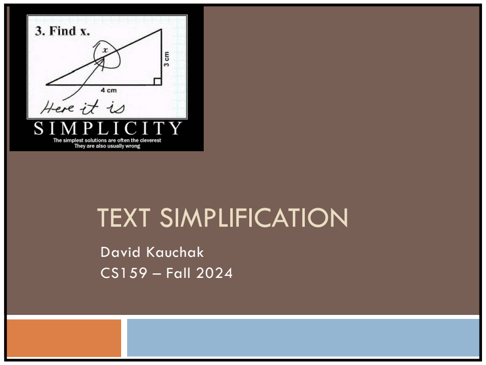
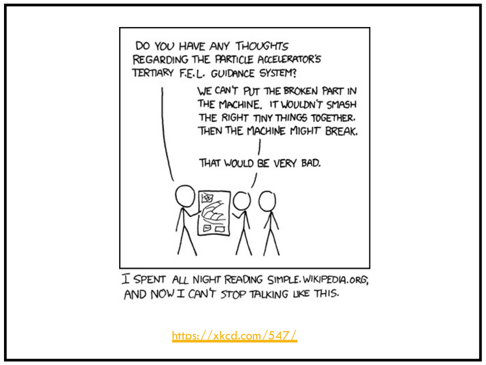

3. Find x.
x
3 cm
4 cm
Here it is
SIMPLICITY
The simplest solutions are often the cleverest
They are also usually wrong

TEXT SIMPLIFICATION

David Kauchak
CS159 – Fall 2024

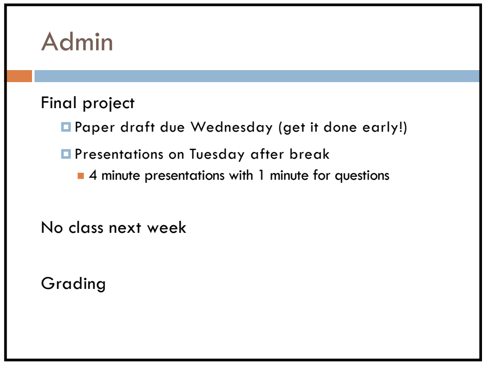
1


DO YOU HAVE ANY THOUGHTS REGARDING THE PARTICLE ACCELERATOR'S TERTIARY F.E.L. GUIDANCE SYSTEM?
WE CAN'T PUT THE BROKEN PART IN THE MACHINE. IT WOULDN'T SMASH THE RIGHT TINY THINGS TOGETHER. THEN THE MACHINE MIGHT BREAK.
THAT WOULD BE VERY BAD.

I SPENT ALL NIGHT READING SIMPLE.WIKIPEDIA.ORG, AND NOW I CAN'T STOP TALKING LIKE THIS.

https://xkcd.com/547/

2



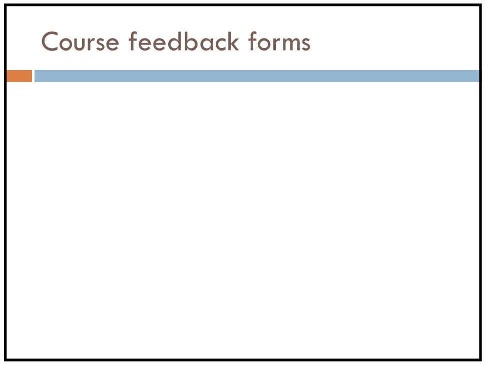
Admin

Final project
- Paper draft due Wednesday (get it done early!)
- Presentations on Tuesday after break
  - 4 minute presentations with 1 minute for questions

No class next week

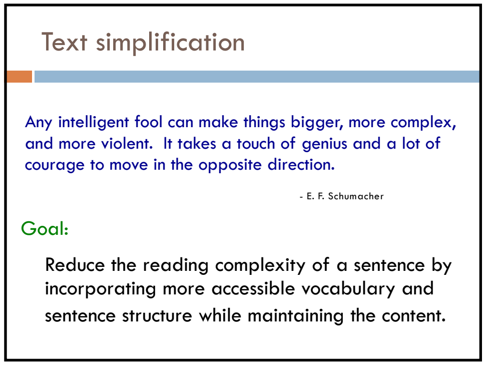
Grading

3



Course feedback forms

4



Text simplification

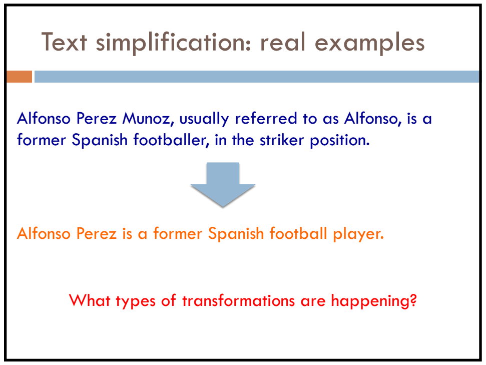
Any intelligent fool can make things bigger, more complex, and more violent. It takes a touch of genius and a lot of courage to move in the opposite direction.

- E. F. Schumacher

Goal:

Reduce the reading complexity of a sentence by incorporating more accessible vocabulary and sentence structure while maintaining the content.

