Patterns of Wisdom: Discourse-Level Style in Multi-Sentence Quotations

Kyle Booten and **Marti A. Hearst**
UC Berkeley
Berkeley, CA 94720
kbooten@berkeley.edu, hearst@berkeley.edu

Abstract

Quotations are kernels not just of wisdom but also of beautiful and striking language. While recent studies have characterized the stylistic features of quotations, we characterize the *order* of stylistic information within quotations. Analyzing a corpus of two-sentence quotations collected from the social network Tumblr, we explore the ways that both low-level features and high-level features tend to occur in either the first or second sentence. Through analysis of examples, we interpret these tendencies as manifestations of rhetorical patterns. Results from a prediction task suggest that stylistic patterns are more prominent in quotations than in a comparison corpus.

1 Introduction

The ancient arts of rhetoric described ways of wielding language to make it particularly persuasive or memorable. Central in this endeavor—a predecessor to modern linguistics (Dolven, 2013)—was the description of rhetorical "tropes" or "figures," patterns of language from the level of the phoneme (as in the case of rhyme or alliteration) to higher-level syntactic and even logical structures (Peacham, 1954). In the figure of *epistrophe*, successive clauses end with the same words. In *pysma*, the speaker (or writer) launches a series of sharp and vehement questions; this and other rhetorical figures describe language at the level of discourse—that is, the relationship between linguistic elements across sentences.

Recent studies of quotations have described what makes certain fragments of text more memorable than others (Danescu-Niculescu-Mizil et al., 2012; Guerini et al., 2015); this ongoing project can be seen as a contemporary, empirical investigation of the rhetorical arts. Inspired by the way that classical rhetoricians described complicated and high-level linguistic patterns, the present study takes the novel step of analyzing the way linguistic elements are sequenced within quotations. After describing a minimalistic set of stylistic features designed to capture common patterns in quotations, we demonstrate that some features tend to occur in certain positions within a quotation. We investigate whether quotes are more predictable than other genres in a "Quote Ordering Task" in which the goal is to distinguish the correct version of a quotation from one whose sentences have been reversed.

2 Related Work

2.1 Style of Memorable Language

Danescu-Niculescu-Mizil et al. (2012) used a variety of features to distinguish popular movie quotations from unmemorable lines from the same movie. By modeling word and part-of-speech sequences of quotes and non-quotes, they found that quotations tended to use less common words but that these words were placed in more common syntactic patterns. The researchers evocatively hint at some "common syntactic scaffolding" that structures quotations, and we build on this finding by characterizing these patterns. They also found that quotations tend to contain linguistic features that make them more "generalizable," such as tendency toward indefinite over definite articles.

Other researchers have attempted to analyze quotations in ways that evoke the traditional figures

Proceedings of NAACL-HLT 2016, pages 1139–1144,
San Diego, California, June 12-17, 2016.

of rhetoric. Exploring the same movie quotes corpus as well as other corpora, Guerini et al. (2015) found that memorable quotes were more euphonic, with more instances of rhyme and alliteration than their non-memorable counterparts. Kuznetsova et al. (2013) developed several methods for quantifying the creativity of word combinations like "disadvantageous peace," and found that quotes were more likely to contain creative combinations than non-quotes. In terms of classical rhetoric, such unexpected word combinations could embody figures such as *oxymoron*.

2.2 High-Level Style

To investigate quotations in a new way, we are inspired by researchers who have analyzed text quality using what may be called "high-level" features — i.e., moving beyond the specific lexical, syntactic, or phonemic properties of sentences to explore the overall structure of sentences or even the discursive relationships between them.

Feng et al. (2012) engineered features from sentences' CFG parse trees to describe their overall structure and classify them in terms of *a priori* rhetorical categories, such as *loose* vs. *periodic*. They found that these features were helpful for authorship classification, and a later study used similar features to predict the success of novels (Ashok et al., 2013). Wible and Tsao (2010) and Gianfortoni et al. (2011) designed n-gram based features to capture local lexico-syntactic sequences within sentences.

Looking at the level of discourse, Louis and Nenkova (2012) found that adjacent sentences "exhibit stable patterns of syntactic co-occurence" — i.e., certain types of sentences tend to follow certain other types of sentences. Furthermore, they demonstrated that sentences with similar communicative purposes are syntactically similar, so syntax can be taken as a proxy for communicative purpose. We build on both of these observations.

