1. Introduction

For my group I chose the community of fountain pen users. Using data from the subreddit [www.reddit.com/r/fountainpens](http://www.reddit.com/r/fountainpens), I show how the community communicates with positive sentiment, frequently gives thanks, values information/advice and detailed responses, indicates willingness to take suggestions and update people on decisions made, uses hedges to to indicate uncertainty and subjectivity, and uses boosters to emphasize emotion and support.

2. Data

r/fountainpens is “A community for fountain pen enthusiasts, from the novice to the collector.” Comments from r/fountainpens were collected from <https://files.pushshift.io/reddit/comments> from the beginning of the subreddit (9/3/10) to 10/31/18, totalling 1237760 comments. In addition to the comment body, other data was included as well (e.g., score (upvotes - downvotes), date of creation, id of what the comment was replying to). A field called is\_submitter (indicating whether the author of the comment is the submitter of the post; in reddit terms, we would say the commenter is the OP) was added in 2017, so comments in 2017 and 2018 have this field. Several subcorpora were created: low scores (score < 1), high scores (score > 6), thanks (comments that include “thank” in the text), thanked (comments for which at least one reply has “thank” in the text), submitter (comments where is\_submitter is true), question (comments with at least one question mark), and questioned (comments for which at least one reply has a question mark in the text). No differences were found in linguistic patterns between submitter, question, and questioned, even though they didn’t all have the same comments.

3. Methods

spaCy was used for tokenization, part of speech tagging, and dependency parsing. Frequency distributions (e.g., word counts), bigrams, and trigrams were calculated from spaCy results. Keyword analysis done using python implementations I wrote of existing algorithms. Sentiment analysis was done using nltk’s Vader. LIWC was used for comparing the main corpus to reference categories (blogs, expressive writing, and twitter) and for comparing subcorpora to the main corpus. I created a list of hedge words based on reading two articles (both of which are in the References section), but didn’t end up using these results much since LIWC and keyword analysis ended up being more relevant. Looking at metadata, creating subcorpora, finding example comments, etc. was done using pandas. Analysis (at least, the parts that I found interesting) ignored case (it treated everything as lowercase). The code I used for analysis can be found at <https://gitlab.com/cmd16/reddit-corpus-linguistics> and <https://gitlab.com/cmd16/spacy_corpus_linguistics_fanfic> .

4. Findings

4.1 Saying “thanks” (thanks subcorpus)

4.1.1 Frequency, word count, and sentiment analysis

People say thanks a lot. In the main corpus, “thanks” ranks 98th out of 125084 words, and it has a frequency of 70811, which normalizes to 1615 occurences per million words. “thank” is also common; its rank is 189 and its normalized frequency is 729 occurences per million words. thank has a lower mean word count than the main corpus (according to LIWC); i.e., the comments in thanks tend to be shorter. Many of the comments are just “thanks” or “thank you”, so this is not surprising. thanks has a more positive sentiment according to both VADER and LIWC. It’s also worth noting that even the main corpus has higher affect, higher posemo, and lower negemo than the reference categories (blog posts, expressive writing, and twitter posts), so r/fountainpens comments in general tend to have more positive sentiment.

4.1.2 What are people thankful for?

Keyword analysis indicates that commenters are thankful for information (info/information, sharing), suggestions (suggestion/s, advice, recommendation/s, input, insight), and detail (detailed). So, in general people in the community want to learn from other people’s knowledge and experience, and they appreciate more detailed responses (rather than having a “get to the point” attitude).

4.1.3 How do people respond with thanks?

Keyword analysis also indicates that a significant number of the comments explicitly say the info/advice was helpful. Other keywords include “definitely”, “will”, and “try”. These show up in the context of the commenter saying that they will definitely try what is suggested, as seen in (1). Other common phrases include “look into” and “check out”. (2) is an example of this. “decided” is another keyword. People say thanks for the recommendations and let everyone know what they decided to do. See (3) for an example of this. It’s also worth noting that thanks has a higher percentage of comments where the commenter is the OP as compared to the corpus as a whole.

1. I’ll try it on a better quality paper tomorrow and let you know the results
2. Well I will definitely look into that
3. Thanks everyone, I decided to get a TWSBI Eco, and I'll keep in mind the JinHao x750 + Zebra G nib when I want to try out a flex nib.

4.2 Comments that people say “thanks to” (thanked subcorpus)

4.2.1 Word count and sentiment analysis

What sort of comments do people say thank you to? I noticed several patterns. Comments in this subcorpus tend to have a higher word count as measured by LIWC, which makes sense given that thanks includes lots of thanks for “detailed” responses. Comments tend to have an overall higher positive sentiment according to VADER (according to LIWC affect, posemo, and negemo are all lower).

