Social networks of lexical innovation

Investigating the diffusion of neologisms on Twitter

Quirin Würschinger
LMU Munich

15th March 2020

Contents

List of Figures

List of Tables

1 Introduction

The coronavirus has recently spread with shocking speed and has tragically affected the lives of people around the world. Its fatal consequences have demonstrated the devastating power of exponential diffusion in social networks. A leading research group of biologists analysing the contagion of Covid-19 have shown that the virus spreads supra-exponentially and that it resembles the diffusion of cultural and linguistic innovations such as internet memes. (editorial staff, 2020) Does this confirm the popular perception that certain cultural and linguistic innovations 'go viral'?

state gist of the paper

Societies constantly evolve, new products and practices emerge and speakers continually invent and adopt new words which diffuse through social networks of communicative interaction. Influential models in sociolinguistics like the S-curve model (J. Milroy, 1992) share fundamental features with economic models (Rogers, 1962) of diffusion and show commonalities between the spread of cultural and linguistic innovations. These models assume that the diffusion of innovations across social networks follows universal trajectories and that rates of spread depend on sociolinguistic dynamics such as network density and the presence or absence of weak ties (Granovetter, 1977). Unlike research on biological and cultural diffusion processes, sociolinguistic research has only fairly recently been provided with data sources that are equally suitable for large-scale, data-based approaches that use social network analyses to study these phenomena empirically.

Social media platforms like Twitter have changed the way we communicate and interact and how information spreads and they offer big amounts of data for empirical research. Sociological research has recently been concerned with pressing issues regarding the impact of online social networks for the spread of hate speech, fake news and the power of 'influencers', bots and institutions on public opinions and elections, which increasingly strain the social fabric. For (socio-)linguists, social media data provide huge amounts of authentic text along with information about speakers and their interactions which opens up new possibilities for the empirical study of language variation and change as well as the diffusion of linguistic innovations.

advantages of going beyond frequency

The size of these data as well as their informal nature allow corpus studies on the use and spread of new words which are often propagated via social media channels. In contrast to modern web corpora, which share these features, the availability of metadata about speakers and their interactions has been particularly important for the advancement of sociolinguistic research. Recent sociolinguistic work on lexical innovation has, for example, used data from social media users to gain insights about general trajectories (Nini, Corradini, Guo & Grieve, 2017) and geographical patterns (Eisenstein, O'Connor, Smith & Xing, 2014; Grieve, 2017, 2018) of diffusion and about factors that influence whether new words spread successfully (Grieve, forthcoming).

Social network analysis analyses interactional patterns between language users and allows to zoom in on the sociolinguistic dynamics of diffusion to see, for examples, whether the adoption of new words remains limited to closely connected sub-communities, whether they reach bigger parts of the speech community and whether certain individuals or

groups influence this process. While the study of social networks has had a long research tradition in sociolinguistics and has shaped current models of diffusion (e.g. J. Milroy and Milroy, 1985), new data sources have only fairly made it feasible to conduct large-scale empirical analyses using this approach. Advances in data and methodology of this kind put computational sociolinguistics in an excellent position to tackle old research questions in new ways, to test long-standing theoretical models empirically and to explore new questions and insights about language and society.

Social network analyses of social media data have been successfully employed in diverse fields, for example to study the spread of diseases (Lu et al., 2018), opinions (West, Paskov, Leskovec & Potts, 2014) and political attitudes (Pew Research Center, 2019). So far the application of similar approaches to the study of lexical innovation remains scarce, although recent work shows promising results in testing mathematical models for the diffusion of new words in social networks. (Goel et al., 2016)

This paper aims to investigate the spread of new words on the social media platform Twitter, using social network analysis to zoom in on the sociolinguistic dynamics of diffusion.

ask some concrete questions: e.g. 'Who uses the word *ghosting*?'

2 Theoretical background

How do new words diffuse and become conventional lexical items in a language system?

2.1 Research perspectives

A substantial body of linguistic research has tackled this question from different **perspectives**. (Schmid, 2016, p. 16)

From a **structural** perspective, main areas of interest include which word-formation processes are involved in forming new words, whether they are formally and semantically transparent, whether they show variation and change in the process of lexicalization and which status the resulting neologisms have in the language system (institutionalization). (e.g. Bauer, 1983; Lipka, 2005

Cognitive perspectives focus on how individuals process and store lexical innovations. Speakers generally use new words when they experience a communicative need to talk about entities or practices that cannot be expressed by their language's inventory of conventional words yet. In order for neologisms to successfully diffuse, speakers need to successfully negotiate their meaning (co-semiosis) in discourse, others need to adopt the behaviour of using these words (co-adaption). Continued exposure and use of new words can then lead to the entrenchment of new words in the mental lexicon of speakers. (Schmid, 2008)

Sociolinguistic perspectives transcend the level of the individual to study the diffusion of new words across speakers. The diffusion of lexical innovations is commonly seen as successful when the majority of the speech community has accepted a new word as a conventional lexical unit that can and is being used in communicative practice.

2.2 The S-curve model

S-curve models of linguistic change (Labov, 2007; J. Milroy, 1992; Nevalainen, 2015) assume universal sociolinguistic dynamics for the diffusion of linguistic innovations.

