The Evolution of TED Talks

A temporal content analysis of TED talk transcripts from 2006-2021

LSE Candidate ID: 10443

Table of Contents

Abstract

TED talks have been touted to be very useful in popularizing scientific research and spreading globally important ideas. The popularity of TED is a consequence of both its content as well as the presentation of its erudite pool of speakers, the latter reinforcing the former. Studies in the past have recognized the importance of understanding the features of the talks using content analysis. This research aims to extend the past literature by performing a temporal analysis of TED's transcripts (obtained from TED's website via web scraping) with the help of descriptive statistical analysis and also use topic modelling to identify the main topics TED covers. It was discovered that recent TED talks have higher linguistic complexity than earlier ones. Pronominal choices of speakers and the tendency to include numerical information to support ideas has seen a decline over the years. Talks relating to human feelings emerged as the most common. The results of this research will help understand the evolution of the contemporary world's most famous public speaking initiative and can be used to shape the future of public speaking.

Introduction

The first TED was conceived in 1984 as a public-speaking initiative and was the brainchild of Richard Saul Wurman when he observed the potential in bringing together the best minds operating in the domains of technology, entertainment and design (TED n.d.). In 2006, TED began publishing its talks online. As of now, TED boasts of having over a billion views on its videos published on both its website as well as on Youtube (TED n.d.).

Over the years, TED has amassed a very large following (Masson 2014, Ludewig 2017). Speakers at TED's conferences have been provided a platform to popularize their research, with some of these speakers even being elevated to "superstar-level" status (Sugimoto et al. 2013). A TED format of condensing complex research into talks under 18 minutes has been recommended as supplements to traditional classroom lectures (Romanelli, Cain and McNamara 2014).

The popularity of TED thus calls for an understanding of why these talks have become such a phenomenon. A listener's proclivity to TED might be related to the content delivered as well as to the style of delivery. Thus, this research is interested in the following research question – "**How has the verbal presentation of content in TED talks evolved over the years (2006-2021)?**". This is further split into 5 narrow research questions (NRQ).

NRQ1: How has the complexity of talks changed over the years?

NRQ2: Have talks become funnier over the years?

NRQ3: Have pronominal choices of speakers changed over the years?

NRQ4: How has the usage of numeric information in talks changed over the years?

NRQ5: What broad topics have been covered in the talks published online from 2006-2016?

The TED talk transcript corpus collected via web-scraping is used for this analysis. The corpus consists of 4000+ TED talks that have been published on the TED website as of 04/04/2021 18:30 IST.

The methods of analysis used includes both descriptive statistical text analysis and topic modelling. The former is used to answer NRQs 1-4 while the latter tackles NRQ5.

Textual analysis reveals that linguistic complexity has increased over the years. Pronominal choices of "we" and "you" have decreased while the usage of "I" remains more or less constant. Speakers who include the audience i.e use "we" or "you" are more inclined to use numerical information in their talks than those who heavily rely on personal experiences i.e use "I". It is also found that speakers who use a lot of "we" tend to make non-humorous talks when compared to those who use a lot of "I" or "you". The most common topics covered by TED included those concerned with human feelings, the environment and health.

The code for this project is hosted at https://github.com/ry05/TED-Talk-Content-Analysis .

Motivation

TED has always been considered to be a highly influential initiative in the world of free education (Sugimoto et al. 2013). However, TED itself is not a catalyst for ground-breaking research. Instead, it only acts as the platform to help propagate research output (Morgan 2014).

Speeches have known to be important elements in bringing about change in how society perceives certain ideas as the delivery influences the adoption of the idea (Roos 2013). At its core, a TED talk is essentially an attempt to bring research into mainstream media using public speaking. Hence, emphasis should be on the *presentation of content* rather than *content that is presented* in TED.

The literature on TED talk presentation includes studies of body language (Chang and Huang 2015, Rost 2018), speaking style (Gallo 2016) and pronominal choices (Tsou, Demarest and Sugimoto 2015). However, there is very limited research in identifying how the verbal presentation has changed over the years. Such temporal analysis will help understand the evolution of the contemporary world's most famous public speaking initiative.

Therefore, this research aims to conduct an analysis of how TED talks are verbally presented and how this has evolved across the years. An understanding of the broad topics that are covered in these talks will help realize what topics matter the most to TED as an organization.

Description of Corpus

The *TED Talk Transcript Corpus* has been used for this analysis. It has been created by collecting data from the official website of TED. The corpus consists of transcripts of 4322 talks(all kinds of TED initiatives) hosted on the website from 27/6/2006 to 5/4/2021. However, some of these had titles such as "None" and "My wish". After removing these less-useful cases, there was a total of 4298 talk transcripts, also called *documents*. An associated dataset with other attributes of the talks including title, duration, number of views, speaker details etc. has also been scraped and collected. The corpus is hosted at https://github.com/ry05/TED-Talk-Content-Analysis/tree/main/data/raw. Table 1 presents some summary statistics about the corpus in terms of basic units of text – syllables, words and sentences. Syllables are important for this analysis as talks are *more often heard than read*.

