Computationally Modeling the Impact of Task-Appropriate Language Complexity and Accuracy on Human Grading of German Essays

Zarah Weiss    Anja Riemenschneider
Pauline Schröter    Detmar Meurers

Department of Linguistics, University of Tübingen
IQB, Humboldt-Universität zu Berlin

14th Workshop on Innovative Use of NLP for Building Educational Applications
Florence, Italy, August 2nd 2019
Introduction

- Complexity and accuracy core components in national educational standards for language arts and literacy (CCSSO 2010; KMK 2012)
- Doubts about teachers’ ability to evaluate complexity and accuracy of texts (CCSSO 2010; Vögelin et al. 2019)
- Assessed manually in German Abitur
  - Official school-leaving state examination
  - Determines admission to university
- Study teachers’ grading behavior in authentic Abitur data
Research Questions and Hypotheses

How do complexity and accuracy influence teachers’
- language performance grades (partial score)?
- content grades (partial score)?
- overall grades (composite score)?

It should be the case that complexity and accuracy
- strongly affect language performance grades
- do not affect content grades
- weakly affect overall grades
## Education System in the U.S. and Germany

<table>
<thead>
<tr>
<th></th>
<th>U.S. System</th>
<th>German System</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Education standard</strong></td>
<td>CCSSO</td>
<td>KMK</td>
</tr>
<tr>
<td><strong>High-stakes testing</strong></td>
<td>repeatedly</td>
<td>final examination</td>
</tr>
<tr>
<td><strong>Qualitative complexity</strong></td>
<td>teachers</td>
<td>teachers</td>
</tr>
<tr>
<td><strong>Quantitative complexity</strong></td>
<td>automatic</td>
<td>teachers</td>
</tr>
<tr>
<td><strong>Automatic Testing industry</strong></td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>
German *Abitur*, Federal States, and the IQB

- *Abitur* = official state examination required for university
- Education is a matter of the German federal states
- The Institute for Educational Quality Improvement (IQB) → monitors schools’ adherence to educational standards → provides an official pool of tasks for the *Abitur* → Includes templates for performance requirements
- States may choose and partially alter tasks from the pool
The Data

- Graded essays from German Abitur in 2017 ($N = 344$)
- Subject: German literature and language examination
- Collected across German states and digitized by the IQB
- Texts respond to one of four different task prompts
  - $2 \times$ interpretation of literature (IL-1, IL-2)
  - $2 \times$ material-based argumentation (MA-1, MA-2)
Task-Effects

- Task prompts request and elicit texts of different length
- Influences correlation of text length and overall grade
- Task-effects are known to influence linguistic complexity (Alexopoulou et al. 2017; Yoon & Polio 2016)
Selecting and Representing Writing Complexity

- Select authentic texts of more and less task-appropriate overall linguistic complexity for the experiment (±ALC)

- Two-fold strategy:
  1. Build document vector representations capturing relevant dimensions of complexity
  2. Create a ranking of these vector representations to identify more and less complex documents

- Separately for each task to account for task-differences
Step 1: Creating Complexity Vectors

Student Essays → Automatic Language Complexity Assessment → Theory-driven Feature Selection → Data-driven Feature Selection → Complexity Vectors

IL-1 Doc 1

Theory-driven Feature Selection

Data-driven Feature Selection

Complexity Vectors

<table>
<thead>
<tr>
<th>Doc 1</th>
<th>Doc 2</th>
<th>...</th>
<th>Doc N</th>
</tr>
</thead>
<tbody>
<tr>
<td>.23</td>
<td>.67</td>
<td>...</td>
<td>.43</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>-.44</td>
<td>.23</td>
<td>...</td>
<td>-.12</td>
</tr>
</tbody>
</table>
Automatic Complexity Assessment

- Automatically extract **320 complexity features** (Weiss 2017)
- Successfully used to assess German readability and L1/L2 development (Weiss & Meurers 2018, 2019, in press)
- Measures of human processing, language use, and lexical, morphological, syntactic, and discourse complexity
- Based on SLA research where Complexity, Accuracy, and Fluency are dimensions of language performance (Bulté & Housen 2012; Wolfe-Quintero et al. 1998)
The Impact of Complexity and Accuracy on Human Essay Grading

Zarah Weiss, Anja Riemenschneider, Pauline Schröter, and Detmar Meurers

Introduction

Outline

Background

The Abitur Data

Our Data

Task-Effects

Complexity Vectors

Building Complexity Vectors

Task-Wise Vector Differences

Similarity-Based Ranking

Experiment

Set-Up

Results

Discussion

Conclusion

References

Appendix

Theoretically-Motivated Complexity Features

► Education standards name examples of welcome writing strategies to make language more complex (KMK 2012)

