Reading demands in secondary school:
Does the linguistic complexity of textbooks increase with grade level and academic orientation of the school track?

Sowmya Vajjala

ALT Brownbag series, Fall 2017, Iowa State University, USA

17 November 2017
Team: Very inter-disciplinary!

- Computational Linguists: Sowmya Vajjala, Detmar Meurers, Maria Chinkina
- Education Researchers: Karin Berendes, Ulrich Trautwein
- Statistician: Wolfgang Wagner
- German as Second Language (GSL?): Doreen Bryant
- two undergrad students Harriet Biederman and Lisa Nassif helped us compile our corpus
- two other undergrad students: Tobias Kolditz, Julia Krings helped us with some of the programming tasks

Initial funding (2013-14): Intramural Research Grants, LEAD Graduate School - University of Tübingen, Germany
What is "text complexity"?

- generally: "inherent difficulty of reading and comprehending a text combined with a consideration of reader and task variables" (CCSSO, 2010)
- considering the text alone, it refers to the properties of language (e.g., grammar, morphology) that make reading comprehension difficult.
Background/Motivation for our work

- Teaching materials have a substantial effect on learning outcome (even in the digital age!)
- Students should read texts such that their linguistic complexity varies as a function of their reading competence. (Common Core Standards in US is an example)
- This is as important for language as it is for content subjects (e.g., Geography, Science etc)
Background/Motivation for our work

- Teaching materials have a substantial effect on learning outcome (even in the digital age!)
- Students should read texts such that their linguistic complexity varies as a function of their reading competence. (Common Core Standards in US is an example)
- This is as important for language as it is for content subjects (e.g., Geography, Science etc)
- This background makes studies on linguistic complexity a problem of significance in education and pedagogy.
- It is also a challenging problem to develop computational approaches that assess text complexity.
Background/Motivation for our work

- Teaching materials have a substantial effect on learning outcome (even in the digital age!)
- Students should read texts such that their linguistic complexity varies as a function of their reading competence. (Common Core Standards in US is an example)
- This is as important for language as it is for content subjects (e.g., Geography, Science etc)
- This background makes studies on linguistic complexity a problem of significance in education and pedagogy.
- It is also a challenging problem to develop computational approaches that assess text complexity.
- However, is the assumption that textbooks are increasingly complex (systematic complexification assumption) true?
Research Questions

- RQ1: Does the linguistic complexity of textbooks increase from grades 5–10? (one would think it does, that is what educational standards say)
- RQ2: Are there differences in linguistic complexity between texts read by academic track versus vocational track students? (may be?)
- RQ3: Are different publishers holding the same standards in terms of linguistic complexity? (they should, if RQ1 is true)
Methods: Corpus

- Geography textbooks from one German state: grades 5–10, across two school tracks - Gymnasium and Hauptschule, 4 different publishers
- OCR + Manual corrections
- Each section in a lesson is treated as one individual text (Why?: Coming in few slides)
- We ended up having 2928 text snippets in total.
- Grades became categories: 5/6, 7/8, 9/10 instead of a continuum of numbers at this stage, because German textbooks are not written for single grade!
Methods: Linguistic Features potentially reflecting complexity

- 165 variables covering: lexical, syntactic, morphological, and discourse aspects of language.
- Extracted with our own code, built on standard German tools for language processing (POS tagger, parser, lists of morphological suffixes, connective words etc)
- Some of these have underlying theories, with others, it is exploratory.
What are some of these features?