Table 1 Summary statistics of the TED talk corpus

Text unit	Mean	Standard deviation	Median	IQR	Range
Syllables	2435.15	1397.03	2359.50	1215.25 to 3300	5 to 14945
Words	1736.73	1022.61	1660	848.50 to 2367.75	1 to 11023
Sentences	86.24	54.46	78.50	41 to 117	1 to 524

1. Total number of documents in corpus = 4298

2. IQR is Interquartile range i.e between 25th percentile and 75th percentile

3. All measures rounded to 2 decimal places

All the 4298 talks however were not used for the analysis as many of these were non-conformant to the selection criteria. For example, the talks with less than 15 sentences were observed to be musical performances. This filtering is depicted in figure 1.

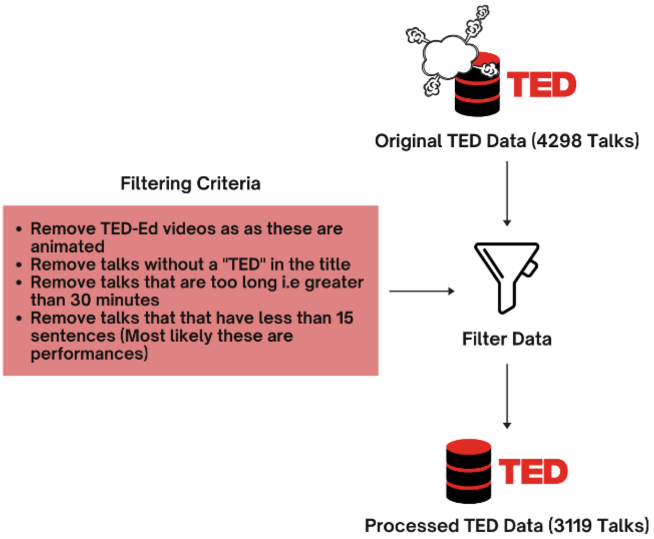

TED
Original TED Data (4298 Talks)
Filtering Criteria
• Remove TED-Ed videos as as these are animated
• Remove talks without a "TED" in the title
• Remove talks that are too long i.e greater than 30 minutes
• Remove talks that that have less than 15 sentences (Most likely these are performances)
Filter Data
TED
Processed TED Data (3119 Talks)

Figure 1. Filtering corpus documents

Description of Methods

It is generally accepted practice to employ quantitative text analysis methods to analyze speech transcripts (Kubát and Cech 2016, 18; Tucker, Capps and Shamir 2020). In the past, TED talks have been subject to quantitative analysis (Tsou, Demarest and Sugimoto 2015). This research employed similar methods and made comparisons across the years. The analysis was performed in 2 phases.

Phase 1 (Descriptive Statistical Text Analysis)

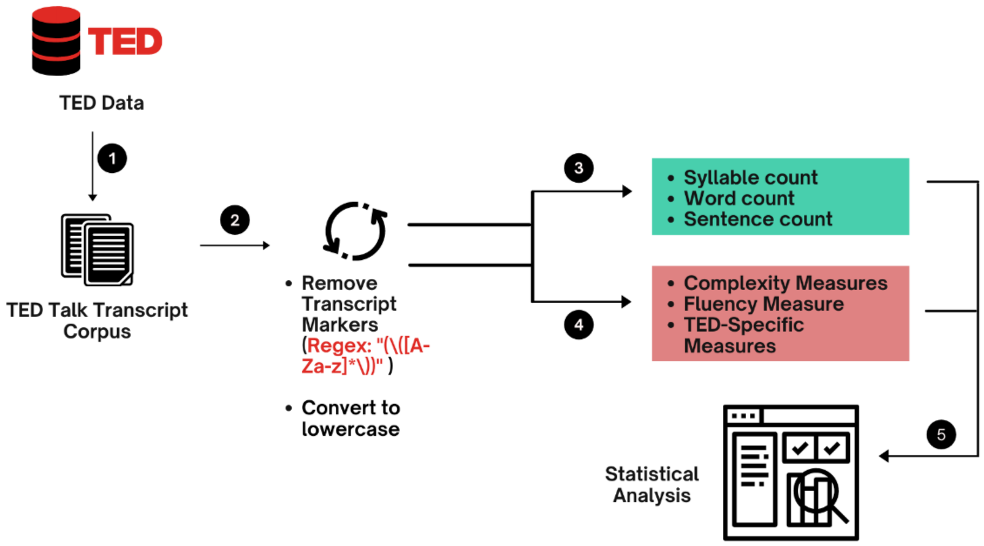


Figure 2. Steps in descriptive statistical text analysis (Steps marked)

Figure 2 describes the steps used in this phase. Transcript markers in step 2 are special formats like "(Laughter)" or "(Applause)" that has been added by TED into their transcripts to describe audience actions. These are not words spoken by the speaker and hence are removed. However, "(Laughter)" is stored separately in order to calculate the LF measure (described below). The measures in the red box are used to understand the verbal presentation of the talks.

