Communal IRC Vocabulary:

A Case Study of Jargon and Slang in an IRC Community

Katherine _____

Jun 25, 2017

1 Abstract

In this paper, i carry out a brief case study of the usage of jargon and slang in the context of an IRC community, determining which words of this type are used most often, measuring their frequency of use by individual community members, and postulating on the results, speculating why certain of these words might differ from others in their commonality and range of adoption.

2 Background

Internet Relay Chat, or IRC, is a chat protocol first conceived and implemented in the late 80s by Jarkko Oikarinen.[1] IRC users connect to one another via *networks*: collections of one or more servers run by independent sources. Once connected, they can join *channels*: chat rooms which allow groups of people to speak together. These channels are generally dedicated to a particular topic, community, or project (e.g. a programming project), and thus, particularly in the case of communities, a general sense of channel-wide camaraderie can usually be assumed.

Since it's creation, IRC has grown to connect hundreds of thousands of users across hundreds of networks worldwide.[2] This broad international scope, coupled with its long history of use, purely textual content, and generally passionate and technically literate userbase, make IRC an especially interesting subject for linguistic research.

3 Discussion

In the interest of simplicity, i chose to limit the scope of this study to a particular mid-sized, tight-knit IRC community channel, `#lainchan`, investigating statistical data on the use of *jargon* (community or domain specific technical terminology) and *slang* (generally informal community or region specific vocabulary) within that channel. These both tending to be community-oriented, i was curious to determine their adoption rates, whether generally shared or isolate to further subgroupings.

Jargon and slang coined within a modern community of this sort are generally referred to as *neologisms* (defined loosely as being either new words or existing words that have been given new meanings[3]). Such words are distinguished from *nonce-words*, which are single-instance coinages local to a one-time utterance or work and having no

[1] Jarkko Oikarinen, "Internet Relay Chat," accessed June 25, 2017, `http://www.kumpu.org/irc.html`.

[2] "IRC Networks - Top 100," accessed June 25, 2017, `http://irc.netsplit.de/networks/top100.php`.

[3] Tony Mcenery Paul Baker Andrew Hardie, *A Glossary of Corpus Linguistics* (Edinburgh University Press, 2006).