<table>
<thead>
<tr>
<th>Feature</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average sentence length</td>
<td>Ratio of genitive nouns to all nouns</td>
</tr>
<tr>
<td>Average number of complex nominals per clause</td>
<td>Ratio of derived nouns to all nouns</td>
</tr>
<tr>
<td>Num. of Adversative and concessive connectors</td>
<td>Third-person personal pronouns</td>
</tr>
<tr>
<td>Passive voice usage ratio</td>
<td>Percentage of difficult words</td>
</tr>
<tr>
<td>Number of Hypernyms per word</td>
<td>Content word overlap</td>
</tr>
</tbody>
</table>

(Full list is in the paper referred in the end)
Analysis Methods-1

RQ1 & RQ2

- text classification with machine learning algorithms
- Goal: Develop a ”model” that can predict the complexity of new texts on the grade scale.
- Ingredients:
  - lots of graded texts (our corpus explained above)
  - programs to extract language features (linguistic features)
  - programs to ”learn” a model that maps feature vectors to grade levels (sequential minimal optimization algorithm)
  - program to use this learned model on new texts (goes through Step 2 and uses the output of Step 3 on new data)
Analysis Methods-2
Interpretable analyses for RQ1 & RQ2

- multi-level regression with restricted, theoretically grounded features (with selected 10 variables out of 165)
- What is it?: hierarchical model used to study problems in which there are multiple levels of variation in data (nested datasets)
- goal: gain an understanding of what contributes to linguistic complexity.
- Why is it relevant here?: complexity can vary at the level of grade, and at the level of school and between publishers (there are other possibilities, we did not explore). So, we need an understanding of the structure of these multiple levels of variation
Analysis Methods-3

RQ3

- ANOVA for unbalanced designs with grade level, school track, and publisher as well as all two- and three-way interactions between the factors.
- Why: explaining which differences are significant.
Results - with classification

- Classification by grade level: (given a text, predict its grade level); random baseline of 33%
  - Overall: 53.7% accuracy
  - Only Gymnasium: 55.7%
  - Only Hauptschule: 53.4%
  - Classification was better for doing only lowest vs highest level instead of all levels.

- Classification by school type: (given a text, predict what school track it belongs to); random baseline: 50%
  - 75-78% for overall, and as well as at each grade level (i.e., if we know we are talking about grade 5/6, does this text suit Gymnasium track or HS track?) scenarios.

- Differences between publishers: 4–18% difference if we train a model with one publisher, and test with another in some cases.
Results - with classification

- Classification by grade level: (given a text, predict its grade level); random baseline of 33%
  - Overall: 53.7% accuracy
  - Only Gymnasium: 55.7%
  - Only Hauptschule: 53.4%
  - Classification was better for doing only lowest vs highest level instead of all levels.

- Classification by school type: (given a text, predict what school track it belongs to); random baseline: 50%
  - 75-78% for overall, and as well as at each grade level (i.e., if we know we are talking about grade 5/6, does this text suit Gymnasium track or HS track?) scenarios.
Results - with classification

Classification by grade level: (given a text, predict its grade level); random baseline of 33%
  - Overall: 53.7% accuracy
  - Only Gymnasium: 55.7%
  - Only Hauptschule: 53.4%
  - Classification was better for doing only lowest vs highest level instead of all levels.

Classification by school type: (given a text, predict what school track it belongs to); random baseline: 50%
  - 75-78% for overall, and as well as at each grade level (i.e., if we know we are talking about grade 5/6, does this text suit Gymnasium track or HS track?) scenarios.

Differences between publishers: 4–18% difference if we train a model with one publisher, and test with another in some cases.
Note: Theory says all these features should increase with grade level, and from vocational to academic track. So, positive co-efficients is a good sign.

<table>
<thead>
<tr>
<th>Linguistic feature</th>
<th>Grade level</th>
<th>School track</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7/8 vs. 5/6</td>
<td>9/10 vs. 7/8</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>p</td>
</tr>
<tr>
<td>Average sentence length (in words)</td>
<td>0.93</td>
<td>.053</td>
</tr>
<tr>
<td>Average word length (in syllables)</td>
<td>0.04</td>
<td>.017</td>
</tr>
<tr>
<td>Average length of longest dependency</td>
<td>0.57</td>
<td>.054</td>
</tr>
<tr>
<td>Average number of complex nominals per clause</td>
<td>0.03</td>
<td>.206</td>
</tr>
<tr>
<td>Root type-token ratio</td>
<td>−0.11</td>
<td>.418</td>
</tr>
<tr>
<td>Modifier variation</td>
<td>0.02</td>
<td>.006</td>
</tr>
<tr>
<td>Ratio of derived nouns to all nouns</td>
<td>0.02</td>
<td>.037</td>
</tr>
<tr>
<td>Ratio of genitive nouns to all nouns</td>
<td>0.01</td>
<td>.002</td>
</tr>
<tr>
<td>Adversative and concessive connectors</td>
<td>0.04</td>
<td>.007</td>
</tr>
<tr>
<td>Third-person personal pronouns</td>
<td>−0.02</td>
<td>.143</td>
</tr>
</tbody>
</table>

