Corpus Analysis of Reddit Conversations Related to the 2020 Olympic Games

Osaze Cuomo
Hyogo University

Reference Data:

Cuomo, O. (2023). Corpus Analysis of Reddit Conversations Related to the 2020 Olympic Games. In P. Ferguson, B. Lacy, & R. Derrah (Eds.), *Learning from Students, Educating Teachers - Research and Practice*. JALT. https://doi.org/10.37546/JALTPCP2022-18

This paper presents the results of a data collection project focused on online Reddit conversations about the 2020 Olympic games. The study explored the use of theme-specific corpora from online sources for use in language education. The data collection and analysis process involved the use of Python scripts to gather 254,267 comments from the most upvoted Reddit posts related to the Olympics in the 2 months around the Games. The collected comments were then processed using NLTK (Natural Language Tool Kit) to identify parts of speech within each comment, and VADER (Valence Aware Dictionary and sEntiment Reasoner) to give each comment a sentiment score. The data was then analyzed for general insights and compiled into a word frequency list which ranked words based on their usage in the collected comments.

本稿では、2020年のオリンピックに関するオンラインRedditの会話に焦点を当てたデータ収集プロジェクトの結果を紹介する。本研究では、オンラインソースからテーマ別のコーパスを言語教育に利用することを検討した。データ収集と分析プロセスでは、Pythonスクリプトを使用して、大会前後の2ヶ月間にオリンピックに関連するRedditの投稿のうち、最も賛同票を得ている投稿から254,267件のコメントを収集した。収集したコメントは、NLTK（Natural Language Tool Kit）を用いて各コメント内の品詞を特定し、VADER（Valence Aware Dictionary and sEntiment Reasoner）を用いて各コメントに感情スコアを付与した。その後、一般的な洞察を得るためにデータを分析し、収集されたコメントでの使用に基づいて単語をランク付けする単語頻度リストにまとめた。

This paper outlines the process and results of a data collection project that used Python scripts to collect and analyze word usage and sentiment in online Reddit conversations relating to the 2020 Olympic Games held in Tokyo, Japan, with the primary goal of creating a word frequency list including sentiment values. The motivation for this project lied in exploring tendencies in vocabulary use around a specific topic and assessing the potential of this methodology in preparing word lists for language assessments, as well as enabling students to conduct their own investigations into language use and gain a more comprehensive understanding of current language trends. English, like all languages, is in a state of constant change, however in recent years the pace of this change has accelerated due to the widespread adoption of digital forms of communication (McCulloch, 2020). It is increasingly common for people to communicate with friends and family, as well as complete strangers, using text and an online interface. This has changed the way people exchange ideas, as conversation with strangers is now something that can be done at any time from any location. Public digital communication leaves behind a trail which gives researchers the ability to measure and track changes in language and communication patterns across online platforms (Dang et al., 2020), such as keyword tracking on Twitter used to measure public sentiment during the COVID-19 pandemic (Kausar et al., 2021). For both the language teacher and student, access to real-world communications can provide insights into specific vocabulary or parts of speech that are in common use, which can help focus limited teaching or study time for a more efficient learning experience. Additionally, language learners often have difficulty understanding the subtext or connotations in using a specific word, and the addition of sentiment scores on word lists may be helpful in gaining a better understanding of where and when words may be appropriate. The creation of corpora from authentic sources is one way educators can account for changing language patterns, adapt to student expectations, and take advantage of new tools to maintain relevance and student interest in an increasingly digital world.