- The **trajectory** of spread is expected to follow an S-curve shape, with low rates of diffusion in early stages, followed by a period of accelerating spread with a tipping point at the mid point in the diffusion curve after which diffusion slows down and the curve flattens towards the end of the diffusion process.
- These temporal trajectories are assumed to correspond to the **sociolinguistic dynamics** of which individuals and groups interact with each other and adopt the target innovation.
  - In the **first stage** of slow diffusion only a small number of early adopters take up the innovative words. The individuals who use the new word typically form dense networks connected by strong ties. The structure of tight-knit communities of potentially like-minded individuals of similar backgrounds facilitates the successful negotiation of meaning (co-semiosis) of new words. High rates of interactions in these communities leads to high rates of exposure for individuals, which fosters co-adaption, entrenchment and usualization of new words in these communities.
  - In cases of successful diffusion the initial stages are followed by an **acceleration in spread** when new words increasingly reach speakers outside these tight-knit communities via weak ties (Granovetter, 1977). Rates of diffusion increase substantially when speakers that are not part of the initial group of early adopters start to accomodate the new words, allowing the innovations to reach a broader spectrum of the speech community.
  - In **later stages**, rates of diffusion slow down again as the majority of the speech community has already adopted the new words while smaller pockets speakers remain reluctant to take up the new words.
- S-curve models have mainly been applied to the **linguistic domains** of phonology and syntax. Fundamental differences between lexemes and linguistic items on other levels such as phonemes and grammatical constructions might affect whether the validity and reliability of such models for *lexical* innovation.
  - For example, grammatical constructions such as the *going to* future used to express a speaker's future intention serve to fulfill relatively abstract communicative needs that remain stable over time. By contrast, on the lexical level, linguistic innovations are typically tied to concrete cultural referents such as products and practices whose conceptual relevance is much more volatile over time. For example, many lexical innovations such as *millennium bug* denoting the fear of a computer crash at the beginning of the new millennium can show high rates of diffusion and become entrenched and conventional among the majority of the speech community. Without continual conceptual relevance in public discourse, however, these words fail pass on to the next generation

of speakers. S-curves are commonly assumed to be found when linguistic innovations compete for **'semantic carrying capacity'** (Nini et al., 2017), however, in many if not most cases of *lexical* innovation the conceptual carrying capacity is far from stable over time which represents a critical deviation from the traditional assumptions behind S-curves in language change.

– However, the generally strong theoretical and empirical basis of the S-curve model for language innovation and change, also from studies on the diffusion of cultural innovations (Rogers, 1962), and the precise formulation of the sociolinguistic dynamics underlying different phases of diffusion still make it an attractive **blueprint** for the empirical study of the sociolinguistic diffusion of lexical innovations.

2.3 Empirical approaches

Empirical approaches studying the diffusion of lexical innovations have only recently become feasible with the advent of new data sources and computational methods.

Earlier work had to rely on **traditional linguistic corpora**. Due to the low-frequency nature of neologisms general linguistic corpora do not allow to study broad ranges of neologisms and thus pose limits to making broad and robust generalizations about the nature lexical innovation. Despite these limitations, case studies on selected neologisms (Hohenhaus, 2006) and studies on specific fiels of neology (Elsen, 2004) managed to shed light on the spread of new words in more specific domains.

The advent of **web corpora** in the 21st century provided researchers with bigger and less formal data to study lexical innovation.

The sheer size of big corpora → bigger samples

monitoring corpora (Davies, 2013)

the nature of web corpus data is particularly suitable for investigating lexical innovations as language on the web is very creative, more informal sources, bigger spectrum of language use new words often first occur on the web and use on the web significantly influences whether these new formations catch on or not.

In particular, a range of tools enabled the creation of specialized corpora for the investigation of neologisms. (Renouf, Kehoe & Banerjee, 2006; Kerremans, Stegmayr & Schmid, 2012; Lemnitzer, n.d.; Gérard, 2017; Cartier, 2017)

social media corpora Grieve, Nini and Guo, 2016; Eisenstein et al., 2014 size nature creative, hotbed authentic language use driving force social network information users community characteristics influencers

2.4 Current framework: the EC-Model (Schmid 2020)

The EC-Model integrates both structural, cognitive and sociolinguistic perspectives on language structure, variation and change.

Entrenchment: individuals

Conventionalization: society, sociolinguistics → focus of this paper

Usualization

Diffusion
definition

2.5 Operationalization

- going beyond frequency
- sociolinguistic dynamics
  - number of users
  - social network characteristics
  - influencers

3 Data and methods

3.1 Data

- Twitter
  - demographics
  - anatomy of a tweet
- corpus
  - longitudinal: retrospective, starting from early attestations
  - big data
  - social network information
- sample
  - basis: bottom-up selection by NeoCrawler (Kerremans & Prokic, 2018, 2)
  - covering candidates from clusters found in earlier empirical work (Kerremans, 2015)
    - no diffusion
    - topical
    - recurrent
    - advanced
  - extension
    - reasonably successful: e.g. technical innovations like *blockchain*
    - sociolinguistically interesting: e.g. political terms such as *covfefe*

