Final Report

Juliane Hanel
799251
Potsdam University
hanel@uni-potsdam.de

Advanced Natural Language Processing (WS 2018/2019)
Dr. Tatjana Scheffler

03. March 2019

Introduction & Motivation

Together with Atreya Shankar and Luis Glaser, I worked on the Advanced Natural Language Processing (ANLP) final project. As avid music fans, we wanted our project to revolve around this general topic and decided to focus on lyrics of popular music pieces published within the past thirty years. Popular music is a part of pop culture, a term that is defined as commercial culture based on popular taste (Oxford Pocket Dictionary, 2009, Pop culture). To us it seemed interesting to look for changes as well as continuity in lyrics of popular songs, as they reflect the general sentiment of the popular taste. Song lyrics are a good medium for sociocultural analysis since most members of society are free to contribute to pop music and can publicly address topics that affect them. An example for this are rap songs:

> “Through rap, many African-American youths are seizing the power to form their own identities and to project them, through the popular media, to a wide audience.“
> (Saunders, 1993, 23)

Song lyrics can express something about a collective mindset or feeling, the societal situation and also the identity of a group of people, majorities as well as minorities, at the time the songs were published.

One of our main inspirations for the project topic was the paper Lyrics-based Analysis and Classification of Music by Fell and Sporleder, 2014. In this paper, Fell and Sporleder presented a novel approach for analysing and classifying music. They chose an exploratory approach and experimented with features, modelling semantic and stylistic properties of song lyrics, and used these features for the three classification tasks: genre detection, distinguishing the best and the worst songs, and determining the approximate publication time of a song. The results showed that lyrics-based feature mining deemed to be relevant for the classification tasks, indicating that the evolution of genres can be observed based on features extracted from song lyrics (Fell and Sporleder, 2014).

In our project, we looked at song lyrics of popular songs of the three major genres: Hip Hop, Pop, and Country. Taking an exploratory approach, we extracted shallow textual features indicating lexical richness and overall complexity of the lyrics, such as *average word length* and *Type-token ratio* (*TTR*) (Baker et al., 2006), as well as more sophisticated features like the *emotional sentiment* of songs. Furthermore, we analysed part-of-speech (PoS) tags of the lyrics, hoping to find tendencies like *egocentrism* by looking at first- and second-person singular pronouns. We also analyzed the *frequency of non-standard words* and extracted the *most frequent words*. Afterwards, we performed a change analysis in order to determine whether significant intragenre changes with respect to the features can be found.

Research Hypotheses

There exists anecdotal evidence that Country music converged to Pop music by incorporating its features.[1][2][3] Looking at the change of *TTR* and *average word length*, we determined whether this shift is evident in our dataset. Lower *Type-token ratio* would for example be caused by repetition, a stylistic device often used in Pop music.

The second hypothesis is derived from Nathan Dewall et al. (2011). A research question in their paper was "Are U.S. Song Lyrics Becoming More Self-Focused Over Time" and their findings revealed that it was indeed the case. We tested this hypothesis by looking at the ratio of first-person to second-person pronouns.

Furthermore, Powell-Morse (2015) found that songs lyrics from all genres became less sophisticated over the course of the past decade. For his analysis, he used tools such as the Flesch-Kincaid grade index[4] and the Gunning Fog index[5]. We want to investigate if this decline in lexical diversity can also be found in our data. We extended the original research by broadening the time frame of the analysis and using the aforementioned measures of *Type-token ratio* and *average word length*.

In summary, we investigated the following three hypotheses for the period 1990-2019:

1. Country music has incorporated mainstream Pop music features.
2. Pop music became more egocentric.
3. Popular music of all genres became less lexically diverse.