An Introduction to Reddit

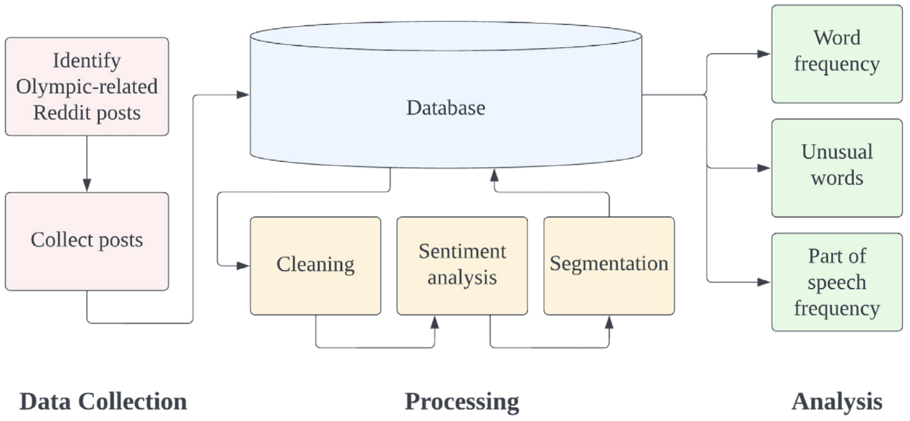
This project used Reddit posts and conversations related to the 2020 Olympics as its primary data source. Reddit is a platform divided into subreddits, or interest groups, where users can post content related to the subreddit and engage in conversation via text comments with other users. Subreddits range from broad (r/news, r/politics) to focused (r/olympics, r/diy). The site is one of the main drivers of English language news and political discourse online (Horne & Adali, 2017) and has emerged as a data source for researchers interested in a variety of fields (Proferes et al., 2021). It is one of the top-20 most visited websites worldwide (Statista, 2021b), with over 300 million posts in 2020, or approximately 830,000 posts per day (Reddit, 2020). Its actively engaged user base provides a constant stream of language data in the form of publicly available, yet anonymous, posts and comments which prompted its selection as a data source for this project.

Using Reddit as a data source is not without its problems, namely that it excludes those who do not use the platform (Hargittai, 2018). The user base skews young and male, with the majority being under 50 years of age and from English-speaking countries, primarily the United States (Statistia, 2021a). There is also a significant, yet unknown number of bots posting and engaging in conversation on the platform. While this is an issue, evidence suggests that human responses to bots reflect the sentiment of the bot's comment, and there is evidence that lexical entrainment is taking place, where humans overlap with bot comments in a similar manner as takes place in human-human conversations (Ma & Lalor, 2020). As bots are now a part of online discourse it makes sense to include their contributions. Despite the issues, Reddit provides a real-time stream of English language data related to a wide variety of topics that can be used to inform language education and learning.

Methodology

The project consisted of three main phases: data collection, processing, and analysis (Figure 1).

Figure 1
Project Workflow



Data Collection

In the data collection phase, the most popular Reddit posts related to the Olympics between July 1, 2021, and August 31, 2021 were identified using pushshift.io (Baumgartner, n.d.). Two types of posts were used: *megathreads* in the subreddit r/olympics, and standard *posts* from any subreddit which contained the word "Olympics" in the title. Megathreads were automatically created daily in the Olympics subreddit as places for users to post near-real-time reactions to events as they happened, and comments in these threads tended to be short and have no replies. Standard posts were created by individual users and posted within any subreddit. Post comments tended to be longer and have more back and forth conversations, sometimes with hundreds of replies to a single comment. Posts and comments can be upvoted or downvoted by users, an upvote indicating the comment or post is a valuable contribution and a downvote being the opposite. In all, the comments from 16 megathreads were collected, one for each day of the Olympics, as well as the top 16 most upvoted Olympic related standard posts. The comments for each post and megathread were collected and downloaded in CSV format using a Python script (Cuomo, 2022).

Processing

The downloaded comments were tagged with metadata, then tokenized, lemmatized, and cleaned before being compiled into a database. Tokenization refers to the process of preparing text to be read by a software program, in this case Python.

> Untokenized: This is an example sentence.
>
> Tokenized: 'This', 'is', 'an', 'example', 'sentence', '.'

In the lemmatization process, word variations were standardized into their base form so that they would be considered the same word, such as 'get' becoming 'got.' The cleaning process included removing punctuation, non-English words, 'stopwords,' or words that do not significantly change the meaning of a sentence and converting uppercase letters to lowercase. All these steps were done using Python scripts to standardize the comments as much as possible for further analysis.

