Is Gen Alpha 'Cooked?' Exploring Linguistic Differences Across Generations

36-468 Final Project

Dominick Robinson

2024-12-11

Table of contents

Abstract

This study investigates linguistic differences across three generations (Gen Alpha, Gen Z, and Millennials) by analyzing Reddit comments. Using metrics such as MATTR, average token length, and emoji usage, this study aims to address whether Gen Alpha exhibits a smaller

vocabulary, uses less sophisticated language, and employs more slang compared to older generations. Results revealed statistically significant differences in lexical diversity, with Millennials demonstrating the highest diversity and Gen Alpha the lowest. Average token length did not differ significantly among generations, suggesting similar linguistic complexity. Emoji usage, however, showed a marked generational divide, with Gen Alpha using emojis most frequently and Millennials least. Keyness analysis further highlighted generational preferences for words and slang, particularly in Gen Alpha, reflecting distinct communication styles. These findings challenge the stereotype that Gen Alpha's vocabulary is inferior, suggesting instead that generational differences in language use are more reflective of evolving digital communication trends.

1 Introduction

In the words of George Orwell, "Every generation imagines itself to be more intelligent than the one that went before it, and wiser than the one that comes after it."

As of late, many people, including Millennials and particularly those in Generation Z (or Gen Z), suggest that children in Generation Alpha (or Gen Alpha) have a "limited vocabulary" that consists largely of slang words and phrases, such as "skibidi toilet," "rizz," and "sigma," among others. Parents magazine (2024) discusses the rise of terms like "brain rot," highlighting their association with Gen Alpha's immersion in online culture (*Brain Rot: The Language of Gen Alpha* 2024). There is often a negative sentiment around this idea, where people either look down on Gen Alpha for their peculiar language use, or attribute this quirk and other language developmental issues to growing up during the COVID-19 lockdowns at a critical stage of development and due to excessive technology use. This idea is explored in a video by School for Humanity that questions the impact of screen time and social media on young minds ("Is Gen Alpha Cooked? A Deep Dive into Language and Development" 2024). Others hold a more positive sentiment that Gen Alpha is no different from its preceding generations, all of which developed their own slang. This view is explored in a piece on Medium, which argues that Gen Alpha's unique linguistic creativity reflects the natural evolution of language rather than a deficit (Zeng 2024).

I question the notion altogether that Gen Alpha has an exceedingly small vocabulary consisting of their unique slang. Therefore, I want to answer the following questions:

1. Does Gen Alpha have a smaller vocabulary than older generations?
2. Does Gen Alpha use less sophisticated language compared to older generations?
3. Does Gen Alpha use their particular slang more than older generations?

I will attempt to answer these questions with statistics by analyzing Reddit comments made by Millennials and those in Gen Z and Gen Alpha. This research is important because it will settle the debate on whether Generation Alpha actually has a smaller, less sophisticated vocabulary and uses more unique slang than Gen Z and Millennials.

2 Data

I used comments from from reddit.com, particularly r/millenials, r/GenZ, and r/GenAlpha as data. I chose these particular subreddits because they are perhaps the only ones that can tell me which generation each commenter is from.

To get the data, I used the official Reddit API to get the top 1000, newest 1000, and most controversial 1000 posts from each subreddit because the API limited me to 1000 posts for each call, and then I obtained all comments from each post.

To categorize the comments as being made by a millenial, a Gen Z, a Gen Alpha, or an undeterminable member, I examined the "flair" of each commenter, if they had one. "Flair" is simply a title that a Redditor can give themself. For example, members in r/GenAlpha can choose "Gen Z" or "2011" as their flair. I categorized each comment based on whether the flair of the commenter contained a birth year or generation name. Comments without these flairs were of indeterminable generation.