Table 2 describes the descriptive measures used in phase 1. The use of "I" indicates that the speaker is speaking of personal experience while "you" might indicate a call to action as well as the inclusion of the audience into the talk. "We" also indicates such inclusion and hints at the author trying to be among her/his audience (Håkansson 2012). Numerical information/word refers to percentage, ordinals, cardinals or numerals (100, "thousands", "percent" etc.). The NIP measure was computed using Named entity recognition (Nadeau and Sekine 2007) and is extracted from phase 2. Henceforth, these descriptive measures would be referred to as *measures.*

Table 2 Descriptive measures used for TED talk analysis

Type of measure	Name of measure	What does it measure	The relevance of the measure
Complexity (Syntactic)	MWS	Median number of words per sentence	Mean used in the past (Lu 2010, 474-496). Median is not affected by outliers like the mean
Complexity (Syntactic)	FKL	Flesch-Kincaid grade level readability measure	Takes syllables into consideration over words (Flesch 1948). Establishes the minimum grade a listener needs to understand the talk
Complexity (Semantic)	MTLD (Measure of Textual lexical diversity)	Textual lexical diversity	Not sensitive to text length (McCarthy 2005, 94-99). Different talks have different lengths
Fluency	SPM	Ratio of number of syllables used to the total length of talk(minutes)	Expresses the fluidity with which a speaker expresses opinion (Kormos and Dénes 2004, 145-164)
TED-specific	LF (Laughter Frequency)	Ratio of number of times the audience "laughs" to the total length of the talk(seconds).	Humour makes audience more comfortable (Presentation Guru 2018)
TED-specific	PM (Pronominal Measures or Pronominals) [Divided into PM_I, PM_You and PM_We]	Ratio of the number of times a speaker says "I", "you" or "we" to the total length of the talk(seconds). For example, if PM_I is 0.05, it is interpreted as "the speaker on average says 'I' in every 100 seconds"	Use of personal pronouns helps to understand whether the speaker is including the audience in their ideas or sharing a personal story (Håkansson 2012)
TED-specific	NIP (Numerical Info Proportion)	Ratio of number of times a speaker uses a numerical word to the total length of talk(seconds). For example, if NIP is 0.05, it is interpreted as "the speaker uses a numerical quantity in speech in every 100 seconds"	Use of numbers or statistical information improves credibility of talk (VirtualSpeech 2017)

TED-specific measures are introduced through this analysis for the first-time

Statistical Analysis

For the statistical analysis, all 3119 talks were divided into 4 groups (1 group = 4 years) by dividing the 16 years (2006-2021). All groups did not have the same number of talks and their within-group variance was heterogenous for all measures. One-way ANOVA (Tae Kyun 2017) was used to check for the effect of time on each of the measures. If the result was statistically significant (5% significance level), the Games-Howell post-hoc test (Games, Keselman and Clinch 1979) was used to compare the groups with each other. Games-Howell is used as the groups are not always normal, have unequal sizes and exhibit heterogeneity of variance.

The post-normalized versions of the pronominals, NIP and LF (i.e before dividing by the duration), referred hereby as *post-norm measures* were used to find correlations between them.

Phase 2 (Topic Modelling)

Topic modelling is a widely used unsupervised, statistical technique that helps understand themes in textual content (Blei 2012, 77). It is used in phase 2 to gain a high-level overview of the broad topics that TED speakers mainly discuss. Latent Dirichlet Allocation (LDA) is a commonly used topic modelling algorithm that works well with large corpuses (Jelodar et al. 2019). While performance tends to be better on corpuses extracted from a specific domain (Korfiatis et al. 2019), it is harder when a wide range of topics are covered such as in the TED talk corpus.

In order to overcome this drawback, Named Entity Recognition (NER) (David and Sekine 2007) was used to extract nouns from the pre-processed transcripts and pass these into LDA. This is useful as nouns are usually what contains topical information in text. In the past, such a noun-only approach to topic modelling has proven to be better than normal LDA (Martin and Johnson 2015). Figure 3 describes this complete process.