5



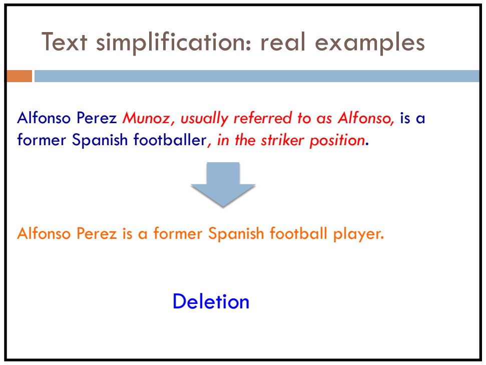
Text simplification: real examples

Alfonso Perez Munoz, usually referred to as Alfonso, is a former Spanish footballer, in the striker position.

Alfonso Perez is a former Spanish football player.

What types of transformations are happening?

6



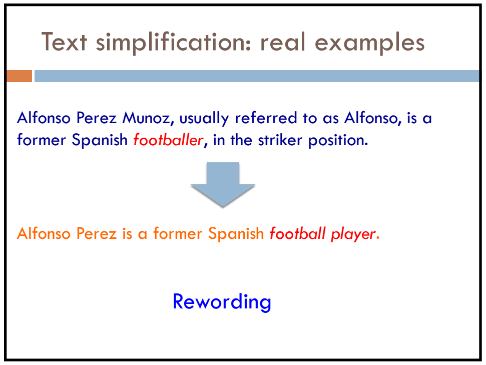
Text simplification: real examples

Alfonso Perez *Munoz, usually referred to as Alfonso,* is a former Spanish footballer*, in the striker position*.

Alfonso Perez is a former Spanish football player.

Deletion

7



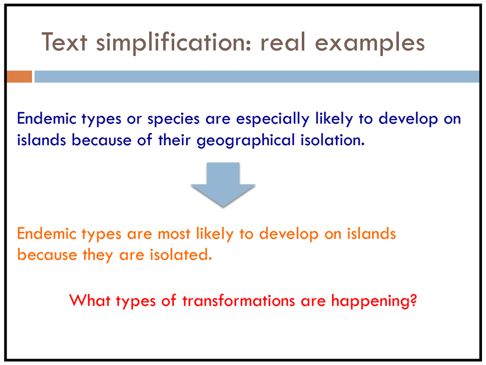
Text simplification: real examples

Alfonso Perez Munoz, usually referred to as Alfonso, is a former Spanish *footballer*, in the striker position.

Alfonso Perez is a former Spanish *football player*.

Rewording

8



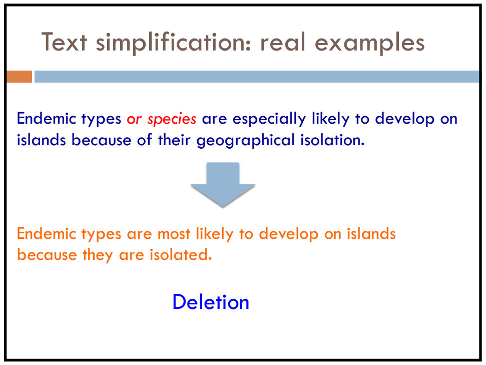
Text simplification: real examples

Endemic types or species are especially likely to develop on islands because of their geographical isolation.

Endemic types are most likely to develop on islands because they are isolated.

What types of transformations are happening?

9



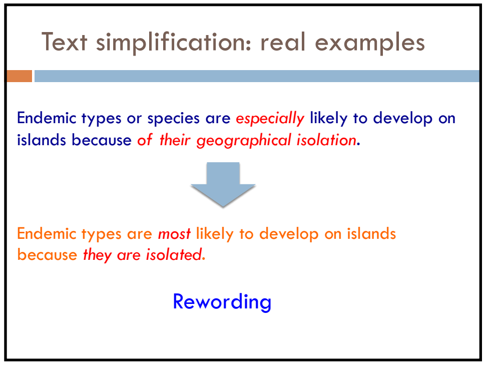
Text simplification: real examples

Endemic types *or species* are especially likely to develop on islands because of their geographical isolation.

Endemic types are most likely to develop on islands because they are isolated.

Deletion

10



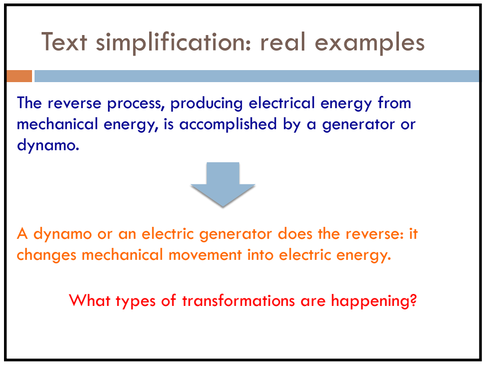
Text simplification: real examples

Endemic types or species are *especially* likely to develop on islands because of *their geographical isolation*.

Endemic types are *most* likely to develop on islands because *they are isolated*.