► Includes argumentation structure, lexical complexity, and syntactic complexity (as well as accuracy)

► Register and norm-appropriateness → academic language (Hennig & Niemann 2013; Snow & Uccelli 2009)

► We identify 75 theoretically-motivated complexity features that are extracted by the system
Theory- and Data-Driven Feature Selection

1. Automatic extraction of 320 complexity features
2. Outlier removal and z-score calculation
3. Calculate the Pearson correlation \( r \) of each complexity feature with essays’ original overall grade \( r_g \)
4. Add theoretically-motivated feature \( f \) ranked by decreasing \( r_g \), if \( f \) correlates
   a. \( \text{abs}(r_g) \geq 0.2; p < 0.05 \) with the overall grade, and
   b. \( \text{abs}(r_f) \leq 0.8 \) with an already added feature
5. Repeat Step 4 for all other features with \( \text{abs}(r_g) \geq 0.3 \)
### Theory- vs. Data-Driven Feature Contribution

<table>
<thead>
<tr>
<th>Task</th>
<th>Theory-Driven</th>
<th>Data-Driven</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>IL-1</td>
<td>20</td>
<td>13</td>
<td>33</td>
</tr>
<tr>
<td>IL-2</td>
<td>32</td>
<td>13</td>
<td>45</td>
</tr>
<tr>
<td>MA-1</td>
<td>13</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>MA-2</td>
<td>9</td>
<td>4</td>
<td>13</td>
</tr>
</tbody>
</table>

- Resulting complexity vectors differ in length
- Most pronounced differences between task objectives (interpretation of literature, material-based argumentation)
- Overall mostly theoretically-motivated features selected
### Zooming in on Complexity Vectors

<table>
<thead>
<tr>
<th>Feature</th>
<th>IL-1</th>
<th>IL-2</th>
<th>MA-1</th>
<th>MA-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTLD</td>
<td>.2014</td>
<td>.4358</td>
<td>.2876</td>
<td>.3361</td>
</tr>
<tr>
<td>Dependent clauses per sentence</td>
<td>.3040</td>
<td>.2528</td>
<td>.2046</td>
<td>-.0380</td>
</tr>
<tr>
<td>Derived nouns per noun phrase</td>
<td>.2394</td>
<td>.4751</td>
<td>.1604</td>
<td>.3301</td>
</tr>
<tr>
<td>Average total integration cost at finite verb</td>
<td>.4093</td>
<td>.4909</td>
<td>.0708</td>
<td>.0308</td>
</tr>
<tr>
<td>Complex noun phrases per noun phrase</td>
<td>.4177</td>
<td>.3186</td>
<td>.1316</td>
<td>-.0353</td>
</tr>
<tr>
<td>Relative clauses per sentence</td>
<td>.3027</td>
<td>.1814</td>
<td>.1381</td>
<td>-.0077</td>
</tr>
<tr>
<td>Dep. clauses w/o conjunction per sentence</td>
<td>.1414</td>
<td>.2460</td>
<td>.0744</td>
<td>.0058</td>
</tr>
<tr>
<td>Conjunctional clauses per sentence</td>
<td>.1632</td>
<td>.2433</td>
<td>.0744</td>
<td>-.0285</td>
</tr>
</tbody>
</table>

- The four vectors include overall 75 unique features
- 18 features generalize across at least three vectors
- Mostly lexical and clausal complexity and nominal style
- Known features of **German academic language**

**The Impact of Complexity and Accuracy on Human Essay Grading**

Zarah Weiss, Anja Riemenschneider, Pauline Schröter, and Detmar Meurers

**Introduction**
- Outline
- Background

**The Abitur Data**
- Our Data
- Task-Effects

** Complexity Vectors**
- Building Complexity Vectors
- Task-Wise Vector Differences
- Similarity-Based Ranking

**Experiment**
- Set-Up
- Results
- Discussion

**Conclusion**
- References
- Appendix
The interpretation of literature task vectors are similar

21/26 features occurring twice are shared by IL-1 and IL-2

Mostly (noun) phrase complexity and human processing

Generalizable characteristics of task objective (interpretation) and type (essay)?
Step 2: Ranking Complexity Vectors

Complexity Vectors

<table>
<thead>
<tr>
<th>Doc 1</th>
<th>Doc 2</th>
<th>...</th>
<th>Doc N</th>
</tr>
</thead>
<tbody>
<tr>
<td>.23</td>
<td>.67</td>
<td>...</td>
<td>.43</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>-.44</td>
<td>.23</td>
<td>...</td>
<td>-.12</td>
</tr>
</tbody>
</table>

