QUANTIFYING LEXICAL NOVELTY IN SONG LYRICS

Robert J Ellis, Zhe Xing, Jiakun Fang, and Ye Wang

School of Computing, National University of Singapore

`{robjellis, xingzhe.cs}@gmail.com; {fangjiak, wangye}@comp.nus.edu.sg`

ABSTRACT

Novelty is an important psychological construct that affects both perceptual and behavioral processes. Here, we propose a lexical novelty score (LNS) for a song's lyric, based on the statistical properties of a corpus of 275,905 lyrics (available at www.smcnus.org/lyrics/). A lyric-level LNS was derived as a function of the inverse document frequencies of its unique words. An artist-level LNS was then computed using the LNSs of lyrics uniquely associated with each artist. Statistical tests were performed to determine whether lyrics and artists on Billboard Magazine's lists of "All-Time Top 100" songs and artists had significantly lower LNSs than "non-top" songs and artists. An affirmative and highly consistent answer was found in both cases. These results highlight the potential utility of the LNS as a feature for MIR.

1. INTRODUCTION

From 2004 through 2013, both U.S. and worldwide Google searches for "lyrics" outnumbered searches for "games", "news", and "weather", as computed by Google Trends[1]. The importance listeners place on song lyrics has motivated several explorations for translating a song's lyric into queryable features: for example, by topic [1], genre [2], or mood [3–6]. All these cited examples have incorporated *word frequency* information: as a key statistic in the computational process. The inverse document frequency (IDF) statistic, for example, is used to identify "diagnostic" terms within a lyric that can be further related to a particular topic, genre, or mood.

In the present paper, we propose using IDF information to derive a quantifiable and queryable feature of song lyrics: a *lexical novelty score* (LNS). "Lexical" refers to properties of individual words, as distinct from their grammatical function or syntactical arrangement. Our LNS is based, in part, on the trimean of IDFs associated with the set of unique words in a lyric. The greater the number of statistically infrequent (i.e.,

"novel") words in a lyric, the higher its IDF trimean.

Why might such a quantification of lexical novelty be useful? A number of answers emerge from the domains of psycholinguistics and psychology. The novelty or unfamiliarity of a stimulus has a direct bearing on basic cognitive processing. For example, words that are statistically infrequent (i.e., have a high IDF) are more difficult to perceive, recognize, and recall than more commonly encountered words (e.g., [7–9]). The affective response *associated* with perceiving novelty, however, is a more complex process. Berlyne [10], for example, extended a classic *inverted–U* relationship first proposed by Wilhelm Wundt [11]: a peak level of perceived pleasantness or "hedonic value" for moderately complex or moderately novel stimuli, and decreased liking for very simple/familiar or very complex/novel stimuli. Such a relationship has been documented across numerous classes of stimuli, including music [12], and can be further modified by an individual perceiver's *preferences* for novelty—a construct that has informed influential models of human personality [13].

Taken together, this evidence suggests that a method to quantify novelty/complexity within song lyrics might find application within the domain of personalized music recommendation. First, generated playlists could be optimized with the "right" level of lyric complexity based upon the user's activity state (e.g., exercising, commuting, or intense studying) [14–15]. Second, by computing the level of lexical novelty in a user's favorite artist, novel artists with a similar level of lexical novelty could be recommended. Third, songs with lyrics that are "not-too-simple" or "not-too-complex" could be used in paradigms supporting native or second language learning [16–17] or language recovery after brain injury [18].

2. RELATED WORK

Methods for translating a text into a single summary statistic or "grade" have been employed in a number of domains. Mid-twentieth century development of *readability metrics*—designed to quantify the ease with which a written text could be comprehended—emerged from the human factors literature (for a review and some context, see [19]), and have come to be widely applied in a variety of natural language settings [20–21]. Readability metrics are simple mathematical transformations of a text's orthographic features: letter

---

count, syllable count, word count, and sentence count.[2] Word frequency information is only rarely incorporated into readability calculations; for example, tallying the number of "difficult" [22] or "unfamiliar" [23] words (as defined by a set of 3000 words), or the "average grade level" of words (from a set of 100,000 words) [24].

By contrast, word frequency information is fundamental to vector space model approaches for text retrieval [25]. The process by which candidate documents are matched to a particular query often involves the use of *term frequency–inverse document frequency* (tf*idf) calculations [26–27]. A useful summary statistic across a set of query terms is their average IDF [28–29]. It should be noted that our proposed idea of a lexical novelty score is distinct from prior uses of tf*idf for *novelty detection* [30], which attempts to detect new information in a "stream" of documents. It is also distinct from *acoustic novelty* audio segmentation methods based on changes in temporal self-similarity [31].