Table 1: Age Ranges for Millenials

Year Range	Generation
1981-1996	Millenial
1997-2009	Gen Z
2010-2011	Gen Alpha

Table 1 shows the age ranges for each generation, which are suggested by (Gutoskey 2021). While Gen Alpha normally goes up to 2024, it only goes up to 2011 here because children under the age of 13 are not allowed to make Reddit accounts, and so I could not accurately identify comments made by those under 13 years of age. I assigned each comment to a generation (or none)

To clean the data, I did the following:

1. Filtered out comments without a generation
2. Detected and filtered English comments
3. Removed unwanted punctuation
4. Retained only alphanumeric characters and emojis
5. Made comments lowercase
6. Trimmed whitespace
7. Removed empty comments

Table 2: Amount of comments by generation and subreddit. Gen Z posted the most comments in each subreddit, followed by Gen Alpha and Millenials across all subreddits.

Generation	Gen Alpha	Gen Z	Millennials	Subreddit Totals
r/GenAlpha	25086 (13.14%)	34815 (18.24%)	903 (0.47%)	60804 (31.85%)

r/GenZ	1947 (1.02%)	107959 (56.55%)	16744 (8.77%)	126650 (66.34%)
r/millennials	1 (0.0%)	521 (0.27%)	2945 (1.54%)	3467 (1.82%)
Generation Totals	27034 (14.16%)	143295 (75.05%)	20592 (10.79%)	190921 (100.0%)

Table 3: Amount of tokens by generation and subreddit. Gen Z provided the most tokens in each subreddit, followed by Millenials and Gen Alpha across all subreddits.

Generation	Gen Alpha	Gen Z	Millennials	Subreddit Totals
r/GenAlpha	352537 (6.52%)	568466 (10.51%)	27620 (0.51%)	948623 (17.54%)
r/GenZ	35150 (0.65%)	3584746 (66.29%)	703063 (13.0%)	4322959 (79.94%)
r/millennials	5 (0.0%)	17611 (0.33%)	118628 (2.19%)	136244 (2.52%)
Generation Totals	387692 (7.17%)	4170823 (77.13%)	849311 (15.71%)	5407826 (100.0%)

After applying the aforementioned procedure to obtain, categorize, and clean the data, I was left with 190,921 comments consisting of 5,407,826 tokens. More details can be seen about each count in Table 3 and Table 2, respectively.

3 Methods

To investigate the difference in vocabulary range between the generations, I used MATTR (Moving Average Type-Token Ratio) from (Brezina 2018), which is a measure of lexical diversity that calculates the average ratio of unique tokens to total tokens over a moving window. Simply, this will calculate the total number of unique words being used, which is, by definition, the expansiveness of a demonstrated vocabulary. I chose MATTR over TTR (Type-Token Ratio) because TTR is sensitive to differences in sizes of the corpuses being compared, while MATTR specifically accounts for differences using a sliding window. I chose a sliding window of size 1000 tokens.

To investigate the difference in linguistic complexity, I tested whether there is a statistically significant difference in the average length of tokens used (longer average token length can indicate more complex words being used more often) between each generation. I chose this metric over others because online comments, such as those found on Reddit, often lack punctuation (e.g., commas), which leaves certain analyses that depend on correct punctuation off the table. (Brezina 2018) also recommends this method for testing linguistic complexity.

Lastly, to investigate the prevalence of slang in each generation, I tested whether there is a statistically significant difference in the proportion of emojis that each generation used in order to explore a potential difference between generations in this expressive style, which could signal a shift if it exists. I also perform keyness keyword analysis to investigate the words and possibly slang that is unique to each generation.

4 Results

4.1 Keyness by Generation

feature	chi2	p	n_target	n_reference
millennial	1152.66	0	637	624
millennials	574.36	0	641	1113
trump	215.41	0	1078	3485
to	195.24	0	23395	113610
boomers	176.57	0	407	1020
their	166.70	0	3009	12424
millenial	125.07	0	92	118
the	113.09	0	31143	156494
folks	107.64	0	164	340
vote	100.70	0	691	2411
genz	98.81	0	263	697
were	87.07	0	2446	10645
recession	85.15	0	77	116
they	83.82	0	5859	27550
congress	82.69	0	106	200