Material

Due to copyright issues, finding a feasible dataset was the first major challenge we encountered. The dataset we planned on using was part of a `Kaggle` challenge[6] and thus publicly available. Unfortunately this dataset is very unbalanced, with most of the lyrics stemming from early 2000s. Consequently, we have created an original method to collect our own dataset. The data mining procedure was undertaken by Luis. Using Wikipedia lists of artists of the respective genres [7], we queried the name of each artist listed on the website Genius.com[8], a website where users can add song lyrics. The resulting information (`Genius ID, song title, artist, lyrics, year, genre, Genius URL`) was stored in a `SQLite` database. After scraping, we gathered a total of 399,342 songs. However, they included many duplicates, non-English lyrics or had incomplete metadata. Moreover, some of them where in fact not lyrics but excerpts from other online publications like a ”song” by Slavoj iek and crossword puzzles from the New York Times. Due to the inconsistency of the collected lyrics, it became evident that we had to employ heuristics to clean the data. Our research focused on English song lyrics, so we used language detection[9] to filter out songs in other languages. Furthermore, we removed duplicate songs and aligned the different annotations users made. In preparation for the feature extraction, we PoS-tagged the lyrics. Further cleaning was performed after the first run of the feature extraction, as we noticed that the language detection did not remove all foreign songs. Since the Non-Standard Word percentage indicates the percentage of words not listed in the English dictionary, we decided that we can remove all songs with more than 40% non-standard words without losing valuable data. Some of the song lyrics were also not listed yet, instead they contained In Progress or similar formulations and track lists of upcoming albums. After manual inspection, we found that these are the songs with a *Type-token ratio* of more than 85% and removed them.

[1] https://abcnews.go.com/WNT/story?id=130057&page=1

[2] https://www.npr.org/sections/therecord/2018/03/20/594037569/how-the-sound-of-country-music-changed

[3] https://people.com/country/locash-country-music-fans-more-open-minded/

[4] https://www.usingenglish.com/glossary/flesch-kincaid-index.html

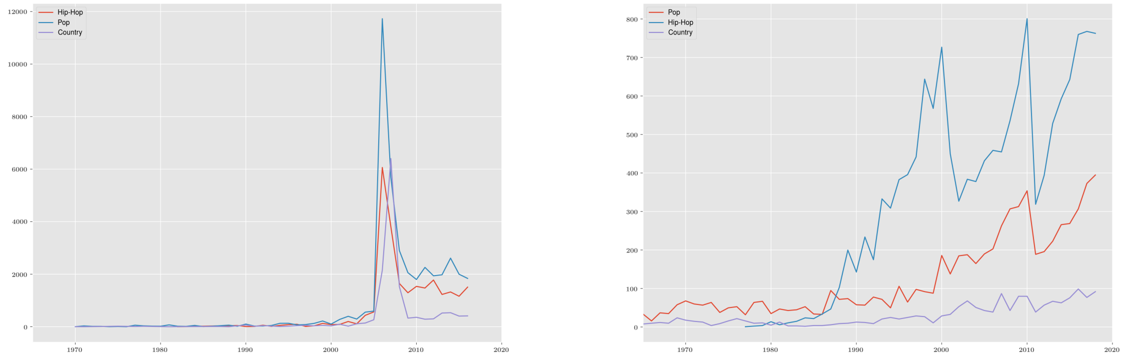
[5] http://gunning-fog-index.com/

[6] https://www.kaggle.com/rakannimer/billboard-lyrics

[7] e.g. https://en.wikipedia.org/wiki/List_of_hip_hop_musicians

[8] https://www.genius.com

[9] https://pypi.org/project/langdetect/



(a) Data distribution of Kaggle dataset.

(b) Data distribution of our dataset.

Approach

Once we had gathered the lyrics and stored them in the database, the individual tasks were performed. For both the feature extraction and the statistical analysis, the programming language `Python` was used. In order to explore the outlined features, the tasks were divided between Atreya and myself. Atreya gathered information about the *most frequent words* and the *sentiment* of song lyrics. I focused on four different features, the first one being the *Type-token ratio* of a song, which is the proportion of unique words to all words. The second feature was *average word length*. The next feature I handled was the *egocentrism* scale. For this, I calculated the ratio of first- to second-person singular pronouns in the lyrics. The higher the resulting number, the more egocentric a song. The last feature I investigated was the *non-standard word* ratio. A non-standard word is a word that does not appear in the English dictionary. In order to obtain this ratio, I divided the number of non-standard words in a song by the number of total words. Information about whether a word exists or not was obtained using the check function from the module PyEnchant[10]. The obtained information was saved in the `SQLite` database.

Statistical Evaluation

I was responsible for performing the change analysis of the numerical features `Average Word Length, Non-Standard Words, Type-token Ratio, Sentiment, Egocentrism.` For the evaluation, we split the data into three subgroups and tested for significant changes in adjacent groups (Figure 3).

