Digital Humanities in the Nordic Countries, 7-9 March 2018, University of Helsinki

The Nordic Tweet Stream: A dynamic real-time monitor corpus of big and rich language data

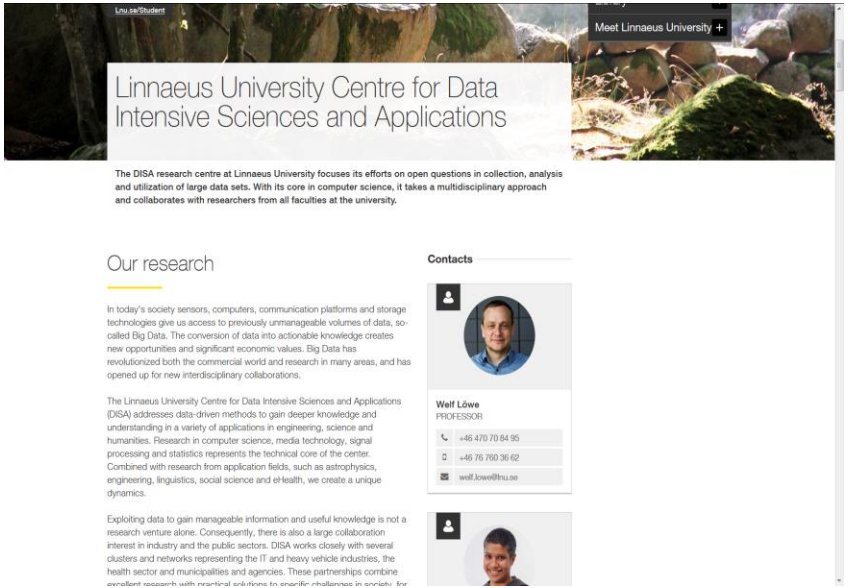
Mikko Laitinen, Jonas Lundberg, Magnus Levin, Rafael Martins

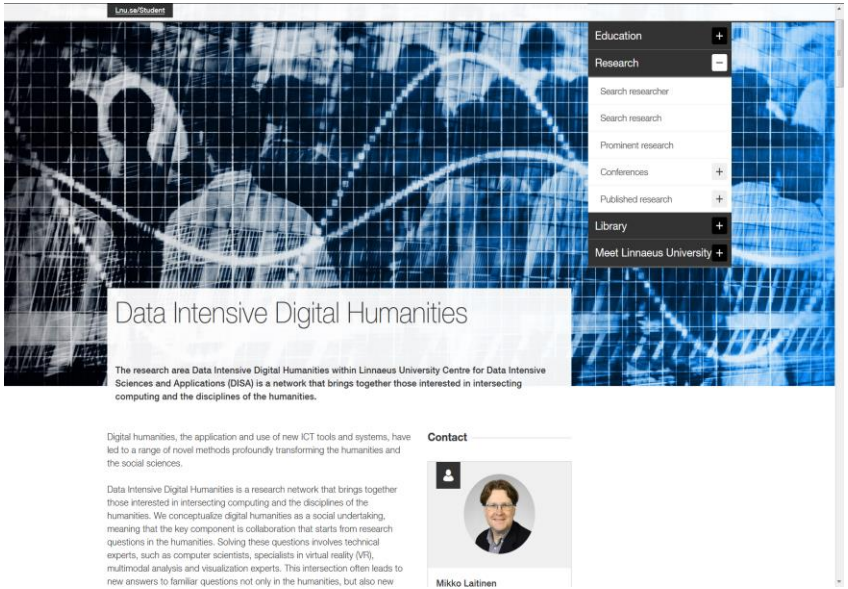
Contact: mikko.laitinen@uef.fi

Objectives

- To present our data streaming
- To contextualize this project within sociolinguistics
- To present some recent activities

Data intensive sciences and applications (DISA) and digital humanities: 5-y. project at Linnaeus University and partners at UEF