3.2 Method: social network analysis

- basis for networks: interactions between users
  - mentions

  - retweets
- network structure
  - nodes: users
  - edges: interactions

4 Sample

4.1 Neologism candidates

- previous categorization (Kerremans, 2015)
  - no diffusion
  - topical
  - recurrent
  - advanced
- dimensions
  - overall degree of diffusion (synchronic): successful vs. unsuccessful: usage frequency, degree centralization
    - no success
    - limited
    - advanced
  - temporal dynamics of diffusion (diachronic)
  - stability: stable vs. topical coefficient of variation
  - trend
  - diffusion: increasing degree of diffusion
  - centralization: decreasing degree of diffusion
  - speed

4.2 Case studies

- criteria
  - covering clusters of neolgism candidates
  - frequency counts comparable
- cases (Kerremans, 2015)
  - no diffusion: *microflat*
  - limited
    - topical: *poppygate*
    - centralized: *alt-left*

        * decreasing: *solopreneur*
    - advanced diffusion:
        * advanced: *upcycling*
        * increasing: *hyperlocal*

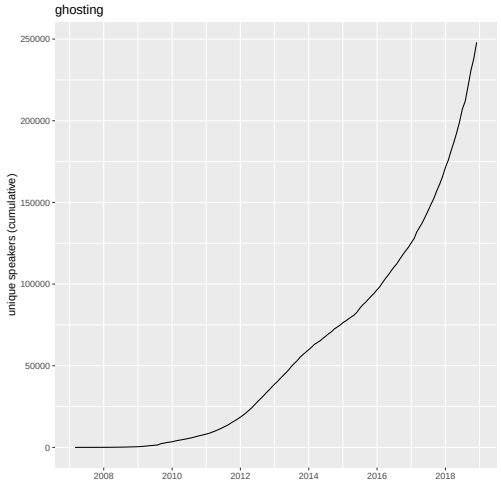
5 Frequency

- Frequency is usually taken as the measure to approximate the degree of conventionality and entrenchment of linguistic entities.
- There are a number of underlying assumptions.
    - speech community: frequency = conventionality
        * high freq. = majority of speakers show entrenchment
    - individuals: exposure → entrenchment
- problems
    - temporal dynamics (e.g. *millenium bug*)
    - output != input (Stefanowitsch & Flach, 2017)
    - high freq != many speakers
    - many speakers != 'majority of the speech community'

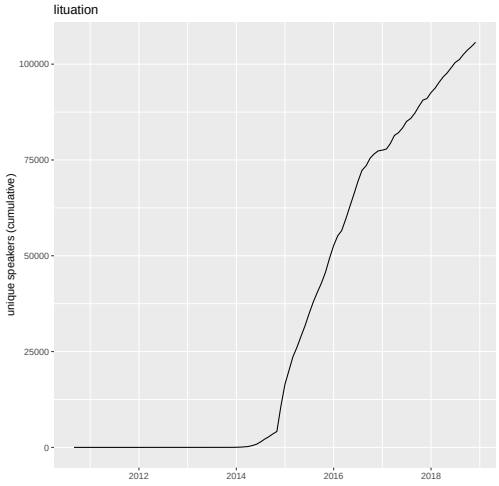
5.1 Cumulative counts

Figure 1: Cumulative frequency counts for case study words.

