#3

Data-intensive sociolinguistics using social media

Mikko Laitinen

Masoud Fatemi

Abstract

This article looks into using large-scale social media data in SSH research and in particular in studies of language variation and change.
It presents a study that investigates the role of social networks in linguistic variability. Previous studies have convincingly shown that networks in which people are connected to each other in loose ties tend to contribute positively to language change. Conversely, networks in which people are closely connected to each other inhibit change. This conclusion is, however, based on small datasets from small networks, and this study tests if the difference is diluted when network size is closer to human average. The results from 3,935 networks suggest this to be the case. Towards the end, the article suggests numerous ways in which large-scale social media data and the use of data intensive methodologies could be increased and encouraged in SSH research.

1. Introduction

This article discusses how large-scale social media data and data-intensive methodologies could be more effectively used in fundamental research. Various social media applications offer a promising and extremely large source of data for a range of disciplines in social sciences and the humanities (SSH) today. This is especially true in the study of language variation and change, which concentrates on the processes of how change emerges from usage and how novel forms spread through communities, and special attention is also put on how various demographic and contextual factors condition variability (Rissanen 2009; Szmrecsanyi 2017). The benefits are clear since social media substantially expands the empirical basis and potential pools of evidence. According to informal estimates, nearly 60% of the world population use various social media applications, and the share in developed countries is substantially higher, with Northern Europe leading in the portion of digital population with over 90% (Statista 2023). Given that many applications have several hundreds of millions of users, it is, in theory, possible to improve empirical validity of variationist studies considerably. Social media data could for instance be used to correct biases in SSH research in general, such as the elite bias in traditional sampling (Groves & Couper, 1998). Alternatively, sociolinguists interested in variability can base their observations on language use of very large populations (Laitinen & Lundberg 2020) or ensure that empirical data also includes language varieties that are under-represented in traditional text corpora (Tyrkkö et al. 2021). Another benefit is that many applications function so that user-generated material, texts, and multimodal material, can be collected in real time and in quantities that were impossible to imagine 20 years ago. Researchers can for instance acquire a detailed overview in regional dialects. Huang et al. (2016) present a dialect survey in which there is material from every county in the mainland USA. Lastly, many applications also contain metadata that can be used to describe and contextualize these user-generated material, so that in some cases it is possible to know when and where for instance certain language forms are used or to be able to trace the emergence of new words almost real time (Grieve et al. 2016; 2018).

It is clear that the importance of social media should not be overestimated and any data and their quality for research should be approached critically. This article argues that despite the potential and benefits and some pioneering studies, data from social media applications are underused at least in variationist linguistics. When such data are used, they are mainly treated as massive collections of text (i.e. traditional corpora), and studies mostly rely on feature-based listings of linguistic forms. That is, research tends to ask similar questions than when using traditional empirical material.

This article presents a study that uses social media data by accounting for the fact that the purpose of social media is to form networks and communities in which people are connected in variable ways. The study deals with the role of social networks in language variation and change. Various studies since the 1970s have convincingly shown that networks play a crucial role in how linguistic change spreads into communities (Milroy 2004). The work carried out has operationalized networks as the sum of contacts and connections that an individual has, and networks in which people live can vary in terms of how strongly connected individuals are. The usual way of characterizing networks sees networks on a continuum in which one end consists of loosely connected individuals (acquaintances) while the other end has close-knit networks of friends and family for instance. Empirical evidence has convincingly shown that weak-tie environments are open to external influences and facilitate change, whereas strong ties lead to norm-enforcing communities that resist change.

This conclusion is, however, based on evidence from small datasets obtained through ethnographic observations from networks and communities of 30–50 individuals (Milroy & Milroy 1992: 4). What is more, most of the evidence comes from traditionally close-knit settings in urban blue-collar settings or from small peripheral rural communities. Given that social media connects large numbers of people networks from a range of settings, the weak-tie hypothesis could be with larger datasets and with data-intensive methodologies. Social media data with large scale and scope could be used as testing ground for the network theory.

In what follows, I will present a study carried out in an interdisciplinary group of sociolinguistics and computer scientists. The study uses a dataset from over 233,000 people, who created some 1.8 million connections between them. These connections are directly observable in the data and can be linked with the nearly 4.8 billion words of text generated by the account holders. Section 2 provides the theoretical background for social network analysis in sociolinguistics. Section 3 shows how my group in the University of Eastern Finland uses interactional data obtained from a directed graph network (i.e. follower patterns, replies, and mentions, etc.) to construct a social network index. In Section 4, the results of a data-intensive case study are presented in which the results suggest that the main conclusion of social network hypothesis in sociolinguistic might require re-thinking. The conclusion outlines future research in data-intensive studies of social media.

2. Social networks in sociolinguistics

Networks are our aggregate connections through which people are grouped on the basis of frequency and quality of their interactional behavior (Milroy 1987). Borrowed from sociology to sociolinguistics in the 1960/70s, the concept has offered a powerful way to understand how relationships that we contract with others can reach through social and geographic spaces to link individuals.

Social network analysis transpires from the idea that individuals establish interpersonal ties of varying strengths to form communities around an ego. Such ego networks are formed around any ego node that forms an anchor for establishing a set of ties that connect people, hence the term ego network. Networks have been defined using both frequency and quality of ties as the key criteria. Frequency points to how dense or loose a network is depending on how extensively people (network nodes) are linked with each other. A network can be maximally dense, when every node knows each other. Tie quality is measured through how multiplex or uniplex ties in a network are. If nodes only know each other through one role, the tie is uniplex, and when through multiple channels (work, hobbies, organizations, etc.), the tie is multiplex. In Figure 1 below, the ego is the yellow center node, and the blue lines indicate one-way connections, while the black mark reciprocal connections. As can be visually confirmed, the network on the left is characterized by loose connections, and the one on the right is stronger.

Importantly for the study of language variation and change, network characteristics influence the rate at which innovations are adopted by individuals (Milroy 2004). A range of empirical studies has shown that strong networks tend to maintain and support local norms and provide resistance to the adoption of competing norms from the outside. Conversely, weak and uniplex ties are important channels for outside influence as people in such situations tend to accommodate to each other linguistically. Contact situations with weak ties therefore contribute positively to how incoming and new features spread.

