Inducing Lexical Style Properties for Paraphrase and Genre Differentiation

Ellie Pavlick
University of Pennsylvania
epavlick@seas.upenn.edu

Ani Nenkova
University of Pennsylvania
nenkova@seas.upenn.edu

Abstract
We present an intuitive and effective method for inducing style scores on words and phrases. We exploit signal in a phrase’s rate of occurrence across stylistically contrasting corpora, making our method simple to implement and efficient to scale. We show strong results both intrinsically, by correlation with human judgements, and extrinsically, in applications to genre analysis and paraphrasing.

1 Introduction
True language understanding requires comprehending not just what is said, but how it is said, yet only recently have computational approaches been applied to the subtleties of tone and style. As the expectations on language technologies grow to include tailored search, context-aware inference, and analysis of author belief, an understanding of style becomes crucial.

Lexical features have proven indispensable for the good performance of most applications dealing with language. Particularly, more generalized characterizations of the lexicon (Brown et al., 1992; Wilson et al., 2005; Feng et al., 2013; Ji and Lin, 2009; Resnik, 1995) have become key in overcoming issues with lexical sparseness and in providing practical semantic information for natural language processing systems (Miller et al., 2004; Rutherford and Xue, 2014; Velikovich et al., 2010; Dodge et al., 2012). Most work on stylistic variation, however, has focused on larger units of text (Louis and Nenkova, 2013; Danescu-Niculescu-Mizil et al., 2012; Greene and Resnik, 2009; Xu et al., 2012) and studies of style at the lexical level have been scant. The few recent efforts (Brooke et al., 2010; Brooke and Hirst, 2013b; Brooke and Hirst, 2013a) have been motivated by the need to categorize genre in multiple continuous dimensions and focused on applying standard methods for lexical characterization via graph propagation or crowdsourcing.

We propose a simple and flexible method for placing phrases along a style spectrum. We focus on two dimensions: formality and complexity. We evaluate the resulting scores in terms of their correlation with human judgements as well as their utility in two tasks. First, we use the induced dimensions to identify stylistic shifts in paraphrase, allowing us to differentiate stylistic properties in the Paraphrase Database (PPDB) with high accuracy. Second, we test how well the induced scores capture differences between genres, and explore the extent to which these differences are due to topic versus lexical choice between stylistically different expressions for the same content. We show that style alone

<table>
<thead>
<tr>
<th>Formal/Casual</th>
<th>Complex/Simple</th>
</tr>
</thead>
<tbody>
<tr>
<td>jesus/my gosh</td>
<td>great/a lot</td>
</tr>
<tr>
<td>18 years/eighteen</td>
<td>cinema/a movie</td>
</tr>
<tr>
<td>respiratory/breathing</td>
<td>a large/a big</td>
</tr>
<tr>
<td>yes/yeah</td>
<td>music/the band</td>
</tr>
<tr>
<td>decade/ten years</td>
<td>much/many things</td>
</tr>
<tr>
<td>1970s/the seventies</td>
<td>exposure/the show</td>
</tr>
<tr>
<td>foremost/first of all</td>
<td>relative/his family</td>
</tr>
<tr>
<td>megan/you there</td>
<td>matters/the things</td>
</tr>
<tr>
<td>somewhere/some place</td>
<td>apr/ap</td>
</tr>
<tr>
<td>this film/that movie</td>
<td>journal/diary</td>
</tr>
<tr>
<td>full/a whole bunch</td>
<td>the world/everybody</td>
</tr>
<tr>
<td>otherwise/another thing</td>
<td>burial/funeral</td>
</tr>
<tr>
<td>father/my dad</td>
<td>rail/the train</td>
</tr>
<tr>
<td>recreation/hobby</td>
<td>physicians/a doctor</td>
</tr>
</tbody>
</table>

Table 1: Paraphrases with large style differences. Our method learns these distinctions automatically.
does differentiate between genres, and that the combined indicators of style and topic are highly effective in describing genre in a way consistent with human judgements.

2 Method

We focus on two style dimensions: formality and complexity. We define formal language as the way one talks to a superior, whereas casual language is used with friends. We define simple language to be that used to talk to children or non-native English speakers, whereas more complex language is used by academics or domain experts.

