Project Dialogism:
Toward a Computational History of Vocal Diversity in English-Language Fiction

Adam Hammond
San Diego State University

Julian Brooke
University of Melbourne
Introduction: Investigating Dialogism at Scale

→ Our aim: to develop methods to permit us to study literary dialogism computationally and at scale, in order to answer questions such as:

- Which literary texts and authors are the most dialogic?
- How did dialogism develop chronologically (and how does its development correlate to political events)?
- How did dialogism develop geographically? (Cf. Moretti)
- Which genres are most dialogic?
Introduction: Investigating Dialogism at Scale

This talk focuses on the methods we’ve developed to approach these questions, and closes with some results and some theoretical reflections.

- Six-dimensional approach to quantifying literary style
- GutenTag
- Calculating dialogism
- Preliminary results
- Theoretical reflections: Are we really measuring dialogism?
The Six-Dimensional Approach to Literary Style

1. Objectivity (words that project a sense of disinterested authority, e.g. “invariable,” “ancillary”)

2. Abstractness (words denoting concepts that cannot be described in purely physical terms, and which require significant cultural knowledge to understand, such as “solipsism” and “alienation”)

3. Literariness (words normally found in traditionally literary texts such as “wanton” and “yonder”)

The Six-Dimensional Approach to Literary Style

4. Colloquialness (words used in informal contexts such as “booze” and “crap”)

5. Concreteness (words referring to events, objects, or properties in the physical world, such as “radish” and “freeze”)

6. Subjectivity (words that are strongly personal or reflect a personal opinion, such as “ugly” and “bastard”)
The Six-Dimensional Approach to Literary Style

Process:

- Human annotators review list of 900 words carefully chosen for stylistic properties
- We use this information to derive stylistic information for all words in 2010 DVD image of Project Gutenberg (> 24k texts)
- This data can be used to produce stylistic profiles for any span of text, i.e., a span of character speech
The Six-Dimensional Approach to Literary Style

Sample stylistic profiles for characters in Virginia Woolf’s *To the Lighthouse*:

<table>
<thead>
<tr>
<th>Character</th>
<th>Unique Words</th>
<th>Objective</th>
<th>Abstract</th>
<th>Literary</th>
<th>Colloquial</th>
<th>Concrete</th>
<th>Subjective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mrs. Ramsay</td>
<td>805</td>
<td>0.07</td>
<td>0.24</td>
<td>0.00</td>
<td>0.03</td>
<td>−0.22</td>
<td>0.03</td>
</tr>
<tr>
<td>Mr. Ramsay</td>
<td>70</td>
<td>0.09</td>
<td>0.58</td>
<td>0.27</td>
<td>0.01</td>
<td>−0.49</td>
<td>0.00</td>
</tr>
<tr>
<td>William Bankes</td>
<td>248</td>
<td>−0.01</td>
<td>0.19</td>
<td>0.03</td>
<td>0.14</td>
<td>−0.17</td>
<td>0.08</td>
</tr>
<tr>
<td>Lily Briscoe</td>
<td>1485</td>
<td>0.06</td>
<td>0.17</td>
<td>0.03</td>
<td>−0.02</td>
<td>−0.15</td>
<td>0.01</td>
</tr>
<tr>
<td>James Ramsay</td>
<td>540</td>
<td>−0.06</td>
<td>0.03</td>
<td>0.08</td>
<td>0.06</td>
<td>−0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>Cam Ramsay</td>
<td>381</td>
<td>−0.10</td>
<td>0.04</td>
<td>0.04</td>
<td>0.10</td>
<td>−0.06</td>
<td>0.00</td>
</tr>
<tr>
<td>Charles Tansley</td>
<td>138</td>
<td>−0.07</td>
<td>0.21</td>
<td>−0.07</td>
<td>0.22</td>
<td>−0.23</td>
<td>0.05</td>
</tr>
</tbody>
</table>
The Six-Dimensional Approach to Literary Style

For detailed explanations and discussions, see:


GutenTag is an NLP-driven tool for digital humanities research in the Project Gutenberg corpus.

The high-level goal of the project is to create an ongoing two-way flow of resources between computational linguists and digital humanists, allowing computational linguists to identify pressing problems in the large-scale analysis of literary texts, while giving digital humanists access to a wider variety of NLP tools for exploring literary phenomena. GutenTag is intended to be a standalone software tool for non-programmers, but the source code is also available.