Rewording

11



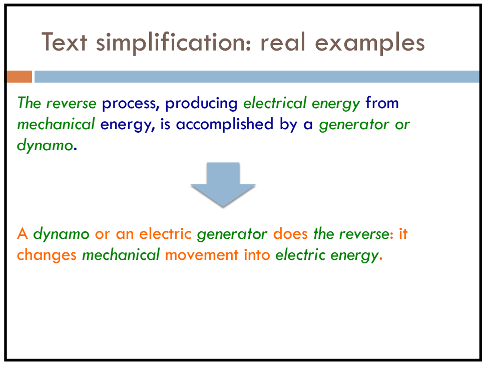
Text simplification: real examples

The reverse process, producing electrical energy from mechanical energy, is accomplished by a generator or dynamo.

A dynamo or an electric generator does the reverse: it changes mechanical movement into electric energy.

What types of transformations are happening?

12

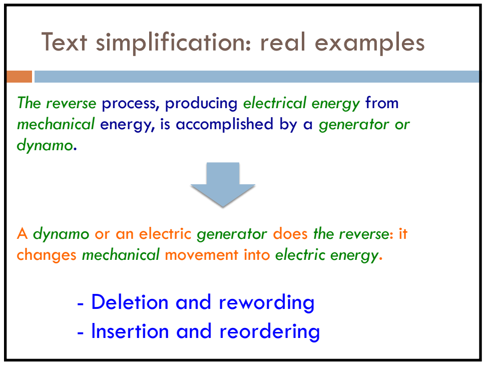


Text simplification: real examples

The reverse process, producing *electrical energy* from *mechanical* energy, is accomplished by a *generator or dynamo*.

A *dynamo* or an electric *generator* does *the reverse*: it changes *mechanical* movement into *electric energy*.

13



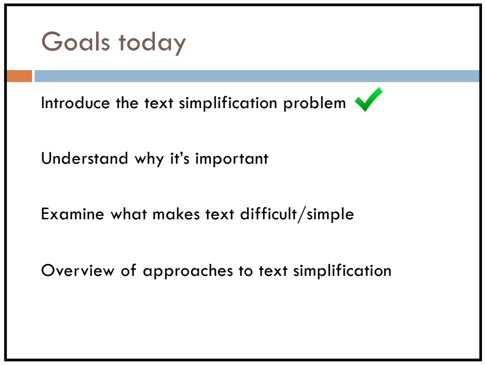
Text simplification: real examples

The reverse process, producing *electrical energy* from *mechanical* energy, is accomplished by a *generator or dynamo*.

A *dynamo* or an electric *generator* does *the reverse*: it changes *mechanical* movement into *electric energy*.

- Deletion and rewording
- Insertion and reordering

14



Goals today

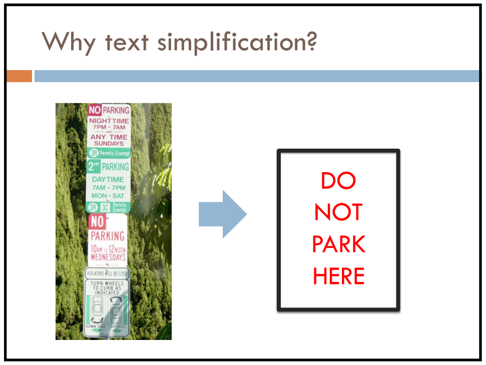
Introduce the text simplification problem ✓

Understand why it's important

Examine what makes text difficult/simple

Overview of approaches to text simplification

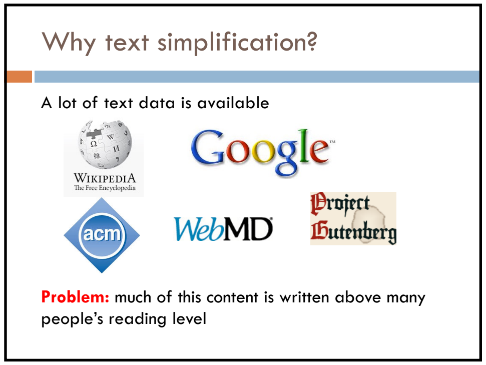
15



Why text simplification?

DO
NOT
PARK
HERE

16

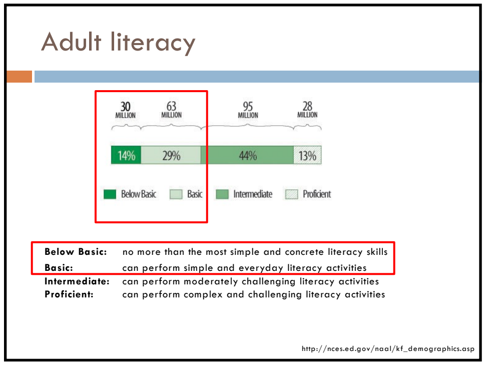


Why text simplification?

A lot of text data is available

Problem: much of this content is written above many people's reading level

17

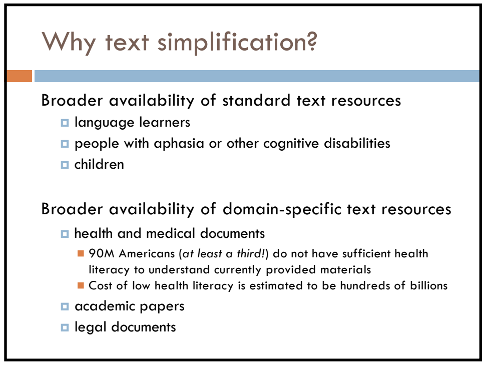


Adult literacy

Below Basic: no more than the most simple and concrete literacy skills
Basic: can perform simple and everyday literacy activities
Intermediate: can perform moderately challenging literacy activities
Proficient: can perform complex and challenging literacy activities

http://nces.ed.gov/naal/kf_demographics.asp

18



Why text simplification?