T1	T2	T3
≤ 2000	≥ 2001 and ≤ 2010	≥ 2011

Figure 2: Publication years for different time steps

The first step of the evaluation was a one-way analysis of variance (ANOVA). ANOVA reveals whether there are statistically significant ($p \leq 0.05$) differences between the means of three or more independent groups (J. Quirk, 2016). This method requires post-hoc testing since the results only indicate that a significant difference exists, not which specific groups differed. For post-hoc testing, we decided to perform one-tailed Welch t-tests since we were interested in the affected groups as well as the effect direction (Lu and Yuan, 2010). Consequently, I performed a one-tailed Welch t-test between group T1 and T2 as well as between group T2 and T3 for the features that ANOVA found to be significantly different. If the test revealed a significant effect ($p \leq 0.05$), the polarity of the t-value gave indication of the size of the observed effect, a negative t-value indicating Group 1 has a smaller mean than Group 2 and a positive t-value indicating the opposite. Furthermore, I calculated the effect size for the significant

[10] https://pypi.org/project/pyenchant/

results. The effect size is needed because we were interested not only whether the differences between two groups are significant but also how big these differences are.

> “With a sufficiently large sample, a statistical test will almost always demonstrate a significant difference, unless there is no effect whatsoever, that is, when the effect size is exactly zero; yet very small differences, even if significant, are often meaningless.“
> (Sullivan and Feinn, 2012)

The effect size measure we used is Cohen‘s d (APA Dictionary of Psychology, 2007, Cohen’s d).

The ANOVA revealed significant changes in 11 out of 15 genre/feature combinations. The post-hoc tests for correlations between T1 and T2 showed that 11 out of 15 features had a significantly different mean between the two time steps: 2 out of 5 in country, 5 out of 5 in Hip Hop, and 3 out of 5 in Pop (Figure 4). However, Cohen‘s d measures only indicated a significant effect size ($d \geq 0.2$), for 3 correlations. Post-hoc testing for the time steps T2 and T3 indicated significant changes for 9 features and significant effect size for 2 correlations (Figure 5).

	Average Word Length	Non-Standard Words	TTR	Sentiment	Egocentrism
Pop	***	***	***†	***	/
Country	/	/	***†	**	/
Hip Hop	***	***†	**	***	***

Figure 4: Intragenre differences for measured features between T1 and T2

	Average Word Length	Non-Standard Words	TTR	*sentiment*	Egocentrism
Pop	/	**	***	/	/
Country	/	/	***†	**	/
Hip Hop	***	***	***†	***	***

Figure 5: Intragenre differences for measured features between T2 and T3

* = $p \leq 0.05$
** = $p \leq 0.01$
*** = $p \leq 0.001$
† = Cohen‘s $d \geq 0.02$

A detailed overview and visualizations of the presented results can be found in the Appendix along with visualizations of Atreya’s Frequent Word Analysis.

Interpretation

Hypothesis 1: *Country music has incorporated mainstream Pop music features.*
The relevant features for this hypothesis are *Type-token ratio* and *sentiment*, as they reflect the complexity and the mood of a genre. The mean of the *TTR* for pop song lyrics is generally lower than the one for Country songs at all time steps. However, the decline is more salient in Country music - there is a significant decrease from T1 to T2 and also from T2 to T3, indicating a constant deterioration of *TTR* over the past decades. In addition, significant change ($p \leq 0.01$) to a more negative *sentiment* can be found in all time steps, indicating that Country music shifted away from its traditional mood. However, Cohen‘s d indicated that the effect size for this finding is not large ($d \leq 0.02$). Regarding the other features, the genre Country does not show significant changes. Nonetheless, the decline in *Type-token ratio* in all time steps partly backs up the hypothesis that Country lyrics started to incorporate common Pop music features, as the lyrics became less complex (and thus more suitable for the mass) over time (Figure 6).

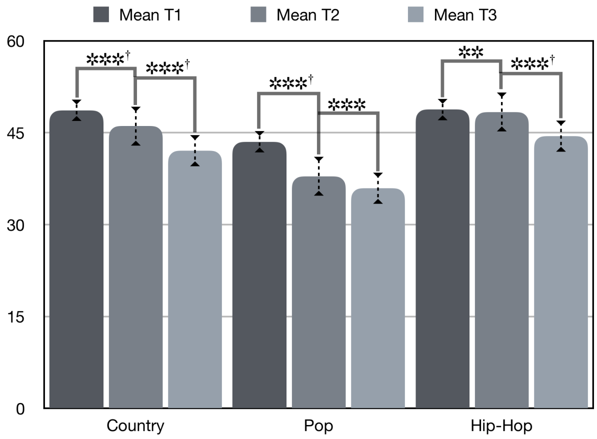


Figure 6: Results of change analysis of *TTR* in percent for all genres

Hypothesis 2: *Pop music became more egocentric.*
According to our findings, Pop music did not become more egocentric. In fact, the means of all three time steps were roughly the same, the statistical testing did not reveal that they differed significantly from each other (Figure 7). Thus, we have to reject the hypothesis and assume that, contrary to anecdotal evidence, Pop music lyrics did not start to incorporate more first- than second-person singular pronouns.