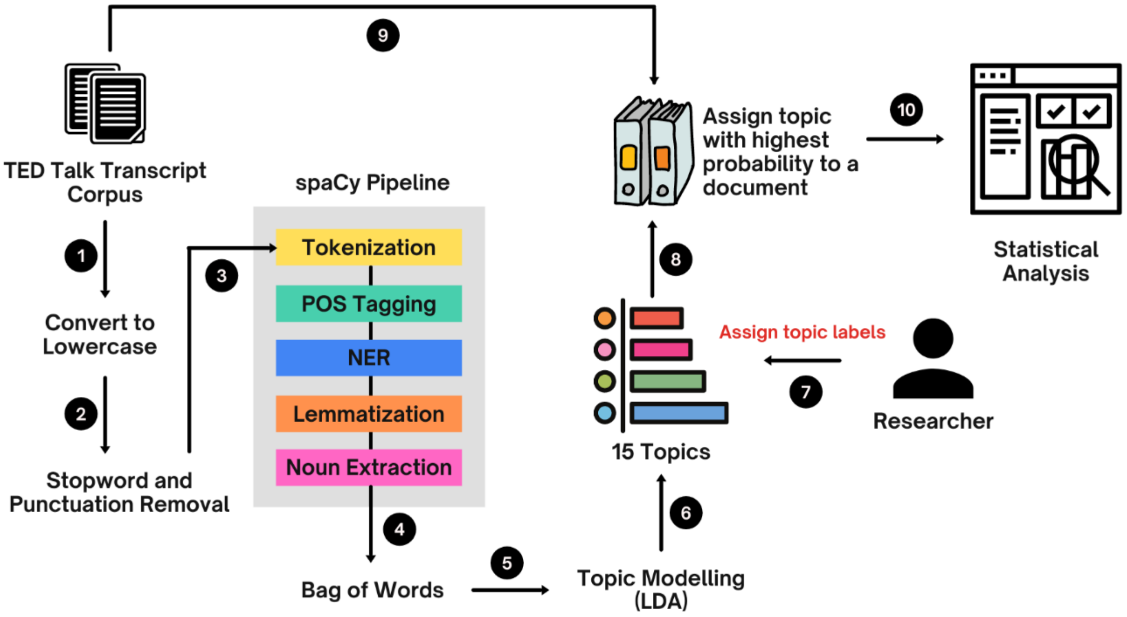


Figure 3. Steps in topic modelling approach (Steps marked)

The spaCy pipeline is a language processing pipeline built with the help of the natural language processing library, spaCy (spaCy n.d.). Tokenization, POS (Part-of-speech) tagging, NER (Named entity recognition) and Lemmatization were pre-built. Noun Extraction was added into the pipeline. Its purpose is to extract all tokens identified as nouns by the POS tagging component. The NER component was used to identify tokens that are numerical information and results in the computation of NIP.

15 topics were identified and labelled with the intervention of the researcher. The 3119 documents were assigned a specific topic label based on which topic was the most frequent in the document.

Results

The results of the analysis are presented in this section in two separate subsections, one for each phase of analysis.

Phase 1 (Descriptive Statistical Text Analysis)

A. Distribution of the descriptive measures

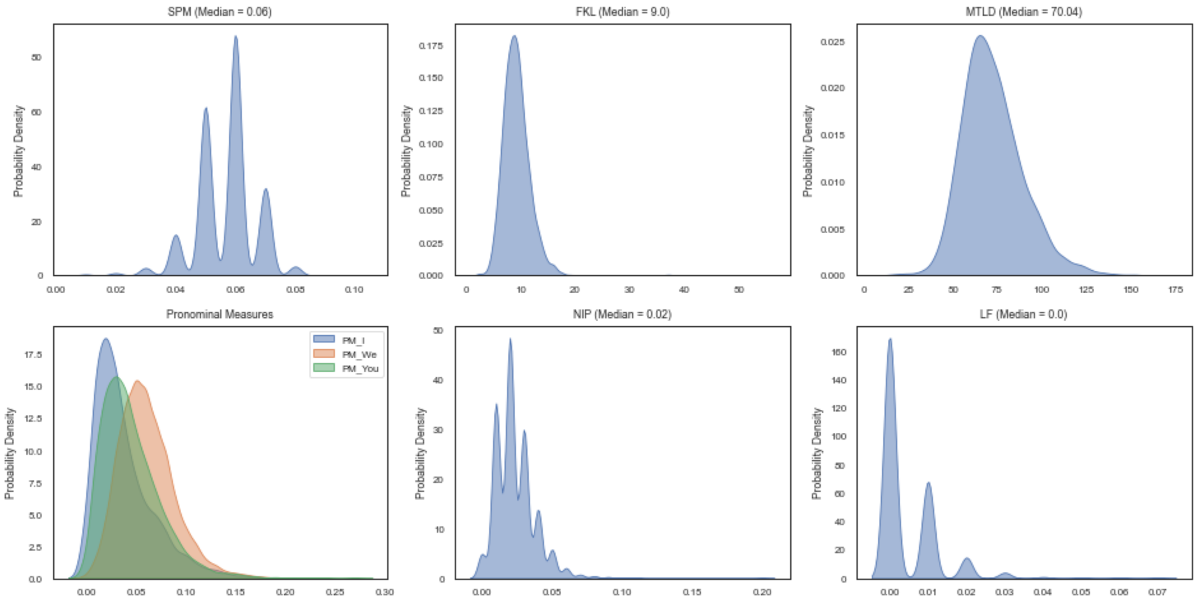


Figure 4. Distribution of descriptive measures (X-axes represent the values of measures)

An *average* (as in median) TED talk can be comprehended by a 9th grader. In general speakers show a slightly higher tendency to use “we” than “you” or “I”.

B. Checking for statistical significance of time effects on the measures

Table 3 displays the results of the one-way ANOVA statistical test to identify if there were significant differences between the 4 time periods across the 9 measures considered. From the ANOVA it is evident that in all measures except PM_I, there is a statistically significant effect of time on at least one measure at the 5% significance level. The Games-Howell post-hoc was conducted on the other 8 measures and table 4 describes the results.