This finding builds on Granovetter's (1973: 1365) observation that "only weak ties may be local bridges". More people can be reached through weak ties, but not all weak ties serve this function, "only those acting as bridges between network segments" (1983: 229). To explain this somewhat counterintuitive observation, Granovetter (1973) argues that close-knit networks encourage local cohesion and norm-enforcing communities in which adopting an innovation is risky. Rocking the boat in a dense network tends to be socially risky, and evolution has taught that such behavior is best avoided. Conversely, loose-knit networks with individuals already on the social fringes are more susceptible to external innovations in the first place. This has been explained by suggesting that people on the fringes have less to lose socially, and rocking the boat effect is not as considerable as in strong-tie environments. Lastly, weak ties may be expected to be more numerous among mobile individuals and are thus more likely to contribute to the diffusion of an innovation. Individuals with weak ties across networks are not necessarily restricted to one network, but can reach through several.

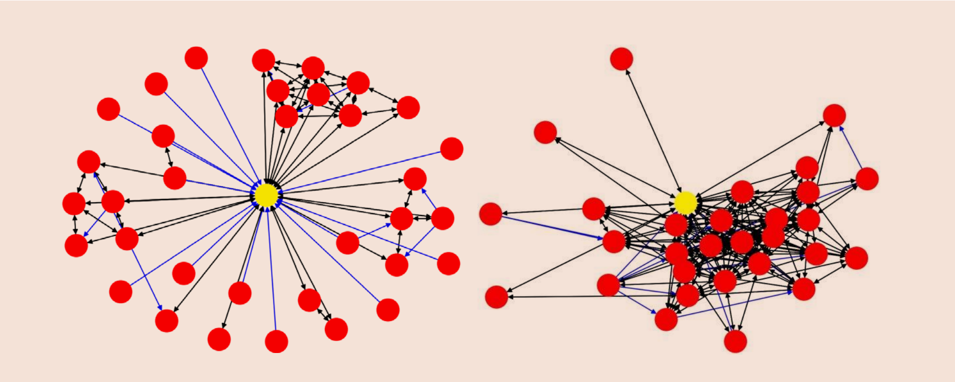


Figure 1. Two ego networks of varying strength (left=looser, right=stronger)

In variationist sociolinguistics, network ties have been operationalized in various ways (Milroy 2004). To illustrate, in the Belfast study, ties were measured using five indicators to establish how dense and complex a particular tie was. These indicators consisted of (a) having membership in a locally-based group, (b) having ties with at least two households in the neighborhood, (c) sharing a workplace with two or more individuals from the neighborhood, (d) sharing a workplace with same-sex individuals from the neighborhood, and (e) being involved in voluntary activities with individuals from the same workplace. The responses resulted in a network strength scale, which formed an independent variable, and these values were compared to the dependent (phonological) variables. The results show that the individuals with strong network ties with the local community also exhibited the highest share of local, vernacular speech, and "that a close-knit network has an intrinsic capacity to function as a norm-enforcement mechanism, to the extent that it operates in opposition to larger scale institutional standardising pressures" (Milroy & Milroy 1985: 359).

A large body of variationist sociolinguistic literature exists in which the network-based approach has been applied to small contemporary communities (Milroy 2004). Milroy & Milroy (1978) use 46 speakers from three urban, blue-collar Belfast communities, and the network ties were established through a participant observation process in which a researcher was introduced to a community by means of a friend-of-a-friend technique. Of these, 12 had network scores qualifying them as weak tie individuals.

Carefully constructed personal networks are obviously important, but the availability of social media data also forces us to ask if the model holds when tested with considerably larger networks. Findings in social anthropology have estimated that an average network size is larger than a few dozen individuals. Dunbar (1992: 469) has suggested that the neocortex size in general and the number of neocortical neurons impose a cognitive upper limit on an individual's information-processing capacity. These limit "the number of relationships that an individual can monitor simultaneously" to around 150 nodes (Dunbar 2023). McCarty et al. (2001) estimate the size of average networks to be substantially larger (c. 290) but having a maximum, while they also recognize that network sizes for various subpopulations can be substantially larger (clergy, politicians, and diplomats, etc.).

Recent findings from other disciplines that use social media suggest that network sizes may indeed play a more substantial role than previously thought. A study on the relationship between trust on social media and network size surveyed 6,383 Facebook Groups users (Ma et al. 2019). The observations show that people trust private groups more than they do public groups, which is to be expected. However, the differences between group types disappear once the group size exceeds a certain threshold (c. 150 members). The findings suggests that when networks become larger, individuals are no longer be able to perform the mental reasoning of who actually is in the group and who is not, and therefore the difference between the levels of the network types.

A critical reader might ask if real world ties and social media ties are indeed different. Various studies suggest that online communities have similar structural characteristics to off-line networks. Dunbar et al. (2015) for instance show that online networks closely resemble, both in size and structure, those found in face-to-face contact. Gonçalves et al. (2011) investigate the size of digital networks and observe them to be in the range of 100–200 nodes. This also seems to hold in highly multilingual networks, as observed by Laitinen & Lundberg (2020) using data in which the main language of a network was controlled.

3. Enriching big social media data

The material used in the study was tailor-made and collected in close collaboration with computer scientists in an ongoing digital humanities research infrastructure project, DARIAH-FI, funded by the Research Council of Finland. Since the objective was not only to collect a sizeable sample of texts generated by account holders, but also to ensure that we would be able to acquire interactional network information of everyone who belongs to an ego network in addition to their texts, the work turned out to be time consuming, taking roughly eight months. The data are from microblogging application Twitter (currently knows as X) and were collected using the academic API, which has now (Aug 2023) been closed down. The data are stored in UEF servers. According to the European Union's legislation (Digital Single Markets Directive 2019/790), born-digital data, such as this, can be stored and used for fundamental research. Similarly, the Finnish copyright legislation was amended in spring 2023 to enable text and data mining.

Obtaining data suitable for sociolinguistic analyses through the API tends to result in noisy data through processes that are not transparent for researchers (i.e. the algorithm behind the interface for collecting large datasets is a dark box). Our objective was to obtain a large sample of what we call genuine human accounts, we therefore established a set of filters to ensure that we only acquire material from certain ego types but exclude certain groups with unusual network sizes and interactional patterns. A case in point is for instance a celebrity who does not follow anyone in their circle but has two million followers. Other types that were excluded consists of organizations and businesses. Filtering was implemented using the following steps:

First, we limited the number of contacts (both friends and followers) to anywhere between 20 and 500. This ensured that we can exclude recently established accounts with little textual content, but also accounts with unusually large networks. Second, we only included accounts that had not been verified, aimed at excluding celebrities for instance. Third, to exclude travelers and tourists, the two geolocation information in the metadata needed to match; the profile location and the tweet location in the form of a GPS code needed to be identical. Fourth, we set limits to the number of messages per account to 3–12,000. In the initial rounds of data collection, a sizeable portion of the accounts had only one message, meaning that they were one-timers who had signed up for the service, but had discontinued using it. Alternatively, the initial dataset also contained a number of accounts with extremely high number of messages, suggesting these are automatically message-generating software robots. The upper limit of 12,000 messages per year means sending c. 33 messages per day. Lastly, we checked the account lists manually to exclude a few dozen accounts that had not been excluded by that point to ensure that all the accounts are operated by a named person, whether the name is real or not is beyond the scope here.