Sentiment analysis on each lemmatized comment was conducted using VADER (Valence Aware Dictionary and sEntiment Reasoner), a rule-based sentiment analysis tool which uses a pre-built lexicon consisting of words and phrases annotated with their respective sentiment scores to gain insight into the writer's sentiment or emotion from a given piece of text (Hutto & Gilbert, 2014). These scores were converted from numerical values into *positive*, *negative*, and *neutral* assignments which were then appended to the database (Real Python, 2022). Each row of the database shows a full comment, its lemmatized form, as well as sentiment scores and other identifying information (Table 1).

Analysis

Using the database, the number of unique words were counted from the lemmatized and cleaned comments, then segmented based on the following attributes: sentiment (positive, negative, neutral), and post type (megathread or post) using Python. The word list was compared with other corpora to identify words which were more common in the Olympic conversations than in daily use. The frequency of different parts of speech was also counted using Natural Language Toolkit (NLTK) and Python (*Language Processing and Python*, n.d.).

Python is a programming language, widely used by programmers, researchers and data scientists, with extensive support for data analysis, visualization, and natural language processing. It is considered to be one of the more readable and user-friendly coding languages, as such it has a large user base, comprehensive documentation, and

Table 1

Comment database from Reddit Olympic Posts (3 of 254,267 rows)

Comment	Lemma	Post_Code	Post	Negative	Neutral	Positive	Compound	Sentiment	Day	Day_of_Week	Date
Probably shouldn't be tbh. For example Slovenia will be going for Pogacar and Roglic - the other riders are only there to support. Similarly Belgium will be going for WvA and Remco.	probably tbh example slovenia go pogacar roglic rider support similarly belgium go wva remco	Megathread	Megathread Day 1	0.0	0.917	0.083	0.4019	Positive	1	Saturday	2021-07-24
I can see the empty stage. That's it.	see empty stage	Megathread	Megathread Day 1	0.231	0.769	0.0	-0.2023	Negative	1	Saturday	2021-07-24
How long to go in the cycling? Wanna know when I should switch back...	long cycling wan na know switch back	Megathread	Megathread Day 1	0.0	1.0	0.0	0.0	Neutral	1	Saturday	2021-07-24

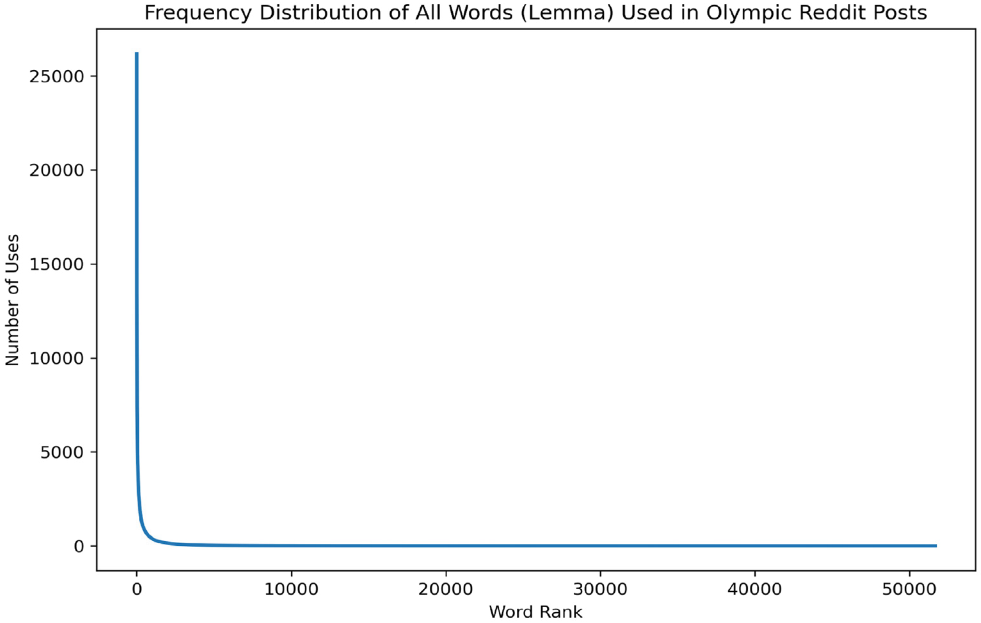
a wide range of tutorials and free learning resources available. It has a vast ecosystem of libraries, essentially add-ons that enable additional functionality, such as pandas for basic data analysis, NumPy for mathematical analysis, and NLTK for natural language processing. Python also facilitates the automated collection of large datasets from the web using APIs, in this case Reddit using PMAW (Pushshift Multithread API Wrapper). Python's power, combined with its relative ease of use, made it a good choice when the project began in 2021. However, data collection and analysis tools are rapidly changing, and the exact methods used in this project will likely have been superseded by other options at the time of this paper's publication.