3. DATASETS AND PREPROCESSING STEPS

3.1 Word frequency tables

The two key data sources for the proposed IDF-based lyric LNS are a lyrics corpus and a look-up table of document frequencies (DFs). Word frequencies *could* be estimated from the lyrics corpus itself. However, such an operation could create a dependency between IDFs and resultant LNSs—or at least necessitate retabulating word frequencies and IDFs as more lyrics were added to the corpus. Word frequency values derived from an *independent* corpus were thus desirable.

Numerous tables of word frequencies have been published (reviewed in [32]): for example, the Brown corpus (1 million words), British national corpus (100M words), Corpus of Contemporary American English (450M words), and Google Books corpus (155 billion words of American English). In the present work, we selected the use word frequency tables derived from the SUBTLEX$_{US}$ corpus [9]; a corpus of subtitle transcripts of 8388 American films and television programs. A list of 74,286 non-stemmed words[5] (46.7M word instances in total) has been compiled, with DFs (from 1 to 8388) and corpus frequencies (from 1 to 2,134,713) tabulated for each word. In addition to being fully and freely available [6], SUBTLEX$_{US}$ word frequencies have the appealing property of being derived from *spoken* source material, which may provide a closer match to the usage patterns in *sung* speech. The IDF of the *i*th word in the SUBTLEX$_{US}$ table was computed as $\log_{10}(8388/DF_i)$.

3.2 Lyrics corpus

Next, we discuss the issue of an appropriate lyrics corpus. The Million Song Dataset [33] is associated with a smaller lyrics corpus (237,662 lyrics) [7], obtained in partnership with musiXmatch[8]. The bag-of-words format used to store each lyric, however, only references the 5000 most frequent word *stems* (the part of a word common to all its inflectional and derivational variants; for example, "government", "governor", "governing", and "governance" are all stemmed to "govern") as computed by the *Porter2* stemmer[9]. (In fact, the 5000-item stemmed word list contains more than 1000 *non-English* stems when cross-checked with a 266,447-item dictionary derived from existing dictionary lists[10].) The manner in which word variants are used during communication, however, conveys rich information about the communicator's language facility [35–37]. Furthermore, word variants can have very different IDFs; in SUBTLEX$_{US}$, the four variants of "govern" listed above have IDFs of .74, 1.32, 2.58, and 3.22, respectively. As a result, a LNS derived from word stems would ignore potentially "diagnostic" differences in lexical usage between lyrics.

For this reason, a new lyrics corpus was obtained via special arrangement with LyricFind[11], a leading provider of legal lyrics licensing and retrieval. In addition to the lyrics corpus itself, metadata comprising performing artist, album, lyricist, and license territory information for each lyric was made available. The full corpus contained 587,103 lyrics. After restricting the corpus to lyrics with United States copyright, 389,029 lyrics remained.

3.3 Lyrics pre-processing

A multi-step procedure converted each lyric from its original text format into a bag-of-words format. Each lyric was first "cleaned" using a series of hand-crafted transformation rules (i.e., $x \rightarrow x'$): (1) splitting of compounds (e.g., *half-hearted→half hearted*) or removal of hyphenated prefixes (e.g., *mis-heard→ misheard*); (2) elimination of contractions (e.g., *you'll've→you will have*; *gonna→ going to*); (3) restoration of dropped initial (e.g., *'til→until*), interior (e.g., *ne'er→never*), or final (*tryin'→trying*) letters; (4) abbreviation elimination (e.g., *mr.→mister*); (5) adjustment of British English to American English spellings (e.g., *colour→color*)[12]; and (6) correction of 4264 commonly misspelled words[13].

Each lyric was then cross-checked with the 266,447-item dictionary. Lyrics in which fewer than 80% of

---

[2] For an illustration, www.readability-score.com

[5] The following items in the SUBTLEX$_{US}$ table were excluded from this tally: *'d, 's, 'm, 't, 'll, 're, don, gonna, wanna, couldn, didn, doesn*.

[6] http://expsy.ugent.be/subtlexus/

[7] http://labrosa.ee.columbia.edu/millionsong/musixmatch

[8] www.musixmatch.com

[9] http://snowball.tartarus.org

[10] http://wordlist.aspell.net

[11] www.lyricfind.com

[12] Using http://wordlist.aspell.net/varcon

[13] Using http://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings/For_machines

unique words could be matched to the dictionary were eliminated; 360,919 lyrics remained. After removing duplicate lyrics, the final corpus contained 275,905 lyrics.

A total of 67.6M word instances was present in this set of songs, with 66,975 unique words. Of these items, 51,832 were an exact match with the 74,286-item SUBTLEX$_{US}$ word list; this accounted for 99.7% of the 67.6M word instances in the lyrics corpus. IDFs derived from the SUBTLEX$_{US}$ corpus were generally in agreement with IDFs derived from the LyricFind corpus itself (Pearson's $r = .837$).