Table 3 One-way ANOVA results testing for differences between the 4 time periods

Name of measure	F-statistic	p-value	Statistically significant?
MWS	39.88	<0.005	Yes
FKL	35.79	<0.005	Yes
MTLD	83.35	<0.005	Yes
SPM	18.74	<0.005	Yes
PM_I	0.18	0.91	No
PM_We	5.30	<0.005	Yes
PM_You	36.44	<0.005	Yes
LF	7.26	<0.005	Yes
NIP	30.93	<0.005	Yes

1. The null hypothesis for all measures is that “time does not have a significant effect on any the measure”

2. All measurements rounded to 2 decimal units

Table 4 Results of Games-Howell Post-hoc test

Name of measure	Time period A	Time period B	Difference of means ± Standard error
MWS	1	3	-2.187 ± 0.30
MWS	1	4	-2.154 ± 0.295
MWS	2	3	-1.454 ± 0.203
MWS	2	4	-1.421 ± 0.195
FKL	1	3	-1.250 ± 0.179
FKL	1	4	-1.163 ± 0.178
FKL	2	3	-0.808 ± 0.113
FKL	2	4	-0.721 ± 0.110
MTLD	1	3	-8.154 ± 0.906
MTLD	1	4	-12.073 ± 0.918
MTLD	2	3	-6.186 ± 0.768
MTLD	2	4	-10.105 ± 0.782
MTLD	3	4	-3.919 ± 0.849
PM_We	2	4	0.005 ± 0.001
PM_We	3	4	0.003 ± 0.001
PM_You	1	3	0.011 ± 0.002
PM_You	1	4	0.013 ± 0.002
PM_You	2	3	0.009 ± 0.001
PM_You	2	4	0.012 ± 0.001

LF	1	2	0.002 ± 0.001
LF	1	3	0.002 ± 0
LF	1	4	0.004 ± 0.001
SPM	1	3	0.056 ± 0.001
SPM	1	4	0.055 ± 0.001
SPM	2	3	0.056 ± 0
SPM	2	4	0.055 ± 0
NIP	1	3	0.005 ± 0.001
NIP	1	4	0.007 ± 0.001
NIP	2	3	0.003 ± 0.001
NIP	2	4	0.005 ± 0.001
NIP	3	4	0.002 ± 0.001

1. The table only includes those entries that are statistically significant at the 5% level (p-value<0.05)
2. Time period keys: {1: 2006-2009, 2: 2010-2013, 3: 2014-2017, 4: 2018-2021}
3. Difference of means calculated as Mean in time period A – Mean of time period B
4. All measurements are rounded to 3 decimal places
5. Red indicates comparisons between the oldest and latest TED talks (2006-2009 and 2018-2021)

NOTE: In the upcoming sections, the usage of the phrase "oldest-latest-difference" indicates the difference of the means of time period 1(2006-2009) and time period 2(2018-2021) in table 4.

C. Answering NRQs 1-4

NRQ1: There are statistically significant effects of time on all 3 complexity measures used (MWS, FKL and MTLD). Negative oldest-latest-difference indicates an increase in complexity over time. This is also supported by figure 5. FKL grade level has made a transition from 8th to almost 10th across the years.

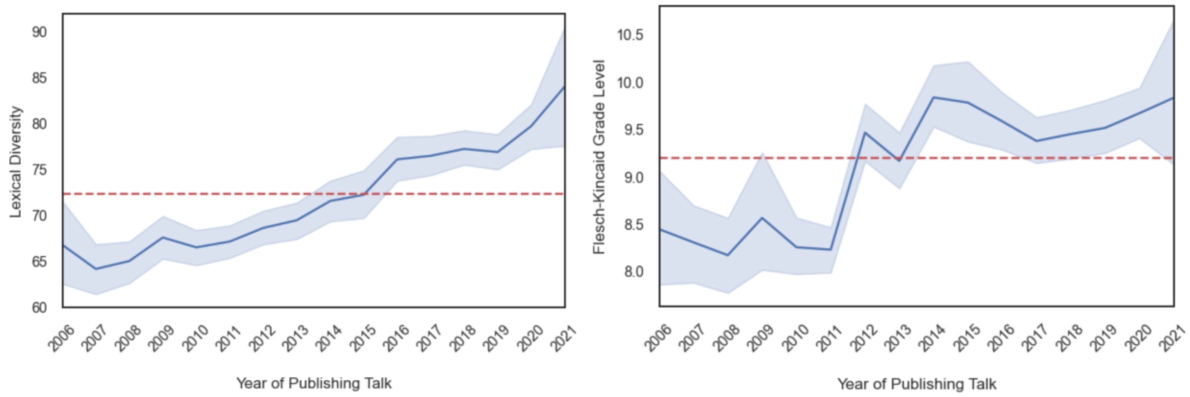


Figure 5. Change in complexity over the years

NOTE: The dark blue line indicates the mean value of the measure for every year. The light blue region represents the 95% confidence interval(CI). The red dotted line is the mean of the measure for all talks.

It is also observed that SPM shows a decrease over time based on its positive oldest-latest-difference value. This could indicate that TED speakers speak slower in recent times compared to older times.

NRQ2: There is statistically significant evidence to consider that in recent times, talks have become less humorous. The sharp dip over the last 2 years in figure 6 might be because of online TED talks where laughter can't be registered by the platform.

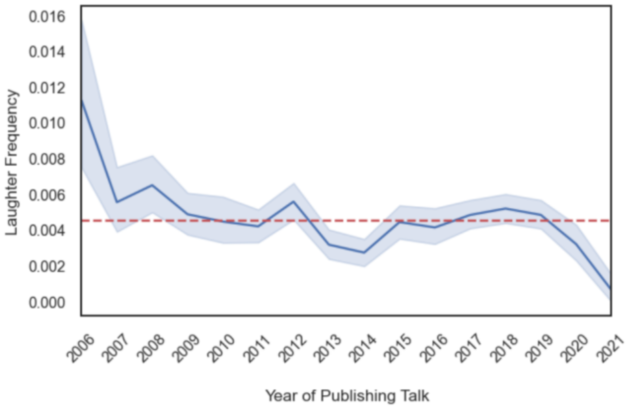


Figure 6. Change in laughter frequency over the years

NRQ3: Statistical evidence suggests that while there is little effect of time on their choice of "I", speakers have in recent times been using lesser "we"s and "you"s.

