Digital social network analysis: Studying diffusion of innovation in online data

in-person course on Thursday (all day) and Friday morning by
Mikko Laitinen

Short CV: Mikko Laitinen is Professor of English Language at the University of Eastern Finland. Currently, he is one of the principal investigators in the national research infrastructure in digital humanities, FIN-CLARIAH, funded by the Research Council of Finland. His research focuses on data-intensive sociolinguistics, on the role social networks in language variation and change, and English as non-native language in the Nordic region.

Course description:
This course focuses on social network analysis in variationist sociolinguistics and uses large-scale social media data as empirical material. It presents a series of studies carried out by an interdisciplinary research group that uses interactional parameters that are observable in social media data. We utilize these interactional parameters as background factors conditioning variability, and our studies develop methods to trace how linguistic innovations spread into digital communities. The larger underlying issue deals with utilizing social media to fuller potential in English linguistics. Much of social media data today tends to be treated only as massive text collections and few approaches take into account the fact that the purpose is to form communities in which people are connected in variable ways.

The studies presented are based on social network theory, which has offered a powerful tool in modeling how linguistic innovations spread into communities (Milroy 1987). Existing work on networks, however, leaves open a number of questions. First, most studies have investigated small networks, whose sizes are substantially smaller than average human networks. In addition, much of the evidence comes from traditionally close-knit urban working-class settings or from peripheral rural communities, and data from highly diverse digital networks could lead to novel insights.

The first lecture first discusses social networks and then presents a study that uses a simple metadata parameter of friends and followers in Twitter data. This study makes use of data from 48,241 user accounts in Sweden, and we compare network sizes with language choice, i.e. we explore whether there is a connection between network properties and the main language used in online communication (Laitinen & Lundberg 2020). We observe that social network size is conditioned by the main language of a user (be that English, Swedish or any other language), and that the sizes of digital networks are actually very close to the average size of human networks, as observed in other fields (e.g. Dunbar 2020).

The second case study introduces a more advanced method (Laitinen, Fatemi & Lundberg 2020). This algorithmic method takes into account not only network size but also network structure. It uses network information that can be quantified and used as proxies in a directed graph network, like Twitter. These network parameters (e.g. density, distance, connectedness and similarity) are adopted from the graph theory but have not previously been used in the corpus-based study of variation and change. The method enables us to label any network with a network score, which is an estimate of whether people in a network are mainly connected to each other through close-knit ties or if the network is characterized by loose-knit ties.

In the third part, I will present a study, which is currently under review. It uses a massive dataset of 3,935 randomly selected ego networks from the US/UK. These networks contain data from 233,774 individuals. In the empirical part, we utilize all the texts from all the individuals in the networks (nearly 4.8 billion words) and make use of the algorithmic method for establishing network strength scores. We test if the weak-tie hypothesis, which has been observed in numerous previous studies, also holds in the digital domain. The case study investigates how two linguistic features that are currently undergoing change in English are conditioned by network properties. The empirical study concentrates on one

orthographic feature (contractions of negatives (e.g. not >n't) and verbs (e.g. we will > we'll)) and one grammatical structure (NEED to + V), both of which are undergoing frequency increases in English, but are driven by differing forces of colloquialization and grammaticalization. Our observations using evidence from data-intensive methods may lead to rethinking the role of social networks in language change.

The fourth part is based on an on-going study that seeks to add a novel angle to previous studies on lexical innovation (e.g. Grieve et al. 2017). We use the same 4.8 billion-word material from UK and US and trace emerging lexis in it, but rather than focusing on lexis itself, we explore from what types of networks these novel items emerge and how network properties condition how they spread.

Towards the end, I will concentrate on how large social media data could be more extensively used in English linguistics in future. This discussion concentrates on the need to enrich large data with social background information. It is clear that the network parameter is just one background factor that could be added to big social media data to make them richer in terms of background information. For future studies, it would be ideal to be able to predict other social parameters, such as users' age, gender or social layer. Methodologically this means developing new ethically sustainable ways of enriching born-digital data, but also increasing interdisciplinary collaboration between sociolinguistics and experts in data mining/artificial intelligence. Similar efforts are currently undertaken in medicine for instance (cf. Mitsuyama et al. 2023 on using artificial intelligence to predict age using chest radiography).

References:

Dunbar, Robin. 2020. Structure and function in human and primate social networks: Implications for diffusion, network stability and health. *Proceedings of the Royal Society A*. 476A.

Grieve, Jack, Andrea Nini and Diansheng Guo. 2017. Analyzing lexical emergence in American English online. *English Language and Linguistics* 21: 99-127.

Laitinen, Mikko & Jonas Lundberg. 2020. ELF, language change and social networks: Evidence from real-time social media data. In Anna Mauranen & Svetlana Vetchinnikova (eds.), *Language Change: The Impact of English as a Lingua Franca*, 179–204. Cambridge: Cambridge University Press.

Laitinen, Mikko, Masoud Fatemi & Jonas Lundberg. 2020. Size matters: Digital social networks and language change. *Frontiers in Artificial Intelligence* 3:46. doi: 10.3389/frai.2020.00046.

Milroy, Lesley. 1987. *Language and Social Networks*. 2nd ed. Blackwell.

Mitsuyama, Yasuhito et al. 2023. Chest radiography as a biomarker of ageing: artificial intelligence-based, multi-institutional model development and validation in Japan. *Lancet Healthy Longev* 4: e478–86. doi 10.1016/S2666-7568(23)00133-2.