UEF // University of Eastern Finland

Linnæus University

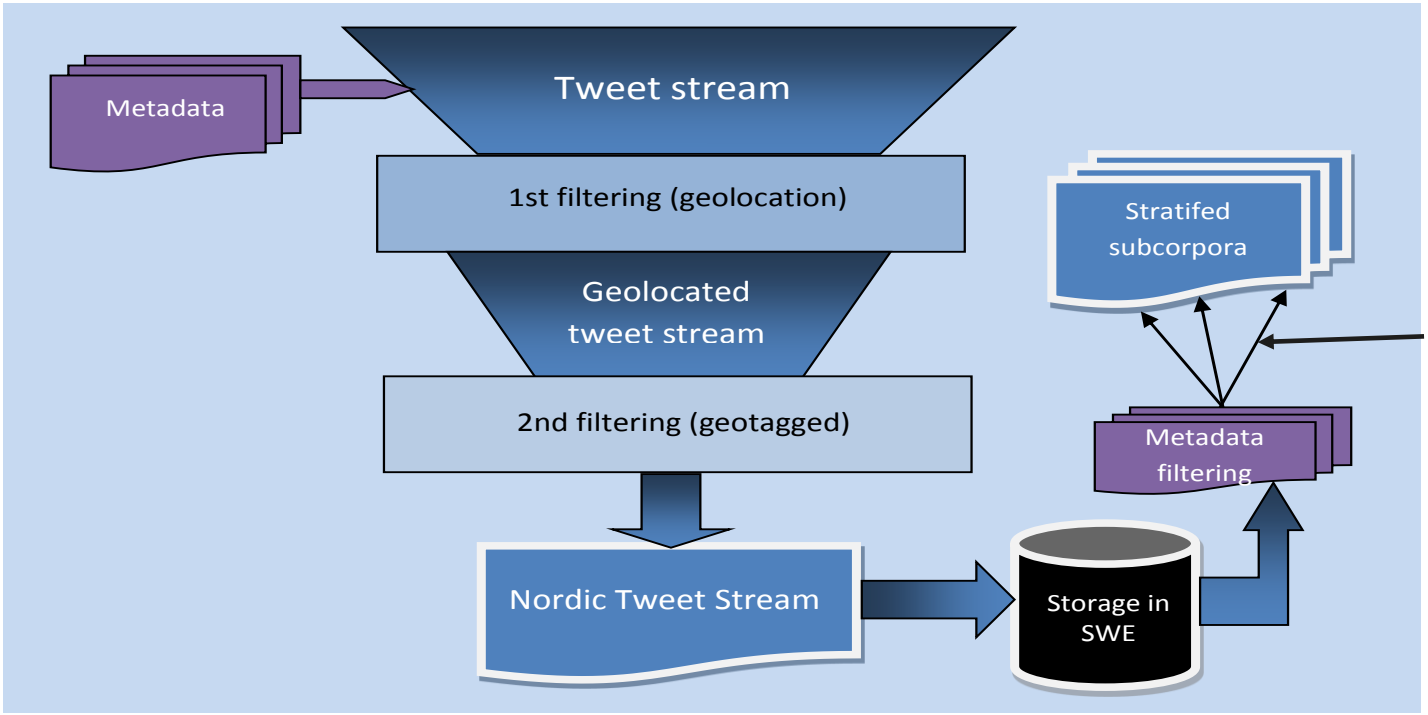
Our team

- Mikko Laitinen coordinator
- Linnaeus
  - Dr. Jonas Lundberg (comp. science)
  - Dr. Rafael Martins, Dr. Ilir Jusufi, Dr. Aris Alissandrakis (Media technology)
  - Prof. Jukka Tyrkkö, Ass. Prof. Magnus Levin (English linguistics)
- UEF
  - Mr. Jehad Aldahdood (database, graphic search interface designer)
  - Initial collaboration with Prof. Pasi Fränti's Machine Learning group at UEF

The Nordic Tweet Stream (NTS)

- real-time data stream
- access to geo-located tweets in five Nordic countries (DEN, FIN, ISL, NOR, SWE)
- streaming initiated in April 2016 (entire dataset from 6 Nov 2016)
  - 26 Feb 2018: 11,657,987 messages from 283,811 user accounts
  - 1,434 unique locations
- Nearly 150 million tokens of text (c. 350,000 tokens added per day)
- over 0.6 billion metadata points