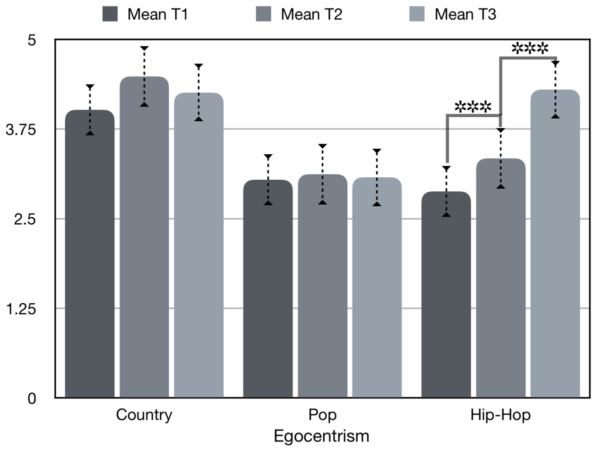


Figure 7: Results of change analysis of Egocentrism ratio in percent for all genres

Hypothesis 3: *Popular music of all genres became less lexically diverse.*
The feature most relevant for this hypothesis is the *Type-token ratio*. Another feature that can be considered for this hypothesis is the *average word length*.
Our findings show significant decrease in *TTR* in all genres. The genre Country shows changes in all time steps, while Pop changes significantly from T1 to T2 and Hip Hop from T2 to T3. A visualization can be found in Figure 6. Regarding the *average word length*, both Pop and Hip Hop show a significant decrease, which seems salient but cannot be considered due to the small effect size. Consequently, the feature *average word length* cannot contribute to the hypothesis evaluation. However, the overall negative change in *Type-token ratio* in all genres indicates that lyrics of all genres became less lexically rich - the vocabulary became less varied and thus less sophisticated over time. This corresponds to the Andrew Powell-Morse‘s findings, who reported that song lyrics became less sophisticated in the last decade (Powell-Morse, 2015). Our findings also expand Powell-Morse‘s results, since we are looking at a different measure of lexical richness and extract information from three decades of popular music instead of just one, showing that the downward trend in sophistication has been existing at least since the 1990s.

Reflection

We were able to back up two out of three initial hypotheses. Our findings suggest that the evolution of music genres can be observed by means of feature extraction from song lyrics. Due to the limited scope of our project, we were only able to investigate six features in three major genres with respect to their evolution in the past 30 years. However, extending the research by evaluating more features, more genres or a larger time frame might bring a valuable insight to the evolution of popular music culture.

As for my learning outcome, I came to realize that the programming itself is just a small part of the actual project work. Deciding on a feasible topic, finding related research and formulating hypotheses was the first major step. When Atreya, Luis and I decided on the project idea, we had many ideas and expectations. Finding a common denominator and not being overly ambitious while doing so was a good experience. The fact that we all come from a different academic background lead to fruitful discussions and different perspectives. Whenever we faced challenges, e.g. during data collection, we were able to quickly resolve the problems both from the technical as well as from the interpersonal side.

I learned that communication is the key to successful group work. Being able to discuss ideas and problems with my group members helped me immensely. At no point I hesitated to ask twice if I did not understand something, which was a very good feeling. This good working environment made organizational issues like splitting up the work and setting up deadlines for sub-tasks easy and enjoyable. Before starting the project, I was not familiar with the version control system `Git`. With the help of Atreya and Luis, I quickly learned the commands and advantages of this system and realized why it is such a valuable skill to have.