Figure 1: Keyness metrics for Millenials. Millenials seem to talk more about 'Trump,' 'Boomers,' 'recession,' and 'Congress' more than Gen Alpha and Gen Z. Interestingly, Millenials said 'folks' more than the other generations, seeming to suggest that it is a slang word for them.

feature	chi2	p	n_target	n_reference
countries	71.90	0	1423	234
lmao	62.83	0	2015	386
trans	58.17	0	1453	259
college	56.82	0	2247	453
government	55.16	0	2263	460
should	50.06	0	5324	1266
uk	48.73	0	416	42
people	47.17	0	25540	6901
sex	45.79	0	1766	354
not	45.67	0	32081	8771
politics	44.59	0	1834	373
	42.34	0	948	163
american	38.30	0	1917	406
woman	37.78	0	1099	205
having	37.20	0	2989	685

Figure 2: Keyness metrics for Gen Z. Gen Z talked more about 'countries,' 'trans,' 'college,' 'sex,' 'UK,' 'American,' and more frequently used the crying laughing emoji than Gen Alpha and Millenials.

feature	chi2	p	n_target	n_reference
alpha	4087.86	0	1518	3462
gen	2961.13	0	2773	12117
skibidi	2055.13	0	609	1119
downloader	1948.11	0	150	0
i	1603.25	0	12956	116897
toilet	1457.68	0	480	979
	1280.41	0	679	2065
brainrot	1223.32	0	359	653
bro	1001.59	0	645	2246
dmca	974.04	0	75	0
savevideo3232donate	974.04	0	75	0
savevideomessage	974.04	0	75	0
furries	956.86	0	258	429
	934.97	0	234	358
homework	887.04	0	152	140

Figure 3: Keyness metrics for Gen Alpha. Gen Alpha used words like 'skibidi,' 'toilet,' 'brainrot,' 'bro,' 'furries,' and more frequently used the skull and fire emojis more than Gen Z and Millenials.

Figure 1, Figure 2, and Figure 3 display the keyness information. Millenials seem to talk more about Trump, Boomers, recession, and Congress more than Gen Alpha and Gen Z. Interestingly, Millenials said "folks" more than the other generations, seeming to suggest that it is a slang word for them. Gen Z talked more about countries, trans, college, sex, UK, American, and the crying laughing emoji than Gen Alpha and Millenials. Gen Alpha used words like skibidi, toilet, brainrot, bro, furries, and the skull and fire emojis more than Gen Z and Millenials. These are all common slang words commonly and stereotypically associated with Gen Alpha.

Each generation seemed to talk about themselves or other generations .

4.2 Metrics by Generation and Subreddit

Generation	MATTR	Avg Token Length	Emoji Proportion
Gen Alpha	0.448960	4.175915	0.003913
Gen Z	0.447041	4.348474	0.001512
Millennial	0.467708	4.423748	0.000712

Figure 4: MATTR, average token length, and proportion of tokens that were emojis for each generation.

As shown in Figure 4, MATTR and average token length increases from youngest to oldest generation, while proportion of emojis used descreases from youngest to oldest generation.

The mean MATTR, calculated with a window size of 1000, was 0.448 (SD = 0.084504) for Gen Alpha, 0.447 (SD = 0.104411) for Gen Z, and 0.468 (SD = 0.113258) for Millennials. The mean average token length was 4.18 (SD = 3.728583) for Gen Alpha, 4.35 (SD = 3.728583) for Gen Z, and 4.42 (SD = 2.429536) for Millennials. The mean emoji proportion was 0.0039 (SD = 0.036681) for Gen Alpha, 0.0015 (SD = 0.025900) for Gen Z, and 0.0007 (SD = 0.018214) for Millennials.

Metric	Test Statistic	p-value
MATTR	F(2, df) = 2261.27	<.0001
Avg Token Length	F(2, df) = 1.34	.262
Emoji Proportion (Chi-Square)	^{2}(2, N = 5,415,826) = 1792.18	<.001

Figure 5: MATTR, average token length, and proportion of tokens that were emojis for each generation. MATTR and emoji usage showed statistically significant differences

An ANOVA was conducted to examine differences in average token length across generations. The results were not statistically significant, $F(2,df)=1.34$,$p=.262$, suggesting that generation does not significantly influence average token length.