Inference of Ideal Complexity Vector

Similarity-based Complexity Vector Ranking

Identification of Comparable +-ALC Documents

Essay Selection

IL-1 +ALC

IL-1 -ALC
Inferring Task-Wise Ideal Complexity Vectors

- Reference vector identifying the polarity of the correlation
- Assign maximal and minimal feature values to feature dimensions of appropriate and inappropriate complexity
- Positive correlations with original overall grade → 1
- Negative correlations with original overall grade → 0

<table>
<thead>
<tr>
<th>Grade Correlations IL-1</th>
<th>IL-1'</th>
</tr>
</thead>
<tbody>
<tr>
<td>Root TTR = .3140</td>
<td>1</td>
</tr>
<tr>
<td>Word in dlexDB = -.3367</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
### Document Ranking

- Force each feature in the complexity vectors to range from 0 to 1 using min-max scaling
- Calculate Manhattan distance between each document vector and its corresponding ideal vector
- Rank documents task-wise by increasing distance
- Rank from more to less appropriate language complexity ($\pm ALC$)
Document Selection and Further Processing

- Consider only documents with a medium overall grade → often more difficult to rate and avoids ceiling/floor effects
- Select texts of comparable length from top and bottom rank → 16 documents selected (2 +ALC and 2 -ALC per task)
- Manual extraction of punctuation, spelling, and grammar errors by the IQB to assess text accuracy
Teachers Participating in Essay (Re-)Grading

- 33 subjects (14 female, 9 male, 0 diverse)
- Age $\mu = 46.4 \pm 8.7$ years; range $= [34; 65]$
- Teaching experience $\mu = 19.9 \pm 9.1$ years; range $= [5; 38]$
- Graded Abitur at least twice, mostly more than 8 times
Provided Materials and Grading Set-Up

▶ Each text was graded by 16 teachers
▶ Mail with 8 texts without original grades (50:50 ±ALC)
▶ Grading at home with Abitur scale: 0 (worst) to 15 (best)
▶ Grading template with content and language requirements
▶ Best approximation of real-life Abitur grading
Evaluation

- Linear mixed regression model for each grade
- Response variable: language, content, or overall grade (re-)assigned by teachers in the experiment
- Predictor variables: $\pm ALC$ and z-scores of $\frac{\sum \text{errors}}{\text{word}}$
- Random intercept for task (IL-1, IL-2, MA-1, MA-2)
## Results: Influence on Language Performance Grades

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Inter.)</td>
<td>6.989</td>
<td>0.561</td>
<td>12.468</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>+ALC</td>
<td>1.374</td>
<td>0.368</td>
<td>3.732</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Error</td>
<td>-1.992</td>
<td>0.211</td>
<td>-9.459</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

- +ALC texts get higher language performance grades
- More errors lead to lower language performance grades
- This confirms our expectations as complexity and accuracy are components of language performance
Results: Influence on Content Grades

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Inter.)</td>
<td>6.138</td>
<td>0.772</td>
<td>7.948</td>
<td>0.003</td>
</tr>
<tr>
<td>+ALC</td>
<td>0.614</td>
<td>0.393</td>
<td>1.562</td>
<td>0.120</td>
</tr>
<tr>
<td>Error</td>
<td>-1.265</td>
<td>0.227</td>
<td>-5.586</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

- No evidence that complexity influences content grading
- More errors lead to lower content grades
- Punctuation, spelling, and grammar errors individually show the same kind of influence
- This **partially violates our expectations** as complexity and error rate are conceptually unrelated to content quality
Results: Influence on Re-Assigned Overall Grades

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Inter.)</td>
<td>6.460</td>
<td>0.696</td>
<td>9.278</td>
<td>0.002</td>
</tr>
<tr>
<td>+ALC</td>
<td>0.703</td>
<td>0.359</td>
<td>1.962</td>
<td>0.051</td>
</tr>
<tr>
<td>Error</td>
<td>-1.518</td>
<td>0.208</td>
<td>-7.316</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

- Marginally significant impact of +ALC on overall grades
- More errors lead to lower overall grades
- Corresponding to the results for the partial grades the impact of error rate is over-proportionally strong
Discussion

Complexity

- Language performance grades successfully reflect differences in quantitative complexity
- Grades experienced teachers assign to ecologically valid texts are not unduly influenced by complexity differences
- Earlier findings for teachers in training do not carry over (Vögelin et al. 2019)

Accuracy

- Accuracy influences all grades – even when it is irrelevant
- This is a problematic issue for German Abitur
- Confirms Rezaei & Lovorn (2010); Cumming et al. (2002)
Conclusion & Outlook