Note. The degrees of freedom referring to the cluster level sample size (number of books) were as follows: $df(7/8$ vs. $5/6) = 22$, $df(9/10$ vs. $7/8) = 21$, $df(9/10$ vs. $5/6) = 21$. The $p$-values refer to one-tailed tests. Statistically significant $p$-values are printed in bold. Results were based on a total of 2,928 texts.
<table>
<thead>
<tr>
<th>Linguistic feature</th>
<th>$\eta^2$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Full model</td>
<td>$G$</td>
<td>$S$</td>
<td>$P$</td>
<td>$G \times S$</td>
<td>$G \times P$</td>
<td>$S \times P$</td>
<td>$G \times S \times P$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average sentence length (in words)</td>
<td>.096</td>
<td>.021</td>
<td>.044</td>
<td>.018</td>
<td>.001</td>
<td>.002</td>
<td>.010</td>
<td>.002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average word length (in syllables)</td>
<td>.100</td>
<td>.058</td>
<td>.020</td>
<td>.006</td>
<td>.001</td>
<td>.002</td>
<td>.000</td>
<td>.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average length of longest dependency</td>
<td>.088</td>
<td>.020</td>
<td>.041</td>
<td>.015</td>
<td>.002</td>
<td>.003</td>
<td>.011</td>
<td>.002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average number of complex nominals per clause</td>
<td>.065</td>
<td>.013</td>
<td>.013</td>
<td>.006</td>
<td>.001</td>
<td>.011</td>
<td>.004</td>
<td>.003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Root type-token ratio</td>
<td>.194</td>
<td>.005</td>
<td>.019</td>
<td>.045</td>
<td>.022</td>
<td>.063</td>
<td>.006</td>
<td>.002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modifier variation</td>
<td>.066</td>
<td>.006</td>
<td>.012</td>
<td>.017</td>
<td>.004</td>
<td>.013</td>
<td>.006</td>
<td>.002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ratio of derived nouns to all nouns</td>
<td>.111</td>
<td>.071</td>
<td>.014</td>
<td>.006</td>
<td>.002</td>
<td>.005</td>
<td>.003</td>
<td>.002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ratio of genitive nouns to all nouns</td>
<td>.064</td>
<td>.025</td>
<td>.011</td>
<td>.003</td>
<td>.003</td>
<td>.003</td>
<td>.003</td>
<td>.003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adversative and concessive connectors</td>
<td>.034</td>
<td>.006</td>
<td>.004</td>
<td>.006</td>
<td>.000</td>
<td>.004</td>
<td>.007</td>
<td>.003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Third-person personal pronouns</td>
<td>.036</td>
<td>.006</td>
<td>.008</td>
<td>.001</td>
<td>.005</td>
<td>.005</td>
<td>.001</td>
<td>.003</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note.**  $G = \text{Grade Level (5/6, 7/8, 9/10)}$; $S = \text{School Track (Vocational Track, Academic Track)}$; $P = \text{Publisher (four levels)}$. Bold-faced effects are statistically significant ($p < .05$). The $p$-values were estimated with robust maximum likelihood models (SAS Glimmix procedure with EMPIRICAL option). $\eta^2$ and semipartial $\eta^2$ (based on Type III sums of squares) were estimated with the SAS GLM procedure. Degrees of Freedom: $df(G) = 2$, $df(S) = 1$, $df(P) = 3$, $df(G \times S) = 2$, $df(G \times P) = 6$, $df(S \times P) = 2$, $df(G \times S \times P) = 4$. Results were based on a total of 2,928 texts.
Back to our original Research Questions