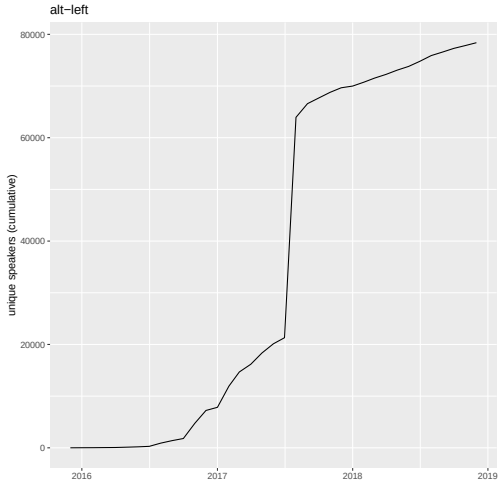
(a) *ghosting*



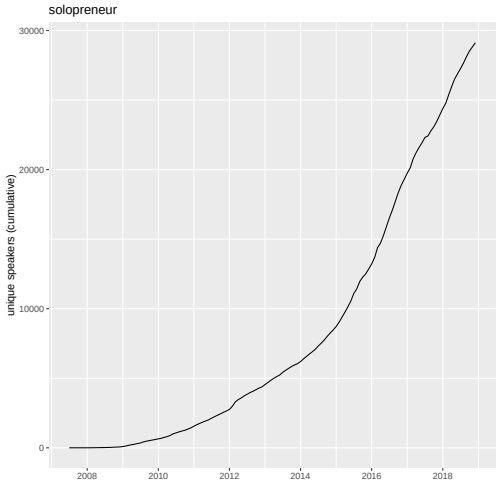
(b) *lituation*



(c) *alt-left*



(d) *solopreneur*



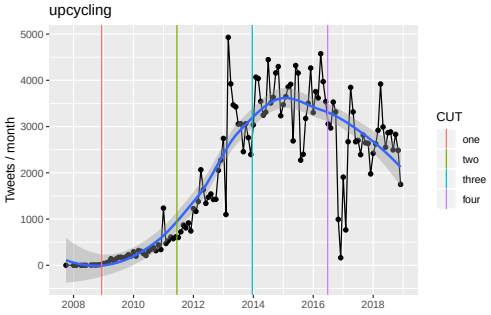
5.2 Temporal dynamics

5.2.1 Coefficient of variation

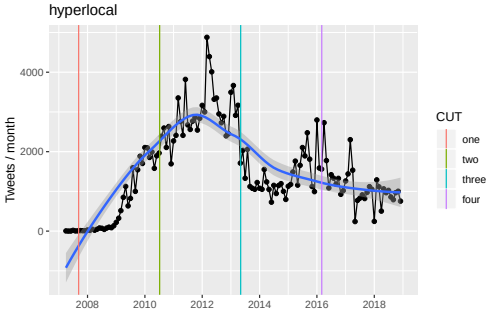
- most volatile candidates: *poppygate*, *burkini* etc.
- will be excluded from in-depth analyses

Figure 3: Temporal dynamics in usage frequency for case studies.

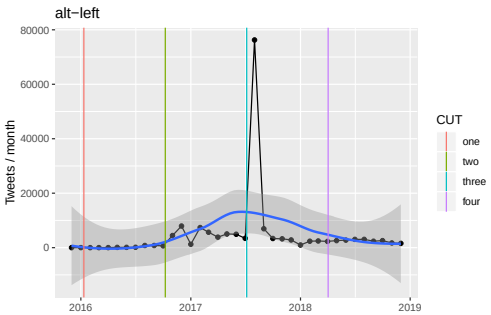
(a) *upcycling*



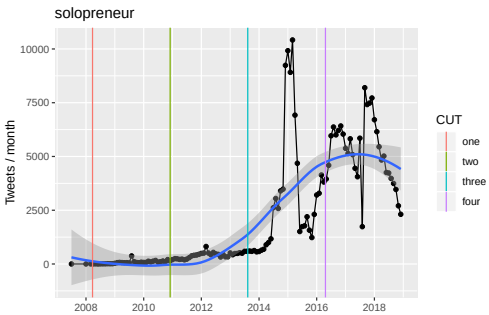
(b) *hyperlocal*



(c) *alt-left*



(d) *solopreneur*



6 Social network analysis

I will go beyond frequency and look into the sociolinguistic dynamics more closely

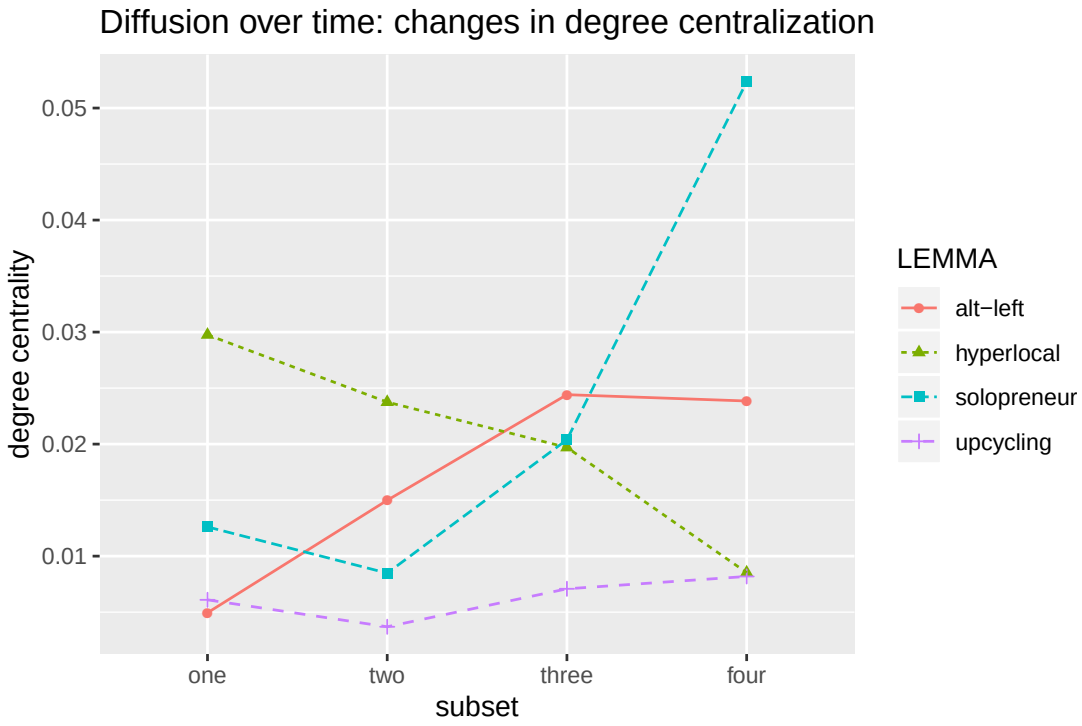
- sociolinguistic dynamics of diffusion over time
- sociolinguistic conventionality status of neologism