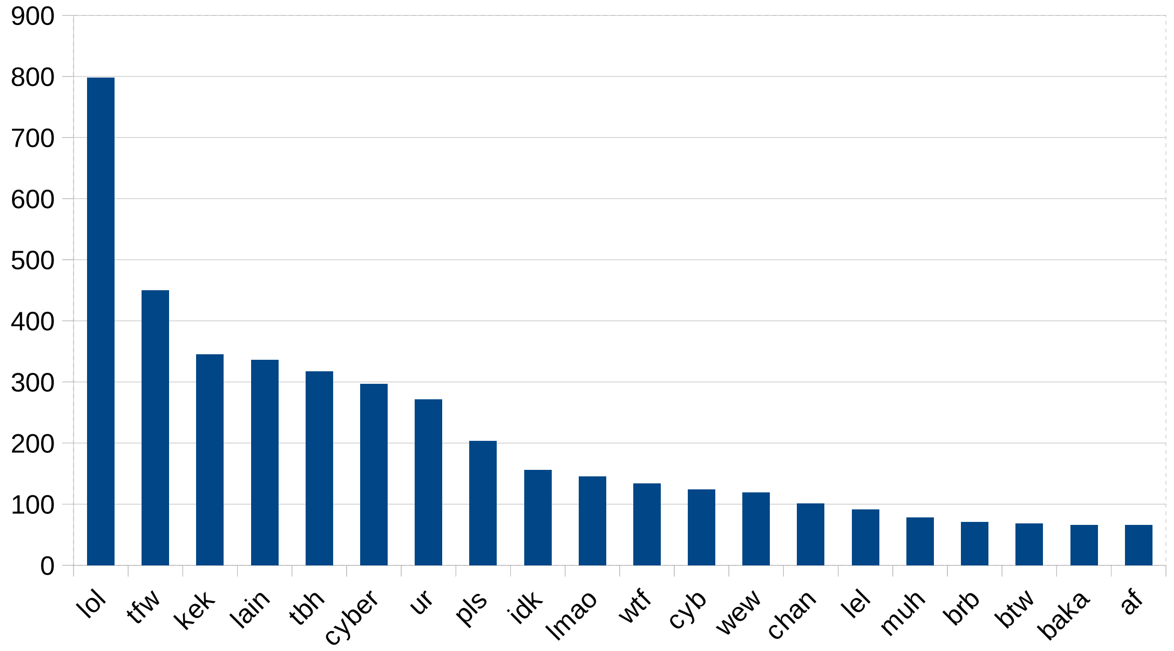

Figure 1: top 20 words in `#lainchan`, by frequency of occurence over the past 1.5 years

lasting scope, and thus, when isolating potential jargon or slang words within a community, it is important to determine first the degree of frequency of use and define some lower bound for what is to be accepted.[4]

4 Method

The first piece of my implementation is an archive of text spoken in `#lainchan` over the past several years, which was automatically written by my IRC client (this backlog is considered an *archive* rather than a *corpus* as it was collected passively and has not been actively selected or processed[5]). This archive was then fed to a script written in Perl, which collected words found and their frequency of use. This step was not entirely trivial, as several habits of IRC users, such as the tendency to mention names and paste URLs, introduce potential non-word clutter which had to be stripped. This task was solved largely through the use of *regular expressions*, which provide a simple, deterministic way of defining pattern matches.

Listing 1: link and word regular expressions

```
(http[s]?://\S*|\S*\..{1,2})
((?:\w+\')*\w+)
```

From the resulting words, i selected the most frequently occurant *lexemes*: the base of a group of words which share meaning but differ in form due to conjugation etc.[6] This reduction process is known as *stemming*, and it is non-trivial, as words forms are not derived in a completely systematic way. Consider the cases of “writing”, derived from “write”, and “winning”, derived from “win”: one cannot rely on a simple rule such as “append the -ing suffix to produce a gerund” as it will fail in both cases for different reasons. neologisms further complicate stemming, as their resultant forms cannot be determined using a database approach, being newly emergent. For these reasons, lexeme isolation required manual

[4] Daphné Kerremans, *A Web of New Words: A Corpus-Based Study of the Conventionalization Process of English Neologisms* (Peter Lang, 2015).

[5] Paul Baker, *A Glossary of Corpus Linguistics.*

[6] Ibid.