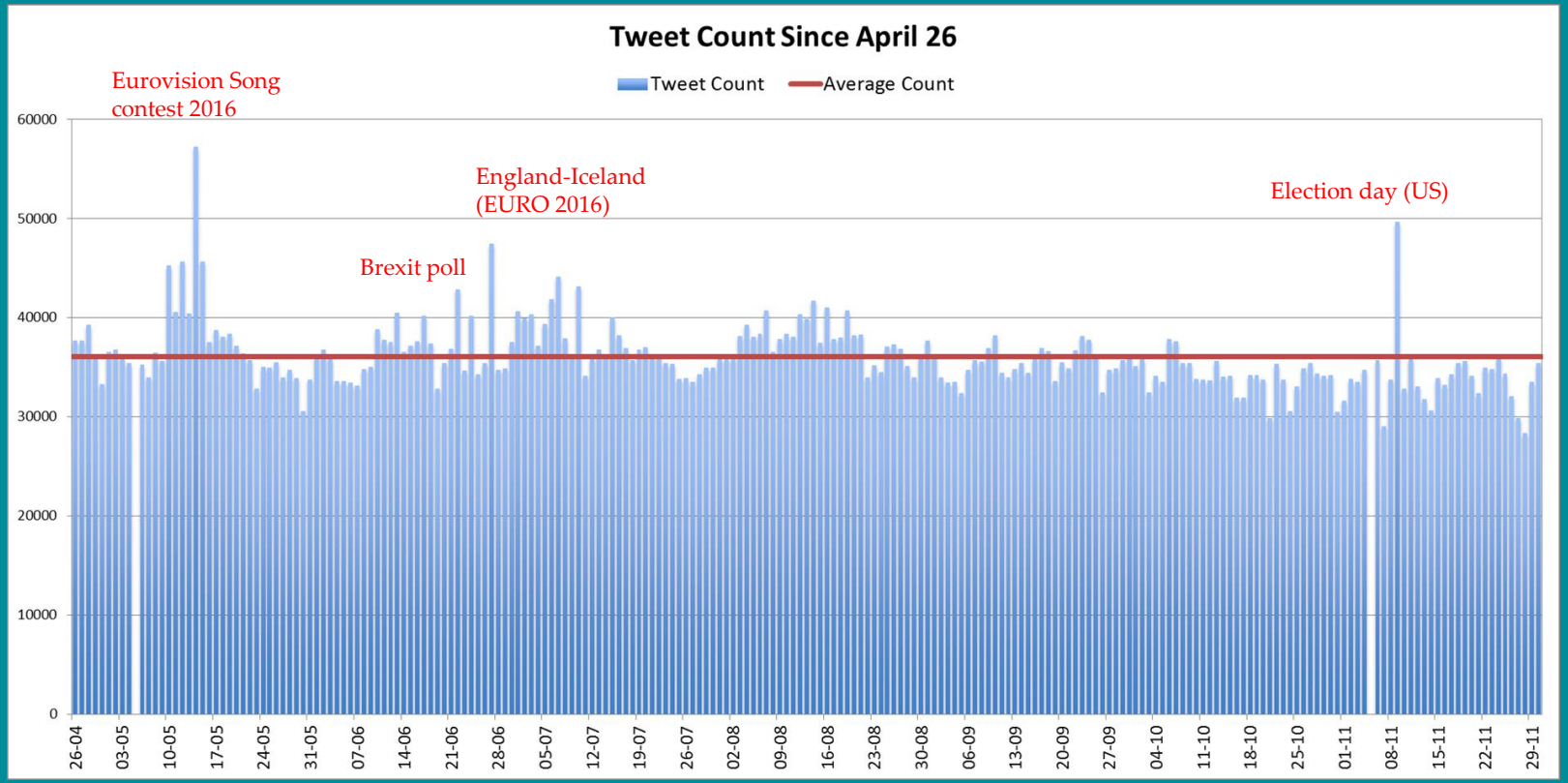
Data streaming



Supervised machine learning to exclude bots (a text-based classifier and a profile-based system developed by Lundberg et al. 2018)

Operational for English bots (23.2%) and Swedish (1.2%)