I also learned things about managing big datasets and how to work with `SQLite` databases. The programming routine and experience I gained while working on the project is also a good learning outcome. Since the work was divided between all of us and the tasks partly depended on each other, I had to pay even more attention to code design and documentation than usual. This taught me not to be negligent and I will keep up this routine for my future programming assignments, even if I do not work in a group. Being responsible for the statistical analysis and thus for choosing the appropriate tests was a good reminder of the knowledge about statistical methods I acquired during my Bachelors degree. Determining significant changes in the data and visualizing them was a challenging but worthwhile task, as the visualization of the final results speaks for itself and makes the findings easily comprehensible. Thanks to our exploratory approach, my group was able to make use of several techniques presented in the ANLP class, ranging from text processing to sentiment analysis. This facilitated our work greatly and it was very rewarding to be able to combine our acquired knowledge from the ANLP class and put it into practice. All the aforementioned things lead to the conclusion that this hands-on experience of designing and implementing a project from scratch a valuable experience. Together with Atreya and Luis, I did not only manage to fulfill my learning goals but even went beyond this and mastered unforeseen issues easily.

References

APA Dictionary of Psychology. 2007. *APA Dictionary of Psychology*. American Psychological Association.

Paul Baker, Andrew Hardie, and Tony McEnery. 2006. *A glossary of corpus linguistics*. Edinburgh University Press.

Michael Fell and Caroline Sporleder. 2014. Lyrics-based analysis and classification of music. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 620–631, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Thomas J. Quirk. 2016. *One-Way Analysis of Variance (ANOVA)*. Springer.

Zhenqiu Lu and Ke-Hai Yuan. 2010. *Welch's t test*. Thousand Oaks, CA: Sage.

C Nathan Dewall, Richard Pond, Jr, W. Keith Campbell, and Jean Twenge. 2011. Tuning in to psychological change: Linguistic markers of psychological traits and emotions over time in popular U.S. song lyrics. *Psychology of Aesthetics Creativity and the Arts*, 5:200–207.

Oxford Pocket Dictionary. 2009. *The Oxford Pocket Dictionary of Current English (4th Edition)*. Oxford University Press.

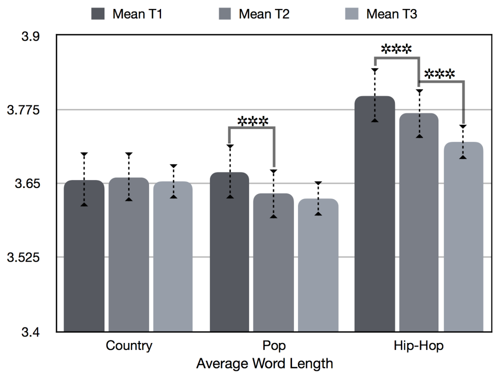
Andrew Powell-Morse. 2015. Lyric intelligence in popular music: A ten year analysis.

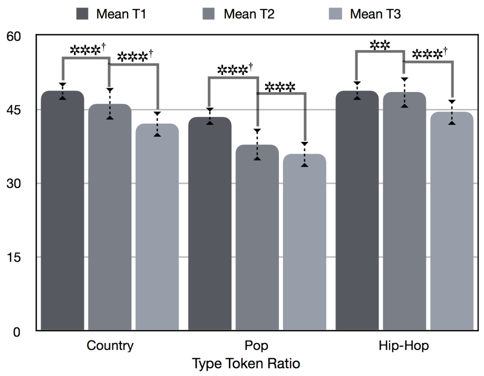
Ralph Saunders. 1993. *Kickin' Some Knowledge: Rap and the Construction of Identity in the African-American Ghetto*, volume 10. Arizona Anthropologist.

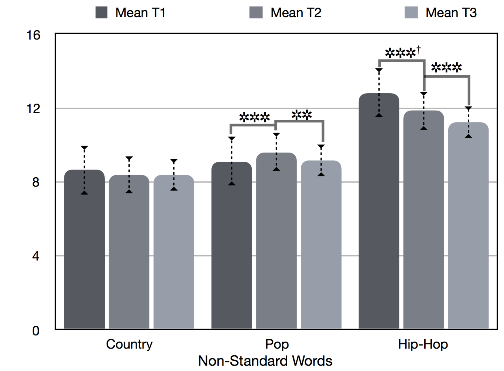
Gail Sullivan and Richard Feinn. 2012. Using effect size or why the P value is not enough. *Journal of graduate medical education*, 4:279–82.