Results

The final database consisted of 4,421,426 words from 254,267 Reddit comments, which included 51,677 unique character strings. The character strings are predominantly "words," however non-traditional words such as *biohack*, *wowowow*, and *150k* were also included in the total number of words. Reddit usernames take many forms and were included in the final word list. More than half the words in the word list (26,954) were used just once, and only 10,300 words were used 10 times or more. The most frequent 2,000 words account for 82.67% of the word list, while the most frequent 200 words account for 42.95% (Figure 2).

Figure 2

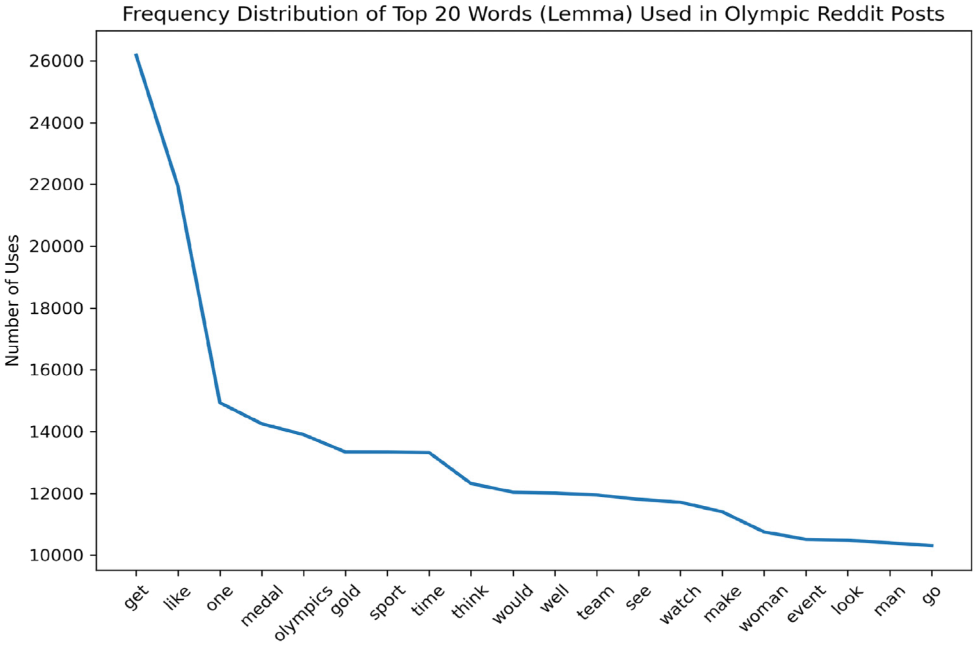
Frequency Distribution of All Words Used in Olympic Reddit Posts



The top 20 most frequently used words accounted for 12.09% of the entire word list (Figure 3).

Figure 3

The Top 20 Words Used in Olympic Reddit Posts



Singular nouns, adjectives, and plural nouns were the three most frequently used parts of speech (Figure 4). The parts of speech abbreviations are explained in Table 2.