Real-time access



Using the NTS

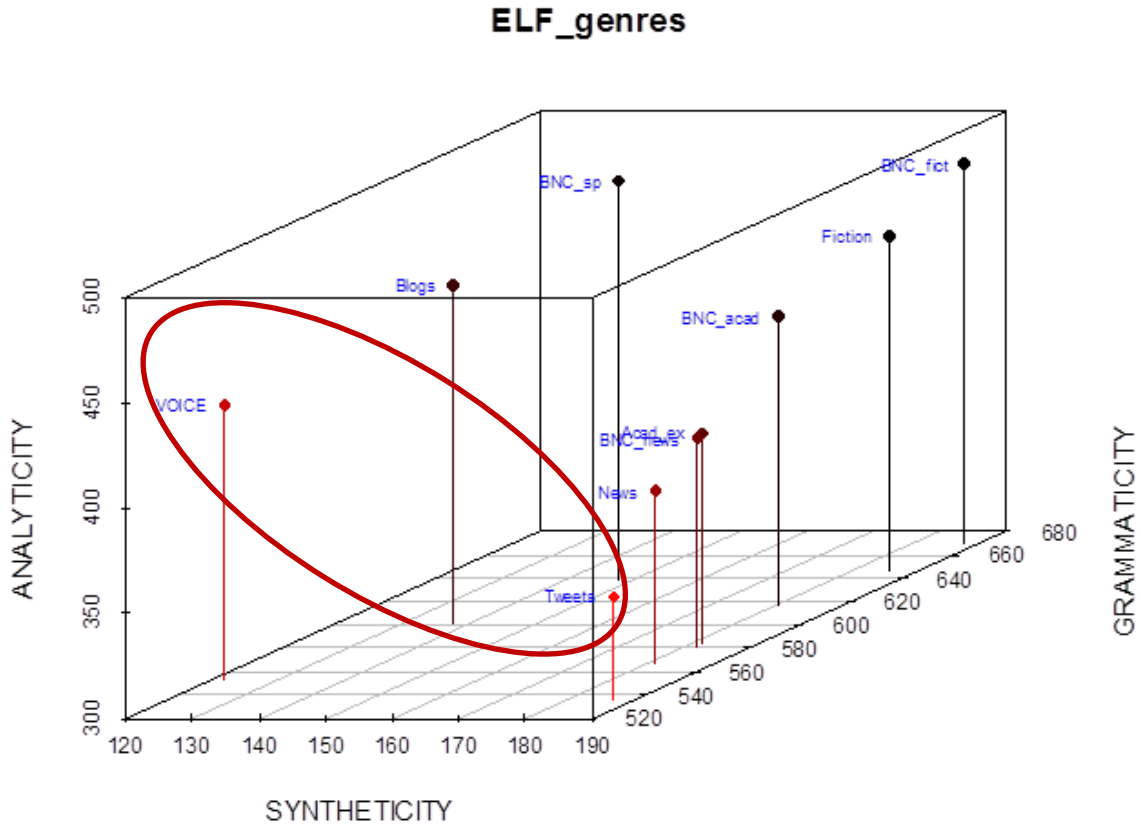
- Material for our studies of English as lingua franca (ELF) (e.g. Mauranen 2012; Mauranen et al. 2015)
- Charting new ways of analyzing ELF
  - Quantitative study of grammar
  - Language change in ELF
- Exploring how new evidence could offer new answers to familiar questions in language change and ELF, and (hopefully) lead to entirely new questions

Previous research

- Coats (2015, 2016, 2017a, b): Finnish Twitter English, gender stratification
- Eisenstein et al. (2014): regional diffusion of lexis
- Scheffler (2014): A German Twitter snapshot
- Barbaresi (2016): Austrian Twitter corpus
- Huang et al. (2016): American English dialect areas
- Goncalves et al. (2017): Americanization in Twitter

An overview of recent research activities

A. English tweets and macro-level genre characteristics (Laitinen in press)

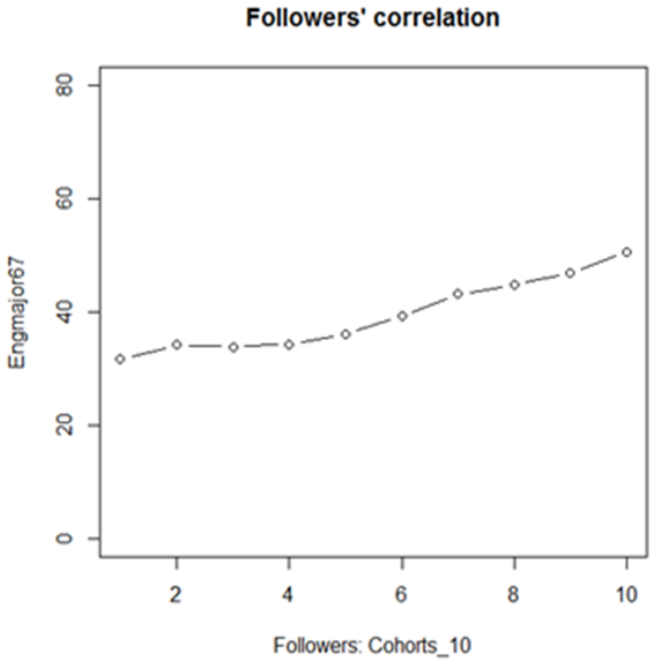


Typological profiling (Szmrecsanyi 2009)