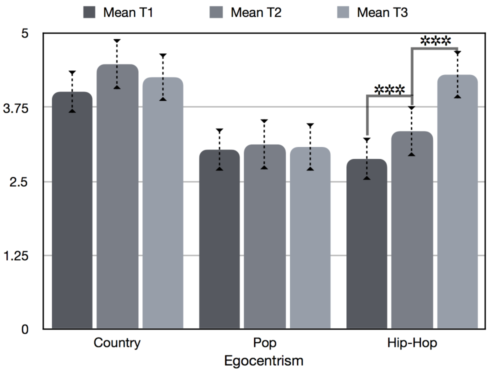
Appendix

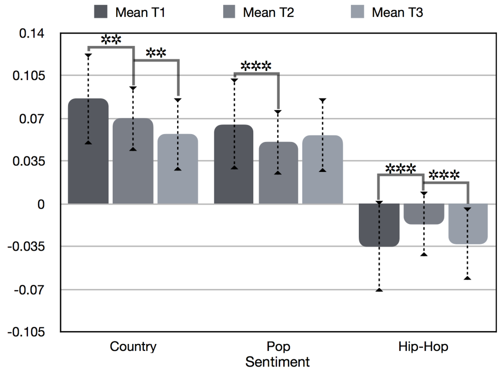
Visualization of Statistical Analysis









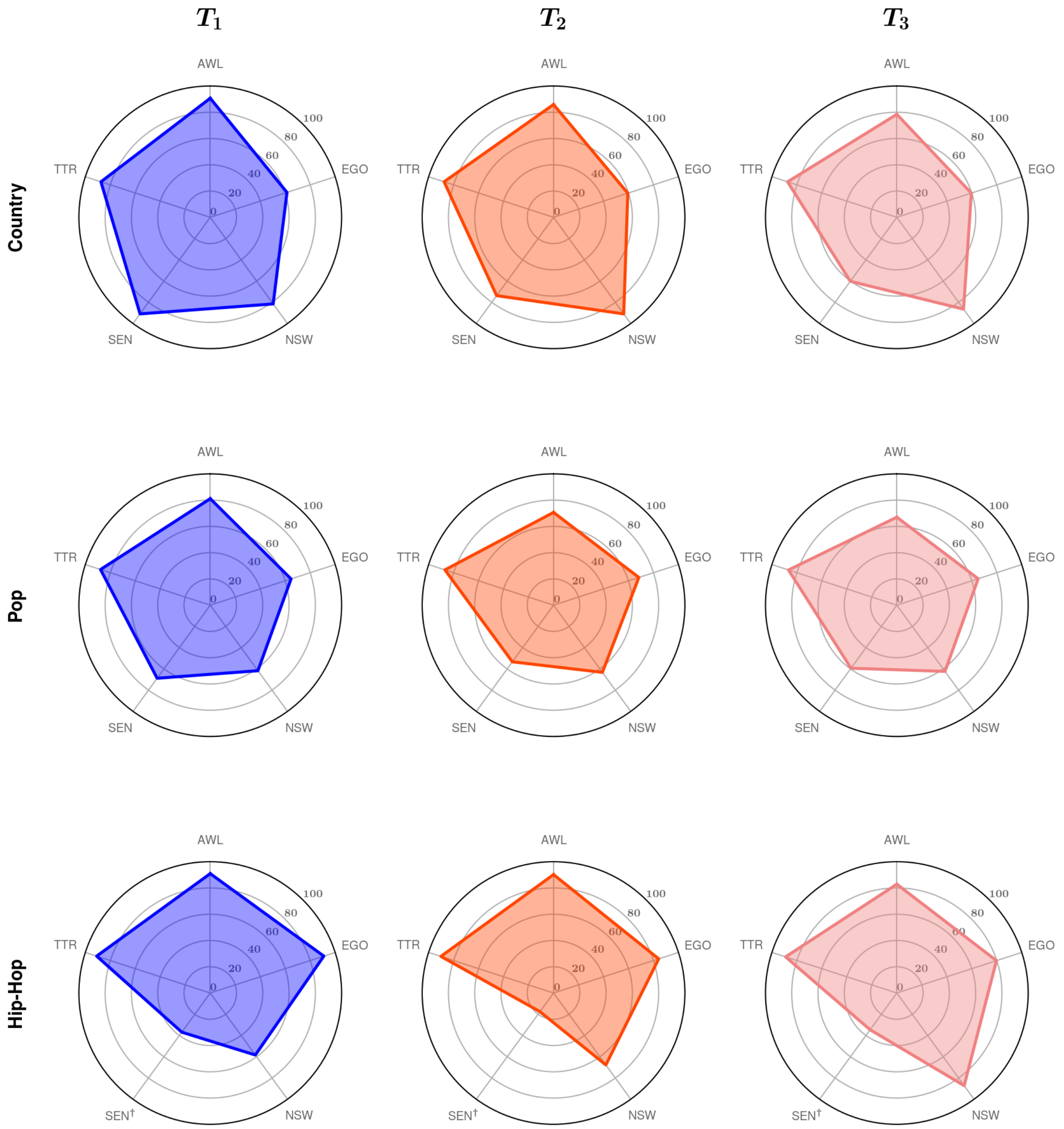


** = $p \leq 0.05$
** = $p \leq 0.01$
*** = $p \leq 0.001$
† = Cohen‘s $d \geq 0.02$

Results of Statistical Analysis

Genre	Group Size	Feature	ANOVA $F-$value	ANOVA $p-$value	$T_1 \rightarrow T_2$			$T_2 \rightarrow T_3$		
					$t-$value	$p-$value	$d-$value	$t-$value	$p-$value	$d-$value
Country	572	AWL	0.107	8.98e−01	-0.294	3.84e-01	-0.0174	0.459	3.23e-01	0.02716
		TTR	51.885	**1.33e-22**	3.831	**6.72e-05**	**0.2265**	6.314	**1.95e-10**	**0.37333**
		NSW	0.627	5.35e-01	1.021	1.54e-01	0.0603	-0.108	4.57e-01	-0.00636
		EGO	0.680	5.07e-01	-1.205	1.14e-01	-0.0713	0.536	2.96e-01	0.03170
		SEN	9.445	**8.33e-05**	2.361	**9.19e-03**	0.1396	2.093	**1.83e-02**	0.12377