Figure 4

Frequency Distribution of Top 10 Parts of Speech Used in Olympic Reddit Posts

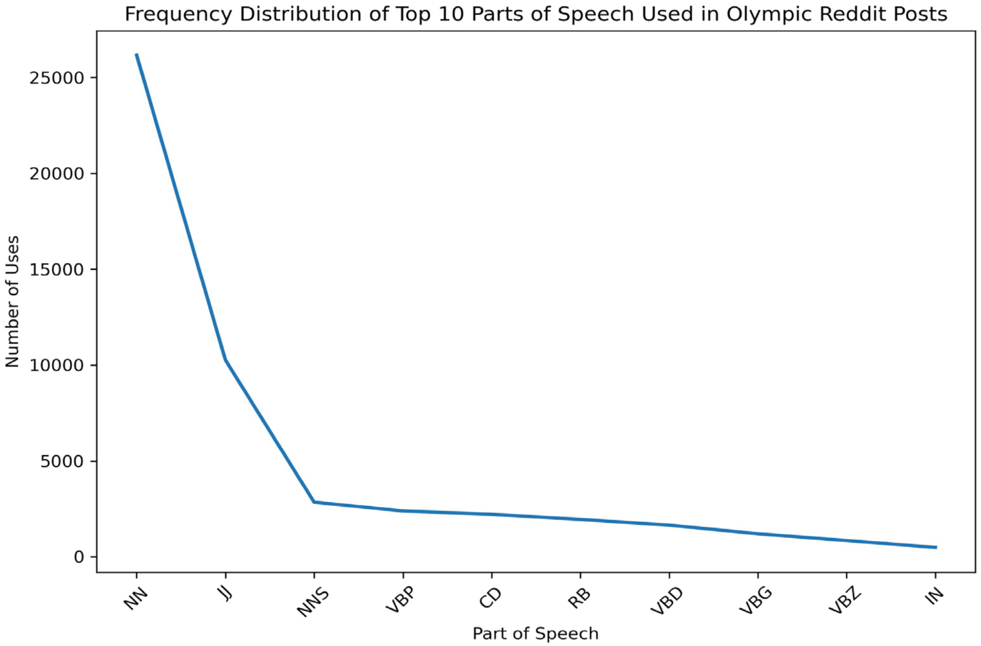


Table 2

Part of Speech Codes and Meanings

Part of Speech	Meaning
NN	Noun, singular or mass
JJ	Adjective
NNS	Noun, plural
VBP	Verb, non-3rd person singular present
CD	Cardinal number
RB	Adverb
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VVBZ	Verb, 3rd person singular present
IN	Preposition or subordinating conjunction

Sentiment scores gave an indication of the general sentiment of each comment in the Reddit conversations used to compile the overall word list. Looking at the average VADER sentiment score, as described above (Table 1), gives an idea of the general tone of conversation within each post (Figure 5), with full titles for each post (Table 3).