For more information, visit www.projectgutenbtag.org.
GutenTag (www.projectgutentag.org)

→ Open-source software tool for computational research in Project Gutenberg corpora (USA, >44k texts; Australia, 2.5k texts; Canada, 1.5k texts)

→ Why PG? Because it’s big, it’s clean, it’s public domain, and it’s free

→ GutenTag allows users to quickly build large, customized worksets, relying on PG metadata, derived metadata, and a built-in genre classifier

→ For instance, one can quickly collect all prose fiction published between 1880 and 1950, excluding collections
GutenTag (www.projectgutentag.org)
GutenTag (www.projectgutentag.org)

- Can export as plain text or TEI XML
- The latter uses sophisticated ruled-based system to produce structural tags, distinguish narration from character speech, generate lists of characters, and associate spans of speech with specific characters
- Also uses our own literature-specific NER system, LitNER, which outperforms leading NER systems on literary texts
Margaret—<said who="#Margaret">“ I understand , ”</said>said<persName corresp="#Margaret">Margaret</persName>—<said who="#Margaret">“ at least , I understand as much as ever is understood of these things . Tell me now what happened on the Monday morning . ”</said></p>
<p><said who="#Helen">“ It was over at once . ”</said></p>
<p><said who="#Margaret">“ How , <persName corresp="#Helen">Helen</persName>? ”</said></p>
<p><said who="#Helen">“ I was still happy while I dressed , but as I came downstairs I got nervous , and when I went into the dining-room I knew it was no good . There was<persName corresp="#Evie">Evie</persName>— I can n't explain — managing the tea-urn , and<persName corresp="#Mr._Wilcox">Mr. Wilcox</persName>reading the Times . ”</said></p>
<p><said who="#Margaret">“ Was<persName corresp="#St._Paul">Paul</persName>there ? ”</said></p>
<p><said who="#Helen">“ Yes ; and<persName corresp="#Charles_Wilcox">Charles</persName>was talking to him about stocks and shares , and he looked frightened . ”</said></p>
<p>By slight indications the sisters could convey much to each other .<persName corresp="#Margaret">Margaret</persName>saw horror latent in the scene , and<persName corresp="#Helen">Helen</persName>’s next remark did not surprise her .</p>
<p><said who="#Helen">“ Somehow , when that kind of man looks frightened it is too awful . It is all right for us to be frightened , or for men of another sort — father , for instance ; but for men like that ! When I saw all the others so placid , and<persName corresp="#St._Paul">Paul</persName>mad with terror in case I said the wrong
GutenTag (www.projectgutentag.org)

```xml
<particDesc ana="character_list">
  <listPerson>
    <person xml:id="Margaret">
      <persName>Margaret</persName>
      <addName>Meg</addName>
      <addName>Madge</addName>
    </person>
    <person xml:id="Helen">
      <persName>Helen</persName>
      <sex>F</sex>
    </person>
    <person xml:id="Charles_Wilcox">
      <persName>Charles Wilcox</persName>
      <addName>Charles</addName>
    </person>
    <person xml:id="Henry">
      <persName>Henry</persName>
      <sex>M</sex>
    </person>
    <person xml:id="Mr._Wilcox">
      <persName>Mr. Wilcox</persName>
      <sex>M</sex>
    </person>
    <person xml:id="Leonard_Bast">
      <persName>Leonard Bast</persName>
      <addName>Mr. Bast</addName>
      <addName>Bast</addName>
      <addName>Len</addName>
      <addName>Leonard</addName>
    </person>
  </listPerson>
</particDesc>
```
GutenTag (www.projectgutentag.org)

Try it yourself in downloadable and online beta versions at http://www.projectgutentag.org/

See also:


Calculating Dialogism

→ Initial idea: for each style, treat each character as datapoint and calculate weighted variance (weighted by relative proportion of speech by each character) to produce number indicating stylistic variation across characters for that dimension, and average across styles for overall number

→ In practice, unequal spans of text for characters produced unreliable results (short spans tended to produce extreme stylistic results)
Calculating Dialogism

> Revised approach: base metric on stylistic distances between the narrator and two clusters of characters.

> Clusters are formed by grouping the speech of characters with similar styles.

> Parameters include:

  - Minimum words necessary to include character
  - Sample size for direct comparisons, and number of times to compare samples (this helps get useful results when clusters are of different sizes)
Calculating Dialogism

→ Clusters for E. M. Forster’s *Howards End* with sample size set to 1000 and minimum character words set to 200:

- Narrator
- Characters 1: Margaret, Helen, Tibby, Henry
- Characters 2: Mrs. Wilcox, Dolly, Mrs. Munt, Evie, Miss Avery, Charles Wilcox, Miss Schlegel, Mr. Wilcox, Leonard Bast
For *Howards End*, the algorithm found that the two characters groups were strongly differentiated (p < 0.01) in 5 of 6 dimensions — abstract, objective, colloquial, concrete, and subjective — with the most significant distinctions (p < 0.0001) in colloquial and concrete.

- Characters 1: Margaret, Helen, Tibby, Henry
- Characters 2: Mrs. Wilcox, Dolly, Mrs. Munt, Evie, Miss Avery, Charles Wilcox, Miss Schlegel, Mr. Wilcox, Leonard Bast
Preliminary Results

→ Workset composed of GutenTag matches for prose fiction (no collections) published between 1880 and 1950 in PG USA (3608 results), Australia (838), and Canada (565). Texts shorter than *Heart of Darkness* excluded, as well as one long collection of novels.