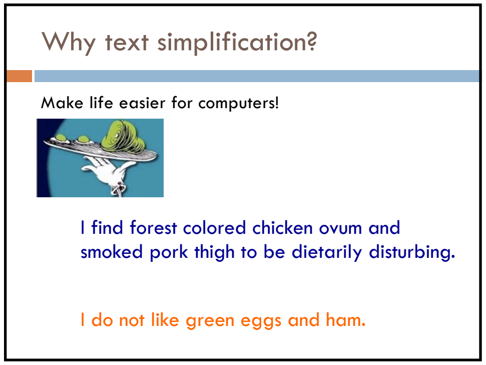
Broader availability of standard text resources

- language learners
- people with aphasia or other cognitive disabilities
- children

Broader availability of domain-specific text resources

- health and medical documents
  - 90M Americans (*at least a third!*) do not have sufficient health literacy to understand currently provided materials
  - Cost of low health literacy is estimated to be hundreds of billions
- academic papers
- legal documents

19



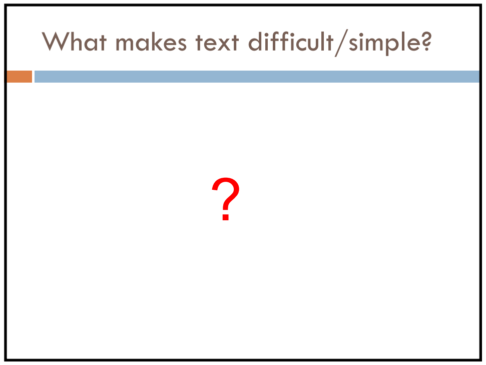
Why text simplification?

Make life easier for computers!

I find forest colored chicken ovum and smoked pork thigh to be dietarily disturbing.

I do not like green eggs and ham.

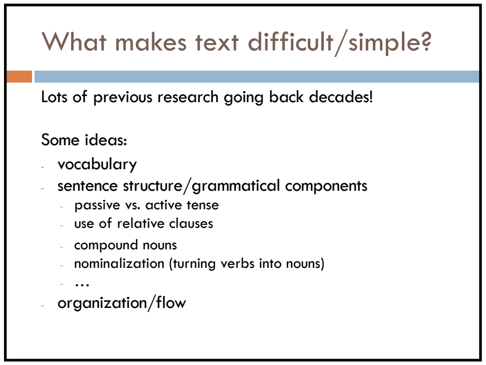
20



What makes text difficult/simple?

?

21



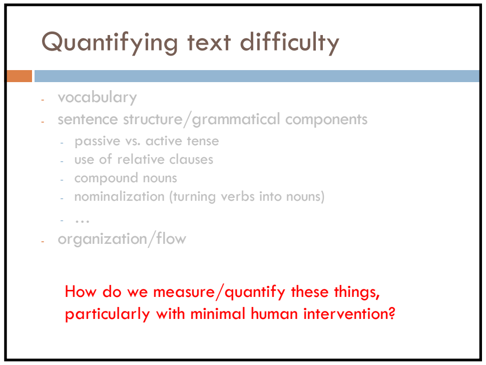
What makes text difficult/simple?

Lots of previous research going back decades!

Some ideas:

- vocabulary
- sentence structure/grammatical components
  - passive vs. active tense
  - use of relative clauses
  - compound nouns
  - nominalization (turning verbs into nouns)
  - …
- organization/flow

22

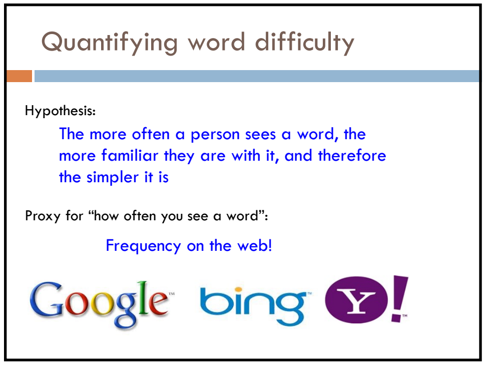


Quantifying text difficulty

- vocabulary
- sentence structure/grammatical components
  - passive vs. active tense
  - use of relative clauses
  - compound nouns
  - nominalization (turning verbs into nouns)
  - …
- organization/flow

How do we measure/quantify these things, particularly with minimal human intervention?

23



Quantifying word difficulty

Hypothesis:

The more often a person sees a word, the more familiar they are with it, and therefore the simpler it is

Proxy for "how often you see a word":

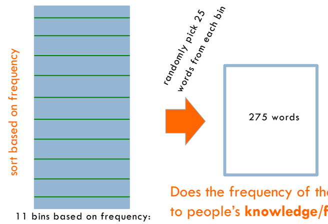
Frequency on the web!

Google
bing
Y!