- RQ1: Does the linguistic complexity of textbooks increase from grades 5–10? (one would think it does, that is what educational standards say)
- RQ2: Are there differences in linguistic complexity between texts read by academic track versus vocational track students? (may be?)
- RQ3: Are different publishers holding the same standards in terms of linguistic complexity? (they should, if RQ1 is true)
Conclusions-1

- RQ1: Does the linguistic complexity of textbooks increase from grades 5–10?
  ⇒ We have classification accuracies way above chance (50% for all grades, 75% between lowest and highest grades)

- RQ2: Are there differences in linguistic complexity between texts read by academic track versus vocational track students?
  ⇒ Yes, it seems so. Classification with different linguistic features gave good prediction accuracies (75%) at all grade levels.

- RQ3: Are different publishers holding the same standards in terms of linguistic complexity?
  ⇒ No. There are significant differences between publishers in terms of linguistic features, and in terms of prediction accuracies.
Conclusions-1

- **RQ1**: Does the linguistic complexity of textbooks increase from grades 5–10?
  ⇒ We have classification accuracies way above chance (50% for all grades, 75% between lowest and highest grades)

- **RQ2**: Are there differences in linguistic complexity between texts read by academic track versus vocational track students?
  ⇒ Yes, it seems so. Classification with different linguistic features gave good prediction accuracies (75%) at all grade levels.
RQ1: Does the linguistic complexity of textbooks increase from grades 5–10?
⇒ We have classification accuracies way above chance (50% for all grades, 75% between lowest and highest grades)

RQ2: Are there differences in linguistic complexity between texts read by academic track versus vocational track students?
⇒ Yes, it seems so. Classification with different linguistic features gave good prediction accuracies (75%) at all grade levels.

RQ3: Are different publishers holding the same standards in terms of linguistic complexity?
RQ1: Does the linguistic complexity of textbooks increase from grades 5–10?
⇒ We have classification accuracies way above chance (50% for all grades, 75% between lowest and highest grades)

RQ2: Are there differences in linguistic complexity between texts read by academic track versus vocational track students?
⇒ Yes, it seems so. Classification with different linguistic features gave good prediction accuracies (75%) at all grade levels.

RQ3: Are different publishers holding the same standards in terms of linguistic complexity?
⇒ No. There are significant differences between publishers in terms of linguistic features, and in terms of prediction accuracies.
Results provide only partial support for systematic complexification assumption.
Conclusions-2

- Results provide only partial support for systematic complexification assumption.
- Textbooks are not developed systematically in terms of evolving language uses, and there are differences between publishers.

Research in reading is also needed to integrate texts and reader characteristics.
Conclusions-2

- Results provide only partial support for systematic complexification assumption.
- Textbooks are not developed systematically in terms of evolving language uses, and there are differences between publishers.
- More understanding about what makes a text easy or difficult for readers is needed for textbook makers.
Results provide only partial support for systematic complexification assumption.

Textbooks are not developed systematically in terms of evolving language uses, and there are differences between publishers.

More understanding about what makes a text easy or difficult for readers is needed for textbook makers.

Research in reading is also needed to integrate texts and reader characteristics.
Limitations

- Only textbooks from one country, one state, one subject, and written at one point of time.
- Did not take into account reader characteristics, different tasks for which they read.
- We do not have any reading competence data about these texts - we just assumed that textbooks get complex as grade increases (common assumption all complexity research makes).
- We did not consider topic. Our exploration was only about the linguistic properties of text itself.
- Computational tools are also not 100% accurate
Eventual Goal

aid decisions on choosing the right reading materials: ”when (i.e., for which age groups or grade levels) to give what (i.e., the kind and complexity of linguistic features) to whom (i.e., which school track or for good vs. poor readers)”
For more details:

Thank you!

Questions?
contact: sowmya@iastate.edu