Figure 5

Average Sentiment Score by Post

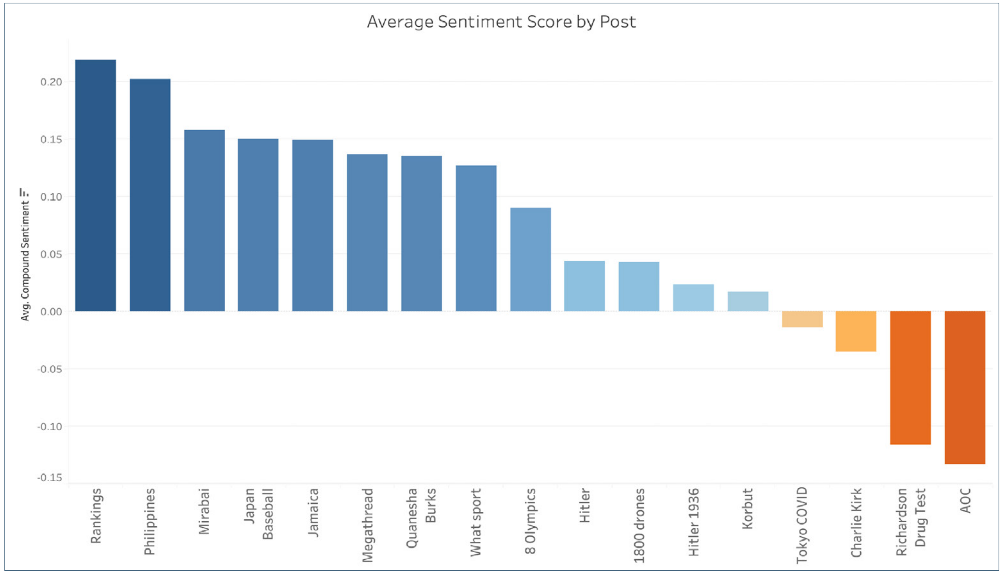


Table 3

Post Codes and Full Titles

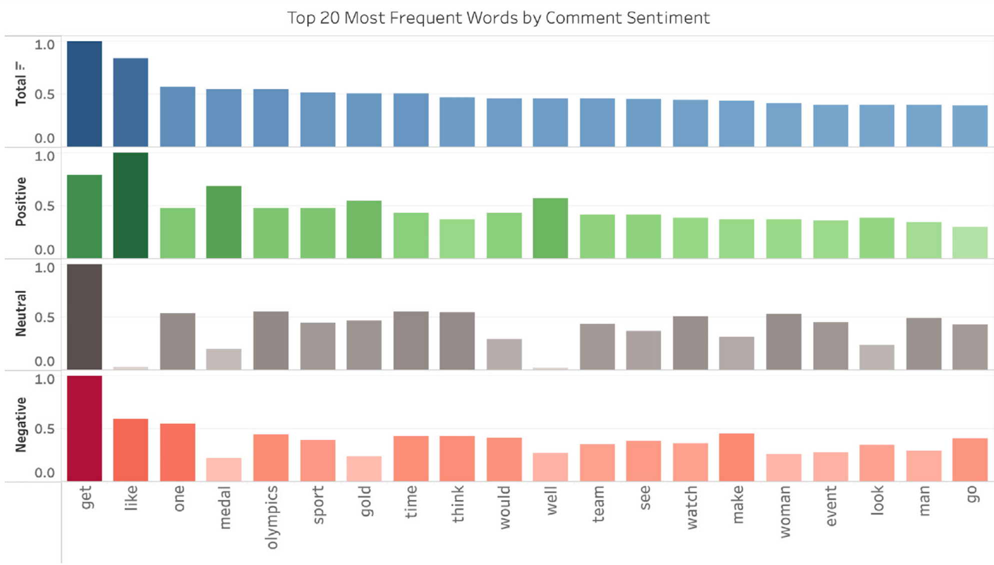
Post Code	Post Title
Rankings	The #1 rankings are now settled. The United States of America will be #1 in Golds, #1 in Silver, #1 in Bronze, and #1 in Total Medals in the 2020 Tokyo Olympics.
Philippines	Philippines just won its first Gold medal in the Olympics
Mirabai	India's Mirabai Chanu at her home in Manipur after winning Silver Medal in the Tokyo Olympics.

Post Code	Post Title
Japan Baseball	[Nightengale] Japan wins their first Olympic gold medal in baseball and Team USA takes the silver medal after Japan's 2-0 victory in thrilling final. They won't meet again in the Olympics until 2028 in Los Angeles with breakdancing scheduled to replace baseball and softball in 2024 in Paris.
Jamaica	Jamaica take Gold, Silver and Bronze in Women's 100m final. Tokyo Olympics 2021.
Megathread	Megathread (Day 1 - 16)
Quanesha Burks	Quanesha Burks - from McDonalds to the Tokyo Olympics in Long Jump!!
What sport	What sport should be in the Olympics but isn't?
8 Olympics	8 F**king times in Olympics. Take a bow
Hitler	Hitler being high on meth in the 1936 Olympics
1800 drones	1800 drones above the National Stadium in Tokyo, Japan form a globe at the Tokyo2020 Olympics opening ceremony
Hitler 1936	Hitler watching 1936 Olympics high on dexamphetamine
Korbut	"Korbut Flip" by Olga Korbut in Munich 1972 Olympics.
Tokyo COVID	Tokyo Covered Up Arrival of Deadly New COVID Variant Just Before the Olympics
Charlie Kirk	Charlie Kirk goes after the Simone Biles for pulling out of the olympics, calling her a "Selfish Sociopath"
Richardson Drug Test	Sha'Carri Richardson fails drug test for marijuana, could miss Olympics.
AOC	Alexandria Ocasio-Cortez says cannabis prohibition is a 'racist and colonial policy' and condemns Olympics ban of Sha'Carri Richardson

Normalized sentiment scores for the top 20 most frequently used words in the word list show whether the word is more often used in positive, negative, or neutral comments in the posts collected (Figure 6).