24

Validating frequency hypothesis

Google unigrams: ~13M



sort based on frequency

randomly pick 25 words from each bin

275 words

11 bins based on frequency: 1%, 10%, 20%, …, 100%

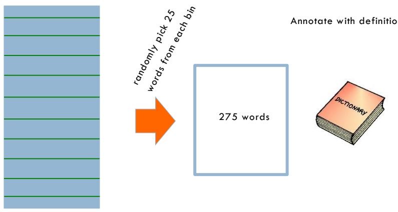
Does the frequency of these words relate to people's **knowledge/familiarity** with these words?

25

Validating frequency hypothesis

Google unigrams: ~13M

Annotate with definition



randomly pick 25 words from each bin

275 words

11 bins based on frequency: 1%, 10%, 20%, …, 100%

26

Validating frequency hypothesis

marmorean:

a) crimson-and-grey songbird that inhabits town walls and mountain cliffs of southern Eurasia and northern Africa

b) of or relating to or characteristic of marble

c) the most common protein in muscle

d) a woman policeman

27

Validating frequency hypothesis

marmorean:

a) crimson-and-grey songbird that inhabits town walls and mountain cliffs of southern Eurasia and northern Africa

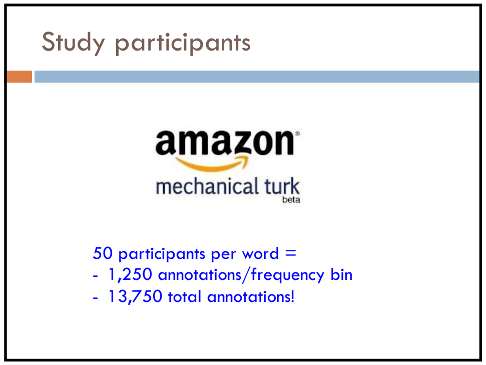
b) of or relating to or characteristic of marble

c) the most common protein in muscle

d) a woman policeman

random definitions from other words in data set

28

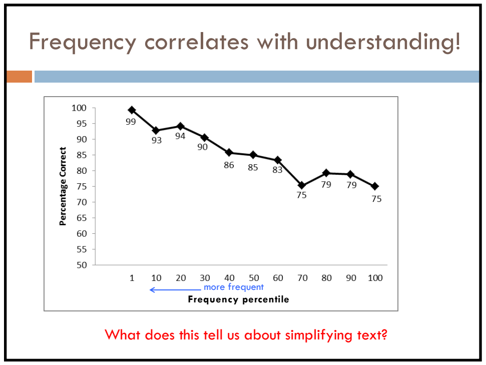


Study participants

amazon mechanical turk beta

50 participants per word =
- 1,250 annotations/frequency bin
- 13,750 total annotations!

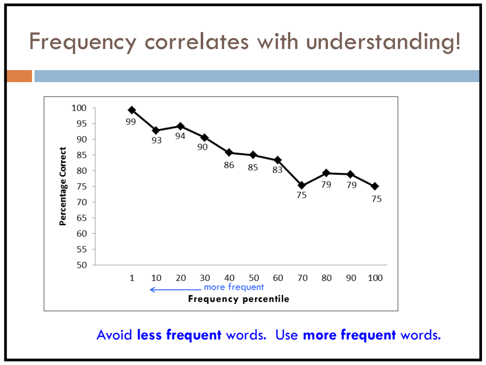
29



Frequency correlates with understanding!

What does this tell us about simplifying text?

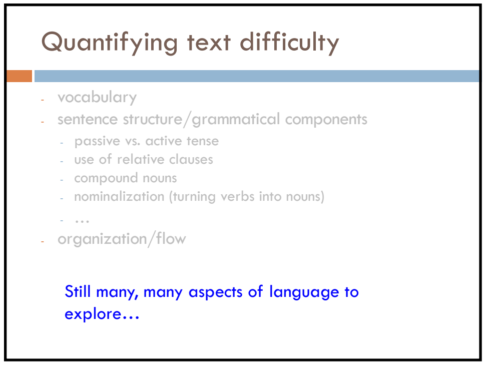
30



Frequency correlates with understanding!

Avoid **less frequent** words. Use **more frequent** words.

31



Quantifying text difficulty

- vocabulary
- sentence structure/grammatical components
  - passive vs. active tense
  - use of relative clauses
  - compound nouns
  - nominalization (turning verbs into nouns)
  - …
- organization/flow

Still many, many aspects of language to explore…

32

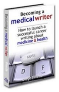
Goals today

- Introduce the text simplification problem ✓
- Understand why it's important ✓
- Examine what makes text difficult/simple ✓
- Overview of approaches to text simplification

33

Spectrum of solutions

Focus on these types of approaches today



amazonmechanical turk Artificial Artificial Intelligence

writer assist tools/resources

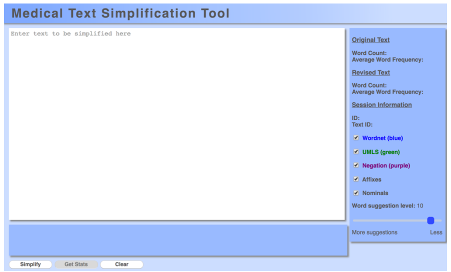
- readability formulas
- simple word lists
- flag difficult text sections
- simplification thesauruses
- rule-based with human verification
- …

Google ~~Translate~~ Simplify

manual ⟷ semi-automated ⟷ fully automated

34

Writer assist tool



Medical Text Simplification Tool

Enter text to be simplified here

Original Text

Word Count:
Average Word Frequency:

