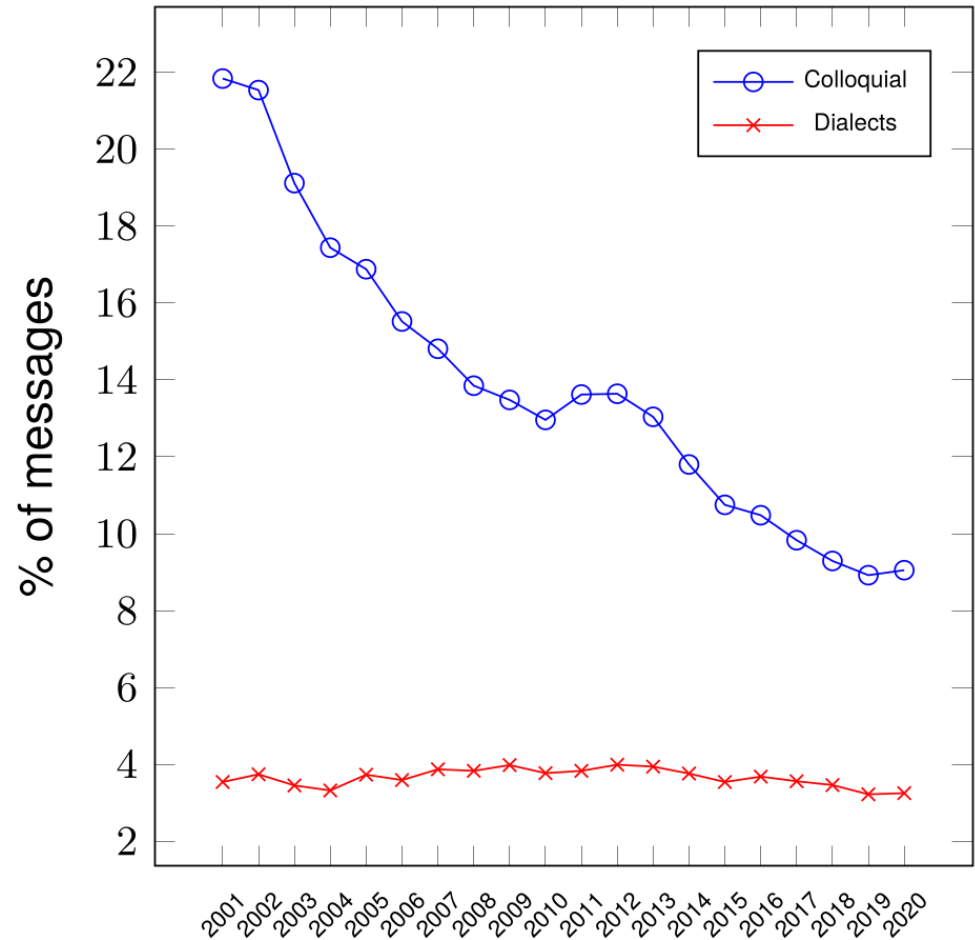
Model	S24	+ Web	+ SKN
Balanced			
NB	0.45±0.01	0.25±0.01	0.59±0.01
SVM	0.89±0.02	0.90±0.01	0.91 ±0.02
fastText	0.87±0.01	0.86±0.02	0.87±0.01
HeLI	0.86±0.02	0.59±0.01	0.76±0.03
FinBERT	0.76±0.23	0.21±0.00	0.21±0.00
Ensemble	0.89±0.00	0.89±0.02	0.89±0.02
Random sample			
NB	0.06±0.00	0.73±0.00	0.75±0.01
SVM	0.72±0.02	0.86 ±0.01	0.86 ±0.00
fastText	0.77±0.02	0.85±0.00	0.85±0.01
HeLI	0.80±0.02	0.37±0.02	0.62±0.01
FinBERT	0.58±0.37	0.73±0.00	0.73±0.00
Ensemble	0.73±0.02	0.85±0.00	0.86 ±0.01

Model	S24	+ Web	+ SKN
Balanced			
NB	0.56±0.01	0.55±0.01	0.10±0.00
SVM	0.80±0.01	0.81±0.01	0.78±0.01
FastText	0.74±0.02	0.75±0.02	0.70±0.00
HeLI	0.68±0.02	0.68±0.02	0.63±0.03
FinBERT	0.81±0.01	0.81±0.01	0.78±0.01
Ensemble	0.82 ±0.01	0.82 ±0.01	0.78±0.00
Random sample			
NB	0.86±0.01	0.79±0.02	0.00±0.00
SVM	0.84±0.01	0.83±0.01	0.74±0.00
FastText	0.76±0.01	0.76±0.01	0.71±0.01
HeLI	0.72±0.01	0.73±0.01	0.80±0.01
FinBERT	0.85±0.02	0.87 ±0.01	0.82±0.02
Ensemble	0.85±0.01	0.85±0.01	0.81±0.00



Variation and change

- Of 93.7M messages, 90.8M are in Finnish, 15.9M in non-standard Finnish and 3.5M in dialectal Finnish
- Most messages thus in standard Finnish and most non-standard messages in colloquial style
- Usage of the colloquial style is in steady decline
- Traditional dialects are not used much but constantly





Olli Kuparinen
Faculty of Information Technology and
Communication Sciences, Tampere University
Department of Digital Humanities,
University of Helsinki