Pop	2272	AWL	12.925	**2.50e-06**	3.884	**5.22e-05**	0.1152	0.916	1.80e-01	0.02718
		TTR	218.898	**7.30e-93**	14.839	**5.76e-49**	**0.4403**	5.363	**4.30e-08**	0.15911
		NSW	6.396	**1.68e-03**	-3.289	**5.07e-04**	-0.0976	2.861	**2.12e-03**	0.08489
		EGO	0.104	9.01e-01	-0.466	3.21e-01	-0.0138	0.259	3.98e-01	0.00769
		SEN	8.615	**1.83e-04**	4.197	**1.38e-05**	0.1245	-1.635	5.11e-02	-0.04850
Hip-Hop	4835	AWL	109.494	**6.35e-48**	5.786	**3.72e-09**	0.1177	8.973	**1.71e-19**	0.18250
		TTR	238.885	**8.43e-103**	1.885	**2.97e-02**	0.0383	17.225	**8.28e-66**	**0.35033**
		NSW	141.212	**1.83e-61**	10.183	**1.57e-24**	**0.2071**	6.594	**2.25e-11**	0.13411
		EGO	97.936	**5.64e-43**	-4.972	**3.38e-07**	-0.1011	-8.517	**9.48e-18**	-0.17321
		SEN	41.440	**1.13e-18**	-8.574	**5.76e-18**	-0.1744	7.359	**1.01e-13**	0.14966

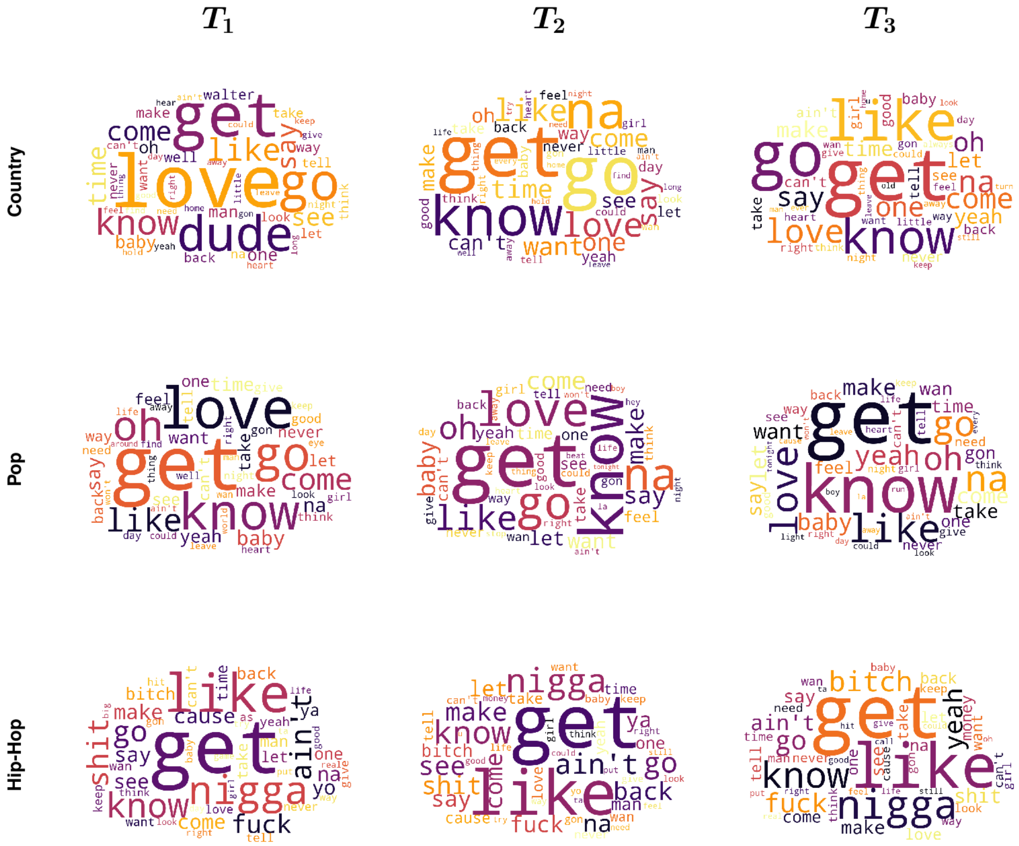
Normalized Visualization of Statistical Analysis