The data collection focused on accounts from the US and the UK, and it aimed at capturing ego networks from both inner cities in large metropolitan areas and from sub-urban and rural areas (see the next section). It resulted in a dataset of 3,935 ego networks (Table 1 below) with over 305 million messages. Altogether, these networks contain 233,774 other account holders (nodes), making this a massive dataset. The networks vary in size, the largest containing 439 accounts, and the mean is 60 nodes. If we compare these to networks used in prior ethnographic studies, these networks are larger than small networks that can be captured with manual methodology. The networks are closer in size to average human networks (Dunbar 2020) but fall below the average estimates mentioned above. Indeed, 75% of the networks have ≤85 nodes, which is understandable, given that people have off-line relationships to master as well.

The texts generated by those in the network are collected for each user and connected to unique identifiers and to the entire networks, so that we can extract either all messages sent by one user or all the messages in a network.

The dataset contains nearly 4.8 billion tokens from over two hundred thousand accounts. Is this a lot? The answer is both yes and not at all, since it has been estimated that an average English-speaker utters around 16,000 words a day (Mehl et al. 2007), and the dataset therefore represents material that is theoretically equivalent to nearly 700 years of spoken production of an individual. However, the dataset represents only slightly over one day of spoken production of 233,774 people (233,774*16,000=c. 3.7 bil.).

Equally important to sheer size is the richness in the dataset, since it includes the connections between all the accounts in a network. We have interactional informa-

	Egos	Messages	Nodes	Tokens	Mean length	Mean size	Median size
US urban	2,037	166,031,396	123,031	2,445,454,135	14.7	61	52
US sub/rural	958	71,882,986	56,453	1,163,547,607	16.2	59	52
UK urban	538	37,475,449	30,168	619,904,835	16.5	57	49
UK sub/rural	402	30,288,492	24,122	554,319,126	18.3	60	53
Total	3,935	305,678,323	233,774	4,783,225,703	15.6	60	52

Table 1. The basic frequencies in the social media datasets

tion of who is a friend with whom, whose messages are shared, replied, or quoted, and how many times an account interacts with the others. Altogether, there are nearly 1.8 million interactions in these networks.

What is done next in pre-processing the material involves using the interactional information to calculate network strength labels for each network. This algorithmic process has been presented in detail in Laitinen, Fatemi & Lundberg (2020) and Laitinen & Fatemi (2022). The process is multidimensional and takes into account six interactional properties in the networks. These properties are:

- Frequency of communication between nodes
- Network density (mean values and the spread of values of betweenness centrality)
- Connectedness of nodes in a network
- Distance between nodes (closeness centrality)
- Similarity of nodes when measured in the number of shared friends and followers

To establish network labels, we calculate the mean of the six values for each network. The process results in network scores (NS) that can theoretically range between 0 and 1. If the mean is 0 (practically impossible), it indicates a maximally loose-tie network, while 1 indexes a maximally close-knit network, where every node is connected to each other.

The results section contains two parts, the first of which looks into the outcome of the algorithmic model and an overview of the kinds of network our dataset contains. In Section 4.2, we take a closer look at how linguistic variation is conditioned by these networks. It presents a study that tests the weak-tie hypothesis in language change using data-intensive methods and large evidence from social media.

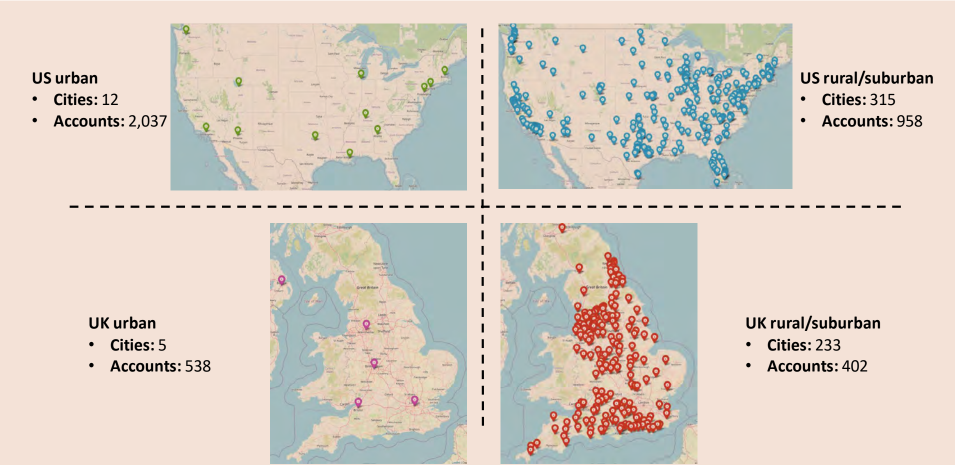


Figure 2. Geographic distribution of the ego networks

4. Results

4.1. What types of networks emerge from big data?

We collected data from both urban centers and sub-urban/rural to cover as much potential variation as possible in the ego networks. Given that the local administrative systems are slightly different, the work involved manually selecting a set of urban locations first (large urban counties in the US, such as the Dallas County, and metropolitan boroughs and counties in the UK, such as the metropolitan borough of Birmingham). Then, we expanded the collection to surrounding sub-urban and rural areas to cover as many geographic regions as possible. Figure 2 visualizes the locations, listing also the basic details for the number of locations and the ego networks (accounts) collected from them.

The second step consisted of collecting all the messages of the ego and the nodes in each network. In this step, we also collected all the interaction in each network and this information forms the basis for estimating how weak or strong a network is. The outcome of the algorithmic method is illustrated in Figure 3. The left side shows how the NS scores are spread in the total dataset, the mean being 0.47 and visualized in the blue vertical line. Most of the networks (90%) are centered between 0.37 and 0.61, and the observations are not normally distributed. Kolmogorov-Smirnov normality test results in p-value which fails to reject the null hypothesis (D=0.61, p<0.0001). This is also visually confirmed in the Q-Q-plot in the appendix. The visualization on the right illustrates the differences in the NS scores in the four subsets. We first observe the uneven number of observations, so that the US urban dataset with its 2,037 ego networks is by far the largest.