4. LYRIC-LEVEL LEXICAL NOVELTY SCORE

4.1 First-pass LNS: IDF$_{TM}$

A first-pass LNS for a lyric was defined as the trimean of SUBTLEX$_{US}$-derived IDFs (IDF$_{TM}$) associated with the set of w unique words in that lyric (w_u):

$$TM = \frac{Q_1 + 2Q_2 + Q_3}{4}, \tag{1}$$

where Q_1, Q_2, and Q_3 are the first quartile, second quartile (median), and third quartile, respectively. The trimean is an outlier-robust measure of central tendency [37]. For example, a low-frequency variant of a common word not "corrected" during the cleaning step would yield a spuriously high IDF; the trimean (but not the arithmetic mean) is robust to this kind of outlier.

The higher a lyric's IDF$_{TM}$, the more low-frequency (i.e., novel) words it contains. Figure 1 plots IDF$_{TM}$ as a function of w_u for all 275,905 lyrics (using $\log_{10}$ scaling on the x-axis). Observed w_u values range from 12 to 895.

A few illustrative cases are highlighted on Figure 1. The highest IDF$_{TM}$ (= 2.3212; LyricID 1142131; marked ①) is "Yakko's World" from the cartoon *Animaniacs*. (Example text: "There's Syria, Lebanon, Israel, Jordan / Both Yemens, Kuwait, and Bahrain / The Netherlands, Luxembourg, Belgium, and Portugal / France, England, Denmark, and Spain".) The lowest IDF$_{TM}$ (= 0.0016; LyricID 53540; marked ②) is "You Don't Know" by Killing Heidi. ("I can see you / And you don't have a clue / Of what you've done / And there's no reason / For what you've done to / Done to my ...".)

Lyric ③ (LyricID 786811; "One More Bite of the Apple" by Neil Diamond) has the same w_u as ① (= 153), but a much lower IDF$_{TM}$ (= 0.0804), indicating lower lexical novelty: "Been away from you for much too long / Been away but now I'm back where I belong / Leave while I was gone away / But I do just fine". Lyric ④ (LyricID 78427; "Revelation" by Blood) has nearly the same w_u as ② (24 vs. 23) but a much higher IDF$_{TM}$ (= 1.5454), indicating higher lexical novelty ("Writhe and shiver in agonies undreamable / Wriggling and gasping / Anticipating the tumescent / Revelation of the flesh").

Finally, cases ⑤ (LyricID 335431; "The Tear Drop"

by Armand van Helden) and ⑥ (LyricID 1452671; "Sunshine" by Bow Wow) both have $w_u = 195$, but very different IDF$_{TM}$ values (1.8464 vs. 0.1378). High lexical novelty is present in ⑤ ("A buttress breaching barrage blast / A tumultuous thunderbolt tirade / An annihilating eradicating avalanche of absolute absolution"); low lexical novelty is present ⑥ ("What you hear me talkin' 'bout / You just ain't gonna find out / Walkin' around in somebody's club / Now she's sayin' her house").
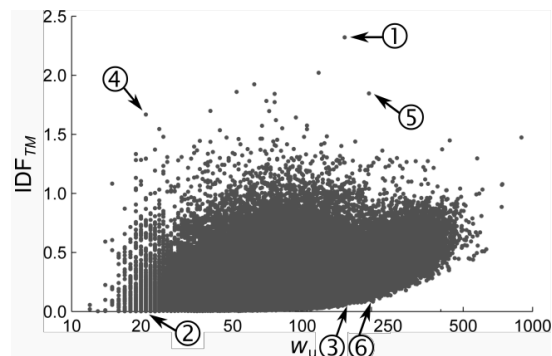

Figure 1. Scatter plot of unique words (w_u) versus IDF$_{TM}$.

A clear relationship is visible between w_u and IDF$_{TM}$ (Pearson's $r = .477$): as w_u increases, so does the minimum observed IDF$_{TM}$. This can be attributed to statistical patterns present in natural language. Specifically, a small number of words account for a large percentage of total word instances; a phenomenon which follows Zipf's law (e.g., [38]). In the SUBTLEX$_{US}$ corpus, for example, 10 words (*you, i, the, to, a, it, that, and, of, what*) account for 24.3% of all 46.7M word instances. Because IDF$_{TM}$ is derived from the set of *unique* words in a lyric, as w_u increases, so too must the number of lower-frequency (i.e., higher-IDF) words, causing the IDF$_{TM}$ to rise. Such a pattern would manifest for any L-estimator (mean, median, midhinge, etc.).