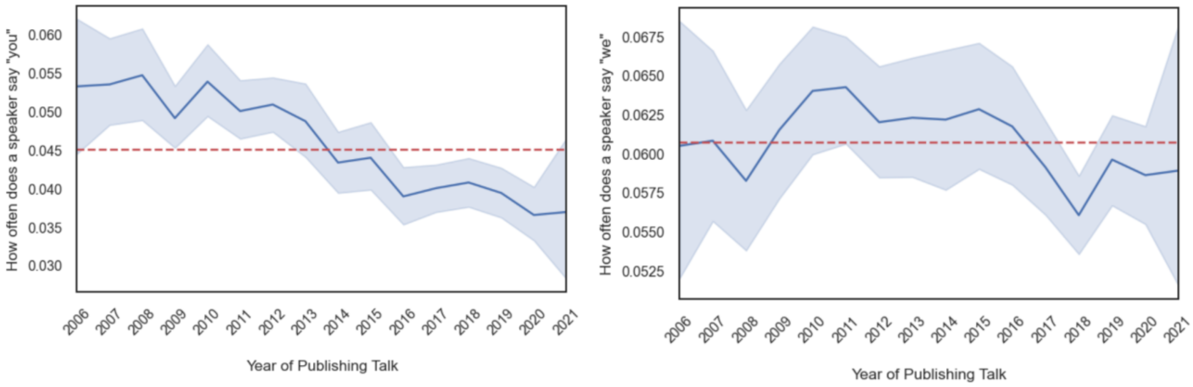


Figure 7. Change in pronominal measures over the years

The very wide CIs of PM_We and the non-existence of oldest-latest-difference for it suggest against making any strong assumptions of its trend, however pronominal choices involving "you" are on the decline as in figure 7.

NRQ4: There is statistical evidence to support the idea that in recent times speakers have reduced their involvement with numerical information in their talks. Figure 8 and a positive oldest-latest-difference support this while the wide CIs do cast some doubt.

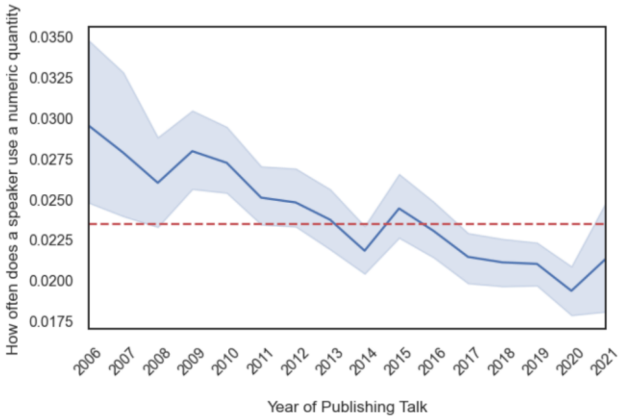


Figure 8. Change in NIP over the years

D. Correlation between the measures

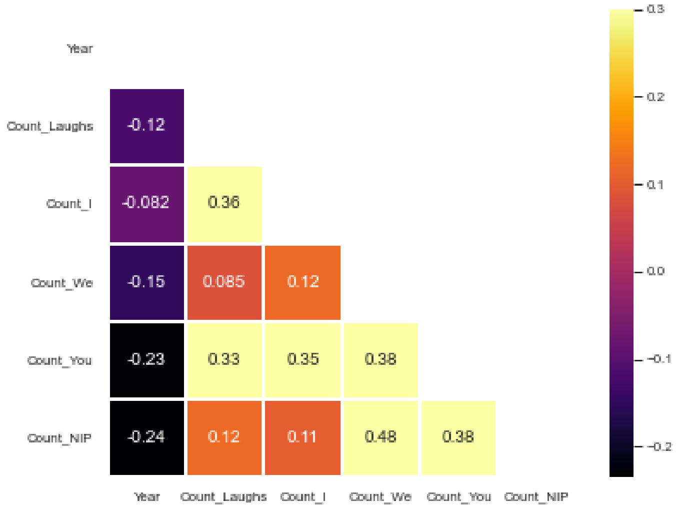


Figure 9. Correlation matrix

Figure 9 represents the Pearson's correlation (Sedgwick 2012) between the post-norm measures (see Description of methods phase 1).

