

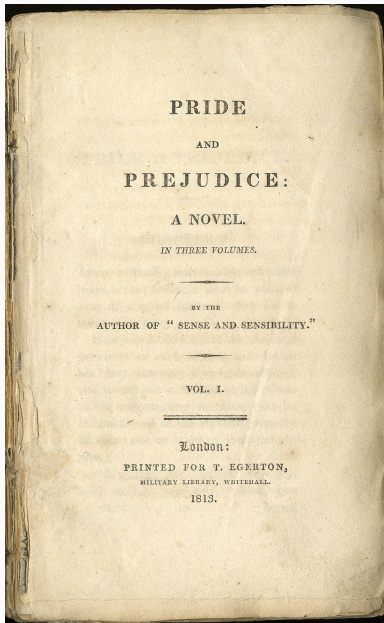
JANE AUSTEN'S PRIDE & PREJUDICE

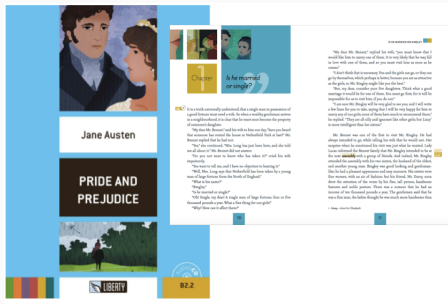
A COMPUTATIONAL STUDY OF TEXT ADAPTATION

Emily Franzini & Marco Böhler
Digital Humanities Conference 2017, Montréal, Canada



JANE AUSTEN'S PRIDE & PREJUDICE



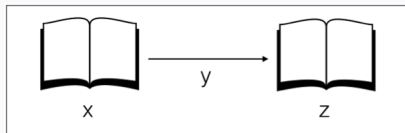


Definition:

Graded readers are “**simplified books** written at **varying levels of difficulty** for second language learners”, which “cover a huge range of genres ranging from adaptation of classic works of literature to original stories, to factual materials such as biographies, reports and so on” [Waring 2012].

To computationally analyse the process Y and classifying the changes:

- Do the changes follow strict rules?
- Do they form patterns?
- Can they be computationally reproduced?



TYPES OF CHANGES

1. Structural changes:

- I do not wish to be too **hasty**.
- We must not **conceal** it.

2. Cognitive changes:

- ... **Soon after this event**, Elizabeth received a visit...

3. Structural & cognitive changes:

- Elizabeth is **exceedingly handsome**.

THE POWER OF COGNITIVE CHANGES

Stylistic analyses of the original novel compared to an automatic text simplification (ATS) and to a human-made graded reader.

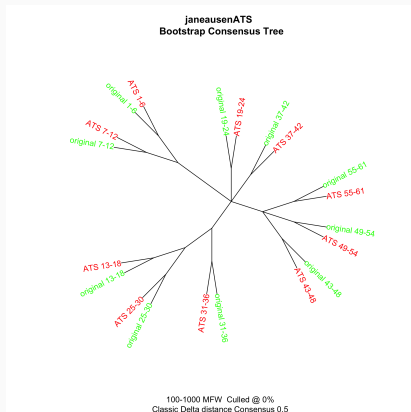


Figure 1: Dendrogram of the ON compared to ATS.

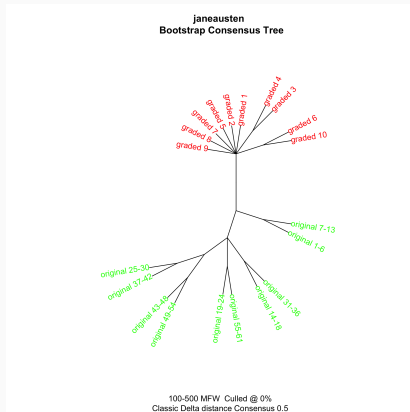
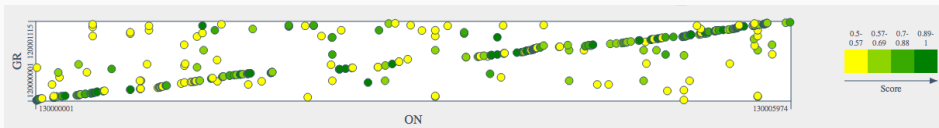


Figure 2: Dendrogram of the ON compared to the GR.

