Bob Dylan has just celebrated his 72nd birthday! He was also recently made an honorary member of the American Academy of Arts and Letters, a prestigious position he now shares with all of these acclaimed artists. As someone who is a huge fan of Bob Dylan and also of corpus linguistics, I would like to celebrate his life's work so far by sharing with you some results from a mini corpus project I have recently been working on.

I thought a corpus analysis of Bob Dylan's lyrics would be interesting because as well as being regarded as one of the most acclaimed and influential songwriters in recent American history, Dylan has been writing songs and releasing albums for more than 50 years. His continued productivity and influence on popular music is celebrated within arenas of light literary scholarship, biographical documentation and also in hardcore academic research (Dunlap, 2006; Gearey, 1988; Shelton, 2011). Even historical scholars, cultural sociologists and literary commentators (Marshall, 2003; Wilentz, 2011; Rogovoy, 2009) continue to adulate the immense impact of his songs on 20th and 21st century culture. Such commentators are careful not to restrict their dissection and admiration of his music and lyrics to the so-called '1960s golden period" of Dylan's creative successes. Whitwell (2008: 469), in her research paper which explores the emotional content of Dylan's lyrics, notes that we should all be aware that "his career is still in progress and his is not to be regarded only in the past tense". Whether you're a huge fan or not, I think this comment succinctly summarizes the consensus among scholars, critics and fans alike that Dylan has sustained his creative and expressive panache for writing moving and emotive poetry throughout his life.

With such an expansive catalogue of work, I thought it would be cool to get a broad snapshot of the lexical composition of Dylan's lyrics over time. Firstly I created a corpus of all of Bob Dylan's lyrics which were available in part from a now defunct webpage [AUTHOR WAS ALERTED TO THIS ON APRIL 30th 2015]. I then grouped the lyrics and put them into bins based on the years that they were released. In total, I counted 32 studio albums of original lyrics and music. I didn't include bootlegs, live songs or live albums as this would have duplicated songs within the corpus. I also made a point of trying to include songs only penned by Dylan himself. The corpus spans 50 years of lyrical material; Dylan's first album, *Bob Dylan*, was released in 1962 and the latest album, *Tempest*, was released in 2012. Here is a list of all the albums alongside the year they were released and Dylan's age at the time of the album's release. This may help when interpreting the graphs below, which only have the year marked on the x-axis and not the album names.

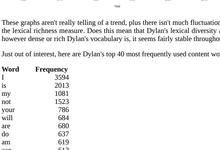
Year	Album	Age
1962	*Bob Dylan*	21
1963	*The Freewheelin' Bob Dylan*	22
1964	*Another Side Of Bob Dylan; The Times They Are A-Changin'*	23
1965	*Bringing It All Back Home; Highway 61 Revisited*	24
1966	*Blonde On Blonde*	25
1967	*John Wesley Harding*	26
1969	*Nashville Skyline*	28
1970	*New Morning; Self Portrait*	29
1973	*Pat Garrett And Billy The Kid*	32
1974	*Planet Waves*	33
1975	*Blood On The Tracks; The Basement Tapes*	34
1976	*Desire*	35
1978	*Street Legal*	37
1979	*Slow Train Coming*	38
1980	*Saved*	39
1981	*Shot Of Love*	40
1983	*Infidels*	42
1985	*Empire Burlesque*	44
1986	*Knocked Out Loaded*	45
1988	*Down In The Groove*	47
1989	*Oh Mercy*	48
1990	*Under The Red Sky*	49
1997	*Time Out Of Mind*	56
2001	*Love And Theft*	60
2006	*Modern Times*	65
2009	*Together Through Life*	68
2012	*Tempest*	71

Dylan's vocabulary: lexical diversity and richness

In total there are 98, 299 words and 24, 064 word types (different words) in the Bob Dylan corpus. In linguistics the *token total* denotes the total number of words altogether, and the *type total* denotes the total number of different words. For example in the following line, "*How many roads must a man walk down, before you call him a man?*", there are 14 tokens and 12 types (*a* and *man* are repeated twice). The type/token ratio is useful as it can tell you a little bit about the lexical diversity of a text, which basically indicates the amount of unique words a text contains.