A Chi-Square test was conducted to determine if there is a significant difference in emoji usage proportions among generations (Gen Alpha, Gen Z, and Millennials). The results showed a statistically significant difference in emoji usage among generations, $2(2,N=5{,}415{,}826)=1792.18$, $p<.001$. Emoji usage decreased as age of generation increased.

5 Discussion

5.1 Main Findings

This analysis explored the differences in vocabulary size, complexity, and emoji usage across three generations—Gen Alpha, Gen Z, and Millennials—using Reddit comments as data. Gen Alpha exhibited the lowest lexical diversity (MATTR = 0.448), while Millennials had the highest (MATTR = 0.468). The results of the ANOVA confirmed statistically significant differences in MATTR across generations, suggesting that Millennials demonstrate a slightly broader vocabulary in their online communication. Differences in average token length were minimal and statistically insignificant. This finding suggests that the length of words used online does not vary significantly by generation. Emoji usage significantly decreased as the age of the generation increased. Gen Alpha used emojis at the highest rate (0.39%), followed by Gen Z (0.15%), and Millennials (0.07%). A Chi-Square test revealed significant differences in emoji usage among generations, indicating that younger generations favor emojis more in their communication.

Alternatively, Gen Alpha did use the slang stereotypically associated with them (e.g., "skibidi," "toilet," "bro," "brainrot") statistically significantly more than the other generations.

5.2 Implications

These findings support the stereotype that Gen Alpha possesses a "limited vocabulary." However, while they exhibit slightly lower lexical diversity, this does not necessarily imply a lack of linguistic sophistication. On the contrary, the lack of statistically significant differences in average token length fails to support the notion that Gen Alpha exhibits less linguistic sophistication compared to their older counterparts. Further, the high use of emojis and unique slang terms reflects a shift in communication style rather than a reduction in linguistic sophistication. Besides, it was perhaps to be expected that older generations have slightly higher lexical diversity, as they have lived longer and had more opportunities to learn more words.

5.3 Limitations and Next Steps

There are a few limitations in the data to be wary of, however. First, it may be the case that certain keywords appeared because they were the topic for discussion in the subreddits. For

example, keywords relating to generation (e.g., "boomers," "millenials") appeared, likely because they were topics for discussion in a subreddit about generations. Similarly, it may be the case that slang commonly associated with Gen Alpha appeared because Gen Alpha Redditors were discussing stereotypes about the slang use. This issue could be dealt with by obtaining a larger sample size of comments, because it would ensure that more commenters from different generations are talking commenting on posts about these topics in each subreddit.

Another limitation was the inability to identify posts by children under the age of 13. This leaves only a 2 year window out of 14 possible years for Gen Alpha, which severely limits the information about Gen Alpha. If this study is replicated in a few years, a greater proportion of Gen Alpha will be eligible, which could reveal new insights.

Lastly, the data being gathered from Reddit limited the options available for measuring linguistic sophistication. Because Reddit comments often lack correct punctuation, certain methods recommended by (Brezina 2018) could not be used. Gathering data from other sources may allow for an expanded analysis.

6 Acknowledgments

I used ChatGPT to help with coding. I found it helpful because I am not the best at R. All ideas and interpretations of results are my own.

Works Cited

Brain Rot: The Language of Gen Alpha. 2024. https://www.parents.com/brain-rot-2024-oxford-word-of-the-year-8754608?utm_source=chatgpt.com.

Brezina, V. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge University Press.

Gutoskey, Ellen. 2021. "What Are the Age Ranges for Millennials and Generation z?" 2021. https://www.mentalfloss.com/article/609811/age-ranges-millennials-and-generation-z.

"Is Gen Alpha Cooked? A Deep Dive into Language and Development." 2024. https://www.youtube.com/watch?v=pHRVrl4Wr7Q.

Zeng, Forrest. 2024. "Gen Alpha Is Not Cooked." https://medium.com/%40forrestzeng/gen-alpha-is-not-cooked-98abf170eda4.