3 Data-sets

Quotations 1. We gathered a data-set of quotations from the social network Tumblr (Chang et al., 2014). Like other social networks, Tumblr has become a place where users frequently share quotations. In fact, users have the option of using a "quote" data-type, which provides separate text-entry fields for the quotation and its source. We gathered quotes via Tumblr's API. Users sometimes use the "quote" data-type to share messages that are not actually quotations, most often brief personal musings. To minimize such non-quotes in our corpus, we retained only those quotes that users themselves had described with the hashtag "#quote." Since we were interested primarily in discourse-level schemas of quotations, we retained only quotations that were exactly two sentences long. In the pre-processing of this and other corpora we removed repeats within and between corpora. We also removed quotations containing quotations (i.e. reported speech), as they were used in highly inconsistent ways in the Tumblr data. (n=4237)

Quotations 2. A test set. We repeated the same steps described above for Tumblr quotes gathered with the "#quotation" hashtag. (n=1846)

Non-Quotes 1. As a comparison corpus, we gathered two-sentence paragraphs from the Brown corpus (Francis and Kucera, 1979). After the same preprocessing steps, we were left with a collection of such paragraphs. (n=1846)

Non-Quotes 2. Again mining the Brown Corpus, we also gathered sequential pairs of sentences randomly chosen from paragraphs longer than two sentences. (n=1846)

4 Features

Take the following quotation from our data set, attributed to Philip Roth: "You cannot observe people through an ideology. Your ideology observes for you."[1] In simple terms, this quotation is a negative statement ("can*not*") followed by a positive one. Yet the opposite pattern can likewise appear in quotations, as in this one attributed to William Blake: "Great things are done when men and mountains meet. This is not done by jostling in the street." Some quotations begin with questions; others end with them. Some (like the one by Roth) begin with a generic "you," while others deploy this pronoun in the second sentence. We describe each sentence of each quotation in terms of the following features meant to capture general lexical and syntactic pat-

[1] This quote also exemplifies *antimetabole*, in which words in the first clause appear reversed in the second.

Feature	#S1	#S2	χ^2	NQ2?
Highest χ^2				
It	60	**175**	56.3	x
But	18	**93**	50.7	x
it	266	**436**	41.2	x
And	33	**93**	28.6	x
They	7	**44**	26.8	x
Other Unigrams				
People	**21**	2	15.7	
?	**120**	66	15.7	
simply	3	**20**	12.6	
Do	**38**	13	12.3	
Love	**19**	3	11.6	
When	**43**	22	6.8	
n't	**262**	209	6.0	
not	**212**	170	4.6	
What	**45**	28	4.0	
High-Level				
CC + NP + VP .	11	**68**	41.1	x
WHNP + SQ + S + .	**39**	9	18.7	
NP + VP + .	**883**	748	11.2	
WHADVP + SQ + .	**18**	5	7.3	
CC + PP + , + NP + VP + .	2	**12**	7.1	
CC + SBAR + , + NP + VP + .	5	**16**	5.8	
IN + NP + VP + .	2	**10**	5.3	
S + , + S + CC + S + .	**5**	0	5.0	
INTJ + , + NP + VP + .	**7**	1	4.5	
CC + NP + ADVP + VP + .	2	**9**	4.5	

Table 1: Features that preferentially occur in a sentence position, sorted by χ^2 value; dominant sentence position is in bold. *Highest* χ^2 represents the top five unigrams. *Other Unigrams* represents other select examples. NQ2 is whether feature is also ordered in the same way in the comparison corpus ("x" if this is the case).

terns in quotations, regardless of what other more classical rhetorical figures they may contain.

Unigrams. As a baseline feature, we note in which sentence, 1 or 2, a unigram occurs.

High-Level Syntax. Feng et al. (2012) found that the top level of syntactic parse trees (in the case of Stanford PCFG Parser's output, which we also used (Klein and Manning, 2003), two levels beneath `ROOT`) provided a useful feature for authorship identification. For instance, the sentence "Forgotten is forgiven." can be represented by the construction *NP + VP + .*, a noun phrase followed by a verb phrase followed by punctuation. This feature, they argue, provides an interpretable representation of the general syntactic structure of a sentence. We directly employ this feature.