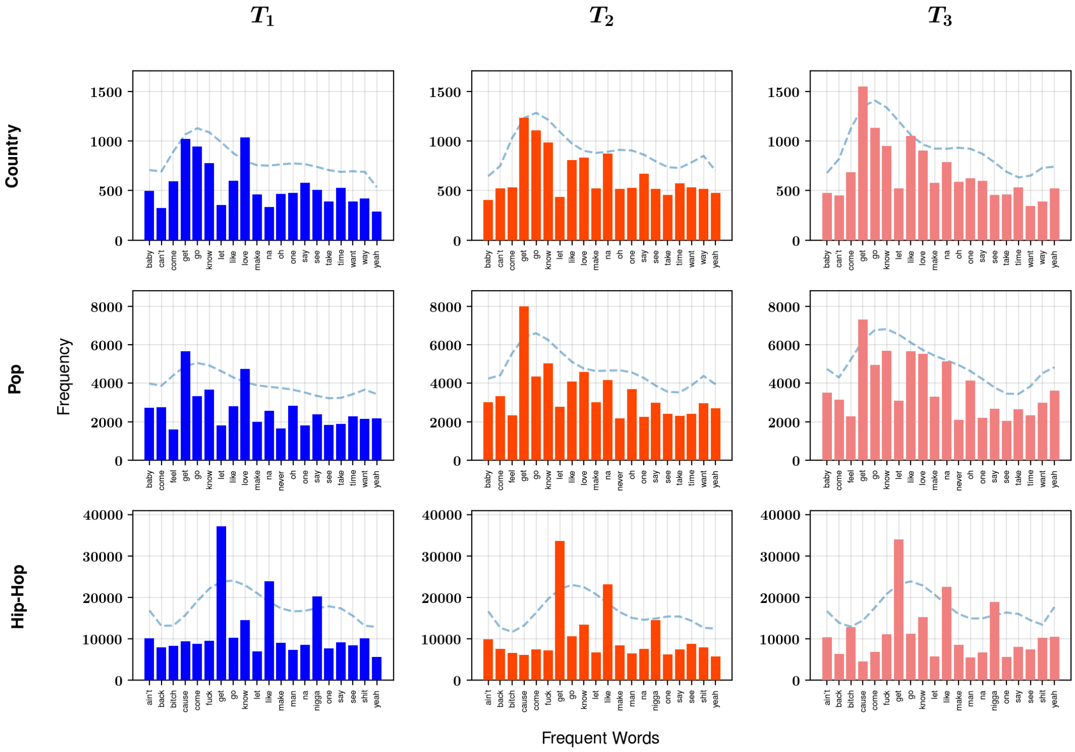
SEN	Sentiment Value
SEN†	Sentiment Value (negative)
TTR	Type-token Ratio
EGO	Measure for Egocentrism
NSW	Non-standard Words
AWL	Average Word Length

Legend for spider plot.

Results of Frequent Word Analysis



Word clouds of the top 50 most frequent words for each genre.



Visualization of the frequency of the top 20 most frequent common-words across the time steps.