Detecting Multiple Transitions in Literary Texts

¹Nuette Heyns       ²Menno van Zaanen

¹North-West University, South Africa
²South African Centre for Digital Language Resources, South Africa

nuette.heyns@gmail.com, menno.vanzaanen@nwu.ac.za

LREC 2022
Aim

- Distant reading
- Identifying the high level structure in literary text (Genette et al. 1980)
  1. Changes in location
  2. Different perspectives
  3. Variation in the timeline
- The reason for identifying transitions in literary text
  1. Writing styles
  2. Structural differences
  3. Plot types
## Aim

- Extending previous work (Heyns et al. 2021)
  1. Increase dataset
  2. Identify multiple transitions
- Introduce new system
- Compare new system with previous system
Related work

- Text segmentation/topic shifting
  1. Similarity methods e.g. c99 algorithm Choi et al. (2001)
  2. Lexical chain methods e.g. TextTiling
- The difference
  1. News articles vs Literary text
  2. Topic boundaries vs text transitions
  3. Number of transitions is unknown
Data

- **Pre-processing**
  1. Stopword removal using NLTK
  2. Lowercased, punctuation removed and lemmatized using SpaScy

- **Size of the dataset**
  1. 10 text pairs
  2. 25 snippets from each text
  3. 500 words in each snippet
Evaluation data

- The number of boundaries
  1. Single transition: AB
     1. Snippet 1-25 from text A
     2. Snippet 1-25 from text B
     3. Boundary position is 25
  2. Multiple transitions: ABA
     1. Snippet 1-12 from text A
     2. Snippet 1-25 from text B
     3. Snippet 13-25 from text A
     4. Boundary positions are 12 and 37
Extended Single Transition Identification system
Multiple Transition Identification system
Evaluation

- Extrinsic evaluation method
  1. Root mean squared error (RMSE)

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{n} (p_i - r)^2}{n}}
\]
System comparison

![RMSE results graph](image)

- **STI_Single Transition**
- **STI_Multiple Transitions**
- **MTI_Single Transition**
- **MTI_Multiple Transitions**
Future work

- Experiments on larger texts
- System that works well on both datasets
- Human annotated data
Detecting Multiple Transitions in Literary Texts

1Nuette Heyns   2Menno van Zaanen

1North-West University, South Africa

2South African Centre for Digital Language Resources, South Africa

nuette.heyns@gmail.com, menno.vanzaanen@nwu.ac.za

LREC 2022


References I
### Data extraction

<table>
<thead>
<tr>
<th>Source text</th>
<th>Sentence:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dostoevsky (1866)</td>
<td>…sink upon grass instantly fall asleep morbid condition brain dream often singular actuality vividness extraordinariness resemblance reality time monstrous image create set whole picture fill detail delicate unexpectedly artistically consistent dreamer artist like Pushkin Turgenev even can never invent wake state such sick dream always remain make powerful impression</td>
</tr>
<tr>
<td>Collins (1861)</td>
<td>ruddy face suddenly turn pale leave circus determine find go behind red curtain walk round outside build waste time find door apply admission last come sort passage tattered horsecloth hang outer entrance you come said shabby lad suddenly appear inside mr blyth take lad pocket money greedily valentine hastily enter passage...</td>
</tr>
</tbody>
</table>