A more informative statistic could be obtained if the IDF$_{TM}$ of a lyric with w unique words were compared against a large distribution of simulated IDF$_{TM}$ values obtained from repeated random draws of w unique words from the set of lyrics that had more than w unique words. This procedure is formalized next.

4.2 Scaling IDF$_{TM}$: Monte Carlo simulations

Consider two lyrics, one with IDF$_{TM} = 0.25$ and $w_u = 50$, and the other with IDF$_{TM} = 0.5$ and $w_u = 200$. Two scaling distributions of simulated IDF$_{TM}$ values were created using a 10,000-iteration procedure. To create the scaling distribution for $w_u = 50$, on each iteration, a single lyric was randomly selected from the set of 239,225 lyrics with $w_u > 50$. The full set of words in that lyric (including repeated words) was randomly permuted, the first 50 unique words pulled, and the IDF$_{TM}$ of those words was taken. To create the scaling distribution for $w_u = 200$, a similar procedure was performed, using the set of 15,124 lyrics with $w_u > 200$. Figure 2 presents an

empirical cumulative distribution function (ECDF) of these two scaling distributions. The "scaled IDF_{TM}" is defined as the percentile P (i.e., the y-axis value on the ECDF, multiplied by 100) where $x = IDF_{TM}$. In the above example, when $IDF_{TM} = 0.25$ and $w_u = 50$, $P = 85.8$. By contrast, when $IDF_{TM} = 0.25$ and $w_u = 200$, $P = 10.3$. This can be interpreted as follows: with a longer lyric ($w_u = 200$ vs. $w_u = 50$), the likelihood of obtaining an $IDF_{TM} > 0.5$ *by chance* (i.e., $100 - P$) is much higher (89.7% vs. 14.2%); that is, it is a *less novel* occurrence.
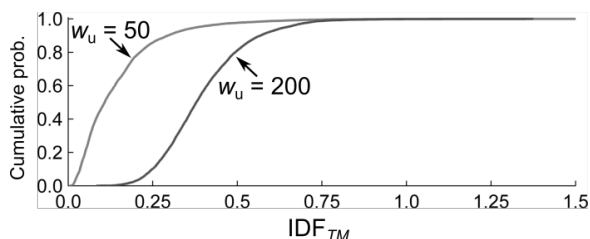

Figure 2. ECDFs of simulated IDF_{TM} values for two representative values of w_u.

To scale the full set of IDF_{TM} values, the above simulation was modified in the following manner. First, the range of target w_u values was capped at 275, thus reserving 5228 lyrics with $w_u > 275$ to create the scaling distribution for $w_u = 275$. Second, the set of target P-values was defined as .01 to 99.99 in increments of .01. Third, to accurately estimate the "tails" of P (i.e., values near 0 and 100), many more Monte Carlo iterations at each w_u are needed; thus, the number of iterations was increased from 10,000 to 1 million.

Figure 3 highlights the results of this simulation. A representative set of "iso-probability curves" resulting from the Monte Carlo simulation are superimposed on the scatter plot first shown in Figure 1. A given curve plots the Pth percentile (where $P = \{.01, 10, 50, 90, 99, 99.9, 99.99\}$) of simulated IDF_{TM} values across the set of w_u values. $IDF_P \approx 0$ indicates *very low* lexical novelty, $IDF_P \approx 50$ indicates *moderate* lexical novelty, and $IDF_P \approx 100$ indicates *very high* lexical novelty. As expected, the iso-probability curves for low P-values mirror the pattern in the real data: higher IDF_{TM} values as w_u increases.
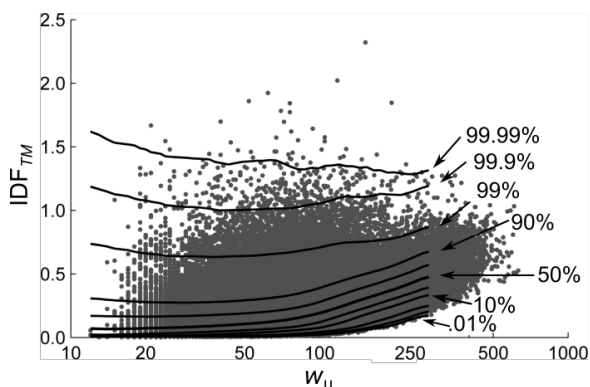

Figure 3. Representative iso-probability curves.