Now, I don't have an author to compare Dylan with and I just wanted to see if there were any trends regarding his lexical diversity over time. For each year represented in the corpus, the mean number of tokens is 3641 and the mean number of types is 891. The token count ranges from 1549 to 8548 words and the type count ranges from 375 to 1831 words. Analysing a longitudinal trend based on numbers this size won't give us anything statistically powerful, but you can still look at a nice plot of Dylan's lexical diversity changing over time (you can click on it to make it bigger):

I also looked at Bob Dylan's lexical density, which is the ratio of content (e.g. nouns, verbs and adjectives) to function words (e.g. prepositions, determiners, conjunctions). Here is a graph of Dylan's lexical density:

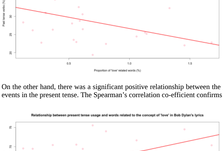


These graphs aren't really telling of a trend, plus there isn't much fluctuation in the values for both lexical measures. Just for the hell of it I ran a Spearman's correlation test which confirmed that there is no significant trend for the type token ration and the lexical richness measure. Does this mean that Dylan's lexical diversity and richness has remained consistent throughout his artistic output? There is no telling how his lexical usage compares to other songwriters (I haven't gotten that far yet). But however dense or rich Dylan's vocabulary is, it seems fairly stable throughout his life.

Just out of interest, here are Dylan's top 40 most frequently used content words, the left column shows you the word, and the right column shows you the raw frequency:

Word	Frequency
I	3594
is	2013
my	1081
not	1523
your	786
will	684
are	680
do	637
am	619
can	612
be	604
was	597
have	558
well	481
just	451
no	411
know	397
one	371
got	368
out	358
going	347
love	316
see	273
time	270
now	257
say	253
said	252
go	250
where	244
man	243
come	235
could	227
been	223
get	217
back	212
baby	207
oh	206
too	200
never	198

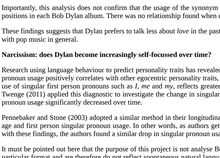
It's also cool to see that Zipf's (1949) law holds for Dylan's entire lyrical output (including function words). Zipf's law states that the frequency of a word is inversely proportional to its rank in a frequency table. In other words, in natural language the most frequent word which is ranked alone in first place (in the Dylan corpus *the* comes out 1st) is twice as frequent as the second most frequent word (*I* is in 2nd place) and three times as frequent as the third most frequent word (*you* is in 3rd place). Essentially, the distribution shows that you get very few high frequency words and lots and lots of low frequency words. Zipf's Law is illustrated by plotting log word frequency against log word rank. The resulting plot should reveal a straight line, which is what you see (straightish) below:



George Zipf was a linguist, but what is cool is that his law holds for lots of other ranked data sets in the social sciences, like income and city population.

Time and tense: As Bob Dylan ages, do his lyrics refer more to the past, present or future?

According to Pennebaker and Stone (2003) a linguistic marker of increasing age is a change in the frequency of future tense and past tense usage relative to present usage. Pennebaker and Stone (2003) aggregated the complete works of 10 authors, playwrights and poets, which includes William Shakespeare, Charles Dickens and George Eliot, and found a significant positive correlation between age and future tense usage such that with increasing age, the authors which were sampled used more future tense lexical verbs. In a separate corpus analysis consisting of text (essays and transcribed interviews) from 3,280 participants, Pennebaker and Stone (2003) found the same positive correlation between age and future tense usage but also a significant negative correlation between age and past tense usage. These results seem counterintuitive; one would expect past tense usage to increase with age. In fact, it appears that young people are more concerned about the past and that older people's language use refers more to the future. I decided to investigate whether Dylan's choice of tense also reflects his increasing age. I managed to pull out of my corpus all occurrences of lexical verbs used in the past, present and future tense. Below is a graph summarising the proportion frequencies of Dylan's past, present and future tense usage per year. Each data point within the same colour scheme represents a different year.