4.2.2 Recommendations

There are a lot of recommendations (as indicated by the keywords “recommend”, “should”, and “suggest”, and second person pronouns). In recommendations, there is also a focus on providing a variety of options. For example, the keyword “can” is used to indicate what is possible, and the keyword “also” is used to mark an alternative option, as seen in the following example:

1. The Kaweco has a converter available now, and you can also syringe-fill cartridges.

4.2.3 Hedging

The language of thanked tends to have more hedging, as indicated by lower certain and higher tentat scores in LIWC. The keyword “might” is used to emphasize the subjectivity of suggestions. (1) is an example of this. The keyword “if” serves a similar purpose by acknowledging that advice may depend on the context. See (2) for an example. Epistemic hedging is also prevalent. “might” can indicate uncertainty, as in (3). The keyword “should” can indicate a recommendation (e.g., (4)), but the phrase “should be” is also frequently used to indicate that the speaker expects their claim to be true but isn’t sure (e.g., (5)). Another example of hedging is the phrase “hope this helps” (and similar variations), which clearly marks a comment as advice/info that is intended to help the poster (or whoever the commenter is trying to address), and displays humility in acknowledging that it may not be helpful. Another example is using the keyword “very” to soften negative sentiments; e.g., (6).

1. You might like the Sheaffer blueblack, actually
2. Just make sure you give them a good amount of time to dry if you’re using cheap paper
3. This might actually be the Signature model
4. You should check it out for your calligraphy
5. The fine should be a pretty close approximation
6. It’s not very flexy at all

4.2.4 Boosting

While hedging is prevalent in thanked, boosting also exists. “very” is used in compliments (e.g., (1)) and to emphasize descriptions (e.g., (2)). “also” can be used to indicate agreement and provide support to what other people have said; e.g., (3).

1. Very nice collection you have there
2. the nib on the VP is very soft
3. I have the matte black also, my fave pen by far.

5. Conclusions

5.1 What does the community value?

Examining these linguistic patterns paints a broader picture of the values of the community. The community values sharing knowledge and experience, and sharing it in detail. It values diversity of opinions, perspectives, and suggestions (as indicated by use of hedging to indicate subjectivity of opinions and by use of providing a variety of suggestions and alternatives). The community values respect and politeness (as indicated by use of hedging and positive (or at least non-negative) sentiment in thanked). Expressing positive emotions is valued, as indicated by more positive sentiment in the main corpus, even more positive sentiment in thanked, and use of boosters for compliments and agreements in thanked. The community values letting people know that their feedback is appreciated, as indicated by saying thanks, by taking people’s suggestions (or at least saying you will), and by keeping people updated as to what you decided to do after soliciting advice.

5.2 An emic perspective on the findings

As a fountain pen user and a member of r/fountainpens for multiple years, I consider myself qualified to provide an insider perspective on why the community has the observed features.

Fountain pens are not widely used by the general public (at least not in the US), so it can be hard to find someone to share your love of fountain pens with. For many users, this is the only outlet they have to talk about fountain pens. With that context in mind, we can see why there is so much positive emotion; this is a place where people can get excited without getting shut down by people who aren’t interested. Because r/fountainpens may be the only support a person has for this hobby, the community of r/fountainpens has solidarity. That solidarity is shown in the positive emotion, the frequent giving of thanks (to encourage people and make them feel valued), and the politeness.

Because this is many people’s only place to talk about fountain pens, this is also a major source for information and advice. You might be able to look things up on the internet or talk to someone at a pen shop, but if you want to discuss something with a variety of users, r/fountainpens is a great place to go. There it is possible to connect with far more people interested in fountain pens than you would likely encounter otherwise, and you can have a conversation (as opposed to a 1-way intake of information), thus getting the context and understanding that you want for your particular situation. This sharing of information and advice is an important part of r/fountainpens since it can be hard to get this experience elsewhere. There are many kinds of fountain pens, inks, etc. so it can be overwhelming trying to figure out everything by yourself. This is particularly true when you recognize there’s a lot of variety and nuance that can be hard to find by yourself (but much easier to see when getting responses from multiple people). Thus, we see the values of sharing knowledge and experience (especially in detail) and of diversity of opinions, perspectives, and suggestions. Tying into both the importance of sharing information/advice and the solidarity of the community is the importance of letting people know their feedback is appreciated. Not only do we appreciate people’s contributions, we also make sure they know it.