→ Total of 4008 texts included in experiment.

→ Parameters as follows:

- Minimum character words: 200
- Word sample size: 1000
- Samples: 50
Preliminary Results

- Output can be filtered in terms of stylistic difference between:
  - Narrator vs. all characters
  - Narrator vs. character cluster 1
  - Narrator vs. character cluster 2
  - Character cluster 1 vs. character cluster 2
**Preliminary Results: Some Interesting Findings**

- Stephen Crane’s *The Red Badge of Courage* has the highest difference in two categories: narration vs. all characters, and narration vs. character group 1.
- Virginia Woolf’s *The Waves* has the twelfth-lowest difference between narration and character group 2.
- Upton Sinclair’s *The Jungle* has the fourth-highest difference between character clusters.
- Zane Grey’s novels appear consistently in top-ten groupings of high stylistic difference in all categories.
Preliminary Results

Top ten results for highest difference between narrator and all characters:

1. The Red Badge of Courage by Stephen Crane
2. Teddy and Carrots Two Merchants of Newspaper Row by James Otis
3. Notes of an Itinerant Policeman by Josiah Flynt
4. Drag Harlan by Charles Alden Seltzer
5. The Ridin’ Kid from Powder River by Henry Herbert Knibbs
6. Strangers at Lisconnel by Jane Barlow
7. The Drift Fence by Zane Grey
8. Sundown Slim by Henry Herbert Knibbs
9. Connie Morgan in Alaska by James B. Hendryx
10. Tales of Lonely Trails by Zane Grey
In *The Red Badge of Courage*, narration is distinguished from character speech at $p < 0.00001$ in all six styles, but character clusters are poorly distinguished (the exception is abstract, where $p < 0.05$).
Preliminary Results

The youth put forth anxious arms to assist him, but the tall soldier went firmly on as if propelled. Since the youth's arrival as a guardian for his friend, the other wounded men had ceased to display much interest. They occupied themselves again in dragging their own tragedies toward the rear.

Suddenly, as the two friends marched on, the tall soldier seemed to be overcome by a terror. His face turned to a semblance of gray paste. He clutched the youth's arm and looked all about him, as if dreading to be overheard. Then he began to speak in a shaking whisper:

I tell yeh what I 'm ' fraid of, Henry— I 'll tell yeh what I 'm ' fraid of. I 'm ' fraid I 'll fall down— an' then yeh know— them damned artillery wagons— they like as not 'll run over me. That 's what I 'm ' fraid of—"</p>

The youth cried out to him hysterically: "I 'll take care of yeh, Jim! I 'll take care of yeh! I swear t' Gawd I will!"</p>

Sure— will yeh, Henry?

The tall soldier beseeched.

"Yes— yes — I tell yeh — I 'll take care of yeh, Jim!" protested the youth. He could not speak accurately because of the gulpings in his throat.

But the tall soldier continued to beg in a lowly way. He now hung babelike to the youth's arm. His eyes rolled in the wildness of his terror. "I was allus a good friend t' yeh, wa'n' t I, Henry? And it ain't much t' ask— is it? Jest t' will me close, Henry?"
Are We Really Measuring Dialogism?

A plurality of independent and unmerged voices and consciousnesses, a genuine polyphony of fully valid voices is in fact the chief characteristic of Dosoteovsky’s novels. (6)

Dostoevsky’s novel is *multi-styled or styleless [...] multi-accented* and contradictory in its values. (15)

— M. M. Bakhtin, *Problems of Doestoevsky’s Poetics*
Are We Really Measuring Dialogism?

From the vantage points provided by pure linguistics, it is impossible to detect [...] any really essential differences between a monologic and a polyphonic use of discourse.

What matters here is not the mere presence of specific language styles, social dialects, and so forth, a presence established by purely linguistic criteria; what matters is the dialogic angle at which these styles and dialects are juxtaposed or counterposed in the work. (182)

— M. M. Bakhtin, Problems of Doestoevsky’s Poetics
Are We Really Measuring Dialogism?

Prose, and especially the novel, is completely beyond the reach of such a stylistics. [...] For the prose artist the world is full of other people’s words, among which he must orient himself and whose speech characteristics he must be able to perceive with a very keen ear. [...] And we, when perceiving prose, orient ourselves very subtly among all the types and varieties of discourse analyzed above. [...] We very sensitively catch the smallest shift in intonation, the slightest interruption of voices in anything of importance to us in another person’s practical everyday discourse. (201)

— M. M. Bakhtin, *Problems of Doestoevsky’s Poetics*
A Closing Thought

Seems a bit too much emphasis on deflation to me, especially if that’s what ends the talk.

— Julian Brooke