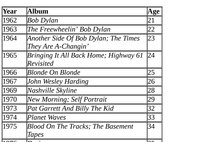


Contrary to Pennebaker and Stone's (2003) findings, there are no statistically significant changes in the proportions of usage of past, present or future tense usages used over time, even though there was a significant difference between the overall frequency each tense was used by Bob Dylan. The total number of past tense occurrences of lexical verbs in the Bob Dylan corpus is 6054, the frequency for present tense lexical verbs is 13264 and the total frequency of future tense occurrences is 1079. The expected frequency for each was 6799. The difference in the frequencies of the three tense types was significant ($\chi^2 = 11041.28$, $df = 2$, $p = < .0001$). This finding is interesting in that it indicates that Bob Dylan does not become progressively concerned with the future as his age increases and nor does he refer less to the past as his age increases. Rather, Dylan is consistent in his reference to events in the past, present or future, even though overall he does prefer to recount events in the present tense compared to referring to events in the past and future tense. Perhaps writing in the present tense is a powerful literary device used (among many other others) that contributes to a song's timeless quality. Here's one example:

> *How many roads must a man walk down, Before you call him a man? How many seas must a white dove sail, Before she sleeps in the sand? Yes, how many times must the cannon balls fly, Before they're forever banned? The answer my friend is blowin' in the wind, The answer is blowin' in the wind.*

Love and tense

Next, I decided to pull out all words that are semantically related to the word *love*. The word list was created using the Corpus of Contemporary American English (COCA) (Davies, 2008). This involved looking up all the lemma forms of all the synonyms related to *love*. By lemma I mean, if a synonym like *devote* shows up in the list, so will all its word forms, like *devotes*, *devoting*, *devotion*, and *devotions*. Surprisingly, two significant relationships were found between the tense type and how much Dylan likes to talk about love. Firstly, there was a significant negative relationship between the frequency of synonyms of the word *love* and the proportion of usage of the past tense. Using a basic statistical model, the Spearman's correlation co-efficient, we can confirm that there is a significant relationship between the frequency of synonyms of the word *love* and the proportion of usage of the past tense, $r_s = -0.48$, $p = < .05$. This means that for each album, the less Dylan refers to the concept of *love*, the more he refers to events in the past tense. Below is a plot with proportion of love related synonyms on the x-axis and proportion of past tense usage on the y-axis. Each data point represents a Bob Dylan album.



On the other hand, there was a significant positive relationship between the frequency of synonyms of the word *love* and the proportion of usage of the present tense. In other words, the more Dylan refers to the concept of *love*, the more he refers to events in the present tense. The Spearman's correlation co-efficient confirms the significant relationship between the frequency of synonyms of the word *love* and the proportion of present tense usage, $r_s = 0.45$, $p = < .05$.

Importantly, this analysis does not confirm that the usage of the synonym of *love* and the past/present tense form of the lexical verb are part of the same local lyrical context, instead the counts simply correlates the total usage of all forms in any positions in each Bob Dylan album. There was no relationship found when correlating future tense usage and the frequency of *love* synonyms, which means that these findings are not trivial.

These findings suggests that Dylan prefers to talk less about *love* in the past tense and more about the concept in the present. It would be interesting to find out whether the preference to refer to love in the present tense over the past tense is consistent with pop music in general.

Narcissism: does Dylan become increasingly self-focussed over time?

Research using language behaviour to predict personality traits has revealed that increased usage of singular personal pronouns (e.g. *I* and *me*) correlates positively with depression (Rude, Gortner, & Pennebaker, 2004). Also, higher singular personal pronoun usage positively correlates with other egocentric personality traits, such as narcissism (Raskin & Hall, 1981; Campbell, Bosson, Goheen, Lakey, & Kernis, 2007). The line of reasoning behind the interpretation of these results is that increased use of singular first person pronouns such as *I*, *me* and *my*, reflects greater self-focus and disconnect from social relationships and communalism. In fact, a recent study of popular U.S. song lyrics from 1980 to 2007, DeWall, Pond, Campbell and Twenge (2011) applied this diagnostic to investigate the change in singular personal pronoun usage across time. DeWall et al. found that in best-selling pop music, first-person singular pronoun usage significantly increased while first-person plural pronoun usage significantly decreased over time.

Pennebaker and Stone (2003) adopted a similar method in their longitudinal project examining the frequency of first person singular usage as a function of age. They found a significant negative correlation coefficient and linear association between age and first person singular pronoun usage. In other words, as authors get older, they display less self-focus in their work. Pennebaker and Stone (2003) propose that this finding reflects a progressive detachment of self-focus over time. Consistent with these findings, the authors found a similar drop in singular pronoun usage over time in their corpus of naturalistic language use made-up of language usage from over 3,000 participants.