5.3 Community feedback on the findings

I shared my work with r/fountainpens to see what people thought about it. Several users chimed in to agree with the theme of kindness in the community. *QuiteARoughCustomer*, who has cancer, said, “My wife made a post here about getting me a nice pen for a gift when I got out of the hospital, and this community was extremely nice and helpful. So nice, in fact, that I felt like they cared enough that they would want to be updated about my condition, which I've now done a couple times since. Each time, I've been overwhelmed by the positive responses. This is definitely the friendliest corner of the internet that I've ever seen, and I've been on the internet since the days when you had to unplug the phone before you logged on.” *A\_Rank\_Amateur* said that ever since learning that fountain pens have “a whole community behind them”, they have been “continually impressed with the kindness, knowledge, and welcoming nature of this group. I tell my friends it's the nicest corner of the internet!” *DudsDeLait* offers a “good example of the warm vibes of this sub”: <https://www.reddit.com/r/fountainpens/comments/k0av0f/just_a_thank_you_to_all_the_lovely_people_of/>. *dagit* suspects this kindness/politeness is “a common trend for subreddits based on relatively niche interests. The worst experiences I've had on the site were in big general reddits. I'm not really sure why that is. I'd say it transcends moderation and goes on to be a fairly general rule.” *dagit* and *Jayna2000* found it interesting that corpus linguistics can be used to find things like kindness, politeness, and enthusiasm. *Jayna2000* echoed the theme of appreciation for detail and asked me to let them know how my paper was received. *manos\_de\_pietro* echoed the part about “having an appreciative audience for one’s hobby, which can be hard to come by IRL”.

5.4 Connection to theories we studied

We can see evidence of the Social Information Processing theory that we talked about in class. Because emotion can be harder to tell online, people are more explicit about using their words to convey emotion. Users also take advantage of the asynchronous communication format of reddit to share updates. This ties into the faceted scheme for computer mediated discourse that Herring proposed. Also part of that scheme is 1-way vs 2-way communication. The 1-way form of communication inherent to reddit allows users to think/research carefully and write the detailed responses that are so often praised.

5.5 Areas for future work

There is a lot more to investigate in terms of the use of boosters. I suspect boosters are used more to emphasize a person’s emotions and opinions that are clearly subjective whereas hedges are used when presenting any opinion/information that is not objective but might appear so. However, I have not done enough research to be able to support or reject this theory.

Another possibility would be to compare this corpus to another corpus. Potential candidates are other subreddits (though I’m not sure what other subreddits would provide an interesting/useful comparison point) and other places that discuss fountain pens (<https://www.fountainpennetwork.com/forum/> might be an interesting reference, though it’s currently unavailable due to a database upgrade).

I made other findings not included in this report that I could explore more and expand into a report. For example, there’s a sort of in-group focus where comments, especially comments with high score, talk about being part of the community (e.g., “Nice pen, and welcome to the community. You’re in for the long run”). I wonder if this in-group focus is an indication of deindividuation. I could also look more at the LIWC summary variables (Clout and Tone in particular might provide additional support for some of the findings I made) now that I understand better what the LIWC summary variables mean.

6. References

6.1 Research references

Herring, S. (2007). A Faceted Classification Scheme for Computer-Mediated Discourse. *Language@Internet*, *4*(1). <https://www.languageatinternet.org/articles/2007/761>

Pennebaker, J.W., Booth, R.J., Boyd, R.L., & Francis, M.E. (2015). Linguistic Inquiry and Word Count: LIWC2015. Austin, TX: Pennebaker Conglomerates (www.LIWC.net).

Takimoto, M. (2015). A corpus-based analysis of hedges and boosters in english academic articles. Indonesian Journal of Applied Linguistics. 5. 90. 10.17509/ijal.v5i1.836.

Vass, H. (2015). A Comparative Analysis of Hedging in a Corpus of Two Written Legal Discourse Genres.

6.2 Data collection tools

Pushshift API: <https://files.pushshift.io/reddit/comments>

Reddit: [www.reddit.com/r/fountainpens](http://www.reddit.com/r/fountainpens)

6.3 Code used for analysis

<https://gitlab.com/cmd16/reddit-corpus-linguistics>

<https://gitlab.com/cmd16/reddit-corpus-linguistics/-/blob/master/Findings.ipynb> (all the findings from analysis that I found interesting)

<https://gitlab.com/cmd16/spacy_corpus_linguistics_fanfic>