Figure 6

Comment Sentiment of Top 20 Most Frequent Words in Reddit Olympic Posts



Discussion

The ability to collect and analyze real-world conversations in close to real-time has several benefits for both the language teacher and student. There are many word lists, or corpora, available such as the News on the Web Corpus (NOW), the Corpus of Contemporary American English (COCA), the New General Service List (NGSL), all of which can help better understand the language to inform teaching decisions and focus limited study time (Cobb & Boulton, 2015). However, theme specific corpora can be helpful not only to uncover keywords which may be frequently used within the theme but not more generally, but also to gain a better understanding of societal trends (Fernández-Cruz & Moreno-Ortiz, 2020).

In the case of the Reddit Olympic Corpus, the words *medal*, *Olympics*, *gold*, and *event* are used more frequently than in daily conversation. Additionally, the words *watch*, *see*, and *look* are all within the top-20 most frequently used words. This information can be used as a starting point in crafting an efficient course of study related to the theme, and the addition of a sentiment score provides students with an idea of the words' subtext when in use. For example, the most frequent word, *get*, is used to convey all sentiments but is slightly more frequently used to express neutral or negative sentiments. In contrast, the word *like* is commonly used for positive sentiments, and though it is rarely used for neutral sentiments it is often used to express negative sentiments. It is then possible to go back to the collected comments and find examples of positive, neutral, or negative comments for any word and analyze why they have been classified in a particular manner, and if that classification matches with expectations.

This all helps to make a connection from the theoretical to the real world. Instead of studying in the abstract, or from situations that may have been relevant in previous eras, students are able to gain insight into modern English using real-world information targeted at a specific theme. The source material can be used as reading materials or writing prompts related to the theme. In the case of the Olympic comments, students might be asked to evaluate the sentiment of the comment and compare their score with the VADER sentiment score. If the scores differ substantially, they could ask for a second opinion, try to justify their score, or try to understand why VADER assigned a different score than theirs. The word frequency list can help learners decide when to spend time remembering the meaning of a new word, or guessing the meaning from context or looking it up and moving on would be a more efficient approach. For example, if a comment in the Olympic theme uses the word *dressage*, contextual clues might not offer much help and learners may wonder if it is a common word worth knowing, or if it is an obscure and rarely used word that they are unlikely to see again and should not devote limited study time towards learning. From a quick search of the Reddit Olympic Corpus, they can see the word was used 956 times, making it rank 451 on the word frequency list. Looking at the COCA or NGSL, one will find that *dressage* is not among the top 2000 words, so a learner interested in the Olympics or equestrian would value the word, but for most learners it would not be worth diverting cognitive resources to add it to their vocabulary. More broadly, the use of online textual data offers a valuable resource for language educators to produce relevant study plans that engage students and evaluate learning based on authentic theme-based language. In this case, the theme was the Olympics, which combine sport, culture, and international relations. The same approach can be used for any theme and using any publicly available digital data source.

However, there are negatives to the approach of building a word list from online sources. It is possible to collect comments from a Reddit post (Cuomo, 2022) and create a corpus (Cuomo, 2021), though it does require some comfort with the programming language Python to modify the scripts for purpose. Each online source would require a different method for collecting text data and creating a corpus, adding a layer of complexity, and potentially discouraging the use of new relevant data sources. In addition, it does take time to collect and analyze the data to gain useful insights, as such it is not something to necessarily use in every study plan. Despite the challenges associated with using theme-based word lists for language study, it is an area where data sources and tools are rapidly evolving and becoming more accessible for students and educators. This progress offers an alternative to a reliance on dated textbooks or the limits of the teacher's intuition, which may be shaped and limited by their background.