General/Abstract Words. Through qualitative analysis of our data, we noticed that many quotation make pronouncements about nouns that might be considered as generalizations or abstractions. For instance: "Peace comes from within. Do not seek it without." In this case, the abstract noun "peace" is the subject of the first sentence. To capture such nouns, we first use the Stanford Dependency Parser (Chen and Manning, 2014) to extract all words in the head position of nominal subject dependencies (excluding stopwords). Using WordNet, we check whether the word's most common synset is both within the hyponym hierarchy of the synset "abstraction.n.06" and within a minimum distance (5) of it[2]; if so, we consider this noun Abstract.[3] The most common such nouns in the Quotations 1 corpus are not necessarily concept words like "peace". For instance, the word "men" appears in this list; many quotations use the word to evoke a generalized (male) subject.[4] We consider a nominal subject to be General if it is within a minimum distance (6) of its root hypernym. As a feature, we observe which (if either) sentence contains more such Abstract or General nouns, normalized by the number of nominal subject dependencies per sentence.

5 Differences Between Sentence Positions

5.1 Feature Comparison

Using balanced subsets of Quotations 1 and Non-Quotes 1 (n=1846), we investigated which feature was more likely to occur in either one of the two sentences' positions within a two-sentence text. For each feature we used a χ^2 test (α=.05) to compare the number of times the feature occurred in first sentences with the number of times the feature occurred in second sentences. Table 1 presents features with a statistically-significant tendency to appear in one sentence or the other, limited to those features that occur at least 5 times and are among the 300 most common features of its type for that corpus. We suggest that this type of analysis can shed light on some of the overarching stylistic strategies of quotations:

Negative-to-Positive: As shown in Table 1, "n't"

[2]We found this number by taking the whole number above the mean of distances to the "abstraction.n.06" synset of a sample of nouns from Project Gutenberg text. For General words we did the same but averaging distances to a noun's root hypernym.

[3]Kao and Jurafsky Kao and Jurafsky (2012) investigated abstractions in poetry using a dictionary of abstract terms.

[4]"The mass of men lead lives of quiet desperation." (H.D. Thoreau)

and "not" were ore likely to occur in the first sentence position than the second sentence position. This tendency suggests that quotations that begin with a negative construction (like the earlier quote by Roth) are more common than those ending with one (like the earlier quote by Blake). Quotes that contain "not" in the first sentence often use the first sentence to make a negative claim about reality, followed by a positive claim. Such quotations tend to use repetitive structures or other types of parallelism, such as *antimetabole*:

> We are not human beings having a spiritual experience. We are spiritual beings having a human experience. (P. de Chardin)

In quotations, "Never" was also more likely to occur in the first sentence than the second sentence, as was "Do"; studying examples revealed that "Do" was very often followed by "not" or "n't." These statistical tendencies point to the ways that Negative-to-Positive constructions also take the form of a negative commandment (e.g., "Never do *X*") in the first sentence, followed either by a positive commandment or an explanation of the reasoning for the commandment:

> Do not worry about your difficulties in mathematics. I can assure you mine are still greater. (A. Einstein)

Cross-Sentence Conjunction: For both Quotations 1 and Non-Quotes 1, the high-level syntax feature with the highest χ^2 value was *CC + NP + VP + ..* This feature tended to occur in the second sentence for both collections; likewise, for both sentences "But" and "And" were more likely to appear in the second sentence. This is not surprising, as coordinating conjunctions mark the "conjunction" relationship of cohesion (Halliday and Hasan, 2014) (i.e. clauses that begin with conjunctions like "But" implicitly refer back to a previous clause). However, for Quotations 1, this syntax pattern was over six times as likely to occur in the second sentence, compared to nearly two times for Non-Quotes 2, a statistically-significant difference (χ^2, $p<.01$). In the Quotations 1 corpus, other high-level features beginning with a coordinating conjunction were also more likely to occur in sentence 1 than 2, including *CC + PP + , + NP + VP + .* and *CC + SBAR + , + NP + VP + ..* For instance:

> Where a goat can go, a man can go. And where a man can go, he can drag a gun. (William Phillips)

Similarly, the high-level syntax pattern *IN + NP + VP + .* was more likely to occur in the second sentences of quotations than the first sentence; this pattern also indicates cohesion:

> One of the most adventurous things left us is to go to bed. For no one can lay a hand on our dreams. (E.V. Lucas)

We note that either of these quotations could be rephrased as a single sentence, such as:

> One of the most adventurous things left us is to go to bed, for no one can lay a hand on our dreams.