4.3 Second-pass LNS: Percentiles

Each IDF_{TM} was mapped to its corresponding IDF_P using nearest neighbor interpolation. IDF_{TM} values below $P = .01$ ($n = 80$) or above $P = 99.99$ ($n = 52$) were set to $IDF_P = 0$ or $IDF_P = 100$, respectively. Figure 4 plots IDF_P as a function of w_u for the final set of 270,677 unique lyrics. The relationship between w_u and IDF_P ($r = -.106$) is much weaker than between w_u and IDF_{TM} ($r = .477$). IDF_P values were roughly uniform (mean = 44.29; standard deviation = 29.70; skewness = 0.255).
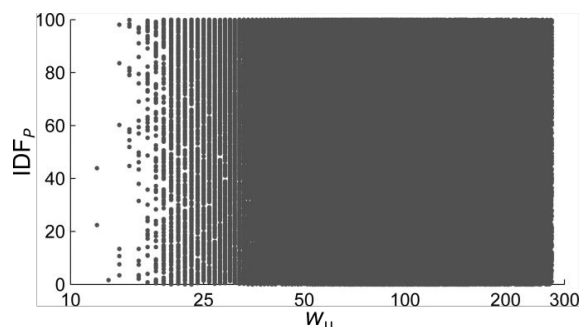

Figure 4. Percentile-transformed novelty scores (IDF_P) as a function of w_u.

Figure 5 presents an ECDF of both IDF_{TM} and IDF_P, highlighting the six lyrics discussed earlier. Compared to IDF_{TM}, IDF_P better differentiates lyrics with high lexical novelty (cases ①, ④, and ⑤) versus low novelty (cases ②, ③, and ⑥).
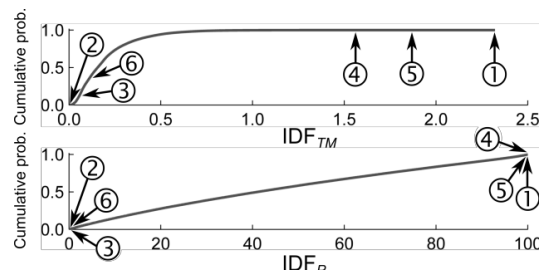

Figure 5. ECDFs for IDF_{TM} (upper) and IDF_P (lower).

5. ARTIST-LEVEL LEXICAL NOVELTY

Having defined IDF_P as the lyric-level LNS, we next sought to characterize lexical novelty at the artist level. Artist information was obtained via LyricFind ArtistIDs, which are distinct for different combinations of individual artists. To increase the specificity of an artist-level score, lyrics recorded by multiple artists (e.g., holiday songs, jazz standards) were excluded. Artists associated with fewer than 10 unique lyrics (λ_u) were deemed to have an insufficient catalog, and were ignored. A final set of 5884 artists (a total of 216,072 lyrics) remained. The trimean of each artist's λ_u IDF_P values was then taken as a simple and intuitive artist-level LNS.

Figure 6 plots artist-level LNS as a function of λ_u; no correlation was present between them ($r = -.009$.) The distribution of values (mean = 43.49; standard deviation = 21.20) was roughly symmetrical (skewness = .459).
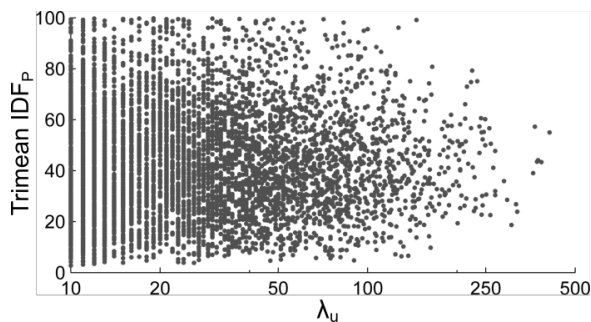
Figure 6. Artist-level LNS as a function of λ_u.

6. BILLBOARD MAGAZINE "TOP" LISTS

Having derived both a lyric-level and an artist-level point estimate of lexical novelty, any number of subsequent analyses may be performed. As an illustrative example, we turn to Billboard Magazine's 2013 ranking of the "All-Time Top 100 Songs"[14] and "All-Time Top 100 Artists"[15]. Rankings were calculated based on overall success on the magazine's "Hot 100" chart, a weekly ranking of the top 100 popular music singles in the United States, published since August 1958 [40–41].

The Top Songs list was determined by Billboard using an inverse point system, with time spent in the #1 position of each weekly chart weighted highest, and time spent in the #100 position weighted lowest. Of the 100 songs on the list, 95 were present in the LyricFind corpus. Lyrics for the remaining five were queried from metrolyrics.com and processed as described in Section 4.

The Top Artists list was determined by Billboard by aggregating all the songs which charted over the course of each artist's career. Of the 100 artists, 98 were among the set of 5884 artists with a valid artist-level LNS; the other two artists had $\lambda_u < 10$.

7. EXPERIMENTAL HYPOTHESES

Two hypotheses were examined, both driven by the assumption that high lexical novelty is less likely to be "chart-worthy". Specifically, we predicted that both lyric-level and artist-level LNSs would be *lower* in the set of Top Songs and Top Artists relative to "non-top" songs and artists in the LyricFind corpus.