- First results from collaboration of computational linguistic and education science research
- Novel methodology to identify task-appropriate language complexity for document selection
- Teachers identify and modularize language complexity but are clearly biased by accuracy across all grades
- Future work will investigate further the link between automatic and human complexity assessment and grading
References


The Impact of Complexity and Accuracy on Human Essay Grading

Zarah Weiss, Anja Riemenschneider, Pauline Schröter, and Detmar Meurers

Introduction

Outline

Background

The Abitur Data

Our Data

Task-Effects

Complexity Vectors

Building Complexity Vectors

Task-Wise Vector Differences

Similarity-Based Ranking

Experiment

Set-Up

Results

Discussion

Conclusion

References

Appendix


The Impact of Complexity and Accuracy on Human Essay Grading

Zarah Weiss, Anja Riemenschneider, Pauline Schröter, and Detmar Meurers

Introduction

Outline

Background

The Abitur Data

Our Data

Task-Effects

Complexity Vectors

Building Complexity Vectors

Task-Wise Vector Differences

Similarity-Based Ranking

Experiment

Set-Up

Results

Discussion

Conclusion

References

Appendix


The Impact of Complexity and Accuracy on Human Essay Grading

Zarah Weiss, Anja Riemenschneider, Pauline Schröter, and Detmar Meurers

Introduction
Outline
Background

The Abitur Data
Our Data
Task-Effects

Complexity Vectors
Building Complexity Vectors
Task-Wise Vector Differences
Similarity-Based Ranking

Experiment
Set-Up
Results
Discussion

Conclusion

References

Appendix

Task Prompts

Interpretation of literature
- IL-1: Interpretation and comparison of poems
- IL-2: Interpretation of novel ending with given focus

Material-based argumentation
- MA-1: Essay on social media and communication
- MA-2: Comment on dialect use in modern societies
  - Based on 7 to 8 materials (essays, statistics, graphics, ...)
  - Word limits of 1,000 and 800 words
## Task Prompts (cont.)

<table>
<thead>
<tr>
<th>Task</th>
<th>Text Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>IL-1</td>
<td>Interpretation of literature</td>
<td>Interpret poem (A) written in the 1950s and compare it with poem (B) written in the 1980s.</td>
</tr>
<tr>
<td>IL-2</td>
<td>Interpretation of literature</td>
<td>Interpret the given excerpt from novel (A). Focus on the conflicts with which the protagonist struggles.</td>
</tr>
<tr>
<td>MA-1</td>
<td>Material-based argumentation</td>
<td>Write a newspaper essay on the influence social media has on our communication. Use around 1,000 words. Include the following materials in your argumentation: 6 essays, 1 poem, 1 statistic.</td>
</tr>
<tr>
<td>MA-2</td>
<td>Material-based argumentation</td>
<td>Write a newspaper commentary on the influence of dialects and sociolects on success in society. Use around 800 words. Include the following materials in your argumentation: 4 essays, 1 interview, 2 graphics.</td>
</tr>
</tbody>
</table>
The Impact of Complexity and Accuracy on Human Essay Grading
Zarah Weiss, Anja Riemenschneider, Pauline Schröter, and Detmar Meurers

Introduction
Outline
Background
The Abitur Data
Our Data
Task-Effects
Complexity Vectors
Building Complexity Vectors
Task-Wise Vector Differences
Similarity-Based Ranking
Experiment
Set-Up
Results
Discussion
Conclusion
References
Appendix
# German Abitur Grading System

<table>
<thead>
<tr>
<th>Grade</th>
<th>Points</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>excellent +</td>
<td>15</td>
<td>100–95</td>
</tr>
<tr>
<td>excellent</td>
<td>14</td>
<td>94–90</td>
</tr>
<tr>
<td>excellent -</td>
<td>13</td>
<td>89–85</td>
</tr>
<tr>
<td>good +</td>
<td>12</td>
<td>84–80</td>
</tr>
<tr>
<td>good</td>
<td>11</td>
<td>79–75</td>
</tr>
<tr>
<td>good -</td>
<td>10</td>
<td>74–70</td>
</tr>
<tr>
<td>satisfying +</td>
<td>9</td>
<td>69–65</td>
</tr>
<tr>
<td>satisfying</td>
<td>8</td>
<td>64–60</td>
</tr>
<tr>
<td>satisfying -</td>
<td>7</td>
<td>59–55</td>
</tr>
<tr>
<td>sufficient +</td>
<td>6</td>
<td>54–50</td>
</tr>
<tr>
<td>sufficient</td>
<td>5</td>
<td>49–45</td>
</tr>
<tr>
<td>sufficient -</td>
<td>4</td>
<td>44–40</td>
</tr>
<tr>
<td>insufficient +</td>
<td>3</td>
<td>39–33</td>
</tr>
<tr>
<td>insufficient</td>
<td>2</td>
<td>32–27</td>
</tr>
<tr>
<td>insufficient -</td>
<td>1</td>
<td>26–20</td>
</tr>
<tr>
<td>failed</td>
<td>0</td>
<td>19–0</td>
</tr>
</tbody>
</table>
## Content Grades and Spelling Error Rate