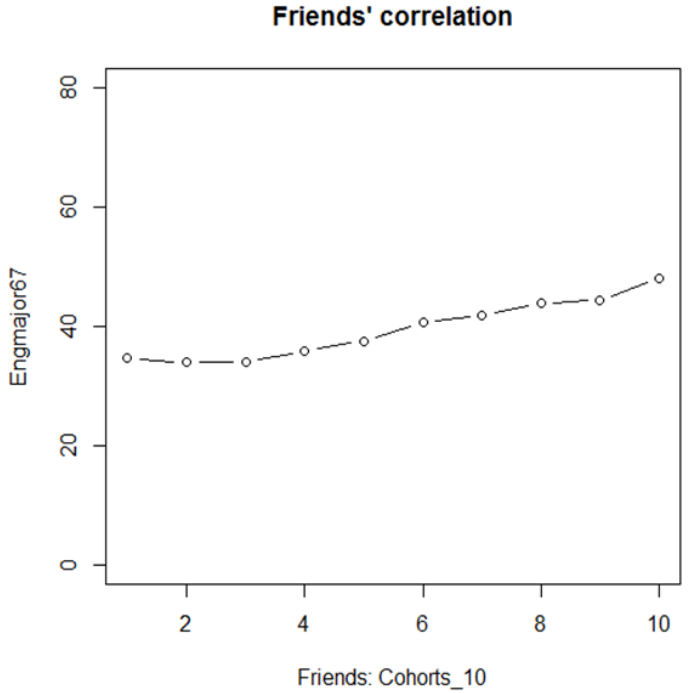
- 11 analytic markers
- 6 synthetic markers