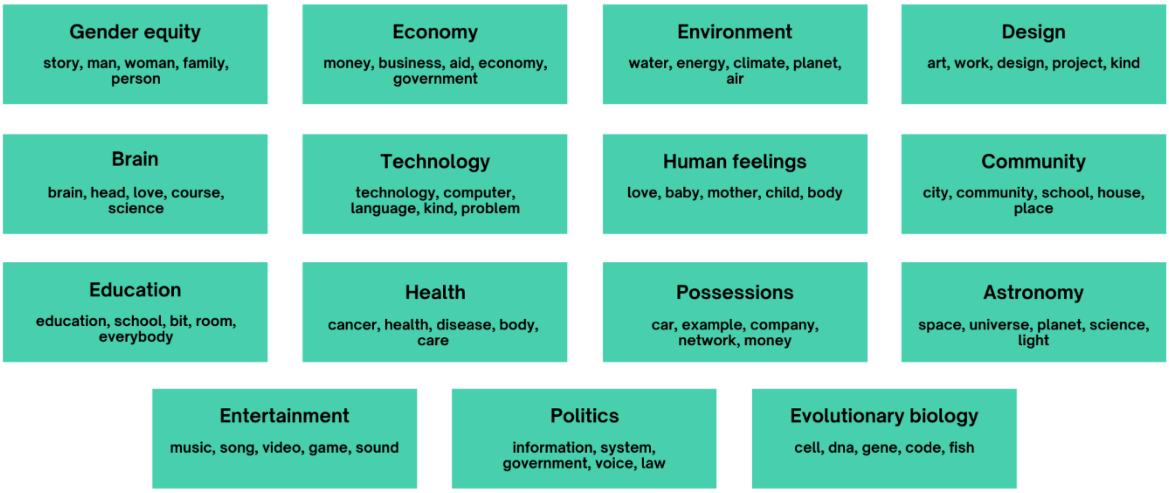
The frequencies of "we" or "you" have much stronger correlations(positive) with the frequency of numerical information in talks than the frequency of "I". This is an indicator of speakers not using as much numerical information when delivering talks motivated by personal experience as they do otherwise.

The correlation between the number of laughs in a talk and the speaker's use of "I" or "you" is almost 400 times the correlation between the former and the speaker's use of "we". This could be because talks with a lot of "we" are poignant, world-changing ideas than don't necessarily have a great scope for humour.

Phase 2 (Topic Modelling)

A. Answering NRQ5

Figure 5 contains the 15 topics that emerged from the topic modelling analysis of TED.



Gender equity
story, man, woman, family, person

Economy
money, business, aid, economy, government

Environment
water, energy, climate, planet, air

Design
art, work, design, project, kind

Brain
brain, head, love, course, science

Technology
technology, computer, language, kind, problem

Human feelings
love, baby, mother, child, body

Community
city, community, school, house, place

Education
education, school, bit, room, everybody

Health
cancer, health, disease, body, care

Possessions
car, example, company, network, money

Astronomy
space, universe, planet, science, light

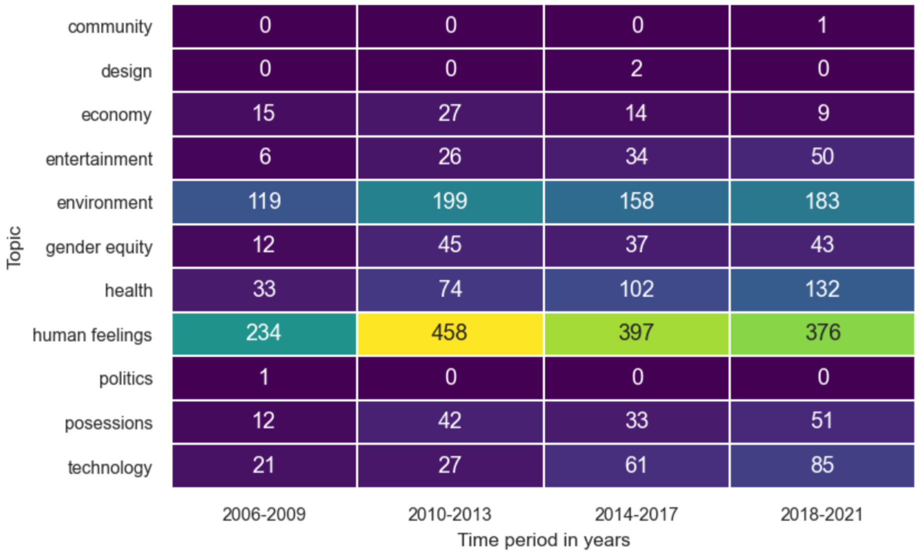
Entertainment
music, song, video, game, sound

Politics
information, system, government, voice, law

Evolutionary biology
cell, dna, gene, code, fish

Figure 10. 15 topics modelled with their researched-labelled names and most relevant words

The above topics were then assigned to each of the 3119 TED talks based on which topic had the highest probability of occurrence. The results were then compared with time as depicted in figure 10.



Topic	2006-2009	2010-2013	2014-2017	2018-2021
community	0	0	0	1
design	0	0	2	0
economy	15	27	14	9
entertainment	6	26	34	50
environment	119	199	158	183
gender equity	12	45	37	43
health	33	74	102	132
human feelings	234	458	397	376
politics	1	0	0	0
posessions	12	42	33	51
technology	21	27	61	85

Figure 10. Most frequent topics across talks

Human feelings, environment and health are the most frequent topics. However, figure 10 is to be read with caution as the results are too dependent on the assigning of topics which is not very reliable as it was performed by a single coder.