6.1 Speaker counts

6.2 Centralization over time

6.2.1 Overview of changes in centralization for case studies.

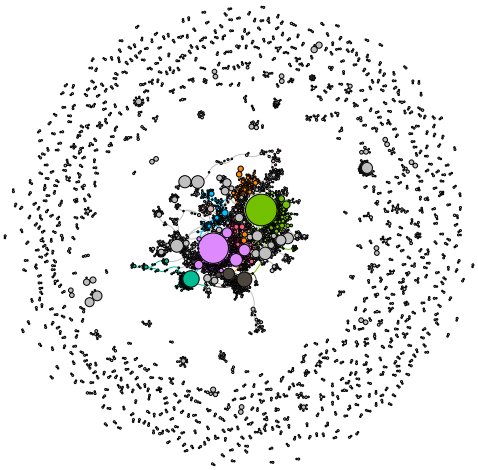
Figure 5: Degree centralization over time for case study words.



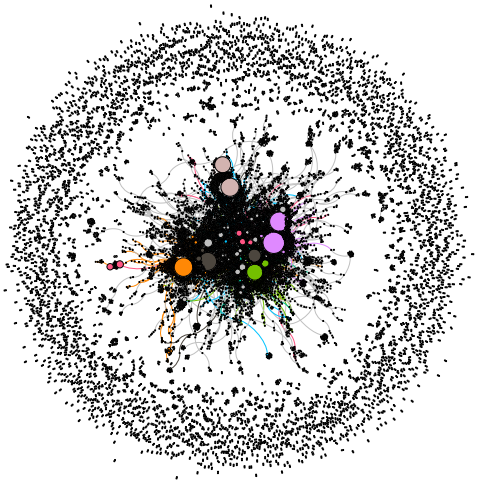
6.2.2 Advanced / from the start: *upcycling*

Figure 6: Social network of diffusion for *upcycling* over time.

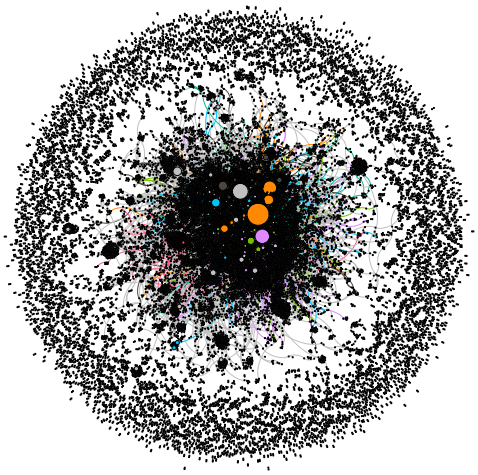
(a) First stage



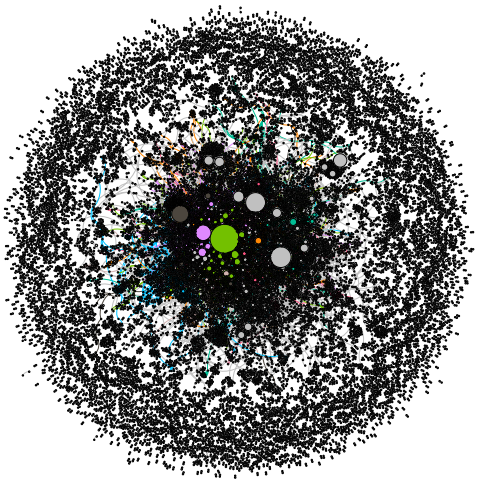
(b) Second stage



(c) Third stage



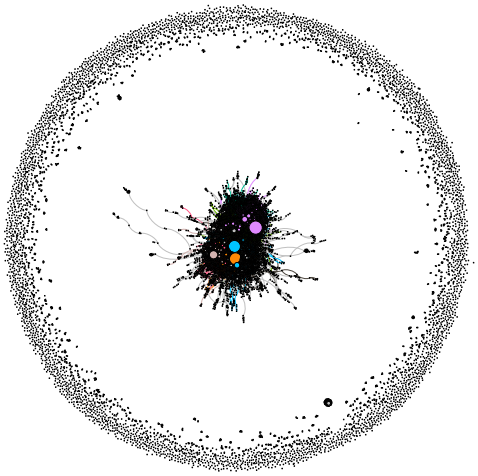
(d) Fourth stage



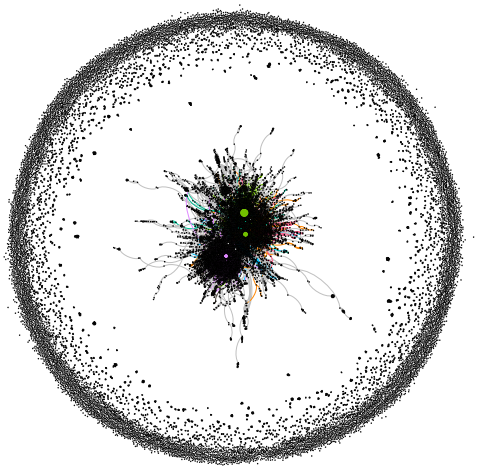
6.2.3 Advanced / increasing: *hyperlocal*

Figure 8: Social network of diffusion for *hyperlocal* over time.