We speculate that there is something stylistically powerful about such sentences in which the second sentence begins with a coordinating or subordinating conjunction. (Perhaps such quotations create a "dramatic" pause for the reader between the sentences.)

Questions: Table 1 shows that high-level syntax patterns that indicate questions, *WHNP + SQ + .* and *WHADVP + SQ + .* occurred more frequently in the first sentence position than the second sentence position, as did the unigram "When." Examining data with the *WHNP + SQ + .* pattern revealed that many of these quotations were actually jokes that take the form of a question/answer dyad.

Sweeping Declarations: For Quotations 1, Abstract Nouns and General Nouns as nominal subjects were more prevalent in the first than in the second sentence (χ^2, $p<.01$). For Non-Quotes 1, General Nouns were also significantly more likely to occur in the first sentence (χ^2, $p<.01$); this was not the case for General Nouns. For quotations only, however, "is" was more likely to occur in the first sentence; likewise, the *NP + VP + .* pattern was also more likely to occur in the first sentences of quotations. Remaining open to other interpretations, we suggest that these facts point to the tendency of quotations to begin with sweeping declarations about "people," "life," "truth," and other broad concepts, kernels of wisdom which the next sentence elaborates or illustrates:

> Love is a trap. When it appears, we see only its light, not its shadows. (P. Coelho)

"Simply": Certain unigrams that tend to occur in a particular sentence can also point to a very specific rhetorical pattern. For instance, the word "simply" was more likely to appear in the second sentence of a quotation than the first. Quotations that manifest this tendency often use this word to emphasize the second sentence's proposition with respect to the first sentence:

I used to dream about escaping my ordinary life, but my life was never ordinary. I had simply failed to notice how extraordinary it was. (R. Riggs)

6 Quote Order Task

We have analyzed stylistic patterns in quotations. However, are these patterns *characteristic* of quotations? To explore this question and to investigate the overall robustness of our features, we define a *Quote Ordering Task*, the goal of which is to distinguish between the original and reversed versions of a quotation. This experiment is in the tradition of tasks for evaluating models of text coherence, such as the one used by Louis and Nenkova (2012). During training, the classifier is shown either the `original` or `reversed` version of a quote. At test time, the classifier must identify each quote as either `original` or `reversed`.

We conducted two experiments. First, using Naive Bayes classifiers, we performed a 5-fold cross-validation test on balanced subsets of three of the four data-sets: Quotations 1, Non-Quotes 1, and Non-Quotes 2 (n=1846 for each). Next we trained on all of Quotations 1 (n=4237) and tested on a separate test set, Quotations 2 (n=1846). In this second test, we trained on both the `original` and `reversed` version of each quote. Table 2 reports results for both tests under various conditions.

For these tests, "high-level features" refers to high-level syntax (of the first sentence, of the second sentence, and both in a sequence), which if either sentence contains more Abstract Nouns, and which if either sentence contain more General Nouns. Combined with unigrams (including stopwords), these features offered slight but not statistically-significant benefit for Non-Quotes 1 in the cross-validation test and slight but not statistically-significant benefit in the second test (testing on Quotations 2). It remains a challenge to integrate such features for classification purposes.

In both tests, however, we were able to predict the order of quotations upwards of 60% of the time. This was not the case for the non-quotes corpora. In cross-validation, the classifier achieved a high score of 62.6% on Quotations 1, 56.0% on Non-Quotes 1, and 52.9% on Non-Quotes 2. The mean top score for Quotations 1 was higher than for the other two collections (two-tailed t-test, $p<.01$). This is evidence that quotations as a genre are more "formulaic" than other textual sequences, their order more easily predicted. We suggest that adherence to latent stylistic patterns is part of what makes quotations seem quotable; as rhetoricians have observed since antiquity, there is power in a pattern.