At the time of writing, GPT-4, an advanced artificial intelligence (AI) language model designed to understand and generate human-like text, promises the ability to access the web and perform data analysis using standard English commands rather than Python or other code (Bubeck et al., 2023). If this capability is implemented, it will dramatically speed up the process of data collection, corpus creation, and visualizing the information in an actionable manner. In addition, the simplified interface would allow educators and students with limited technical prowess to access authentic word lists in near-real time. By capitalizing on these innovative tools, educators can recalibrate their focus and adapt their teaching methods to align with the evolving nature of language use in digital communication. The incorporation of corpora derived from online conversations into language education ensures that both educators and students remain up-to-date with the changing linguistic landscape, ultimately fostering a more effective and engaging learning experience.

Bio Data

Osaze Cuomo works at Hyogo University in the Department of Contemporary Economics. He was born and raised near San Francisco, California, though has spent the last 15 years primarily in Japan, as well as Spain and Thailand. His interests include cross-cultural communication, societal change, and the role of artificial intelligence and digital tools in human communication. <cuomo@hyogo-dai.ac.jp>

References

Baumgartner, J. (n.d.). *Pushshift Reddit search*. Retrieved November 1, 2022, from https://redditsearch.io/?term=olympics

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: early experiments with GPT-4. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2303.12712

Cobb, T., & Boulton, A. (2015). Classroom applications of corpus analysis. *The Cambridge Handbook of English Corpus Linguistics*, 478–497. https://doi.org/10.1017/cbo9781139764377.027

Cuomo, O. (2022, November 11). *Reddit Megathread Comment Collector*. Kaggle. Retrieved from https://www.kaggle.com/osazecuomo/reddit-megathread-comment-collector

Cuomo, O. (2021, December 14). *Reddit Corpus Creator*. Kaggle. Retrieved from https://www.kaggle.com/osazecuomo/reddit-corpus-creator

Dang, T. N. Y., Webb, S., & Coxhead, A. (2020). Evaluating lists of high-frequency words: Teachers' and learners' perspectives. *Language Teaching Research*, 136216882091118. https://doi.org/10.1177/1362168820911189

Fernández-Cruz, J., & Moreno-Ortiz, A. (2020). Building the Great Recession News Corpus (GRNC): A contemporary diachronic corpus of economy news in English. *Research in Corpus Linguistics*, *8*(2), 28–45. https://doi.org/10.32714/ricl.08.02.02

Hargittai, E. (2018). Potential biases in big data: Omitted voices on social media. *Social Science Computer Review*, *38*(1), 10–24. https://doi.org/10.1177/0894439318788322

Horne, B., & Adali, S. (2017). The impact of crowds on news engagement: A Reddit case study. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 751–758. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14977

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *International Conference on Weblogs and Social Media*, *8*(1), 216–225.

Kausar, M. A., Soosaimanickam, A., & Nasar, M. (2021). Public sentiment analysis on Twitter data during COVID-19 outbreak. *International Journal of Advanced Computer Science and Applications*, *12*(2). https://doi.org/10.14569/ijacsa.2021.0120252

Language Processing and Python. (n.d.). Retrieved from https://www.nltk.org/book/ch01.html

Ma, M. C., & Lalor, J. P. (2020). An empirical analysis of human-bot interaction on Reddit. *Proceedings of the Sixth Workshop on Noisy User-Generated Text (W-NUT 2020)*. https://doi.org/10.18653/v1/2020.wnut-1.14

McCulloch, G. (2020). *Because Internet: Understanding the New Rules of Language.* Riverhead Books.

Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying Reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media + Society*, *7*(2), 205630512110190. https://doi.org/10.1177/20563051211019004

Reddit's 2020 Year in Review - Upvoted (2020, December 8). *Upvoted*. Retrieved from https://www.redditinc.com/blog/reddits-2020-year-in-review/

Sentiment analysis: First steps with Python's NLTK Library (2022, September 1). *Real Python.* Retrieved from https://realpython.com/python-nltk-sentiment-analysis/

Reddit usage reach in the United States 2021, by age group. (2021a, June 17). *Statista.* Retrieved from https://www.statista.com/statistics/261766/share-of-us-internet-users-who-use-reddit-by-age-group/

Leading websites worldwide 2021, by monthly visits. (2021b, September 7). *Statista*. Retrieved from https://www.statista.com/statistics/1201880/most-visited-websites-worldwide/