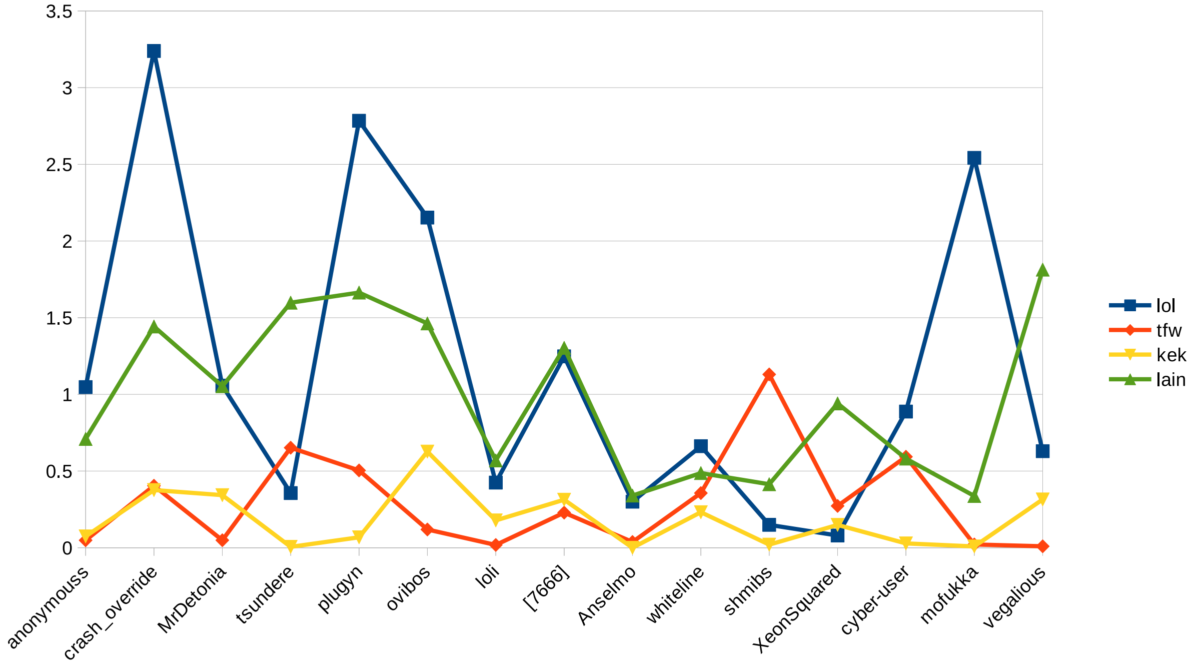

Figure 2: percentage of spoken lines per user containing the top 4 jargon or slang words

intervention, locating likely candidates with fuzzy-finding techniques and merging them by hand.

From these most frequent lexemes, i used the popular `Hunspell`[7] spell checking tool to filter out common dictionary-recognised words and ease the process of isolating jargon and slang words. The archive was then parsed again using a second script, collecting per-user frequency counts for these top-occurent lexemes, and it is from these resultant data i calculated my results.

5 Results

The top 20 most frequently used jargon and slang words can be seen in **Figure 1**. The majority of these (such as 'tfw', 'idk', and the everpresent 'lol') are slang words, with their origins in abbreviations or acronyms. However, they are distinct in their usage and connotative meanings, to the point where the non-abbreviated form, inserted in the same location, would yield a different semantic result. Consider the following usage of 'pls', in which a user named `illu` unintentionally highlights the existence of this distinction:

Listing 2: contrastive usage of slang word 'pls'

```
<illu> pls
<illu> Seriously though please
```

Other words in this list are imports from other languages, such as 'kek' (a westernisation of ケケケ, a Japanese ideophone indicating laughter), or 'baka' (a romanisation of the word 馬鹿), or further variations on other words in the list ('lel' from 'lol' and 'cyb' from 'cyber'), all having taken on meanings distinct from their source words. This highly nuanced connotative variance on existant slang and terminology is likely a result of the purely-textual environment, which does not allow for more conventional side-channel methods of indicating tone etc.

Also over the past 1.5 years, the frequency of usage for the top 4 of these words, among the top 15 users (by lines spoken), is shown in **Figure 2**. This graph displays a surprisingly high degree of variance in use between users. 'lol', for example, occurred

[7] "Hunspell," accessed June 25, 2017, `https://hunspell.github.io/`.

in 3.24% of all lines spoken by most frequent user crash_override, while appearing only 0.08% of the time for least frequent user XeonSquared. This high degree of variance is present for all three of the top words. ‘lain’, on the other hand, had a relatively consistent high-usage across the board (with several dips, but none of the complete zeroing-outs of the other three). In addition, despite being the least frequently used of these four words when taking the community as a whole, it appeared second most often when considering only these top users, with an average appearance rate of 0.98% (lol: 1.17%, tfw: 0.30%, kek: 0.18%). ‘lol’, ‘tfw’, and ‘kek’ are all slang words imported from other communities, where they are still in regular use, whereas ‘lain’, as it is used in the #lainchan IRC channel, is a local jargon coinage. From this, i would tentatively conclude that jargon and slang words of this sort, which have a high adoption rate across the board for most frequent users but appear less often in utterances from less frequent users, can be taken as representative of the community as a whole. As i can’t perform more rigorous testing (being terrible at statistics and mathematics in general), no other useful conclusions can be derived from these data.

References

“Hunspell.” Accessed June 25, 2017. https://hunspell.github.io/.

“IRC Networks - Top 100.” Accessed June 25, 2017. http://irc.netsplit.de/networks/top100.php.

Kerremans, Daphné. *A Web of New Words: A Corpus-Based Study of the Conventionalization Process of English Neologisms.* Peter Lang, 2015.

Oikarinen, Jarkko. “Internet Relay Chat.” Accessed June 25, 2017. http://www.kumpu.org/irc.html.

Paul Baker, Tony Mcenery, Andrew Hardie. *A Glossary of Corpus Linguistics.* Edinburgh University Press, 2006.