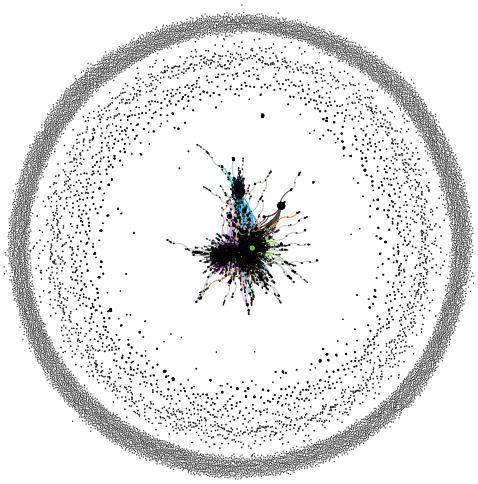
(a) First stage



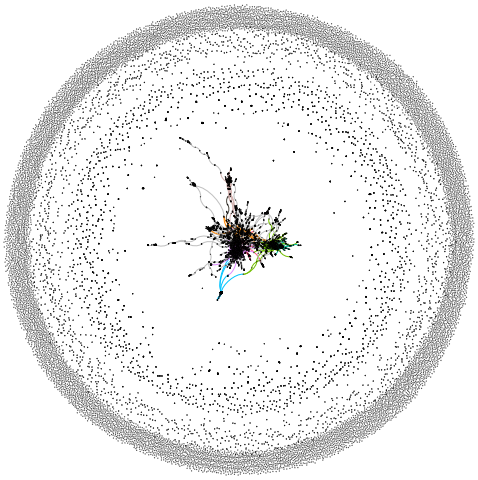
(b) Second stage



(c) Third stage



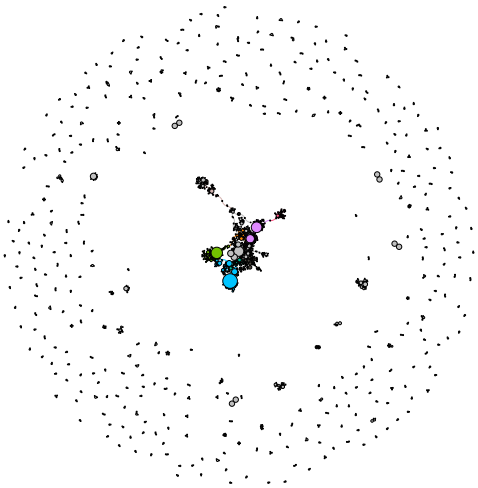
(d) Fourth stage



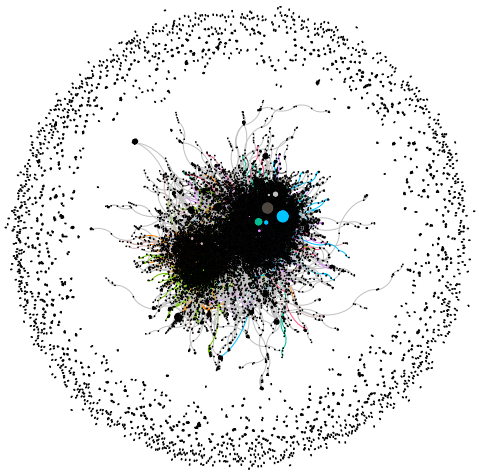
6.2.4 Limited / limited: *alt-left*

Figure 10: Social network of diffusion for *alt-left* over time.