Statistical significance was assessed using a nonparametric two-sample Mann–Whitney (MW) test. A special sampling procedure was implemented to counter-act the bias towards smaller *p*-values when comparing large samples [41]. On each of 10,000 iterations, two samples were drawn. The first sample was always the *n* Top Song or Top Artist LNSs, and the second sample was a random draw (without replacement) of *n* LNSs

from the remaining set of songs or artists (where *n* is 100 for songs and 98 for artists). The distribution of Z-values from the 10,000 MW tests indicates the strength of the difference between the samples: the more negative it falls, the greater our confidence that lexical novelty is *systematically* lower in the set of Billboard items.

8. EXPERIMENTAL RESULTS

8.1 Billboard Top Songs analysis

Figure 7a shows the ECDFs of lyric-level LNS for the set of 100 Top Songs and the remaining 270,582 songs. They are markedly different: LNSs for the Top Songs are "pulled" towards zero, indicating reduced lexical novelty in this set. Consistent with this, the distribution of Z-values (Figure 7b) is strongly negative: 98.4% of MW tests result were significant at $p < .05$, 89.9% at $p < .01$, and 61.1% at $p < .001$. No correlation was present between Billboard's song ranking and a song's LNS ($r = -.148$, $p = .140$).
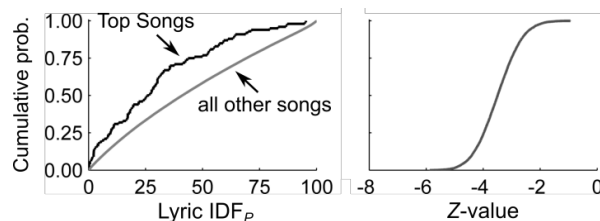

Figure 7. (a.) ECDFs of LNSs for the 100 Top Songs and the remaining 270,582 songs in the corpus. (b.) ECDF of Z-values from the 10,000 MW tests.

8.2 Billboard Top Artists analysis

Figure 8a shows the ECDFs of artist-level LNS for the set of 98 Top Artists and the remaining 5786 artists. As with the Top Songs, LNSs for the Top Artists are pulled towards zero, indicating reduced lexical novelty (i.e., lower IDF_P trimean values) for the set of 98 Top Artists. The Z-value distribution (Figure 8b) is more negative than in the Top Songs analysis: 99.3% of tests were significant at $p < .001$, 95.8% at $p < .0001$, and 85.5% at $p < .00001$. As with the Top Songs, no correlation was present between Billboard's artist ranking and artist-level LNS ($r = -.059$, $p = .564$).
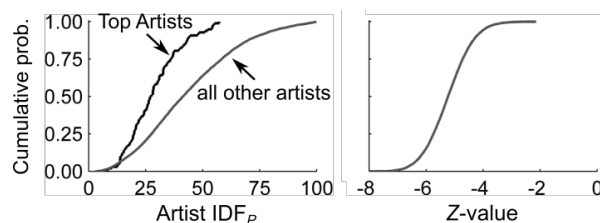

Figure 8. (a.) ECDFs of artist-level LNSs for the 98 Top Artists and the remaining 5786 artists in the corpus. (b.) ECDF of Z-values from the 10,000 MW tests.

---

[14] billboard.com/articles/list/2155531/the-hot-100-all-time-top-songs

[15] billboard.com/articles/columns/chart-beat/5557800/hot-100-55th-anniversary-by-the-numbers-top-100-artists-most-no

9. DISCUSSION

9.1 Summary

Stimulus novelty has influence over perception, memory, and affective response. Here, we define a *lexical novelty score* (LNS) for song lyrics. The LNS is derived from the inverse document frequency of all unique words in a lyric, and is scaled with respect to the number of unique words. Higher-order scores can be easily defined at the level of artists, albums, or genres, creating additional features for filtering operations or similarity assessments.

Although the construct validity of the LNS must be assessed by future user studies (see Section 9.2), a first-pass validation was performed by comparing LNSs associated with Billboard Magazine's "official" lists of the 100 Top Songs and 100 Top Artists with LNSs from random sets of songs and artists. Lexical novelty was significantly lower—in a highly consistent way—for items on the Billboard lists, supporting the broad hypothesis that moderate stimulus novelty is preferred over high stimulus novelty [10–12].

The absence of any significant correlation between Billboard's actual *ranking* of items on the Top Songs or Top Artists lists and our lexical novelty score should not be read as a "strike" against either Billboard's methodology or our own. Rather, we regarded these lists as a source of well-known independent data that enabled us to make *a priori* predictions concerning differences in lexical novelty at the set (rather than the item) level.