We use the Europarl corpus of parliamentary proceedings as an example of formal text and the Switchboard corpus of informal telephone conversations as casual text. We use articles from Wikipedia and simplified Wikipedia (Coster and Kauchak, 2011) as examples of complex and simple language respectively. For each style dimension, we subsample sentences from the larger corpus so that the two ends of the spectrum are roughly balanced. We end up with roughly 300K sentences each for formal/casual text and about 500K sentences each for simple/complex text.

Given examples of language at each end of a style dimension, we score a phrase by the log ratio of the probability of observing the word in the reference corpus (REF) to observing it in the combined corpora (ALL). For formality the reference corpus is Europarl and the combined data is Europarl and Switchboard together. For complexity, the reference corpus is normal Wikipedia and the combined data is normal and simplified Wikipedia together. Specifically, we map a phrase $w$ onto a style dimension via:

\[ \text{FORMALITY}(w) = \log \left( \frac{P(w \mid \text{REF})}{P(w \mid \text{ALL})} \right). \]

We assign formality scores to phrases up to three words in length that occur at least three times total in ALL, regardless of whether they occur in both corpora. Phrases which do not occur at all in REF are treated as though they occurred once.

3 Evaluation

We first assess the intrinsic quality of the scores returned by our method by comparing against subjective human judgements of stylistic properties.

Phrase-level human judgements For each of our style dimensions, we take a random sample of 1,000 phrases from our corpora. We show each phrase to 7 workers on Amazon Mechanical Turk (MTurk) and ask the worker to indicate using a sliding bar (corresponding to a 0 to 100 scale) where they feel each word falls on the given style spectrum (e.g. casual to formal). Workers were given a high-level description of each style (like those given at the beginning of Section 2) and examples to guide their annotation.

<table>
<thead>
<tr>
<th>Low $\sigma$</th>
<th>Formal</th>
<th>Casual</th>
<th>Complex</th>
<th>Simple</th>
</tr>
</thead>
<tbody>
<tr>
<td>exchange</td>
<td>scrutiny</td>
<td>, uh</td>
<td>all that stuff</td>
<td>per capita referendum proportional</td>
</tr>
<tr>
<td>High $\sigma$</td>
<td>his speech in return for of the series</td>
<td>radio are really to move into</td>
<td>mid japan os</td>
<td>possible center of sets</td>
</tr>
</tbody>
</table>

Table 2: Phrases with high and low levels of annotator agreement, measured by the variance of the human raters’ scores ($\text{Low } \sigma = \text{high agreement}$).

We estimate inter-annotator agreement by computing each rater’s correlation with the average of the others. The inter-annotator correlation was reasonably strong on average ($\rho = 0.65$). However, not all phrases had equally strong levels of human agreement. Table 2 shows some examples of phrases which fell “obviously” on one end of a style spectrum (i.e. the variance between humans’ ratings was low) and some other examples which were less clear.

Quality of automatic scores We compute the correlation of our method’s score with the average human rating for each phrase. The results are summarized in Table 4. The log-ratio score correlates with the human score significantly above chance, even matching inter-human levels of correlation on the formality dimension.

4 Applications

We evaluate the acquired style mappings in two tasks: finding paraphrase pairs with differences in style and characterizing genre variation.
agreed → great → sure → yeah
assumes → implies → imagine → guess
currently → today → now → nowadays
most beautiful → very nice → really nice → really pretty
following a → in the aftermath → in the wake → right after
the man who → one who → the one that → the guy that

Table 3: Groups of paraphrases ordered from most formal (left) to least formal (right), as described in Section 4.1.

<table>
<thead>
<tr>
<th>Spearman ρ</th>
<th>Formality</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter-annotator</td>
<td>0.654</td>
<td>0.657</td>
</tr>
<tr>
<td>Log-ratio score</td>
<td>0.655</td>
<td>0.443</td>
</tr>
</tbody>
</table>

Table 4: First row: mean correlation of each rater’s scores with the average of the others. Second row: correlation of our automatic style score with the average human score.