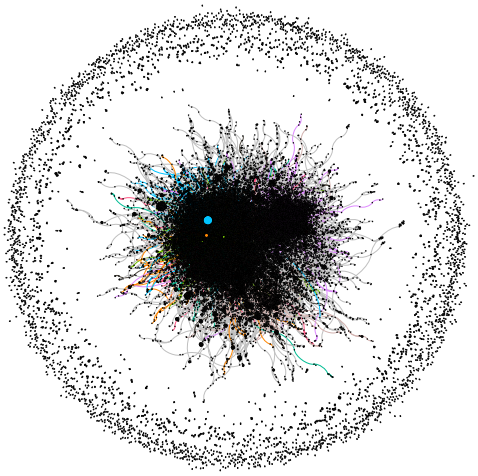
(a) First stage



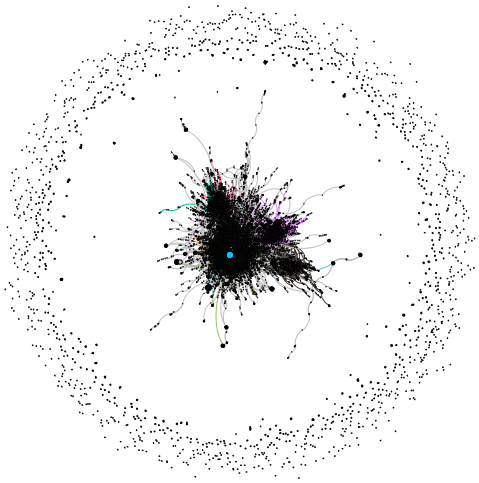
(b) Second stage



(c) Third stage



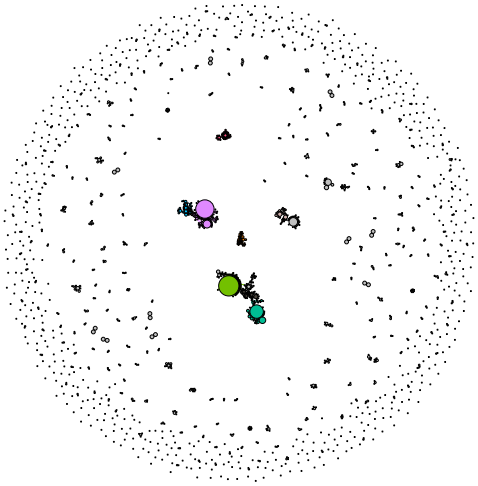
(d) Fourth stage



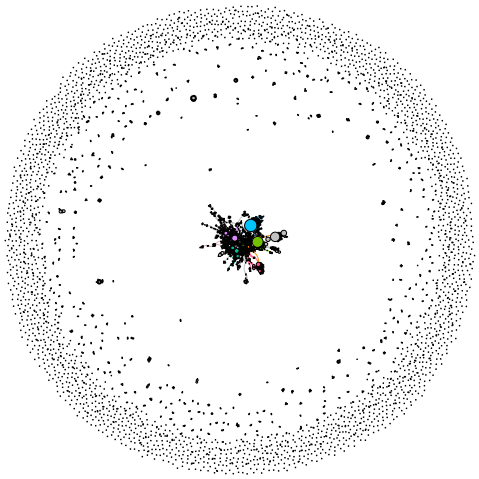
6.2.5 Limited / decreasing: *solopreneur*

Figure 12: Social network of diffusion for *solopreneur* over time.

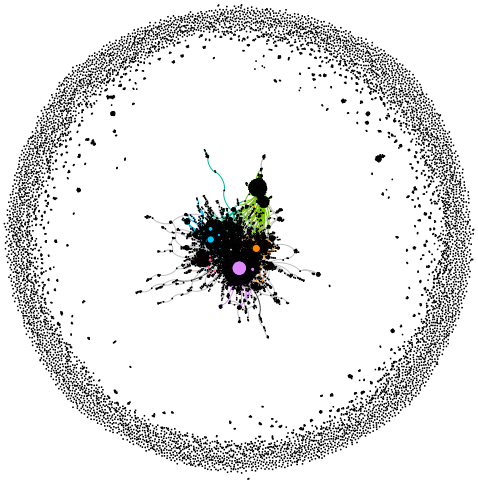
(a) First stage



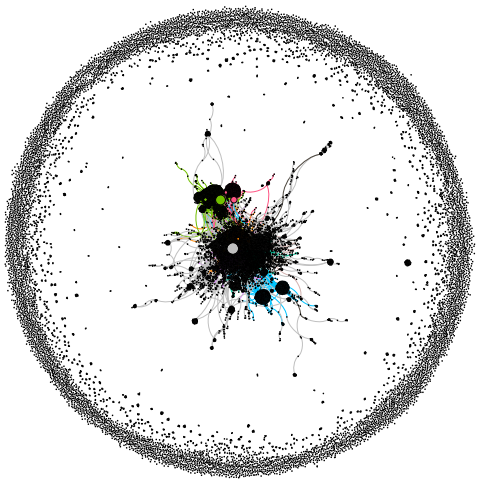
(b) Second stage



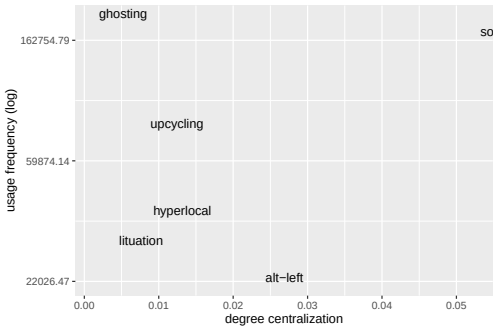
(c) Third stage



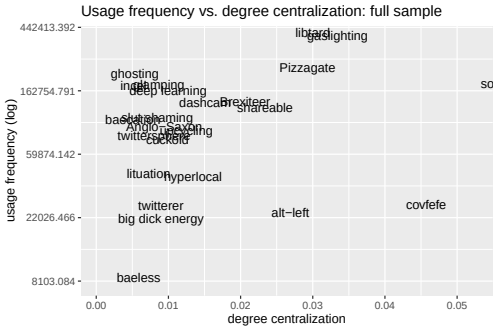
(d) Fourth stage



(a) Full sample.



(b) Case studies.



6.2.6 Biggest changes in the full sample

- increasing diffusion
- centralization

7 Networks vs. frequency

7.1 Plots

7.2 Correlation

7.3 Discrepancies between frequency and centralization

Divergences between frequency and network analysis

- advanced
- topical
- little dispersion
  - political camps:
    - propaganda: *alt-right*, *alt-left*, *covfefe*, *birther*
    - Brexit terms: *Brexiteer*, *Brexiter*, *Brexit*
  - technical

→ case studies

8 Conclusion

- going beyond frequency is important

References

Bauer, L. (1983). English word-formation. Cambridge: Cambridge University Press.