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Inter.)</td>
<td>5.976</td>
<td>0.802</td>
<td>3.335</td>
<td>0.003</td>
</tr>
<tr>
<td>+ALC</td>
<td>0.934</td>
<td>0.444</td>
<td>2.101</td>
<td>0.037</td>
</tr>
<tr>
<td>Spelling</td>
<td>-1.197</td>
<td>0.257</td>
<td>-4.651</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>
Content Grades and Grammar Error Rate

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Inter.)</td>
<td>5.954</td>
<td>0.392</td>
<td>15.172</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>+ALC</td>
<td>0.943</td>
<td>0.379</td>
<td>2.489</td>
<td>0.013</td>
</tr>
<tr>
<td>Grammar</td>
<td>-1.197</td>
<td>0.257</td>
<td>-4.651</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>
## Content Grades and Punctuation Error Rate

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Inter.)</td>
<td>6.484</td>
<td>0.534</td>
<td>12.136</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>+ALC</td>
<td>-0.1016</td>
<td>0.382</td>
<td>-0.266</td>
<td>0.790</td>
</tr>
<tr>
<td>Punctuation</td>
<td>-0.5968</td>
<td>0.1939</td>
<td>-3.078</td>
<td>0.002</td>
</tr>
</tbody>
</table>
Complexity in Second Language Acquisition

- Complexity is an important construct in SLA research
- Language performance = Complexity, Accuracy, Fluency (Bulté & Housen 2012; Wolfe-Quintero et al. 1998)
- Complexity = language elaboration and variety (Ellis 2003)
- Accuracy = native speaker-like error rate (Pallotti 2009)
- Fluency = native speaker-like production rate (Pallotti 2009)
The Impact of Complexity and Accuracy on Human Essay Grading

Zarah Weiss, Anja Riemenschneider, Pauline Schröter, and Detmar Meurers

Introduction
- Outline
- Background

The Abitur Data
- Our Data
- Task-Effects

Complexity Vectors
- Building Complexity Vectors
- Task-Wise Vector Differences
- Similarity-Based Ranking

Experiment
- Set-Up
- Results
- Discussion

Conclusion

References

Appendix

NLP Pipeline
Lexical Complexity

- Describes the elaboration, inter-relatedness, and variation of the lexical system
- Measures vocabulary range and size as well as semantic relatedness
- E.g., type token ratio, lexical density, hyponyms per word
- Bulté & Housen (2014); Wolfe-Quintero et al. (1998)
Syntactic Complexity

- Describes the elaboration and variation of the syntactic domain (often split in clausal and phrasal complexity)
- Measures clausal and phrasal modification and variation
- E.g., % dependent clauses types, NP modifiers per NP
- Kyle (2016); Bulté & Housen (2014); Wolfe-Quintero et al. (1998)
Morphological Complexity

- Describes the elaboration and variation of the morphological system
- Measures derivation, composition, and inflection
- E.g., periphrastic tenses per verb, avg. compound depth
- Pallotti (2015); Bulté & Housen (2014); Hancke et al. (2012)
Discourse Complexity

- More elaborate, inter-related, and varied discourse relations are more complex
- Includes measures of cohesive markers, transition probabilities, co-reference chains
- E.g., connectives per sentence, probability subject of drops
- Origin from theoretical and psycho-linguistic research
- Todirascu et al. (2013); Barzilay & Lapata (2008); Graesser et al. (2004)
Language Use

- Assume that less frequently used or later acquired constructions are more complex
- Includes word or phrase frequency measures from large corpora or age of acquisition (AoA) measures
- E.g., mean AoA per word, mean frq. in dlexDB per word
- Origin from corpus- and psycho-linguistic research
- Paquot (2019); Chen & Meurers (2016); Birchenough et al. (2017)
Human Language Processing

- Measures cognitive complexity through processing times as measures by eye-tracking and reading time
- Includes measures of cognitive load and surprisal
- E.g., maximal DLT integration cost per verb
- Origin from cognitive science, psycho-linguistics, and information theory
- Shain et al. (2016); Gibson (2000)