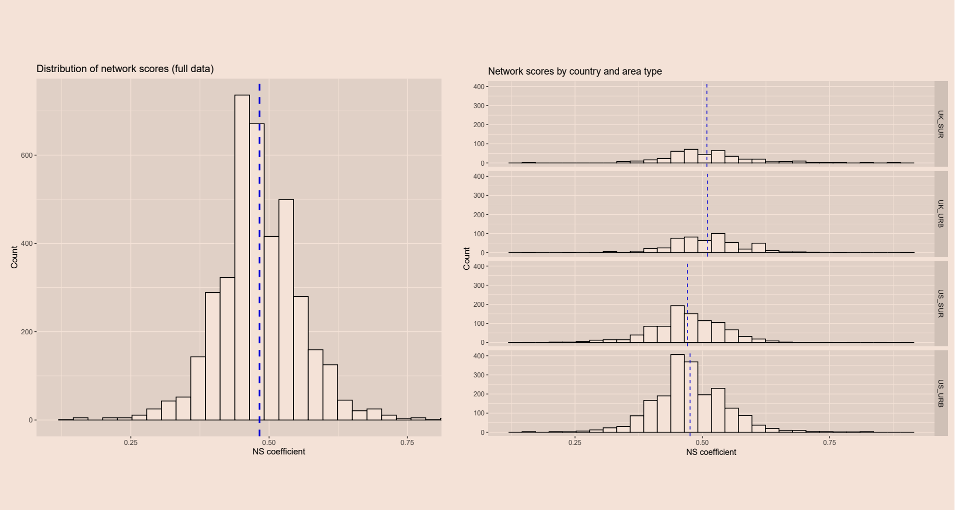


Figure 3. Distribution of the network scores in full data (left) and the US/UK subsets (right)

Another observation concerns the four geographic settings, which show similarities. Each of the settings resembles the others, so that the outcome of the method is nearly identical. There are slight differences between the areas, so that overall the US networks are slightly looser (mean is 0.47) than the networks in the UK dataset (0.51). Somewhat surprisingly in the UK data, the suburban/rural networks are slightly looser than the urban ones. Our initial assumption was that overall people in urban areas would have had looser digital networks in general, given that the data contain nearly four thousand networks, and our data filtering was designed to exclude unusually large networks.

Before moving on to the variables to be investigated in relation to the networks, Figure 4 shows a set of 500 randomly selected networks in relation to their size. It shows that there seems to be no correlation between the NS scores on the y-axis and the number of friends, so the methods of establishing the NS scores is independent of network size. All four subsets show a similar pattern.

Most importantly, the reason why so much space has been devoted to the method is that as a result of the method, we can now treat large social media data as something else than just a massive textual database. This is the focus in the next section, where the NS values are correlated with two dependent variables. They are contractions of negatives and verbs (n't instead of full not; won't instead of will not), and semi-modal NEED to + V-inf.

They are selected based on four criteria. First, the variables must be frequent enough to be studied in microblog messages (see Table 1 showing that the average length of a message is 15.6 tokens). Second, both of them have been shown to be undergoing considerable frequency increases in recent history of English. These changes are also driven by different forces, as contractions are related to what is known as colloquialization, the gradual shift of spoken norms into written language (Leech et al. 2009; Axelsson 1998). According to some estimates, the shift towards contracted forms has been "strongest in American English" (Mair

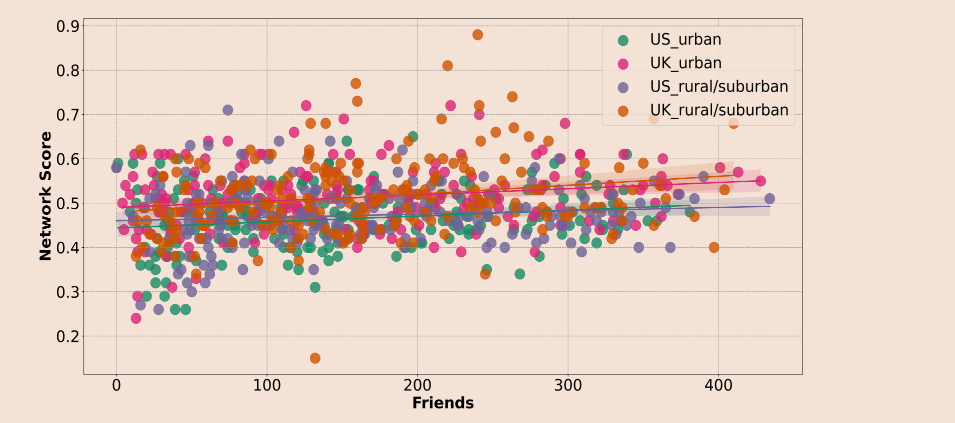


Figure 4. Correlations between NS values and network size

2006: 189). Changes in the semi-modal on the other hand are related to grammaticalization of full lexical items gradually gaining grammatical properties. In the case of NEED to V-inf, recent changes in functions and frequencies have been well documented (Nokkonen 2015; Daugs 2017), and it has been observed to be "spectacularly increasing" according to some estimates (Leech et al. 2009: 94). Thirdly, given that the variables also concern distinct linguistic categories, contractions of orthography and the semi-modals of grammar, it is assumed that they differ in terms of complexity in processing. Contractions are most likely above the level of linguistic awareness and might also be corrected or suggested by proof-reading and text-generating devices today, while NEED to + V-inf is more complex, and not known or directly analyzable by language users in fast-paced digital communication on social media. Lastly, the variable must be automatically retrievable with minimum false positives.

4.2. Networks and ongoing linguistic change

We first retrieve data from small networks to test if the weak tie hypothesis holds in digital networks. If it does, we should be observing statistically significant differences between weak-tie and strong-tie networks so that the normalized frequencies of the dependent variables should be higher in weak-tie environments than in strong-tie networks. We use a conservative threshold value of 60 nodes as the separator between small networks (≤60) and large (>60 nodes). Figure 5 visualizes these observations for both variables (NEED to V-inf on left, and contractions on the right). The y-axis shows normalized frequencies per 100,000 words, and outliers have been removed for simplicity.