B. Social network theory and Twitter metadata: Weak ties facilitate change (Laitinen et al. 2017a)

Evidence from: 199,842 accounts
Followers: truly weak ties



Friends: slightly stronger ties

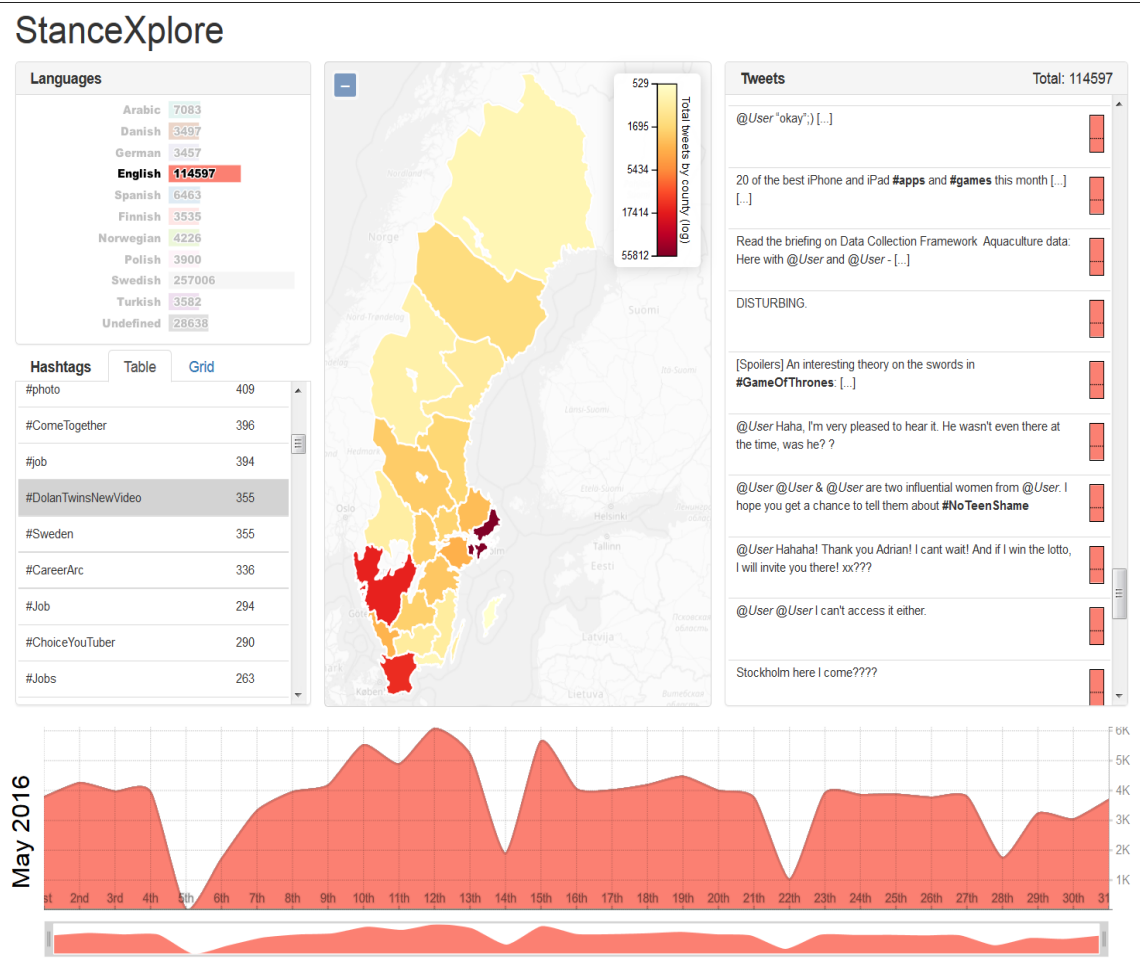


Our data lead us to propose a revision to the social network hypothesis

- “There is a straightforward increase of innovative behavior in the truly weak tie network, but… [our results indicate] that innovations also spread under conditions of stronger networks, given that the network size is large enough. The size of such networks must exceed circa 100–130 individuals according to our observations. It is important to note that if we had restricted ourselves only to small networks, this observation could not have been established.” (Laitinen et al. 2017)