4.1 Differentiating style in paraphrases

Paraphrases are usually defined as “meaning equivalent” words or phrases. However, many paraphrases, even while capturing the same meaning overall, display subtle differences which effect their substitutability (Gardiner and Dras, 2007).

For example, paraphrasing “I believe that we have...” as “I think we got...” preserves the meaning but causes a clear change in style, from a more formal register to a casual one. It has been proposed that paraphrases are rarely if ever perfectly equivalent, but instead represent near synonyms (Edmonds and Hirst, 2002), which contain subtle differences in meaning and connotation.

We test whether our method can tease apart stylistic variation given a set of “equivalent” phrases. We use phrase pairs from the Paraphrase Database (PPDB) (Ganitkevitch et al., 2013). Using the scoring method described in Section 2, we identify paraphrase pairs which display stylistic variation along a particular dimension. We can find pairs \((w_1, w_2)\) in PPDB for which \(\text{FORMALITY}(w_1) - \text{FORMALITY}(w_2)\) is large; Table 1 gives some examples of pairs identified using this method. We can also view paraphrases along a continuum; Table 3 shows groups of paraphrases ordered from most formal to most casual and Figure 1 shows how paraphrases of the phrase money rank along the formality and complexity dimensions. For example, we capture the fact that money is more formal but simpler than the idiomatic expression a fortune.

Pairwise human judgements To evaluate the goodness of our style-adapted paraphrases, we take a random sample of 3,000 paraphrase pairs from PPDB and solicit MTurk judgements. We show workers each paraphrase pair and ask them to choose which of the words is more casual, or to indicate “no difference.” We also carry out the analogous task for the complexity distinction. We take the majority of 7 judgements as the true label for each pair.

In only 9% of the 3,000 paraphrase pairs, turkers decided there was no stylistic difference in the pair, indicating that indeed formality and complexity differences are truly characteristic of paraphrases. In further analysis we ignore the pairs for which the consensus was no difference but note that in further work we need to automate the identification of stylistically equivalent paraphrases.

Automatically differentiating paraphrases Using the human judgements, we compute the accuracy of our method for choosing which word in a pair is more formal (complex). We use the magnitude of the difference in formality (complexity) score as a measure of our method’s confidence in its prediction. E.g. the smaller the gap in FORMALITY,
the less confident our method is that there is a true style difference. Table 5 shows pairwise accuracy as a function of confidence: it is well above the 50% random baseline, reaching 90% for the high-confidence predictions in the complexity dimension.

<table>
<thead>
<tr>
<th></th>
<th>Top 10%</th>
<th>Top 25%</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complexity</td>
<td>0.90</td>
<td>0.88</td>
<td>0.74</td>
</tr>
<tr>
<td>Formality</td>
<td>0.72</td>
<td>0.73</td>
<td>0.68</td>
</tr>
</tbody>
</table>

Table 5: Pairwise accuracy for paraphrase pairs at varying levels confidence. Top 10% refers to the 10% of pairs with largest difference in log-ratio style score. Random guessing achieves an accuracy of 0.5.

4.2 Genre characterization

Now we explore if the dimensions we learned at the sub-sentential level can be used to capture stylistic variation at the sentence and genre level.

Sentence-level human judgements We gather human ratings of formality and complexity for 900 sentences from the MASC corpus (Ide et al., 2010): 20 sentences from each of 18 genres. Recently data from this corpus has been used to study genre difference in terms of pronoun, named entity, punctuation and part of speech usage (Passonneau et al., 2014). We use the data to test a specific hypothesis that automatically induced scores for lexical style are predictive of perceptions of sentence- and genre-level style.

We average 7 independent human scores to get sentence-level style scores. To get genre-level style scores, we use the average of the 20 sentence-level scores for the sentences belonging to that genre.

In human perception, the formality and complexity dimensions are highly correlated (Spearman $\rho = 0.7$). However, we see many interesting examples of sentences which break this trend (Table 6). Overall, inter-annotator correlations are reasonably strong ($\rho \approx 0.5$), but as in the phrase-level annotations, we see some sentences for which the judgement seems unanimous among annotators and some sentences for which there is very little consensus (Table 7). We discuss this variation further in Section 5.