The quantitative patters are clear. Evidence from social media, based on 1,704 small networks with tens of thousands of nodes, supports the prior observations from small data and from ethno-

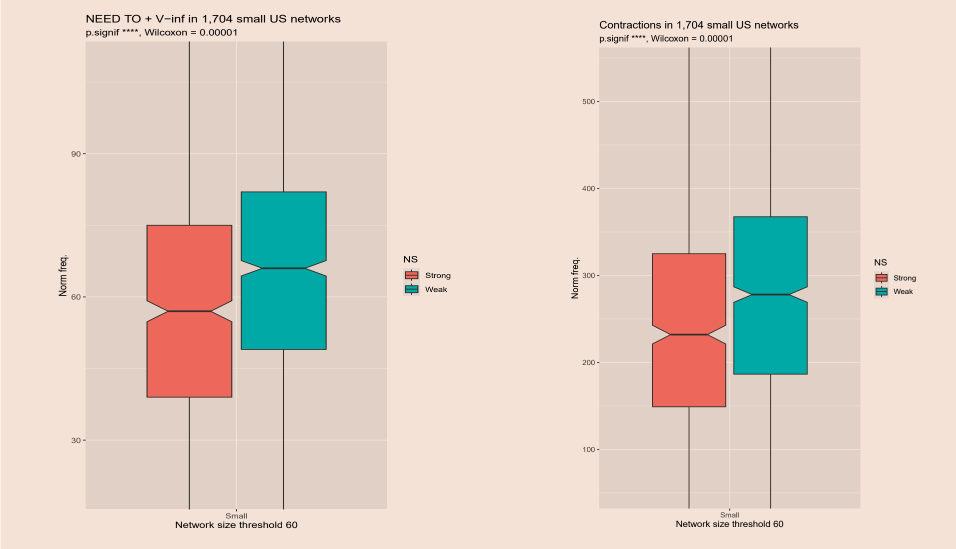


Figure 5. NEED to V-inf and contractions in 1,704 small networks (US)

graphic observations. The boxplots show weak-tie networks (green) and those of strong ties (in red), and the vertical line indicates the median values, which are higher in the weak-tie networks than in strong-tie settings. The notches indicate 95% confidence levels, and the customary practice is to interpret the two environments to be statistically significant if the notches in boxplots do not overlap. The differences between the two environments are also statistically significant when measured using the Wilcoxon rank sum test since the data are not normally distributed. The visualizations only show US networks, but the same patterns are observed in the UK data.

The conclusion is strong. When we investigate digital weak-tie environments, they seem to be favorable settings for incoming linguistic features, more so than strong-tie environments, which systematically show lower frequencies of the incoming variant forms.

Before moving on to comparing small networks with large networks in the data, the importance of these observations ought to be considered. The fact that it is possible to replicate the key finding in extremely large digital datasets strengthens the weak-tie hypothesis. In addition, the findings suggest that data from social media are surprisingly similar and behave similarly to observational data from other settings that have been investigated in previous studies. As pointed out earlier, various studies have suggested that the size and structure on online networks closely resemble offline settings (Gonçalves et al. 2011; MacCarron, Kaski, Dunbar 2016), and these observations suggest that the same applies to networks as conditioning elements in innovation diffusion, though it is clear that more evidence from a larger set of linguistic variables are needed.

The pattern changes when large networks of over 60 nodes are added and compared alongside small networks. Figure 6 shows the total results showing the frequencies of NEED to + V-inf from nearly 4.8 billion words of material from both the UK and US. These observations from 233,774 users suggest that the difference between weak- and strong-tie environments disappears when large networks are observed.

In small networks (on the right), the difference between weak- and strong-tie

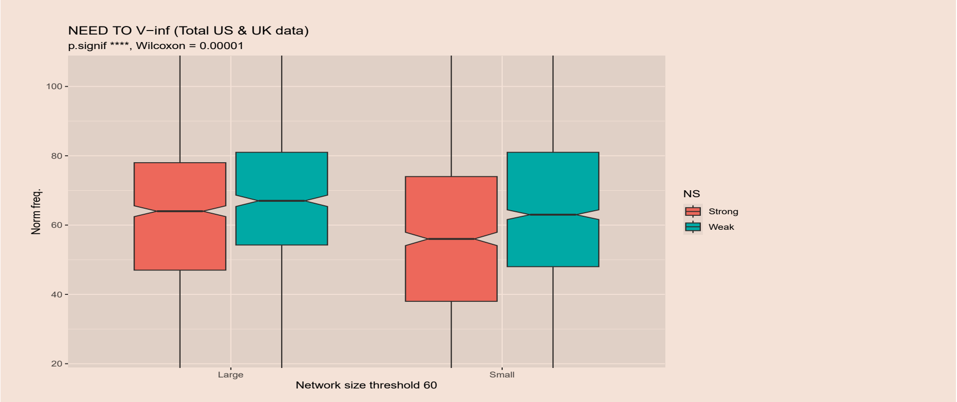


Figure 6. Frequencies of NEED to V-inf in total US and UK data

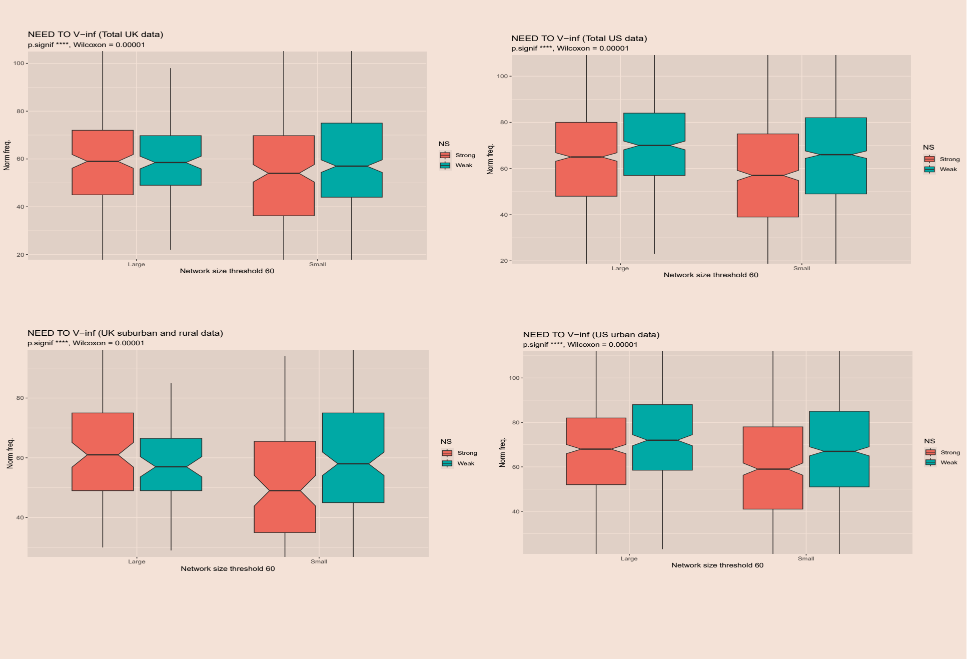


Figure 7. NEED to V-inf in four subsets of network data

networks is considerable, but the difference levels in large networks. The weak-tie environments in both subsets exhibit greater frequencies on the incoming semi-modal NEED to + V-inf, but this difference levels in large networks, as the notches on the left (large networks) overlap. The observed differences between network sizes are statistically highly significant.