Conclusions

The analysis reveals that time has had statistically significant effects on the linguistic complexity, pronominal choices(‘we” and “you) and inclination to using numerical information among TED speakers. It was also realized that speakers who include the audience as a part of their talks were more inclined to use numbers in their speech than those who were speaking of personal experience. Humorous talks contained more “I”s and “you”s than others. Human feelings, environment and health are common topics in TED.

However, it is to be noted that LF as a measure is questionable as it is unclear how TED identifies “laughter” in a talk. Topic modelling errs on reliability as it is limited by the single-coder perspective. Nevertheless, this research is a promising way of looking at the progression of TED over time. Future studies in this domain should be concerned with how to prevent the use of such TED analysis to propagate ethically brittle comparisons of speakers on the basis of their gender or race.

References

Blei, David M. "Probabilistic topic models." *Communications of the ACM* 55, no. 4 (2012): 77-84.

Chang, Yu-jung, and Hung-Tzu Huang. "Exploring TED talks as a pedagogical resource for oral presentations: A corpus-based move analysis." *英語教學期刊* 39, no. 4 (2015): 29-62.

Flesch, Rudolph. "A new readability yardstick." *Journal of applied psychology* 32, no. 3 (1948): 221.

Gallo, Carmine. 2016, "The TED Talk Rule You Should Follow During Presentations". Accessed April 16, 2021. https://www.businessinsider.com/ted-talk-rules-for-presentations-2016-3?IR=T

Games, Paul A., Harvey J. Keselman, and Jennifer J. Clinch. "Tests for homogeneity of variance in factorial designs." *Psychological Bulletin* 86, no. 5 (1979): 978.

Håkansson, Jessica. "The Use of Personal Pronouns in Political Speeches: A comparative study of the pronominal choices of two American presidents." (2012).

Jelodar, Hamed, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey." *Multimedia Tools and Applications* 78, no. 11 (2019): 15169-15211.

Kim, Tae Kyun. "Understanding one-way ANOVA using conceptual figures." *Korean journal of anesthesiology* 70, no. 1 (2017): 22.

Korfiatis, Nikolaos, Panagiotis Stamolampros, Panos Kourouthanassis, and Vasileios Sagiadinos. "Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews." *Expert Systems with Applications* 116 (2019): 472-486.

Kormos, Judit, and Mariann Dénes. "Exploring measures and perceptions of fluency in the speech of second language learners." *System* 32, no. 2 (2004): 145-164.

Kubát, Miroslav, and Radek Cech. "Quantitative Analysis of US Presidential Inaugural Addresses." *Glottometrics* 34 (2016): 14-27.

Lu, Xiaofei. "Automatic analysis of syntactic complexity in second language writing." *International journal of corpus linguistics* 15, no. 4 (2010): 474-496.

Ludewig, Julia. "TED Talks as an emergent genre." *Clcweb-Comparative Literature and Culture* 19, no. 1 (2017).

Martin, Fiona, and Mark Johnson. "More efficient topic modelling through a noun only approach." In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pp. 111-115. 2015.

Masson, Maxime. "Benefits of TED talks." *Canadian Family Physician* 60, no. 12 (2014): 1080-1080.

McCarthy, Philip M. "An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)." PhD diss., The University of Memphis, 2005.

Morgan, Nick. 2014. "What's the problem with TED?" Accessed April 14, 2021. Forbes. https://www.forbes.com/sites/nickmorgan/2014/07/10/whats-the-problem-with-ted/?sh=5a662dce1168

Nadeau, David, and Satoshi Sekine. "A survey of named entity recognition and classification." *Lingvisticae Investigationes* 30, no. 1 (2007): 3-26.

Presentation Guru. 2018. “Using Humour in a Speech or Presentation” Accessed April 17, 2021. https://www.presentation-guru.com/using-humour-in-a-presentation-its-no-laughing-matter/.

Romanelli, Frank, Jeff Cain, and Patrick J. McNamara. "Should TED talks be teaching us something?." *American journal of pharmaceutical education* 78, no. 6 (2014).

Roos, Johan. "The benefits and limitations of leadership speeches in change initiatives." *Journal of Management Development* (2013).

Rost, Madeline M. "TED Talks: What Makes Ideas Worth Spreading?." (2018).

Sedgwick, Philip. "Pearson’s correlation coefficient." *Bmj* 345 (2012).

spaCy. n.d. “Language Processing Pipelines”. Accessed April 10, 2021. https://spacy.io/usage/processing-pipelines

Sugimoto, Cassidy R., Mike Thelwall, Vincent Larivière, Andrew Tsou, Philippe Mongeon, and Benoit Macaluso. "Scientists popularizing science: characteristics and impact of TED talk presenters." *PloS one* 8, no. 4 (2013): e62403.

TED n.d. “TED: Ideas worth spreading” Accessed April 23, 2021. https://www.ted.com.

Tsou, Andrew, Bradford Demarest, and Cassidy R. Sugimoto. "How Does TED Talk? A Preliminary Analysis." *iConference 2015 Proceedings* (2015).

Tucker, Ethan C., Colton J. Capps, and Lior Shamir. "A data science approach to 138 years of congressional speeches." *Heliyon* 6, no. 8 (2020): e04417.

VirtualSpeech. 2017. “Why and How to Bring Statistics Into Your Speech” Accessed April 17, 2021. https://virtualspeech.com/blog/statistics-in-your-speech.