DOT PLOT VIEW OF THE REUSES



Text Re-use Alignment Visualization

X

GR

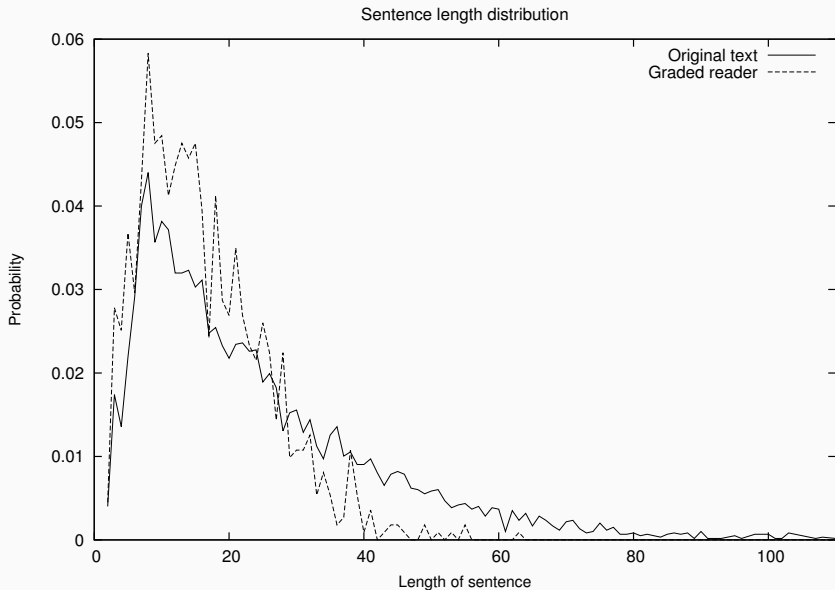
chapter 1 it be a truth universally understand that a single man in possession of a good fortune must need a wife

ON

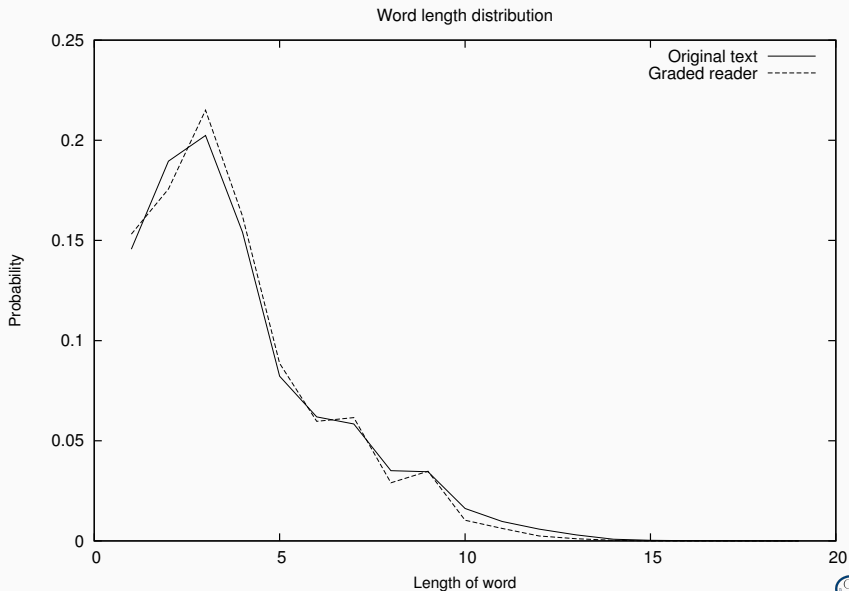
chapter 1 it be a truth universally acknowledge that a single man in possession of a good fortune must be in want of a wife