Revised Text

Word Count:
Average Word Frequency:

Session Information

ID:
Text ID:

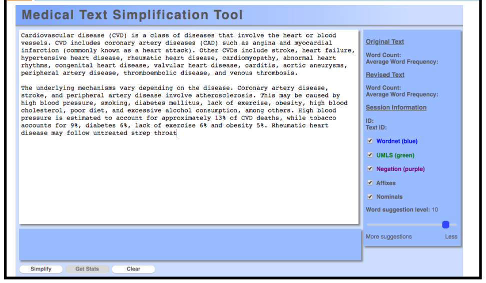
☑ Wordnet (blue)
☑ UMLS (green)
☑ Negation (purple)
☑ Affixes
☑ Nominals

Word suggestion level: 10

More suggestions — Less

35

Writer assist tool



Medical Text Simplification Tool

Cardiovascular disease (CVD) is a class of diseases that involve the heart or blood vessels. CVD includes coronary artery diseases (CAD) such as angina and myocardial infarction (commonly known as a heart attack). Other CVDs include stroke, heart failure, hypertensive heart disease, rheumatic heart disease, cardiomyopathy, abnormal heart rhythms, congenital heart disease, valvular heart disease, carditis, aortic aneurysms, peripheral artery disease, thromboembolic disease, and venous thrombosis.

The underlying mechanisms vary depending on the disease. Coronary artery disease, stroke, and peripheral artery disease involve atherosclerosis. This may be caused by high blood pressure, smoking, diabetes mellitus, lack of exercise, obesity, high blood cholesterol, poor diet, and excessive alcohol consumption, among others. High blood pressure is estimated to account for approximately 13% of CVD deaths, while tobacco accounts for 9%, diabetes 6%, lack of exercise 6% and obesity 5%. Rheumatic heart disease may follow untreated strep throat

Original Text

Word Count:
Average Word Frequency:

Revised Text

Word Count:
Average Word Frequency:

Session Information

ID:
Text ID:

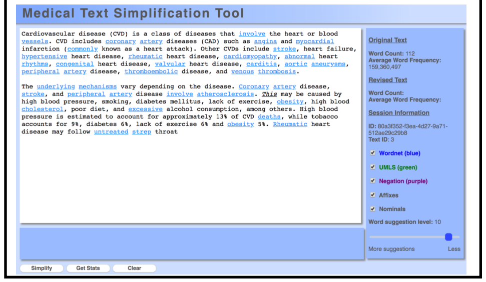
☑ Wordnet (blue)
☑ UMLS (green)
☑ Negation (purple)
☑ Affixes
☑ Nominals

Word suggestion level: 10

More suggestions — Less

36

Writer assist tool



Medical Text Simplification Tool

Cardiovascular disease (CVD) is a class of diseases that involve the heart or blood vessels. CVD includes coronary artery diseases (CAD) such as angina and myocardial infarction (commonly known as a heart attack). Other CVDs include stroke, heart failure, hypertensive heart disease, rheumatic heart disease, cardiomyopathy, abnormal heart rhythms, congenital heart disease, valvular heart disease, carditis, aortic aneurysms, peripheral artery disease, thromboembolic disease, and venous thrombosis.

The underlying mechanisms vary depending on the disease. Coronary artery disease, stroke, and peripheral artery disease involve atherosclerosis. *This* may be caused by high blood pressure, smoking, diabetes mellitus, lack of exercise, obesity, high blood cholesterol, poor diet, and excessive alcohol consumption, among others. High blood pressure is estimated to account for approximately 13% of CVD deaths, while tobacco accounts for 9%, diabetes 6%, lack of exercise 6% and obesity 5%. Rheumatic heart disease may follow untreated strep throat

Original Text

Word Count: 112
Average Word Frequency: 159,360,497

Revised Text

Word Count:
Average Word Frequency:

Session Information

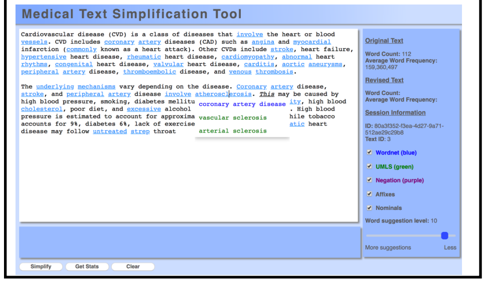
ID: 80a3f352-f3ea-4d27-9a71-512ae29c29b8
Text ID: 3

☑ Wordnet (blue)
☑ UMLS (green)
☑ Negation (purple)
☑ Affixes
☑ Nominals

Word suggestion level: 10

37

Writer assist tool



Medical Text Simplification Tool

Cardiovascular disease (CVD) is a class of diseases that involve the heart or blood vessels. CVD includes coronary artery diseases (CAD) such as angina and myocardial infarction (commonly known as a heart attack). Other CVDs include stroke, heart failure, hypertensive heart disease, rheumatic heart disease, cardiomyopathy, abnormal heart rhythms, congenital heart disease, valvular heart disease, carditis, aortic aneurysms, peripheral artery disease, thromboembolic disease, and venous thrombosis.

The underlying mechanisms vary depending on the disease. Coronary artery disease, stroke, and peripheral artery disease involve atherosclerosis. *This* may be caused by
high blood pressure, smoking, diabetes mellitu ity, high blood
cholesterol, poor diet, and excessive alcohol . High blood
pressure is estimated to account for approxima hile tobacco
accounts for 9%, diabetes 6%, lack of exercise atic heart
disease may follow untreated strep throat

coronary artery disease
vascular sclerosis
arterial sclerosis

Original Text

Word Count: 112
Average Word Frequency: 159,360,497

Revised Text

Word Count:
Average Word Frequency:

Session Information

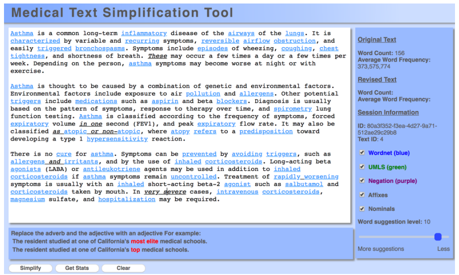
ID: 80a3f352-f3ea-4d27-9a71-512ae29c29b8
Text ID: 3

☑ Wordnet (blue)
☑ UMLS (green)
☑ Negation (purple)
☑ Affixes
☑ Nominals

Word suggestion level: 10

38

Writer assist tool



Medical Text Simplification Tool

Asthma is a common long-term inflammatory disease of the airways of the lungs. It is characterized by variable and recurring symptoms, reversible airflow obstruction, and easily triggered bronchospasms. Symptoms include episodes of wheezing, coughing, chest tightness, and shortness of breath. *These* may occur a few times a day or a few times per week. Depending on the person, asthma symptoms may become worse at night or with exercise.

Asthma is thought to be caused by a combination of genetic and environmental factors. Environmental factors include exposure to air pollution and allergens. Other potential triggers include medications such as aspirin and beta blockers. Diagnosis is usually based on the pattern of symptoms, response to therapy over time, and spirometry lung function testing. Asthma is classified according to the frequency of symptoms, forced expiratory volume *in one* second (FEV1), and peak expiratory flow rate. It may also be classified *as* atopic *or non-*atopic, where atopy refers to a predisposition toward developing a type 1 hypersensitivity reaction.

There is no cure for asthma. Symptoms can be prevented by avoiding triggers, such as allergens *and* irritants, and by the use of inhaled corticosteroids. Long-acting beta agonists (LABA) or antileukotriene agents may be used in addition to inhaled corticosteroids if asthma symptoms remain uncontrolled. Treatment of rapidly_worsening symptoms is usually with an inhaled short-acting beta-2 agonist such as salbutamol and corticosteroids taken by mouth. In *very severe* cases, intravenous corticosteroids, magnesium sulfate, and hospitalization may be required.

Replace the adverb and the adjective with an adjective For example:
The resident studied at one of California's most elite medical schools.
The resident studied at one of California's top medical schools.

Original Text

Word Count: 156
Average Word Frequency: 373,575,774

Revised Text

Word Count:
Average Word Frequency:

Session Information

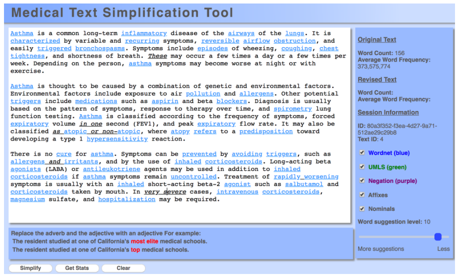
ID: 80a3f352-f3ea-4d27-9a71-512ae29c29b8
Text ID: 4

☑ Wordnet (blue)
☑ UMLS (green)
☑ Negation (purple)
☑ Affixes
☑ Nominals

Word suggestion level: 10

39

How do we identify difficult words?



Medical Text Simplification Tool

Asthma is a common long-term inflammatory disease of the airways of the lungs. It is characterized by variable and recurring symptoms, reversible airflow obstruction, and easily triggered bronchospasms. Symptoms include episodes of wheezing, coughing, chest tightness, and shortness of breath. *These* may occur a few times a day or a few times per week. Depending on the person, asthma symptoms may become worse at night or with exercise.

Asthma is thought to be caused by a combination of genetic and environmental factors. Environmental factors include exposure to air pollution and allergens. Other potential triggers include medications such as aspirin and beta blockers. Diagnosis is usually based on the pattern of symptoms, response to therapy over time, and spirometry lung function testing. Asthma is classified according to the frequency of symptoms, forced expiratory volume *in one* second (FEV1), and peak expiratory flow rate. It may also be classified *as* atopic *or non-*atopic, where atopy refers to a predisposition toward developing a type 1 hypersensitivity reaction.

There is no cure for asthma. Symptoms can be prevented by avoiding triggers, such as allergens *and* irritants, and by the use of inhaled corticosteroids. Long-acting beta agonists (LABA) or antileukotriene agents may be used in addition to inhaled corticosteroids if asthma symptoms remain uncontrolled. Treatment of rapidly_worsening symptoms is usually with an inhaled short-acting beta-2 agonist such as salbutamol and corticosteroids taken by mouth. In *very severe* cases, intravenous corticosteroids, magnesium sulfate, and hospitalization may be required.

Replace the adverb and the adjective with an adjective For example:
The resident studied at one of California's most elite medical schools.
The resident studied at one of California's top medical schools.