6.2 Future directions

The present analyses of Billboard's "Top 100" lists are but one of many analyses that could be performed. Further work could explore differences in lexical novelty among genres, subgenres, or styles (using external sources of metadata, such as Echo Nest [16], Rovi [17] or 7digital [18]); changes in lexical novelty over time (e.g., using lyric copyright date information); or correlations between lexical novelty and other performance-related metrics, such as RIAA-tracked album sales [19].

A potential refinement of our LNS calculation would be to make it sensitive to parts of speech. Numerous English words can serve as multiple parts of speech, often with very different word frequencies. Capturing these usage patterns would, in principle, increase the sensitivity of the LNS. A revised SUBTLEX$_{US}$ table of document frequencies is available that tallies parts-of-speech [42], as are widely used parts-of-speech taggers [20,21], making this modification tractable.

Finally, user studies must be performed to answer whether the proposed LNS *itself* has construct validity. These studies should evaluate, for example, whether lyrics with a high LNS yield longer reaction times and increased effort during a sentence processing task (e.g., as in [43]); or whether lyrics with a moderate LNS receive higher ratings of pleasure or liking than lyrics with either a low or a high LNS.

Together, these future steps will enhance the utility of the LNS in the context of music retrieval and recommendation applications.

10. DATA SET AVAILABILITY

With gratitude to LyricFind, much of the data presented here—lyrics in bag-of-words format; lyric, artist, and album IDs; and lyric- and artist-level lexical novelty scores—is made publically available for the first time: www.smcnus.org/lyrics/.

11. ACKNOWLEDGEMENT

12. REFERECNES

[1]   F. Kleedorfer, P. Knees, and T. Pohle, "Oh Oh Oh Whoah! Towards Automatic Topic Detection In Song Lyrics.," in *Proc. Int. Symp. Music Inf. Retrieval*, 2008, pp. 287–292.

[2]   R. Mayer, R. Neumayer, and A. Rauber, "Rhyme and Style Features for Musical Genre Classification by Song Lyrics.," in *Proc. Int. Symp. Music Inf. Retrieval*, 2008, pp. 337–342.

[3]   C. Laurier, J. Grivolla, and P. Herrera, "Multimodal music mood classification using audio and lyrics," in *Proc. 7th Int. Conf. Mach. Learn. Appl.*, 2008, pp. 688–693.

[4]   X. Hu, J. S. Downie, and A. F. Ehmann, "Lyric text mining in music mood classification," *Am. Music*, vol. 183, no. 5,049, pp. 2–209, 2009.

[5]   M. Van Zaanen and P. Kanters, "Automatic Mood Classification Using TF*IDF Based on Lyrics.," in *Proc. Int. Symp. Music Inf. Retrieval*, 2010, pp. 75–80.

[6]   X. Wang, X. Chen, D. Yang, and Y. Wu, "Music Emotion Classification of Chinese Songs based on Lyrics Using TF*IDF and Rhyme.," in *Proc. Int. Symp. Music Inf. Retrieval*, 2011, pp. 765–770.

---

16 http://developer.echonest.com/docs/v4
17 http://developer.rovicorp.com
18 http://developer.7digital.com/
19 https://www.riaa.com/goldandplatinumdata.php
20 http://ucrel.lancs.ac.uk/claws/trial.html
21 http://nlp.stanford.edu/software/lex-parser.shtml

[7] K. Rayner and S. A. Duffy, "Lexical complexity and fixation times in reading," *Mem. Cognit.*, vol. 14, no. 3, pp. 191–201, 1986.

[8] F. Meunier and J. Segui, "Frequency effects in auditory word recognition," *J. Mem. Lang.*, vol. 41, no. 3, pp. 327–344, 1999.

[9] M. Brysbaert and B. New, "Moving beyond Kučera and Francis," *Behav. Res. Methods*, vol. 41, no. 4, pp. 977–990, 2009.

[10] D. E. Berlyne, *Aesthetics and Psychobiology*. New York: Appleton-Century-Crofts, 1971.

[11] W. Sluckin, A. M. Colman, and D. J. Hargreaves, "Liking words as a function of the experienced frequency of their occurrence," *Br. J. Psychol.*, vol. 71, no. 1, pp. 163–169, 1980.

[12] A. C. North and D. J. Hargreaves, "Subjective complexity, familiarity, and liking for popular music," *Psychomusicology*, vol. 14, no. 1, pp. 77–93, 1995.

[13] M. Zuckerman, *Behavioral expressions and biosocial bases of sensation seeking*. Cambridge university press, 1994.

[14] M. Kaminskas and F. Ricci, "Contextual music information retrieval and recommendation," *Comput. Sci. Rev.*, vol. 6, no. 2, pp. 89–119, 2012.