COMPARISON OF SENTENCE LENGTH



COMPARISON OF WORD LENGTH



DIFFERENCE ANALYSIS FOR PART-OF-SPEECH TAGS

PoS	More frequent in ON	Similar frequency	More frequent in GR
JJS adjective, superlative	X		
JJR adjective, comparative	X		
PDT predeterminer	X		
RBS adverb, superlative	X		
WDT WH-determiner	X		
FW foreign word	X		
: colon	X		
WP\$ WH-pronoun, possessive	X		
NNPS noun, proper, plural	X		
SYM symbol	X		
RP particle		X	
RB adverb		X	
VB verb, base form		X	
TO 'to' as preposition		X	
JJ adjective or numeral, ordinal		X	
NNS noun, proper, singular		X	
CC conjunction, coordinating		X	
PRP\$ pronoun, possessive		X	
NN noun, common, singular		X	
MD modal auxiliary		X	
IN preposition or conjunction, subordinating		X	
DT determiner		X	
VCN verb, past participle		X	
VBG verb, present participle		X	
POS genitive marker		X	
RBR adverb, comparative		X	
EX existential 'there'		X	
UH interjection			X
NNP noun, proper, plural			X
WRB WH-adverb			X
VBD verb, past tense			X
VBP verb, present tense, not 3rd person singular			X
VBZ verb, present tense, 3rd person singular			X
WP WH-pronoun			X
CD numeral, cardinal			X
PRP pronoun, personal			X

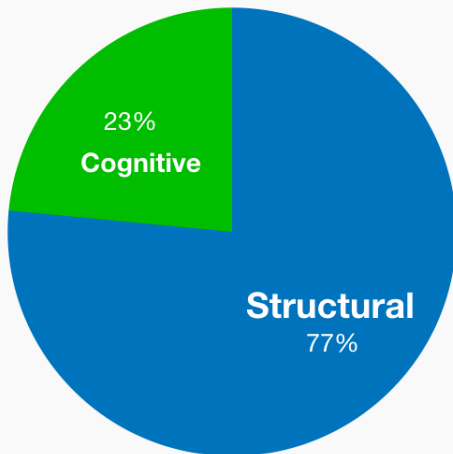
DETAILED DIFFERENCE ANALYSIS

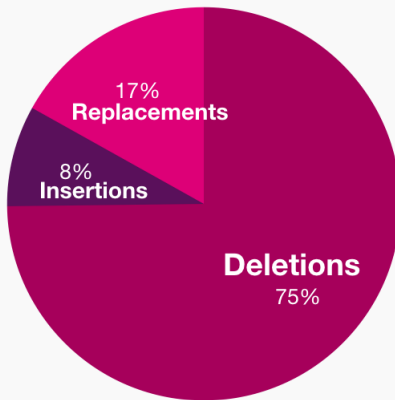
- Sample size of 10% of the Graded Reader

Sentence ID	ORIGINAL NOVEL	Sentence ID	GRADED READER	FUNCTION	SPECIFIC FUNCT	TYPE	KIND	X	SPECIFIC-X	Y
1200001	it	1300001	it							
1200001	is	1300001	is							
1200001	a	1300001	a							
1200001	truth	1300001	truth							
1200001	universally	1300001	universally							
1200001	acknowledged	1300001	understood	repl	sem-rei	struc	simple			
1200001	that	1300001	that							
1200001	a	1300001	a							
1200001	single	1300001	single							
1200001	man	1300001	man							
1200001	in	1300001	in							
1200001	possession	1300001	possession							
1200001	of	1300001	of							
1200001	a	1300001	a							
1200001	good	1300001	good							
1200001	fortune	1300001	fortune							
1200001	must	1300001	must							
1200001	be in want of	1300001	need	repl	sem-rei	cog	hist			
1200001	a	1300001	a							
1200001	wife	1300001	wife							