Feature Set	Q1	NQ1	NQ2	Q2
Unigrams	**62.6**	55.6	**52.9**	63.7
All High-Level Features	57.8	54.1	52.2	58.0
All Features	**62.6**	**56.0**	52.7	**63.9**

Table 2: Performance on Quote Ordering Task for Quotations 1, Non-Quotes 1, and Non-Quotes 2 (5-fold cross-validation, baseline = 50%) and on a separate test set, Quotations 2 (baseline = 52%).

7 Conclusion

We have analyzed linguistic style not merely as the *presence* of features but also the *order* of features across sentences. In quotations, certain words as well as categories of words and syntactic patterns are more likely to appear in the first or second of two-sentence texts. While other genres may also exhibit regularities in the patterning of stylistic information, our results indicate that this stylistic patterning may be especially strong in quotations. Further research could compare a wider variety of genres. Next steps include investigating the relationship to rhetorical goals and running studies with users to determine if they are consciously aware of these stylistic elements when they post quotations. We would like to a better understanding of why people chose to share the quotations they do.

Analyzing discourse-level stylistic tendencies may prove useful for various applications. Bendersky and Smith (2012) demonstrated a method for automatically culling quotations from textual corpora, yet their method was limited to individual sentences. Taking into account the stylistic schemas of quotations could facilitate the gathering of multi-sentence quotations and assist "creative text retrieval" (Veale, 2011) more generally. In the context of social media platforms where quotes circulate, stylistic patterns could also be used to recommend users stylistically-similar quotations to read.

References

Vikas Ganjigunte Ashok, Song Feng, and Yejin Choi. 2013. Success with style: Using writing style to predict the success of novels. *Poetry*, 580(9):70.

Michael Bendersky and David A Smith. 2012. A dictionary of wisdom and wit: Learning to extract quotable phrases. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 69–77.

Yi Chang, Lei Tang, Yoshiyuki Inagaki, and Yan Liu. 2014. What is tumblr: A statistical overview and comparison. *ACM SIGKDD Explorations Newsletter*, 16(1):21–29.

Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *EMNLP*, pages 740–750.

Cristian Danescu-Niculescu-Mizil, Justin Cheng, Jon Kleinberg, and Lillian Lee. 2012. You had me at hello: How phrasing affects memorability. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 892–901. Association for Computational Linguistics.

Jeff Dolven. 2013. Style. In Roland Greene, editor, *The New Princeton Encyclopedia of Poetry and Poetics*, pages 1369–1370. Princeton University Press, Princeton.

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Characterizing stylistic elements in syntactic structure. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1522–1533. Association for Computational Linguistics.

W Nelson Francis and Henry Kucera. 1979. Brown corpus manual. *Brown University*.

Philip Gianfortoni, David Adamson, and Carolyn P Rosé. 2011. Modeling of stylistic variation in social media with stretchy patterns. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 49–59. Association for Computational Linguistics.

Marco Guerini, Gözde Özbal, and Carlo Strapparava. 2015. Echoes of persuasion: The effect of euphony in persuasive communication. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL (NAACL 2015)*, pages 1483–1493.

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. 2014. *Cohesion in english*. Routledge.

Justine Kao and Dan Jurafsky. 2012. A computational analysis of style, affect, and imagery in contemporary poetry. In *NAACL Workshop on Computational Linguistics for Literature*, pages 8–17.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

Polina Kuznetsova, Jianfu Chen, and Yejin Choi. 2013. Understanding and quantifying creativity in lexical composition. In *EMNLP*, pages 1246–1258.

Annie Louis and Ani Nenkova. 2012. A coherence model based on syntactic patterns. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1157–1168. Association for Computational Linguistics.

Henry Peacham. 1954. *The garden of eloquence (1593): a facsimile reproduction*. Scholars' Facsimiles & Reprints.

Tony Veale. 2011. Creative language retrieval: A robust hybrid of information retrieval and linguistic creativity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 278–287. Association for Computational Linguistics.

David Wible and Nai-Lung Tsao. 2010. Stringnet as a computational resource for discovering and investigating linguistic constructions. In *Proceedings of the NAACL HLT workshop on extracting and using constructions in computational linguistics*, pages 25–31. Association for Computational Linguistics.