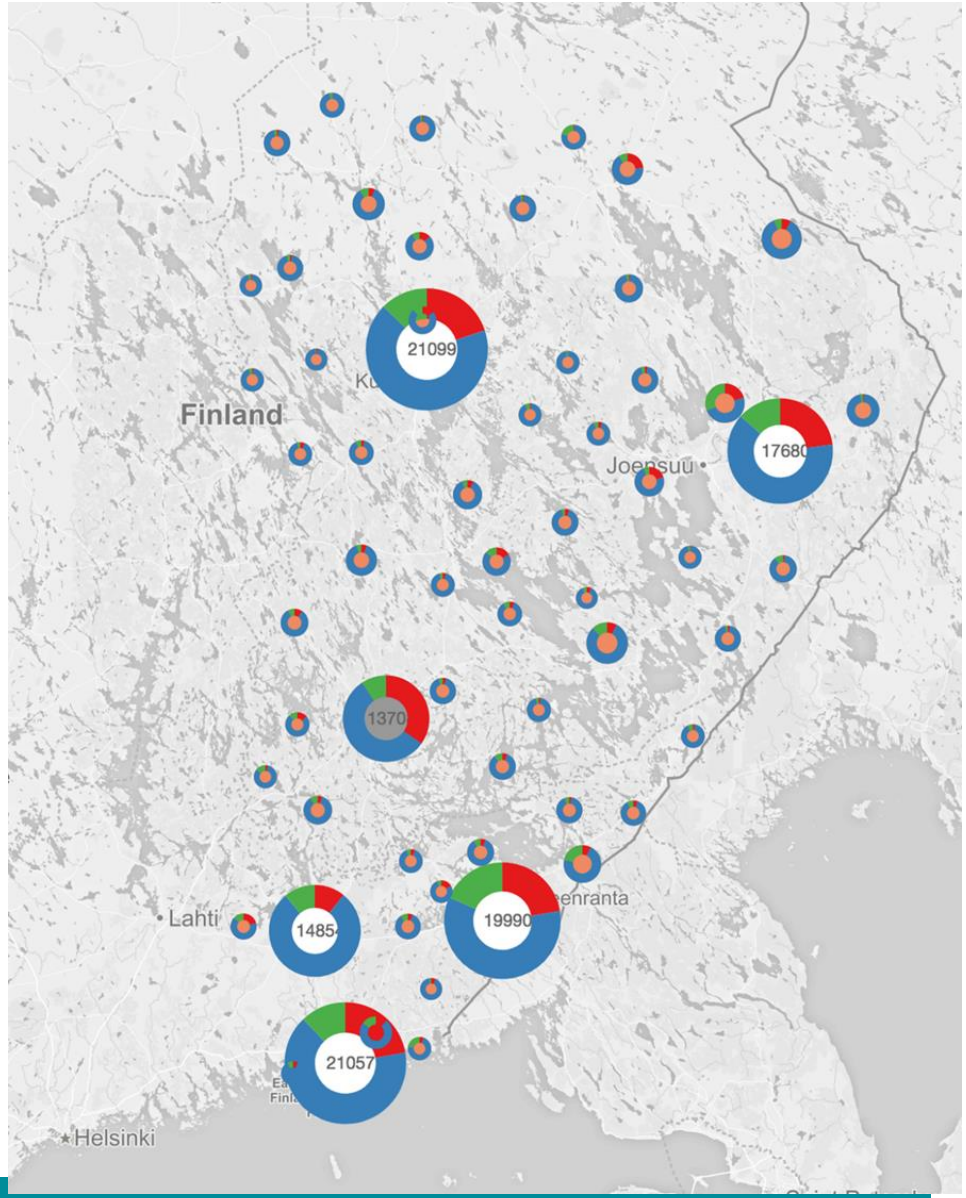
C. Language choice

- Visuals by Rafael Martins (LnU)
- Testing visualizations on the Swedish data

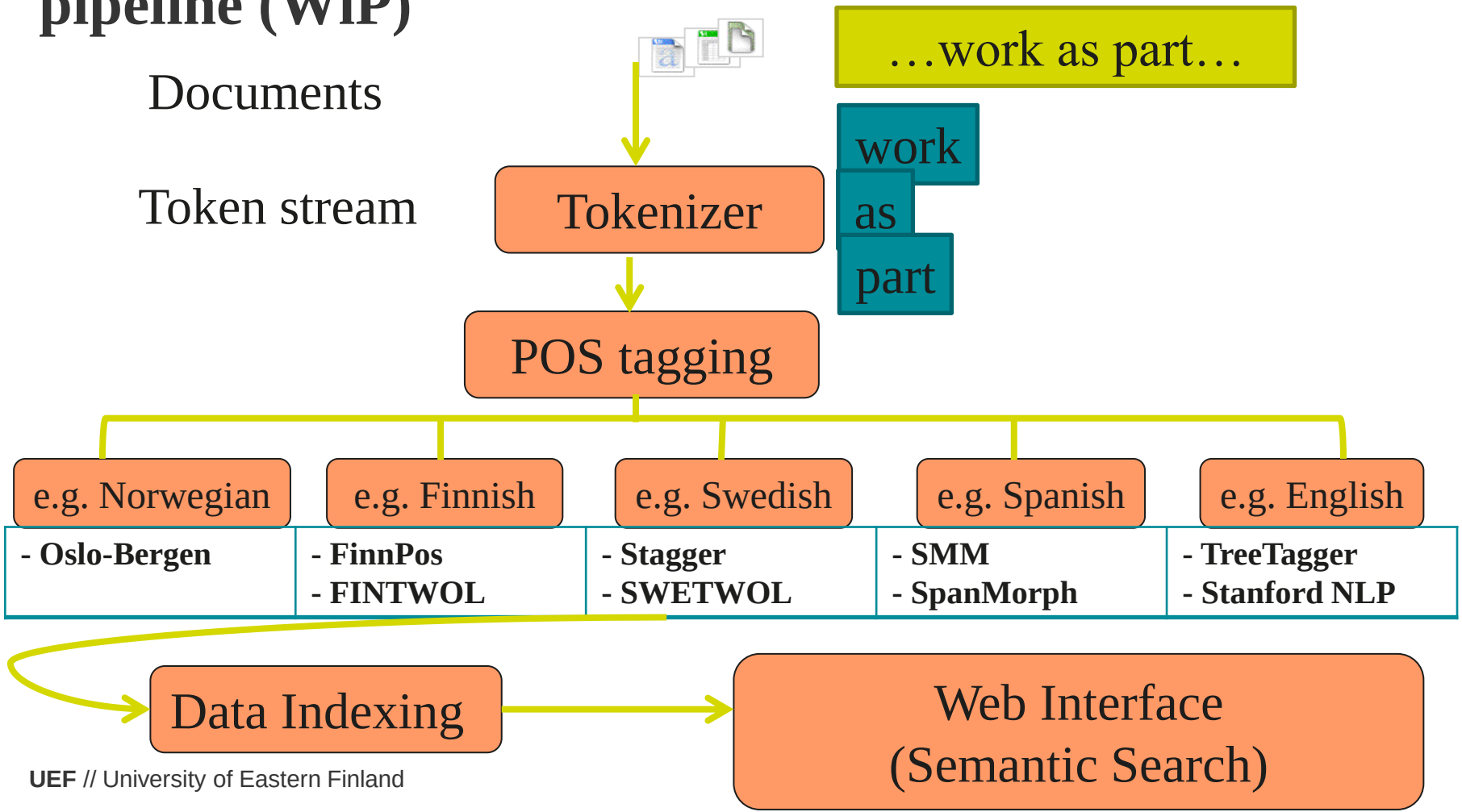


D. Regional language choice (excluding Eng bots)

- 6,397 accounts from 61 municipalities in Eastern Finland (Laitinen, Paulasto & Meriläinen forthc., visuals by Ilir Jusufi)