[15] X. Wang, D. Rosenblum, and Y. Wang, "Context-aware mobile music recommendation for daily activities," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 99–108.

[16] C. F. Mora, "Foreign language acquisition and melody singing," *ELT J.*, vol. 54, no. 2, pp. 146–152, 2000.

[17] K. R. Paquette and S. A. Rieg, "Using music to support the literacy development of young English language learners," *Early Child. Educ. J.*, vol. 36, no. 3, pp. 227–232, 2008.

[18] C. Y. Wan and G. Schlaug, "Music making as a tool for promoting brain plasticity across the life span," *The Neuroscientist*, vol. 16, no. 5, pp. 566–577, 2010.

[19] G. R. Klare, "The measurement of readability," *ACM J. Comput. Doc. JCD*, vol. 24, no. 3, pp. 107–121, 2000.

[20] T. G. Gunning, "The role of readability in today's classrooms," *Top. Lang. Disord.*, vol. 23, no. 3, pp. 175–189, 2003.

[21] G. K. Berland, M. N. Elliott, L. S. Morales, J. I. Algazy, R. L. Kravitz, M. S. Broder, and others, "Health information on the Internet: accessibility, quality, and readability in English and Spanish," *J. Am. Med. Assoc.*, vol. 285, no. 20, pp. 2612–2621, 2001.

[22] J. S. Chall and E. Dale, *Readability revisited: The new Dale-Chall readability formula*. Brookline Books, 1995.

[23] G. Spache, "A new readability formula for primary-grade reading materials," *Elem. Sch. J.*, pp. 410–413, 1953.

[24] M. Milone, "Development of the ATOS readability formula." Renaissance Learning, 2014.

[25] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[26] A. Aizawa, "An information-theoretic perspective of tf–idf measures," *Inf. Process. Manag.*, vol. 39, no. 1, pp. 45–65, 2003.

[27] J. Ramos, "Using tf-idf to determine word relevance in document queries," in *Proc. 1st Inst. Conf. Machine Learning*, 2003.

[28] S. Cronen-Townsend, Y. Zhou, and W. B. Croft, "Predicting query performance," in *Proc. 25th ACM SIGIR*, 2002, pp. 299–306.

[29] L. Zheng, S. Wang, Z. Liu, and Q. Tian, "Lp-norm idf for large scale image search," in *IEEE CVPR*, 2013, pp. 1626–1633.

[30] J. Allan, C. Wade, and Alvaro Bolivar, "Retrieval and novelty detection at the sentence level," in *Proc. 26th ACM SIGIR*, 2003, pp. 314–321.

[31] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in *IEEE Int. Conf. Multimedia Expo.*, 2000, vol. 1, pp. 452–455.

[32] T. McEnery and A. Hardie, *Corpus linguistics: Method, theory and practice*. Cambridge University Press, 2011.

[33] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proc. 12th Int. Symp. Music Inf. Retrieval*, 2011, pp. 591–596.

[34] A. Caramazza, A. Laudanna, and C. Romani, "Lexical access and inflectional morphology," *Cognition*, vol. 28, no. 3, pp. 297–332, 1988.

[35] G. Yu, "Lexical diversity in writing and speaking task performances," *Appl. Linguist.*, vol. 31, no. 2, pp. 236–259, 2010.

[36] A. Xanthos, S. Laaha, S. Gillis, U. Stephany, A. Aksu-Koç, A. Christofidou, and others, "On the role of morphological richness in the early development of noun and verb inflection," *First Lang.*, p. 0142723711409976, 2011.

[37] J. W. Tukey, *Exploratory data analysis*. Reading, MA: Addison-Wesley, 1977.

[38] M. E. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemp. Phys.*, vol. 46, no. 5, pp. 323–351, 2005.

[39] E. T. Bradlow and P. S. Fader, "A Bayesian lifetime model for the 'Hot 100' Billboard songs," *J. Am. Stat. Assoc.*, vol. 96, no. 454, pp. 368–381, 2001.

[40] D. E. Giles, "Survival of the hippest: life at the top of the hot 100," *Appl. Econ.*, vol. 39, no. 15, pp. 1877–1887, 2007.

[41] R. M. Royall, "The effect of sample size on the meaning of significance tests," *Am. Stat.*, vol. 40, no. 4, pp. 313–315, 1986.

[42] M. Brysbaert, B. New, and E. Keuleers, "Adding part-of-speech information to the SUBTLEX-US word frequencies," *Behav. Res. Methods*, vol. 44, no. 4, pp. 991–997, 2012.

[43] A. D. Friederici, "Towards a neural basis of auditory sentence processing," *Trends Cogn. Sci.*, vol. 6, no. 2, pp. 78–84, 2002.