It must be pointed out here that the purpose of this project is not analyse Bob Dylan's lyrics with the aim of drawing conclusions about his day-today language use or language use of the population in general. Music lyrics and poetry are written in a particular format and are therefore do not reflect spontaneous natural language. Moreover, the results of a study on the relationship between Dylan's language use and his ageing is not going to be useful to make any grand claims, as the sample size is one and the number of words contained within the corpus is too small. If anything, it will be useful to examine how Dylan's self-focus fluctuates throughout his career from period to period. Self-focus was calculated using the same method as DeWall et al. which involved subtracting the total frequency of plural pronoun usage from the total frequency of singular pronoun usage. Thus, higher scores reflect greater degrees of self-focus. It is difficult to find out whether overall Bob Dylan is either narcissistic or entirely unselfish without having an adequate baseline with which to make a comparison, but we can see how self-focus fluctuates throughout his career.

According to the basic stats (a Spearman's correlation), Bob Dylan's singular versus plural pronoun usage does not significantly increase with time. Even if we don't know if Dylan's self-focus is abnormally high or abnormally low, this finding suggests that his self-focus, remains stable throughout his life. There might be another system lurking in the data, but it looks like self-focus peaks and troughs from the beginning of his career and ending quite high in the mid-70s, at the release of *Blood On The Tracks*. Self-focus then hits a big low in 1990 and then peaks dramatically in 1997 in the album *Time Out Of Mind*, which was hailed as a return to form for Dylan. I'm not gonna say anymore about this self-focus thing, I'll let you decide what to make of it.

I hope this has been useful to Dylan geeks (Dylanologists) and corpus linguists alike. I know this study could be more empirically robust, but I think that these kinds of quantitative explorations, when implemented with the right amount of scientific rigour, can supplement literary scholarship in general.

I'll sign off this post with a few of my favourite Dylan lines from the song *When The Deal Goes Down*:

> Well the moon gives light, And it shines by night, When I scarcely feel the glow. We learn to live, And then we forgive, O'r the road we're bound to go. More frailer than the flowers, These precious hours, That keep us so tightly bound.
> You come to my eyes, Like a vision from the skies, And I'll be with you when the deal goes down.

References Campbell, W. K., Bosson, J. K., Goheen, T. W., Lakey, C. E., & Kernis, M. H. (2007). Do narcissists dislike themselves "deep down inside?" Psychological Science, 18, 227-229. Davies, M. (2008-) The Corpus of Contemporary American English: 450 million words, 1990-present. Available online at http://corpus.byu.edu/coca/. DeWall, C. N., Pond, R. S., Campbell, W. K., and Twenge, J. M. (2011). Tuning into 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(3), 200-207. Dunlap, J. (2006). Through the eyes of Tom Joad: Patterns of American Idealism, Bob Dylan, and the folk protest movement. Popular Music and Society, 29(5), 549-573. Gearey, A. (1998). *Outlaw Blues: Law in the Songs of Bob Dylan*. Cardozo L. Rev., 20, 1401. Marshall, L. (2007). *Bob Dylan: the never ending star*. Polity. Pennebaker, J. W., & Stone, L. D. (2003). Words of Wisdom: Language Use Over the Life Span. Journal of Personality and Social Psychology, 2003, Vol. 85(2), 291-301. Raskin, R., & Hall, C. S. (1981). The Narcissistic Personality Inventory: Alternate form reliability and further evidence of construct validity. *Journal of Personality Assessment*, 45, 159-162. Rogovoy, S. (2009). *Bob Dylan: prophet, mystic, poet*. Scribner. Rude, S. S., Gortner, E. M., & Pennebaker, J. W. (2004). Language use of depressed and depression-vulnerable college students. Cognition and Emotion, 18(8), 1121-1133. Shelton, R. (2011). *No direction home: The life and music of Bob Dylan*. Backbeat Books. Whitwell, C. (2008). Emotional fluctuations in Bob Dylan's lyrics measured by the dictionary of affect accompany events and phases in his life. Psychological Reports, 102, 469-483. Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley.