Evidence-based Health Text Simplification using Natural Language Processing Algorithms

Gondy Leroy, PhD
Associate Professor

Consulting: gondyleroy@gmail.com
Academic Affiliations:
University of Arizona, gondyleroy@email.arizona.edu
Claremont Graduate University, gondy.leroy@cgu.edu

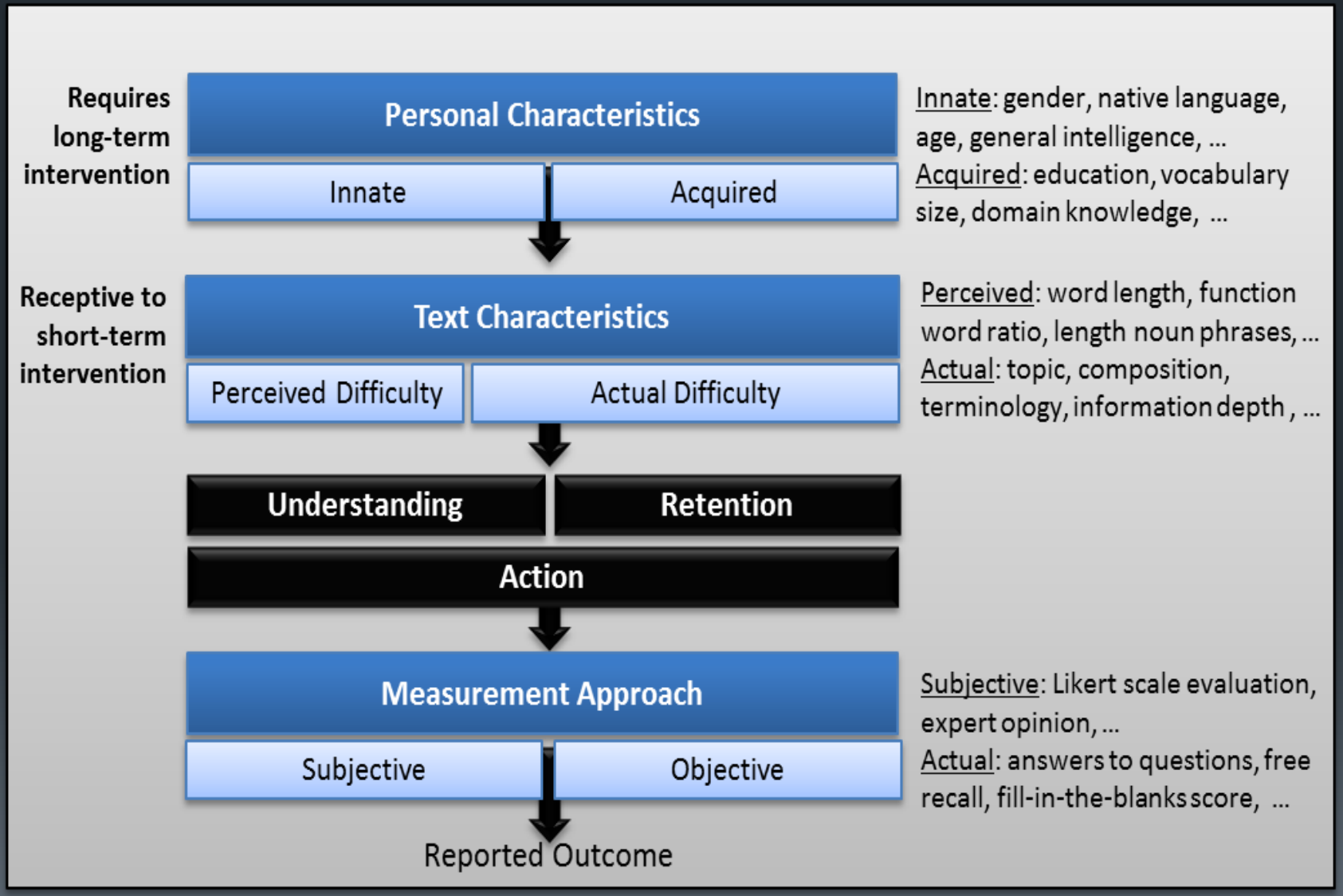
Introduction

Having tools that help write easy-to-understand text would benefit healthcare providers, patients and consumers. Unfortunately, the most popular tools available today are readability formulas do not pinpoint the difficult sections in a text, do not suggest alternative writings and their ratings have not convincingly been shown to correlate with understanding.

Project Overview

We are developing algorithms that process health text, identify difficult sections and suggest easier alternatives. Based on initial corpus analysis, we have identified different grammar, lexical and composition features that differ between easy and difficult text. We test such differences for their impact on both perceived and actual difficulty in large-scale user studies.

Text Simplification



Semi-Automated Text Simplification

- Two constraints (self-imposed)
  - tools to be be useful to all writers without requiring background knowledge in linguistics or education
  - use of the tools should be efficient and effective

- Research Approach
  - STEP 1 = Corpus Analysis to detect features differing between difficult and easy text
  - STEP 2 = Algorithm development
    - Pinpoint difficult sections (words, phrases, …)
    - Propose easier alternatives
  - STEP 3 = User studies

User Studies' Common Elements

Within-subject designs, use Amazon Mechanical Turk services

Perceived Difficulty – measured using sentences

- Independent Variables:
  1. Lexical Simplification, noun phrase complexity, function word ratio …
- Dependent Variable: Perceived difficulty measured with a Likert-scale

Actual Difficulty – measured using texts

- Independent Variable
  1. Lexical Simplification, coherence enhancement, …
- Dependent Variable:
  1. Learning
  2. Understanding
  3. Retention

User Characteristics

- Demographics: age, gender, native language …
- Health Literacy, Stress, Reading Habits, …

Algorithm Development: Lexical Simplification

- Automated identification of difficult words based on term familiarity
  - Based on Google Web Corpus
  - 5,000th most frequent word = threshold
  - Tag each word with lower frequency than threshold
- Automated retrieval and rank-ordering of potential alternatives based on term familiarity
  - Find alternatives in WordNet, Wiktionary, UMLS
  - Include synonyms, hypernymes, definitions, semantic types
  - Rank order based on term frequency
- The writer chooses one of the alternatives to replace the difficult text segment or word

Algorithm Development: Coherence Enhancement

- Local Coherence
  - Increasing use of pronouns for explicitly connections between sentences
    - "The most common causes of morbidity and mortality in the western world can be accounted for by unhealthy patterns of behavior … . Interventions to improve health behavior are sorely needed.":
    - "The most common causes of morbidity and mortality in the western world can be accounted for by unhealthy patterns of …. Interventions to improve ***such*** health behavior are sorely needed."
  - No repetition of prominent antecdent in a coreferential relation
    - "Taking aspirin will reduce the symptoms. The aspirin will help reduce the headache."
    - "Taking aspirin will reduce the symptoms. ***It*** will help reduce the headache."
- Global Coherence
  - Adjustment of spacing based on Gestalt principles: chunking and spacing
    - "…. behavior is required. In this commentary, we explore three relatively new possible roads …. : (1) genetics may influence …. , (2) genetics may tone down …. , and (3) genetics may be ….."
    - "…. behavior is required.
      In this commentary, we explore three relatively new possible roads …. :
      (1) genetics may influence …. ,
      (2) genetics may tone down …. , and
      (3) genetics may be ….."

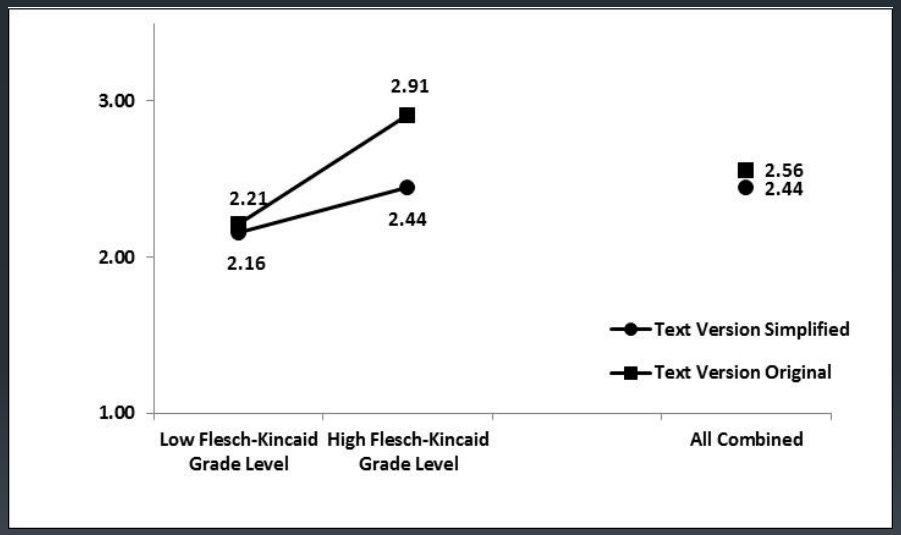
User Study 1: Lexical Simplification

- Algorithm not yet automated
- Study Procedure, N=82
  - Perceived Difficulty – paired sentences & Likert scale
  - Understanding - Text 1 with 4 multiple-choice questions and 1 qualifying question, in random order
  - Understanding - Text 2 with 4 multiple-choice questions and 1 qualifying question, in random order
  - Demographic questions
  - STOFHLA – Part A and B
  - Retention – 30 T/F statements in random order , 15 for each topic

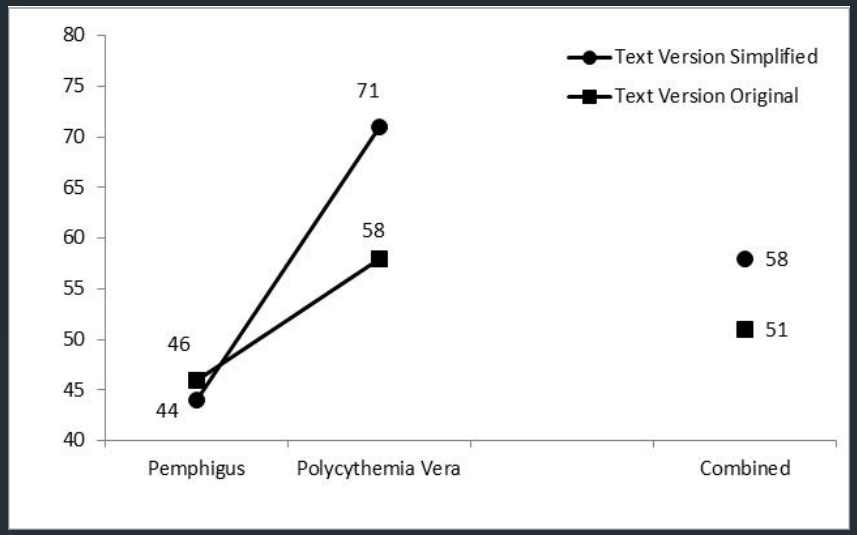
Category		% (N=82)
Gender	Female	45
	Male	55
Race	American Indian or Alaska Native	0
	Asian	17
	Black or African American	4
	Native Hawaiian or Other Pacific Islander	0
	White	80
Ethnicity	Hispanic or Latino	5
	Not Hispanic or Latino	95
STOFHLA	Average	32.9
	Minimum	16
	Maximum	36
Education	Less than high school	2
	High school diploma	35
	Associates	20
	Bachelor	24
	Master	18

User Study 1 Results

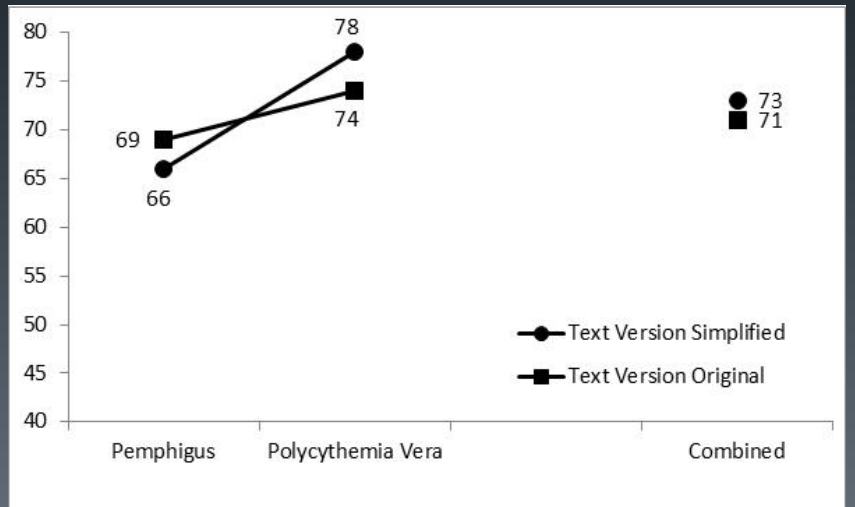
Perceived Difficulty



Actual Difficulty: Understanding



Actual Difficulty: Retention



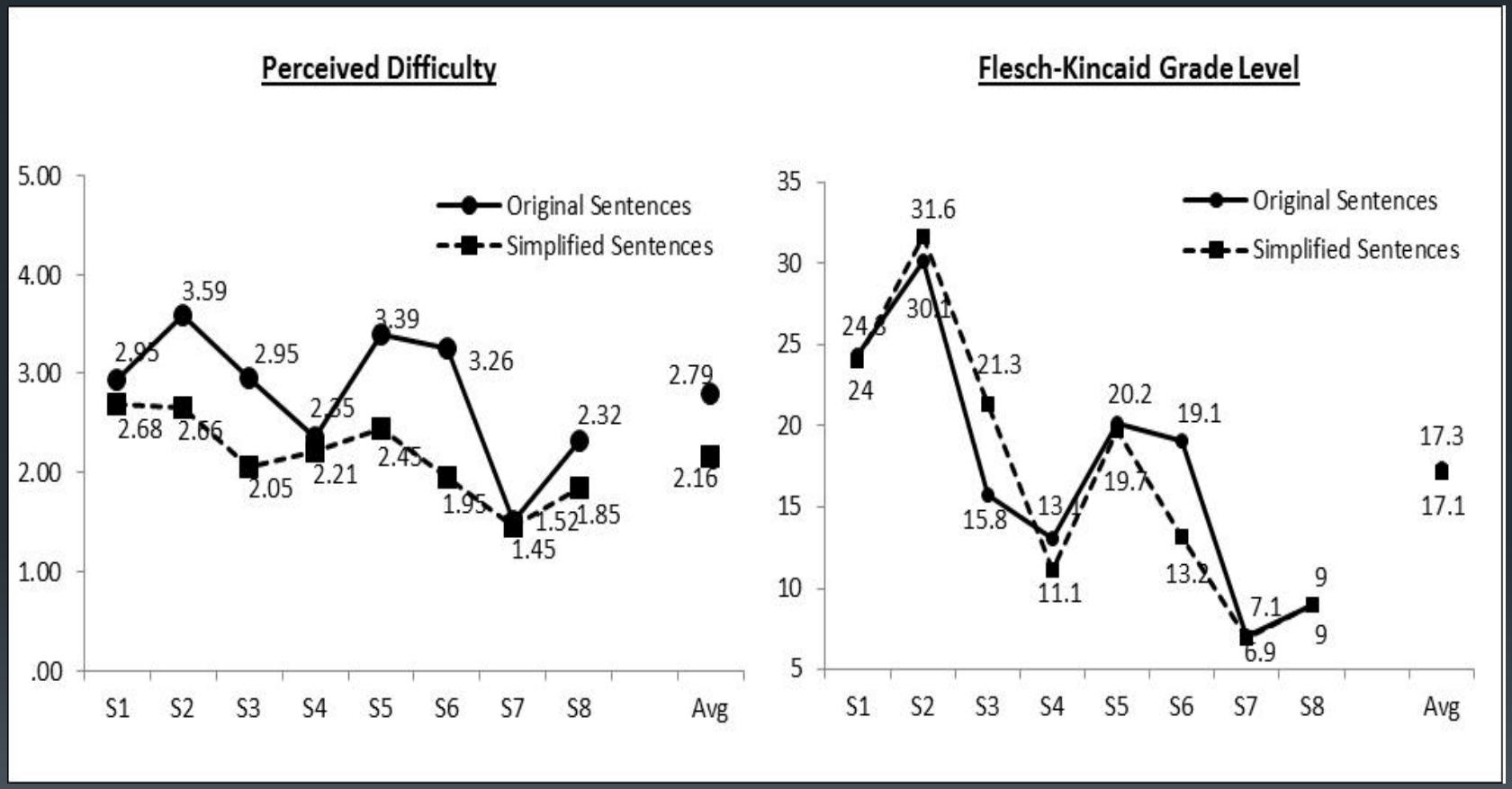
User Study 2: Lexical Simplification & Coherence Enhancement

- Algorithms not yet automated
- Mixed design
  - No coherence enhancement: lexical simplified or not
  - Coherence enhancement: lexical simplified or not

- Study Procedure
  - Actual difficulty measurement of the two documents using the Cloze test
  - Perceived difficulty measure of the eight sentences using a Likert-scale
  - Reader characteristics measurement:
    - Demographic questions about age, gender, race, ethnicity, languages spoken, education level, linguistics or medical knowledge
    - Questions about reading habits, i.e., how often they read books, printed news and magazine, and text online
    - S-TOFHLA
  - Actual difficulty measurement of the same two documents used in section 2 using multiple-choice questions

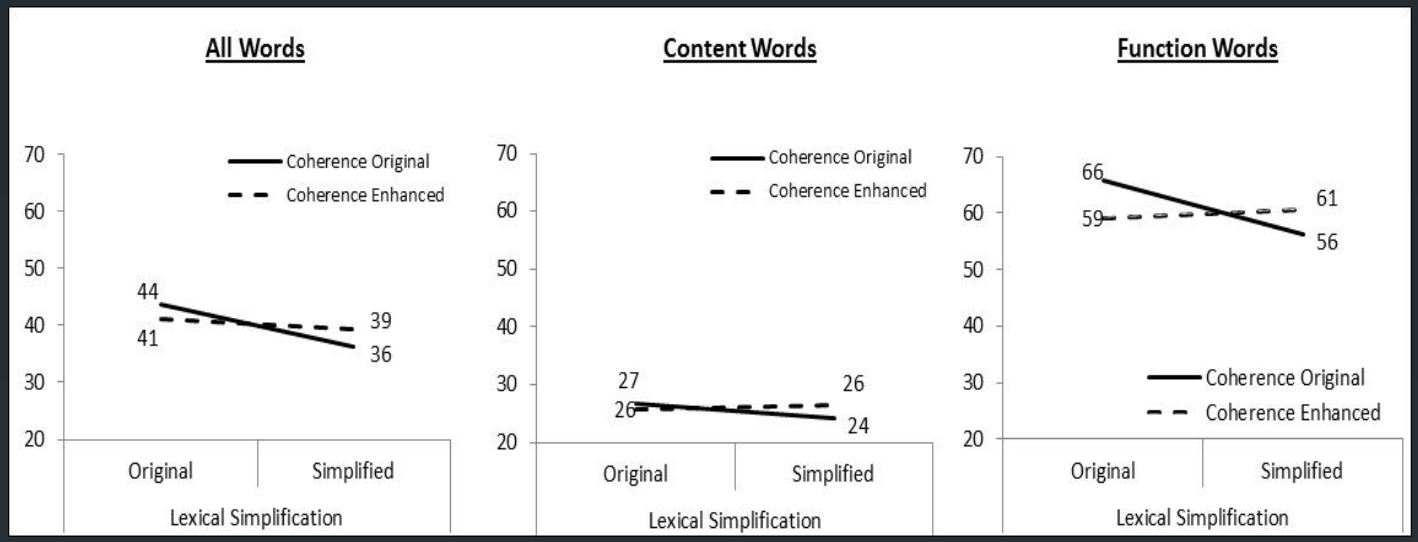
	N = 187	(%)
Age		
20 or younger	13	(7)
21-30	83	(44)
31-40	44	(23)
41-50	22	(12)
51-60	16	(9)
61-70	9	(5)
71 or older	-	-
Gender		
Female	119	(64)
Male	68	(36)
Race (Multiple choices allowed)		
American Indian / Native Alaskan	1	(.5)
Asian	31	(17)
Black or African American	12	(6)
Native Hawaiian or Other Pacific Islander	3	(1.5)
White	142	(76)
Ethnicity		
Hispanic or Latino	14	(8)
Not Hispanic or Latino	173	(92)
Location		
North America	162	(87)
South America	-	
Africa	1	(.5)
Europe	4	(2)
Asia	20	(11)
Australia or Oceania	-	
Education (Highest Completed)		
Less than High School	4	(2)
High School Diploma	68	(36)
Associate's Degree	30	(16)
Bachelor's Degree	57	(31)
Master's Degree	23	(12)
Doctorate	5	(3)
Language Skills (Frequency of Speaking English at Home)		
Never English	2	(1)
Rarely English	10	(5)
Half English	15	(8)
Mostly English	13	(7)
Only English	147	(79)

User Study 2 Results

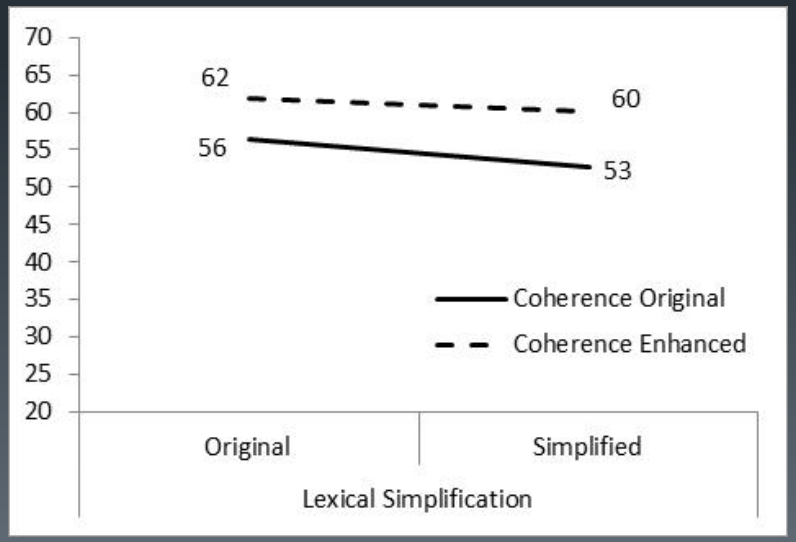


User Study 2 Results

Actual Difficulty: Understanding with Cloze test



Actual Difficulty: Understanding with multiple-choice questions



ANOVAs for: Cloze text (3x) and multiple choice test

Lexical simplification led to worse score for the cloze tests when there was no coherence enhancement
Coherence enhancement led to better multiple choice question answers

User Study 3: Lexical Simplification

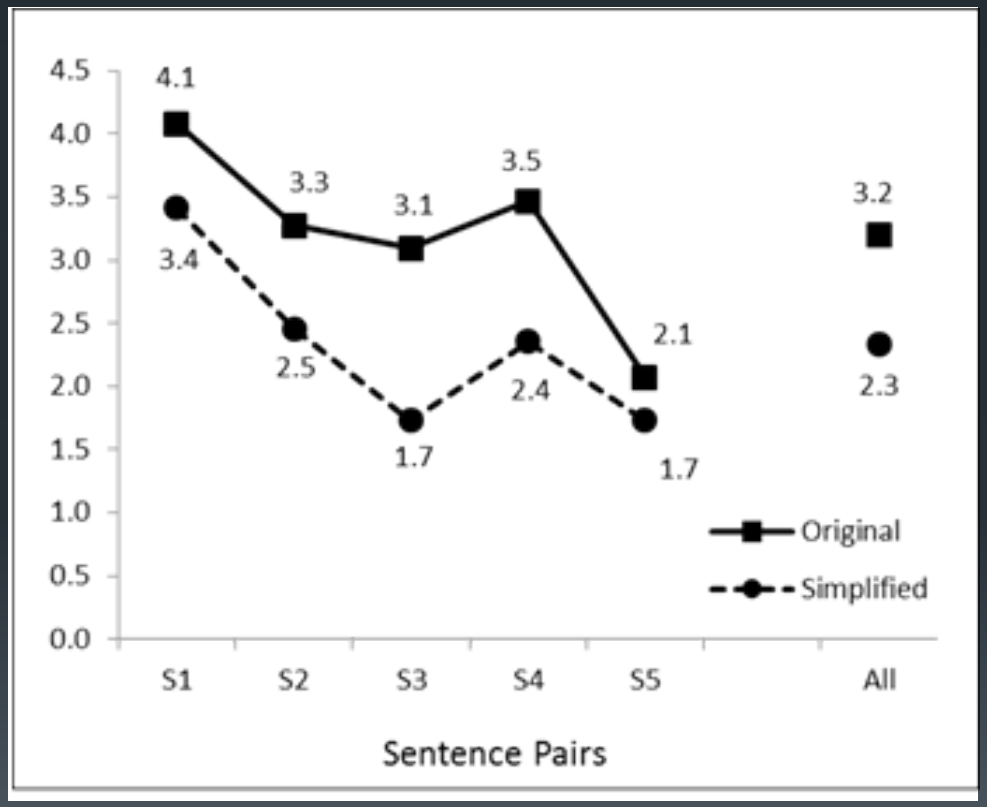
- Automated Algorithm, Independent Writer
- Different Texts (Asthma and Liver Cirrhosis), Measurements and Participants

- Study Procedure
  - Text 1
    - myth-based questions
    - text is shown with new questions
    - Repeat myth-based questions without the text
  - Repeated text 2 (different topic, different version)
  - Individual sentences for perceived difficulty.
  - demographic questions.
  - PSS-10 and S-TOFHLA
  - free recall test for the first and second text

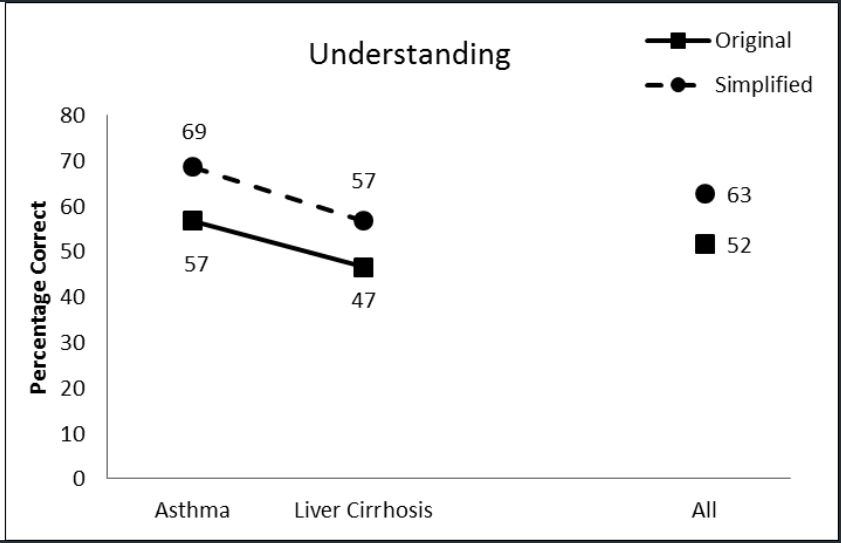
Characteristic		N=99
Age		
	20 or younger	3
	21-30	35
	31-40	24
	41-50	21
	51-60	12
	61-70	4
	71 or older	-
Gender		
	Female	62
	Male	37
Race (Multiple choices allowed)		
	American Indian / Native Alaskan	2
	Asian	7
	Black or African American	5
	Native Hawaiian or Other Pacific Islander	-
	White	88
Ethnicity		
	Hispanic or Latino	7
	Not Hispanic or Latino	92
Education (Highest Completed)		
	Less than High School	1
	High School Diploma	48
	Associate's Degree	16
	Bachelor's Degree	25
	Master's Degree	6
	Doctorate	3
Language Skills (Frequency of Speaking English at Home)		
	Never English	-
	Rarely English	1
	Half English	3
	Mostly English	6
	Only English	89

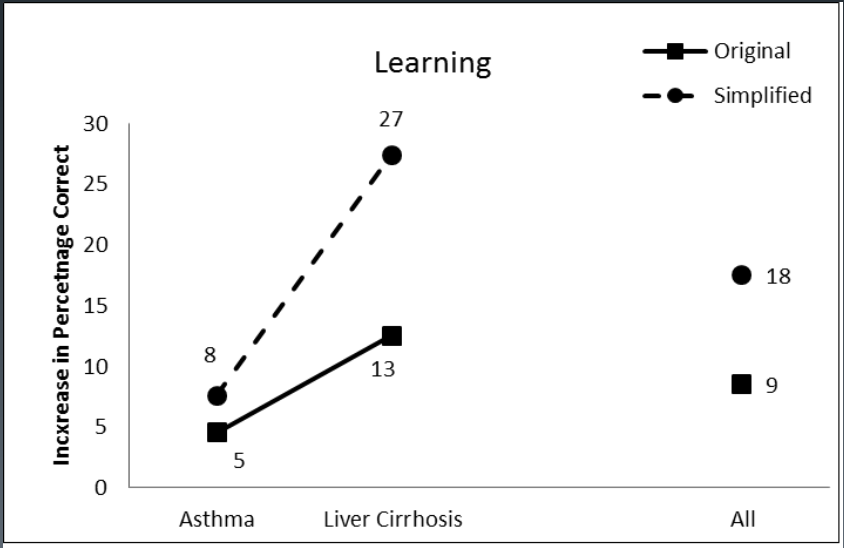
User Study 3 Results

Perceived Difficulty



User Study 3 Results





ANOVAs for understanding and learning

Lexical simplification led to better understanding

Lexical simplification led to better learning (myth based questions)

Asthma document was easier to understand but resulted in less learning

Conclusion & Acknowledgements

- For more information, consulting, speaker engagements: contact GondyLeroy@gmail.com
- Selected Publications
  - G. Leroy, D. Kauchak, O. Mouradi , “A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty”, Int. J. Medical Informatics, In Press.
  - G. Leroy, J.E. Endicott, O. Mouradi, D. Kauchak, and M. Just, "Improving Perceived and Actual Text Difficulty for Health Information Consumers using Semi-Automated Methods", American Medical Informatics Association (AMIA) Fall Symposium, Chicago, IL, November 3-7, 2012.
  - G. Leroy, D. Kauchak and W. Coster, "*Systematic Grammatical Analysis of Easy and Difficult Medical Text*", American Medical Informatics Association (AMIA) Fall Symposium, Chicago, IL, November 3-7, 2012.
  - G. Leroy and J.E. Endicott, Combining NLP with evidence-based methods to find text metrics related to perceived and actual text difficulty," 2nd ACM SIGHIT International Health Informatics Symposium (ACM IHI 2012), Florida, Miami, January 28-30, 2012.
  - G. Leroy and J.E. Endicott, "*Term Familiarity to Indicate Perceived and Actual Difficulty of Text in Medical Digital Libraries*," International Conference on Asia-Pacific Digital Libraries (ICADL 2011) - Digital Libraries -- for Culture Heritage, Knowledge Dissemination, and Future Creation. Beijing, China, October 24-27, 2011.
  - *G. Leroy, S. Helmreich, and J. Cowie, "The Influence of Text Characteristics on Perceived and Actual Difficulty of Health Information", International Journal of Medical Informatics, Vol. 79, Issue 6, Pages 438-449.*
- Academic Team
  - David Kauchak, Ph.D., Middlebury College
  - Melissa Just, Ed.D., Rutgers University
  - Students: James Endicott, Obay Mouradi
- This work was supported by the U.S. National Library of Medicine, NIH/NLM 1R03LM010902-01.