Figure 7 zooms into the various subsets to explore if the US and UK settings diverge and examines if the urban contexts are different from suburban and rural. It shows the results in four settings, the top left visualizes the total UK data, which confirms the pattern (and the working hypothesis) of leveling the role of network strength in large networks. The top right shows the total US data in which the pattern is not as pronounced as in the total UK data, but the trend towards leveling the impact of network ties in large networks is visible.

The bottom half reveals further differences between the two countries. The role of large strong-tie networks seems to be more pronounced in the UK dataset (bottom left visualizes the suburban and rural UK data), but leveling the network strength impact takes place also in the US data (bottom right shows the US urban results), so that the patterns observed in small networks are not are considerable in large networks. The notches in large US urban networks overlap, suggesting that there are not statistically significant differences between large networks.

It is, at this stage, a puzzle why the two geographic settings behave differently, and why the UK results are clearly more advanced in terms of leveling the network

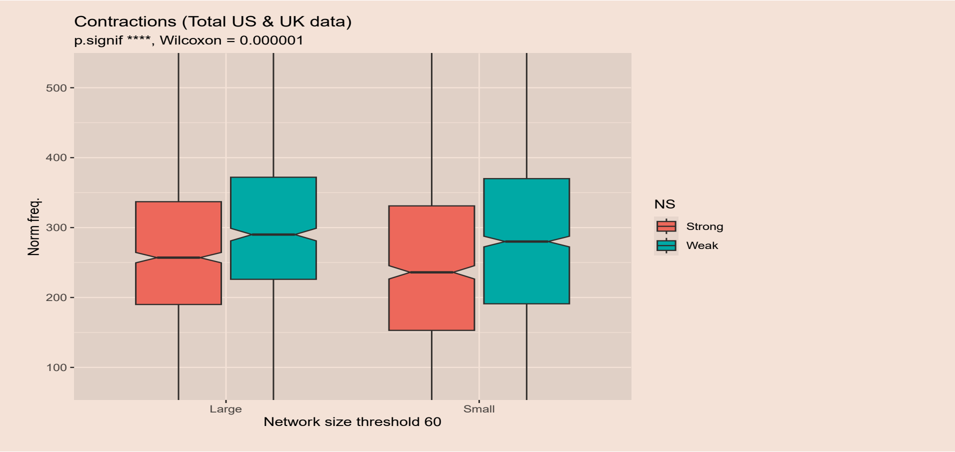


Figure 8. Frequencies of contractions in total US and UK data

strength (the same pattern is observable in the UK urban data). One possible explanation is that the frequency increase of NEED to V-inf has been more considerable in British English (Mair 2015; Leech et al. 2009: Table A5.1), in which the form increased faster than in American English. According to Mair (2015), NEED to V-inf underwent a four-fold increase in British English in from the 1930's until 1990s, while in American English it started to increase only in the 1970s (Daugs 2017). Since we do not (yet) have any other background information of the account holders, it is not possible to claim anything else on the potential role of diachronic depth of the phenomenon. The need for enriching social media with more background variables is discussed in the conclusion section.

The results concerning contractions are presented in Figure 8. The situation is not fully identical in that the difference between small and large networks persists. The difference between weak- and strong-ties that is observable in small networks remains when the network size is increased. This was tested by increasing the threshold to 70, but the same pattern is visible.

When different settings are observed separately, differences start to emerge, and these differences are similar to those observed with the semi-modals in Figure 7 above. That is, the UK contexts show remarkable similarities in which small and large networks are different so that the differences level. If we focus on total UK data (top left in Figure 9), the frequency of contractions in weak-tie networks is conditioned by network type, but in large networks the differences disappear.

Similar to the semi-modals, the leveling takes place in UK settings (bottom left) but persists in the US data. Again, increasing the threshold to 70 does not change the overall result.

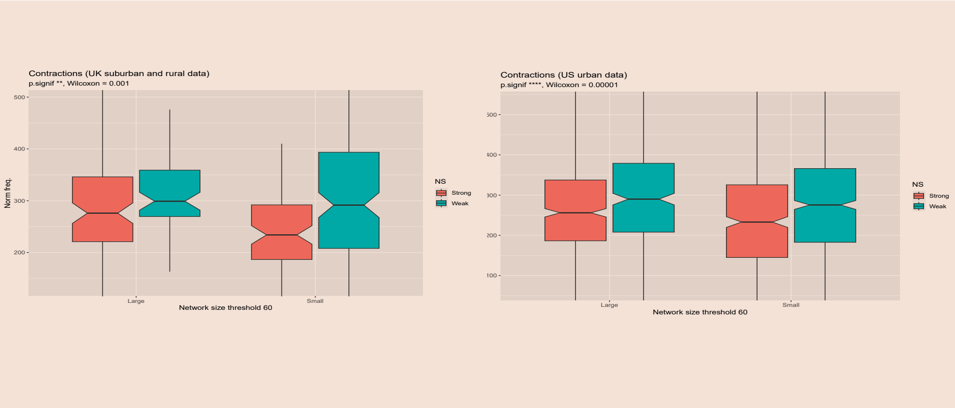


Figure 9. Contractions in four subsets of network data

5. Discussion and outlook

This article has aimed at illustrating the potential of using large-scale social media data and data-intensive methods for SSH research, and particularly for variationist sociolinguistics. The vantage point has been narrow, given the fact that even one subdiscipline of linguistics, i.e. language variation and change is a large field, and it is possible to tackle only a limited number of topics in one study. Therefore, the approach presented here ought to be viewed as programmatic, highlighting the potential of social media rather than being merely a single case study. Given the size and scope of social media applications today, data obtained from them have the potential of providing novel answers to a range of questions in many fields.

When we first zoom into the results here, it is clear that we can access a far greater variety of networks using data-intensive methods than using manual methods. Here, the data set presented aims at correcting the state of affairs that a large part of naturally occurring networks is simply left out of the analysis if researchers rely solely on manual methods. As pointed out in section 4.1., half of the networks used here are larger than 50 nodes, which is generally held to be the upper limit for manually collected networks, at least when participant observation methods are used. It is possible that the ways in which we collected the dataset has influenced the average size of our data, and had we used looser filtering criteria, it is possible that the networks would have been even larger. However, the outcome of the data collection process and the subsequent analysis highlight that digital networks are highly similar to offline networks and ought not to be treated as something completely different (cf. Dunbar et al. 2015).

When it comes to the findings, the results clearly indicate that network size ought to be subjected to further analysis. The results suggest, though not across all settings, that network size plays a substantial role and is therefore a factor to be accounted for in studies that use the weak-tie hypothesis in sociolinguistics.