DETAILED DIFFERENCE ANALYSIS

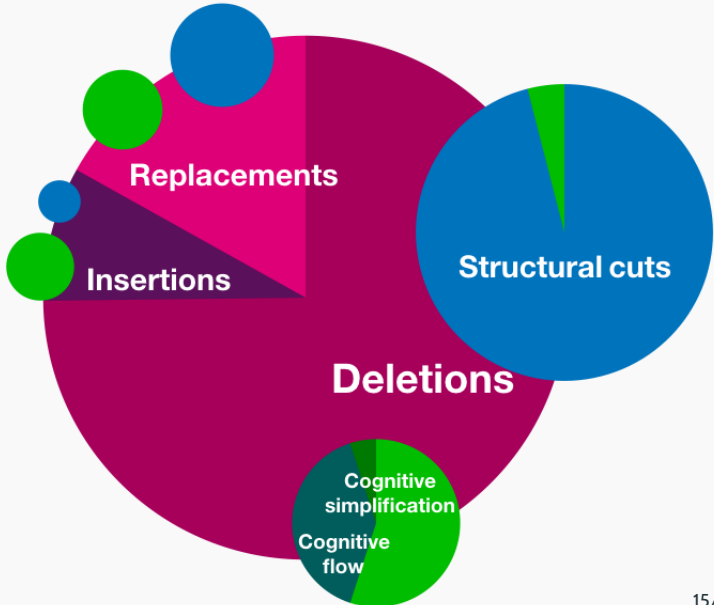
- Distribution by type of change





- Average of 6.73 changes per sentence
- Changes cover more than one third of the text

DETAILED DIFFERENCE ANALYSIS



- Changes of Parts of Speech
- Replacement of multi-word expressions
- Resolution of personal pronouns
- Next step: run the same analysis on other graded readers at different difficulty levels

Speaker

Emily Franzini & Marco Böhler.

Visit us



<http://www.etrp.eu>



contact@etrp.eu

Stealing from one is plagiarism, stealing from many is research.
(Wilson Mitzner, 1876-1933)



SPONSORED BY THE

Federal Ministry
of Education
and Research

The theme this presentation is based on is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License. Changes to the theme are the work of eTRAP.



STATISTICS FOR “DELETION” OPERATIONS

Type	Ratio	cut	flow	hist	simple
Cognitive	11.29%	3.57%	39.29%	1.79%	55.36%
Structural	88.71%	95.91%	0.45%	0.68%	2.95%

Observation 1: Majority (88.71%) of deletion operations are for structural matters.

Observation 2: Most (95.91%) structural deletions are for the sake of “cutting”.

Observation 3: Most (88.71%) cognitive deletions are for the sake of “flow” and “simplification”.

STATISTICS FOR “INSERTION” OPERATIONS

Type	Ratio	cut	flow	hist	simple
Cognitive	87.27%	0.00%	66.67%	0.00%	33.33%
Structural	12.73%	28.57%	14.29%	0.00%	57.14%

Observation 1: Majority (87.27%) of deletion operations are for cognitive matters.

Observation 2: Ratio between cognitive and structural operations are nearly identically flipped for “Deletion” and “Insertion”.

Observation 3: Both for “Deletion” and “Insertion” operations historical spelling (“hist”) plays nearly no role.

STATISTICS FOR “REPLACEMENT” OPERATIONS

Type	Ratio	cut	flow	hist	simple
Cognitive	44.26%	0.00%	38.89%	24.07%	37.04%
Structural	55.74%	1.47%	4.41%	27.94%	66.18%

Observation 1: More balanced ratio between cognitive and structural operations compared to “Deletion” and “Insertion”.

Observation 2: Normalisation of historical variants play a more important role for “Replacement” operations.