![Figure 2: Authors reveal style by choosing casual terms or formal terms for the same concept. Shown is a casual sentence (left) and a formal sentence (right) on the same topic. Alternative paraphrases are ordered casual (top) to formal (bottom).](image)

Automatically characterizing genre The extent to which genre is defined by topic versus style is an open question. We therefore look at two methods for genre-level style characterization, which we apply at the sentence-level as well as at the genre-level.

First, we take the average formality (complexity) score of all words in the text, which we refer to as the “all words” method. Using the style score alone in this way will likely to conflate aspects of topic with aspects of style. For example, the word birthday receives a very low formality score whereas the phrase united nations receives a very high formality score, reflecting the tendency of certain topics to be discussed more formally than others.

![Figure 2: Authors reveal style by choosing casual terms or formal terms for the same concept. Shown is a casual sentence (left) and a formal sentence (right) on the same topic. Alternative paraphrases are ordered casual (top) to formal (bottom).](image)
more phrases to its left than to its right). We then score the overall formality of the text as the proportion of times a formal phrase was chosen when a more casual paraphrase could have been chosen instead. The intuition is captured in Figure 2: when an author is writing about a given topic, she encounters words for which there exist a range of paraphrases. Her lexical choice in these cases signals the style independent of the topic.

Table 8 shows how well our two scoring methods correlate with the human judgements of sentences’ styles. The “all words” method performs very well, correlating with humans nearly as well as humans correlate with each other. Interestingly, when using paraphrases only we maintain significant correlations. This ability to differentiate stylistic variation without relying on cues from topic words could be especially important for tasks such as bias detection (Recasens et al., 2013) and readability (Callan, 2004; Kanungo and Orr, 2009).

Table 7: Style ratings of sentences with high and low levels of human agreement, measured by the variance of the human raters’ scores (Low \( \sigma \) = high agreement).

### Formality Complexity

<table>
<thead>
<tr>
<th>Sent. Genre</th>
<th>Sent. Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter-anno.</td>
<td>0.47</td>
</tr>
<tr>
<td>All words</td>
<td>0.44</td>
</tr>
<tr>
<td>PP only</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Table 8: Spearman \( \rho \) of automatic rankings with human rankings. Genres are the concatenation of sentences from that genre. In “all words,” a text’s score is the average log-ratio style score of its words. In “PP only,” a text’s score is the proportion of times a formal term was chosen when more casual paraphrases existed, effectively capturing style independent of topic.

5 Discussion

Characterization of style at the lexical level is an important first step in complex natural language tasks, capturing style information in a way that is portable across topics and applications. An interesting open question is the extent to which style is defined at the lexical level versus at the sentential level: how strongly are human perceptions of style influenced by topic and context as opposed to by lexical choice? One interesting phenomenon we observe is that inter-annotator correlations are lower at the sentence level (\( \rho \approx 0.5 \)) than at the word- and phrase-level (\( \rho \approx 0.65 \)). Tables 7 offers some insight: for many of the sentences for which human agreement is low, there seems to be some mismatch between the topic and the typical style of that topic (e.g., talking formally about family life, or talking in relatively complex terms about Facebook). When humans are making judgements at the lexical level, such contextual mismatches don’t arise, which might lead to higher overall agreements. Interesting future work will need to explore how well humans are able to separate style from topic at the sentence- and document-level, and how the lexical choice of the author/speaker affects this distinction.

6 Conclusion

We present a simple and scalable method for learning fine-grained stylistic variation of phrases. We demonstrate good preliminary results on two relevant applications: identifying stylistic differences in paraphrase, and characterizing variations between genres. Our method offers a simple and flexible way of acquiring stylistic annotations at web-scale, making it a promising approach for incorporating nuanced linguistic information into increasingly complex language applications.\(^3\)

\(^3\)All human and log-ratio scores discussed are available at [http://www.seas.upenn.edu/~nlp/resources/style-scores.tar.gz](http://www.seas.upenn.edu/~nlp/resources/style-scores.tar.gz)
References


Mary Gardiner and Mark Dras. 2007. Corpus statistics approaches to discriminating among near-synonyms.


Attapol Rutherford and Nianwen Xue. 2014. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In