Cartier, E. (2017). Neoveille, a web platform for neologism tracking. In *Proceedings of the software demonstrations of the 15th conference of the European chapter of the Association for Computational Linguistics* (pp. 95–98). Valencia, Spain: Association for Computational Linguistics.

Davies, M. (2013). Corpus of news on the web (NOW): 3+ billion words from 20 countries, updated every day. Retrieved from https://www.english-corpora.org/now/

editorial staff, E. (2020). Diseases show to interact and spread like internet memes. Retrieved from https://eandt.theiet.org/content/articles/2020/02/diseases-shown-to-interact-and-spread-like-internet-memes/

Eisenstein, J., O'Connor, B., Smith, N. A. & Xing, E. P. (2014). Diffusion of lexical change in social media. *PLOS ONE*, *9*(11), 1–13.

Elsen, H. (2004). Neologismen. Tübingen: Narr.

Gérard, C. (2017). The logoscope: semi-automatic tool for detecting and documenting the contexts of french new words.

Goel, R., Soni, S., Goyal, N., Paparrizos, J., Wallach, H., Diaz, F. & Eisenstein, J. (2016). The social dynamics of language change in online networks. In E. Spiro & Y.-Y. Ahn (Eds.), *Social informatics* (pp. 41–57). Cham: Springer International Publishing.

Granovetter, M. S. (1977). The strength of weak ties. In S. Leinhardt (Ed.), *Social networks* (pp. 347–367). Academic Press.

Grieve, J. (2017). *Geographical patterns of lexical innovation.* Workshop 'Diffusion of Lexical Innovations', LMU Munich.

Grieve, J. (2018). Mapping emerging words in new york city.

Grieve, J. (forthcoming). Natural selection in the modern English lexicon. *English Language and Applied Linguistics.*

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

Hohenhaus, P. (2006). Bouncebackability. a web-as-corpus-based study of a new formation, its interpretation, generalization/spread and subsequent decline. *SKASE Journal of Theoretical Linguistics*, *3*, 17–27.

Kerremans, D. (2015). A web of new words. Frankfurt a. M.: Lang.

Kerremans, D. & Prokic, J. (2018). Mining the web for new words: semi-automatic neologism identification with the NeoCrawler. *Anglia*, (136), 239–268.

Kerremans, D., Stegmayr, S. & Schmid, H.-J. (2012). The NeoCrawler: Identifying and retrieving neologisms from the internet and monitoring ongoing change. (pp. 59–96). Berlin: Mouton de Gruyter.

Labov, W. (2007). Transmission and diffusion. *Language*, *83*(2), 344–387.

Lemnitzer, L. (n.d.). Wortwarte. Retrieved January 14, 2018, from http://www.wortwarte.de/

Lipka, L. (2005). Lexicalization and institutionalization: revisited and extended. *SKASE Journal of Theoretical Linguistics*, *2*(2), 40–42.

Lu, F. S., Hou, S., Baltrusaitis, K., Shah, M., Leskovec, J., Sosic, R., ... Santillana, M. (2018). Accurate influenza monitoring and forecasting using novel internet data

streams: a case study in the boston metropolis. *JMIR Public Health and Surveillance*, *4*(1), e4.

Milroy, J. (1992). *Linguistic variation and change: On the historical sociolinguistics of english*. Oxford: Blackwell.

Milroy, J. & Milroy, L. (1985). Linguistic change, social network and speaker innovation. *Journal of Linguistics*, *21*(2), 339–384.

Nevalainen, T. (2015). Descriptive adequacy of the s-curve model in diachronic studies of language change. VARIENG.

Nini, A., Corradini, C., Guo, D. & Grieve, J. (2017). The application of growth curve modeling for the analysis of diachronic corpora. *Language Dynamics and Change*, *7*(1), 102–125.

Pew Research Center. (2019, October 23). National politics on twitter: small share of U.S. adults produce majority of tweets. Retrieved from https://www.people-press.org/2019/10/23/national-politics-on-twitter-small-share-of-u-s-adults-produce-majority-of-tweets/

Renouf, A., Kehoe, A. & Banerjee, J. (2006). Webcorp: an integrated system for web text search. Rodopi.

Rogers, E. M. (1962). *Diffusion of innovations*. New York: Free Press of Glencoe.

Schmid, H.-J. (2008). New words in the mind: concept-formation and entrenchment of neologisms.

Schmid, H.-J. (2016). *English morphology and word-formation: An introduction* (2nd ed.). Berlin: Erich Schmidt Verlag.

Schmid, H.-J. (2020). *The dynamics of the linguistic system. Usage, conventionalization, and entrenchment*. Oxford: Oxford University Press.

Stefanowitsch, A. & Flach, S. (2017). The corpus-based perspective on entrenchment. In H.-J. Schmid (Ed.), *Entrenchment and the psychology of language learning: how we reorganize and adapt linguistic knowledge* (pp. 101–128). Boston, USA: American Psychology Association and de Gruyter Mouton.

West, R., Paskov, H. S., Leskovec, J. & Potts, C. (2014). Exploiting social network structure for person-to-person sentiment analysis.