The main sociolinguistic result here is that the difference between weak- and strong-tie environments levels for large networks. The finding is closely similar to our initial findings (Laitinen et al. 2018; Laitinen et al. 2020 both of which use much smaller data, and an earlier version of the methodology) and to the findings in other fields of network science (cf. the study by Ma et al. 2019 on trust in Facebook networks).

The observations presented here rely on data from one social media platform, Twitter/X. It has its limitations, but also potential because it is what is known as a directed graph network, where it is possible to observe individual behavior and to know how they actually communicate in their networks. The situation is less ideal in more text-based applications, such as Reddit and Discord for instance. They have other benefits when they offer access to even larger textual data, but they lack the interactional part, so that is it impossible to know how individuals are connected to each other and what types of networks they form.

Before moving on, it is important to highlight that while social media offers substantial potential, it also contains a number of caveats. The first one is that not all applications offer suitable data for research. Being suitable means that such data should contain both user-generated communication and associated metadata. Secondly, it is widely known that certain demographic groups are over-represented on some applications, i.e. young urban males and public figures for instance are over-represented in Twitter/X for instance. Third, social media data are big data, meaning that they are high volume and high velocity material. This means that data handling requires both technical expertise and digital storage capabilities that are not necessarily familiar to SSH researchers. Using social media requires, more often than not, interdisciplinary collaboration in collecting, preprocessing and analyzing primary data. In all, using social media in variationist research means adopting novel ways of working and acquiring competences that are not part of traditional training in the humanities, such as data-intensive methods.

So what can be done to encourage more comprehensive use of large social media and data-intensive methods in SSH in future? The first is step is to ensure that researchers can have access to large-scale social media data. Various large datasets, such as the one used here, are a good start, but building them has recently become more and more difficult. This is related to the fact that much of the work on born-digital data relies on big tech corporations and their APIs. Data availability is increasingly becoming an issue as recent years have witnessed the closing of APIs through which data can be collected (cf. Bruns 2019 on the so-called APIcalypse). Another problem is related to how APIs actually function when extremely large datasets are collected through them. As is widely known, algorithms used by big tech may sometimes be computational dark boxes par excellence, and any data retrieved through them therefore contain considerable uncertainties (NAS, 2019). These uncertainties are related to both data availability in that researchers do not know if such data represent the population in the first place and to changing algorithms in which case a slight change might lead to slight changes in the data collected. Both scenarios clearly limit the potential of large-scale social media in research.

One solution related to problems in data availability of the need for systematically collected and representative benchmark collections, such as the one used here. These sociolinguistic data constants should act as yardsticks that would enable comparisons for any data retrieved through an API. It goes without saying that building representative benchmarks needs to be done carefully, but they act as essential pre-requisites for any activity that aims at ensuring high quality research.

In addition, closely connected with the topic of representative benchmarks is an issue of cross-disciplinary expertise in collecting, handling and using large-scale data from any source. Because of their size and structure, fully utilizing high volume and velocity social media data for instance requires advanced technical competencies that extend beyond competences typically required from SSH researchers. The solution in this study has been to rely on a large nationally funded digital humanities infrastructure that has provided the necessary support functions. In more general terms, such digital research infrastructures are vital for renewing SSH research. They could also contribute positively to reuse of data to ensure comparability and further methodological development. Such infrastructures could not only ensure that researchers would be able to reuse datasets and to replicate empirical studies, but they could also reduce duplicated effort in the development of tools to manage and process these datasets for research use. A concrete issue concerns, for instance, the use of supercomputing facilities and support infrastructures could act as mediators that lower the threshold for using such services in SSH.

One concrete issue related to infrastructures is, for instance, data enrichment, which was done here in the form of adding the NS scores to each ego network. This type of data enrichment could be effectively achieved in interdisciplinary environments. It is widely known that born-digital data from social media are not particularly rich in socio-demographic information, as they contain a limited set of background variables (Laitinen & Fatemi, 2022). This is because of both ethical reasons and for proprietary factors, as social information connected with massive user-generated data is extremely valuable for private enterprises. Moreover, when user-related information is available, it is often self-reported, and thus prone to inaccuracies. In future, we need new ethically sustainable and research-based ways of enriching social media data for research. It would be ideal if in future, we could not only search for certain types of networks using the network parameter, as was done here, but to be able to find ego networks of individuals of certain ages and social layers for example. Adding such social background information could substantially increase empirical validity of studies that use social media data and might also lead to completely new openings at least in variationist sociolingustics.

Lastly, academic institutions, science academies, university networks and European university alliances should, together with decisions makers, ensure sufficient political pressure on big tech and safeguard that use of social media data is in the first place possible. We need sustainable ways of having the best and the most comprehensive data available for fundamental research, and legislative measures in place to guarantee sufficient transparency of what kinds of data are available and in what ways. Such activities extend beyond a single researcher or even a single university, and are instead a matter of governmental-level pressure on the largest digital platform companies that operate as so-called gatekeepers in the internal market. This is in principle guaranteed in the recent European Union regulation (2022/1925) of EU's Digital Markets Act, which came into effect in autumn 2022. It defines digital gatekeepers to be companies that provide "core platform services", such as digital social networking applications or online intermediation services. This legislation acknowledges that gatekeepers currently directly benefit from having access to extremely large amounts of data that they collect as part of their service, and it decrees that these gatekeepers should not undermine the innovation potential of third parties to develop their own services. To what extent this legislation is successful remains to be seen, but at least this legislation contains the basic building blocks to ensure more effective use of social media data in fundamental research. One of its objectives is to ensure that data should be accessible in a format that makes it possible for third parties, such as research institutions, to utilize them.

Various social media data and data-intensive methodologies have substantial potential in sociolinguistics, and while their future importance should not be overestimated, it could be argued that we have only seen the beginning of how they could be used. These first uses have been by researchers who have been able to utilize technical tools for data access and analysis, but similar to the use of text corpora, their full potential will only be measured when their use as primary data sources becomes mainstream.

The Authors

Mikko Laitinen

Mikko Laitinen is Professor of English Language at the University of Eastern Finland in Joensuu. His research focuses on computational and data-intensive digital methods and on the social network theory in language variation and change in English.

Masoud Fatemi

Masoud Fatemi is a doctoral candidate in Computer Science and English linguistics at the University of Eastern Finland and Linnaeus University in Sweden. His research interests lie in machine learning, network analysis, and natural language processing.

References

Axelsson, M. W. (1998). *Contraction in British Newspapers in the Late 20th Century*. Acta Universitatis Upsaliensis.