E. A graphic search interface for the NTS data: pipeline (WiP)



Thank you!

uef.fi

References

Barbaresi, Adrian. 2016. Collection and indexation of Tweets with a geographical focus. Tenth International Conference on Language Resources and Evaluation (LREC 2016), May 2016. *Proceedings of the 4th Workshop on Challenges in the Management of Large Corpora (CMLC)*, (2016) 24–27. <hal-01323274>

Coats, Steven. 2015. Non-standard lexical and grammatical resources in Finland Twitter English. A paper presented at Poznan Linguistics Meeting, 19 September 2015, Poznan, Poland. <http://cc.oulu.fi/~scoats/poznan1handout.pdf>, 22.1.2018.

Coats, Steven. 2016. Grammatical frequencies and gender in Nordic Twitter Englishes. – Darja Fišer and Michael Beißwenger (toim.). *Proceedings of the 4th conference on CMC and social media corpora for the humanities*. Ljubljana: U. of Ljubljana Academic Publishing, 12–16.

Coats, Steven. 2017a. European language ecology and bilingualism with English on Twitter. – Ciara Wigham & Egon Stemle (toim.). *Proceedings of the 5th conference on CMC and social media corpora for the humanities*. Bozen/Bolzano: Eurac Research, 35–38.

Coats, Steven. 2017b. Gender and lexical type frequencies in Finland Twitter English. – Turo Hiltunen, Joe McVeigh & Tanja Säily (toim.). *Studies in Variation, Contacts and Change in English 19: Big and Rich Data in English Corpus Linguistics: Methods and Explorations*. Helsinki: VARIENG. http://www.helsinki.fi/varieng/series/volumes/19/coats/

Gonçalves, Bruno, Lucía Loureiro-Porto, José Ramasco, David Sánchez. 2017. The Fall of the Empire: The Americanization of English. arXiv: 1707.00781.

Eisenstein Jacob, Brendan O'Connor, Noah A. Smith & Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLoS ONE* 9(11): e113114. doi:10.1371/journal.pone.0113114

Yuan Huang, et al. 2016. Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems.* doi:10.1016/j.compenvurbsys.2015.12.003.

Lundberg, Jones, Jonas Nordqvist, Antonio Matosevic. 2018. A. On-the-fly Detection of Autogenerated Tweets, arXix preprint.

Mauranen, Anna. 2012. *Exploring ELF: Academic English Shaped by Non-Native Speakers*. Cambridge: Cambridge University Press.

Mauranen, Anna, Ray Carey & Elina Ranta. 2015. New answers to familiar questions: English as a lingua franca. In Douglas Biber & Randi Reppen (eds.), *Cambridge Handbook of English Corpus Linguistics*, 401–417. Cambridge: Cambridge University Press.

Scheffler, Tatjana. 2014. A German Twitter Snapshot. *Proceedings of LREC,* (2014), 2284–2289.

Szmrecsanyi, Benedikt. 2009. Typological parameters of intralingual variability: grammatical analyticity versus syntheticity in varieties of English. *Language Variation and Change* 21:3, 319–353.

Our publications mentioned

Laitinen, Mikko. In press. Placing ELF among the varieties of English: Observations from typological profiling. In Sandra Deshors (ed.), *Modelling World Englishes in the 21st century: Assessing the interplay of emancipation and globalization of ESL varieties*. (*Varieties of English around the World*). Amsterdam: John Benjamins.

Laitinen, Mikko, Jonas Lundberg, Magnus Levin & Alexander Lakaw. 2017. Revisiting weak ties: using present-day social media data in variationist studies. In Tanja Säily, Minna Palander-Collin, Arja Nurmi, & Anita Auer (eds.), *Exploring Future Paths for Historical Sociolinguistics*, 303–325. Amsterdam: John Benjamins. DOI 10.1075/ahs.7.12lai.

Laitinen, Mikko, Heli Paulasto & Lea Meriläinen. Forthcoming. Monikielinen Twitter: Kielenvalinta ja englannin kielen muutos itäsuomalaisessa tviittivirrassa. In Milla Uusitupa, Leena Kolehmainen & Helka Riionheimo (eds.), Itäsuomalainen monikielisyys. Helsinki: SKS.