Who Really Spreads New Slang Online?

**ABSTRACT:**

It is commonly thought that famous people (politicians, celebrities, activists etc.) are highly influential in defining new terms and introducing them to the public. For this reason, they are often referred to as “influencers”. However, recent studies on social media have found that influencers do not actually have this expected effect on the language of a community. According to this research, the “crowd”, the many ordinary people with few followers, tends to have the larger effect on the popularity of trends on social media. In this project, I intend to confirm this hypothesis and answer the question: Are users that are members of both a group that uses a slang term and a group that does not really responsible for the transmission of new slang into a community, or is slang is introduced through other factors?

I use a computational linguistic approach to answer this question. I select 5 slang terms to study, selection of which is highly linguistically motivated by knowledge of dialects of English. I choose 10 communities on Reddit, a popular website with user-generated content, and write a Python script to obtain and analyze data on the usage of these terms in these communities. Once the data is compiled, I generate an interactive web application that demonstrates the data I collected and the conclusions I draw from it. The final product of this project, and the skills I improve along the way, will make me a much more competitive candidate for industry jobs after graduation.

**Introduction:**

Slang words are often used to measure trends in public life, particularly on social media, and can be used to study how information is transmitted between and within communities. In this project, I compare the role of influencers and ordinary users in the introduction and popularization of new slang words in online communities. It is commonly thought that famous people (politicians, celebrities, activists etc.) are highly influential in defining new terms and introducing them to the public. For this reason, they are often referred to as “influencers”. However, recent studies have found that influencers do not have this expected effect on social media. According to this research, it is the “crowd”, or the many ordinary users with few followers, that tend to have the largest influence on the popularity of social media trends. The purpose of this study is twofold. Firstly, I am interested to see if this research is consistent across social media platforms. Furthermore, I am interested to see if, of the users in the more influential group, it is those that are members of both a community that uses a slang term and one that does not that are responsible for the transmission of new slang into a community. Conducting this study will sharpen my skills in computational linguistics and allow me to apply my linguistics knowledge to shed light on a real-world issue. These skills, and the excellent addition the final product will make to my portfolio of work, will greatly increase my hireability for an industry job in computational linguistics upon graduation.

**Methods:**

I will use data from Reddit to conduct this study. Reddit is a web content rating and discussion forum website, with around 234 million users who generate all of the content on the website. Reddit is composed of numerous forums, known as “subreddits”. These subreddits represent distinct communities on the site, with users who subscribe to a particular subreddit constituting its membership. I first identify 10 subreddits that have varying overlap in membership: *askreddit, askmen, askwomen, askgamers, talesfromtechsupport, talesfromretail, roastme, tifu, askgaybros, asknyc, relationships*. I choose five known, recent slang terms, picked to represent slang from a number of known dialects and subcommunities of English: *thot, stan, periodt, deadass, def*.

I write a Python script to track how many members are using these words in the chosen communities over the past two years, and which of these members are influencers in the community. A member is considered an influencer if they are in the top 5% of members in a subreddit in terms of average score on their posts at a given time. Score is a popularity metric generated by members up-voting and down-voting a post. A higher score means a post is more popular, so members with a higher average score on their posts in a subreddit are more famous in that community. When there is a spike of popularity of a slang term in a subreddit, I will take a more in-depth look at which members are using the term right before the spike.

The Reddit database that I access for this project is already created. My challenge will be how to best query the database to get the necessary information, and how to represent it in the final product. To address this, I will write several versions of the script to access the data differently, and experiment with what model I use to represent the data. With proper structuring of the code, the script, and the model, swapping out communities, terms, and modeling parameters should not be particularly time consuming.

**Tentative Hypotheses:**

If influencers are responsible for the increased popularity of a term, then influencers will use the term more right before the spike in popularity. If it is the crowd that is responsible, then the influencers will use the term more after the spike. I expect that the crowd will have a greater impact on popularity. I also expect that if users are members of two communities, one that uses a slang term and one that does not, then those users will be responsible for introducing the slang of the former into the latter. Once I have completed the study, I will generate an interactive web application that displays the data I collect and outlines the conclusions of my findings.

**Qualifications, Applicable Skills, and Mentorship:**

My work in linguistics greatly informs which communities to choose, based on their known dialects, the slang in use in these dialects, and which terms to use in order to find recent transmissions between the dialects to better model the data. Through my work in computational classes I have gained many skills writing computational linguistic scripts and linguistic data analysis that is applicable to many aspects of this project. Furthermore, which communities and slang terms to choose, as well as the computational portion of this project, are to be supervised by my mentor, Dr. Jason Riggle. Dr. Riggle is an active researcher in computational linguistics, whose research has been in strong collaboration with the Toyota Technological Institute at Chicago, a Toyota-funded research laboratory. Dr. Riggle’s guidance will help me better understand contemporary approaches to the field, as well as assist in troubleshooting methodological issues with the project as they may appear.

**Relevance to Academics and Career:**

I am a second-year Linguistics and Computer Science major, so computational linguistics is at the intersection of my principal interests. This project allows me to apply my knowledge of sociolinguistics, amplified by my training in Computer Science, to help answer a real-world question. This project is an application of the study of language contact using big data, an active area of computational linguistics research in which I am very interested in pursuing at a technological company after graduation. This project will give me a taste of what such research is like, and I intend to continue this type of research for my B.A. Thesis in Linguistics. More saliently, the final product of this project will be a significant addition to my portfolio of work, making me a more competitive candidate for industry jobs after graduation.