Bruns, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, 22(11), 1544–1566, doi: 10.1080/1369118X.2019.1637447

Daugs, R. (2017). On the development of modals and semi modals in American English in the 19th and 20th centuries. In Turo Hiltunen, Joe McVeigh & Tanja Säily (eds.), *Big and Rich Data in English Corpus Linguistics: Methods and Explorations.* (Studies in Variation, Contacts and Change in English, vol. 19). https://varieng.helsinki.fi/series/volumes/19/daugs/ (accessed 15 Oct 2023).

Dunbar, R. (2023). The social brain hypothesis thirty years on: Some philosophical pitfalls of deconstructing Dunbar's number. *Annales Academiae Scientiarum Fennicae* 1/2023, 9–27.

Dunbar, R. (2020). Structure and function in human and primate social networks: Implications for diffusion, network stability and health. *Proceedings of Royal Society A. London.* 476A. doi: 10.1098/rspa.2020.0446.

Dunbar, R. (1992). Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22, 6, 469–493.

Dunbar, R., Arnaboldi, V., Conti, M. & Passarella, A. (2015). The structure of online social networks mirrors those in the offline world. *Social Networks* 43, 39–47. doi: 10.1016/j.socnet.2015.04.005

Gonçalves, B., Perra, N. & Vespignani, A. (2011). Modeling users' activity on Twitter networks: Validation of Dunbar's Number. *PLoS ONE* 6:8: e22656. doi: 10.1371/journal.pone.0022656

Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology* 78:6, 1360–1380.

Granovetter, M. (1983). The strength of weak ties: A network theory revisited. *Sociological Theory* 1, 201–233.

Grieve, J., Nini, A. & Guo, D. (2016). Analyzing lexical emergence in Modern American English online. *English Language and Linguistics* 21:1, 99–127. doi: 10.1017/S1360674316000113

Grieve, J., Nini, A. & Guo, D. (2018). Mapping Lexical Innovation on American Social Media. *Journal of English Linguistics* 46:4, 293–319. doi: 10.1177/0075424218793191

Groves, R., & Couper, M. (1998). *Nonresponse in Household Interview Surveys.* Wiley-Interscience.

Huang, Y., Guo, D., Kasakoff, A. & Grieve, J. (2016). Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems* 59, 244 255 doi: 10.1016/j.compenvurbsys.2015.12.003

Laitinen, M. & Fatemi, M. (2022). Big and rich social networks in computational sociolinguistics. In Paula Rautionaho, Hanna Parviainen, Mark Kaunisto & Arja Nurmi (eds.), *Social and Regional Variation in World Englishes: Local and Global Perspectives,* 166–189. Routledge. doi: 10.4324/9781003227342-9

Laitinen, M., Fatemi, M. & Lundberg, J. (2020). Size matters: Digital social networks and language change. *Frontiers in Artificial Intelligence*, 3:46. doi: 10.3389/frai.

Laitinen, M. & Lundberg, J. (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 University Press.

Leech, G., Hundt, M. Mair, C. & Smith, N. (2009). *Change in Contemporary English: A Grammatical Study*. Cambridge University Press.

Ma, X., Cheng, J., Iyer, S. & Naaman, M. (2019). When do people trust their social groups? In *CHI Conf. on Human Factors in Comp. Systems Proc*. (CHI 2019), May 4–9, 2019, Glasgow. ACM. doi: 10.1145/3290605.3300297

MacCarron, P., Kaski, K. & Dunbar, R. (2016). Calling Dunbar's numbers. *Social Networks* 47, 151–155, doi: 10.1016/j.socnet.2016.06.003

Mair, C. (2015). Cross-variety diachronic drifts and ephemeral regional contrasts: An analysis of modality in the extended Brown family of corpora and what it can tell us about the New Englishes. In Peter Collins (ed.), *Grammatical Change in English World-wide*, 119–146. John Benjamins.

Mair, C. (2006). *Twentieth-Century English: History, Variation, and Standardization*. Cambridge University Press.

McCarty, C., Killworth, P. D., Bernard, H. R., Johnsen, E. C. & Shelley, G. A. (2001). Comparing two methods for estimating network size. *Human Organization* 60:1, 28–39.

Milroy, J. & Milroy, L. (1985). Linguistic change, social networks and speaker innovation. *Journal of Linguistics* 21, 339–384.

Milroy, J. & Milroy, L. (1978). Belfast: change and variation in an urban vernacular. In P. Trudgill, *Sociolinguistic Patterns in British English*, 19–36. Edward Arnold.

Milroy, L. & Milroy, J. (1992). Social network and social class: Toward an integrated sociolinguistic model. *Language in Society*, 21, 1–26.

Milroy, L. (2004). Social networks. In J.K. Chambers, P. Trudgill & N. Schilling-Estes (eds.), *The Handbook of Language Variation and Change*, 549–570. Blackwell.

Milroy, L. (1987). *Language Change and Social Networks*. 2nd edition. Blackwell.

Mehl, M., Vazire, S., Ramírez-Esparza, N., Slatcher, R. B. & Pennebaker, J. W. (2007). Are Women Really More Talkative Than Men? *Science* 317: 5834, 82. doi: 10.1126/science.1139940

NAS. (2019). *Reproducibility and Replicability in Science*. National Academies Press.

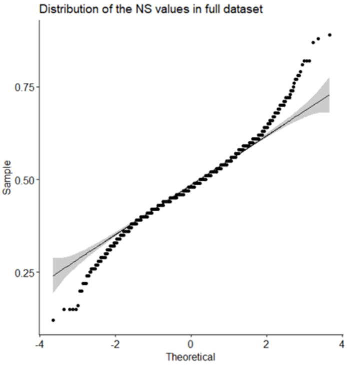
Rissanen, M. (2009). Corpus linguistics and historical linguistics. In A. Lüdeling & M. Kytö (eds.), *Corpus Linguistics: An International Handbook*, 53–68. Mouton.

Statista. 2023. Number of worldwide social media users. Available at https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/. (accessed 19 Aug. 2023).

Szmrecsanyi, B. (2017). Variationist sociolinguistics and corpus-based variationist linguistic: overlap and cross-pollination potential. *Canadian Journal of Linguistics*. 62:4, 685–701. Doi: 10.1017/cnj.2017.34.

Tyrkkö, J., Levin, M. & Laitinen, M. (2021). Actually in Nordic tweets. *World Englishes*, 40(4), 631–649. doi: 10.1111/weng.12545

Appendix:



Appendix 1. Q-Q plot of the distribution of the NS scores in the data.

SUOMALAINEN TIEDEAKATEMIA
FINNISH ACADEMY OF SCIENCE AND LETTERS
ACADEMIA SCIENTIARUM FENNICA