Original Text

Word Count: 156
Average Word Frequency: 373,575,774

Revised Text

Word Count:
Average Word Frequency:

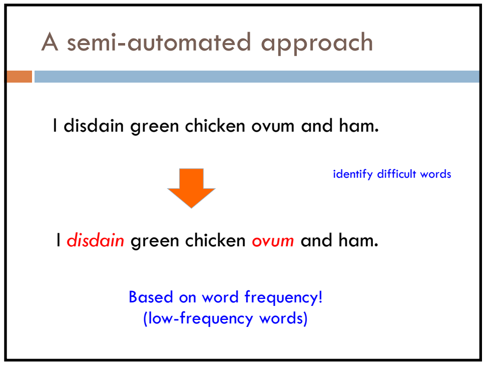
Session Information

ID: 80a3f352-f3ea-4d27-9a71-512ae29c29b8
Text ID: 4

☑ Wordnet (blue)
☑ UMLS (green)
☑ Negation (purple)
☑ Affixes
☑ Nominals

Word suggestion level: 10

40



A semi-automated approach

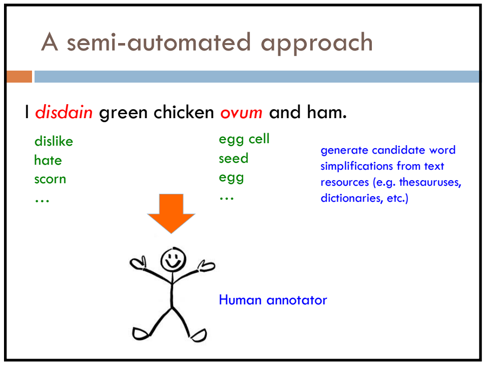
I disdain green chicken ovum and ham.

identify difficult words

I *disdain* green chicken *ovum* and ham.

Based on word frequency!
(low-frequency words)

41



A semi-automated approach

I *disdain* green chicken *ovum* and ham.

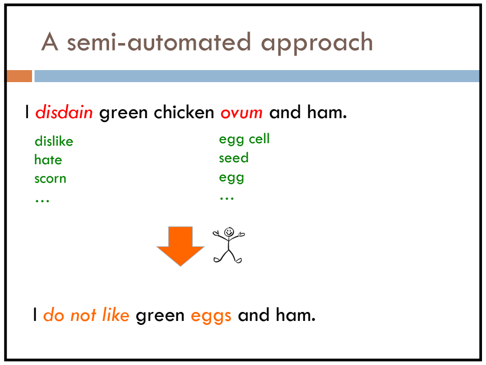
dislike
hate
scorn
…

egg cell
seed
egg
…

generate candidate word simplifications from text resources (e.g. thesauruses, dictionaries, etc.)

Human annotator

42



A semi-automated approach

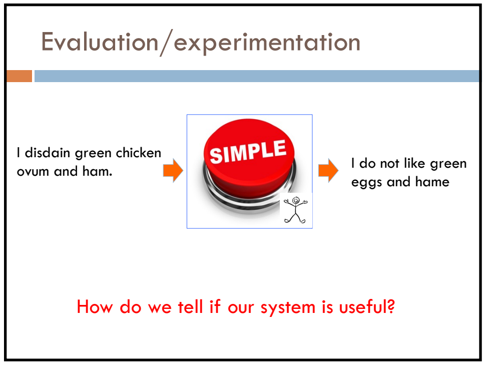
I *disdain* green chicken *ovum* and ham.

dislike
hate
scorn
…

egg cell
seed
egg
…

I *do not like* green eggs and ham.

43



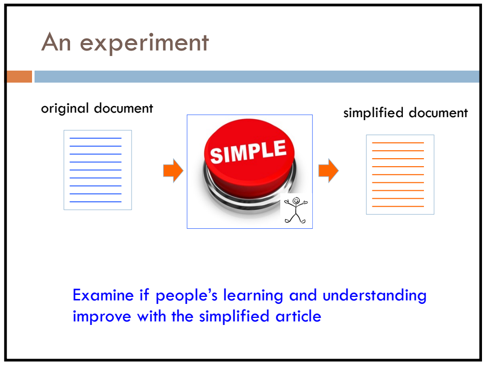
Evaluation/experimentation

I disdain green chicken ovum and ham.

I do not like green eggs and hame

How do we tell if our system is useful?

44



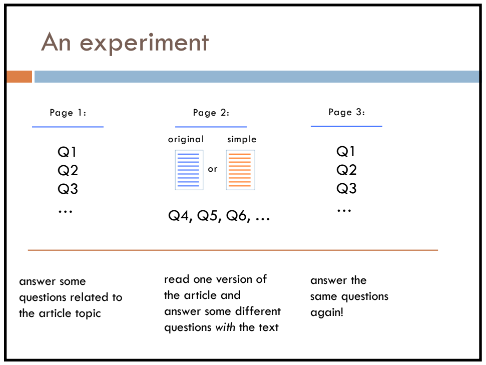
An experiment

original document

simplified document

Examine if people's learning and understanding improve with the simplified article

45



An experiment

Page 1:

Q1
Q2
Q3
…

Page 2:

original or simple

Q4, Q5, Q6, …

Page 3:

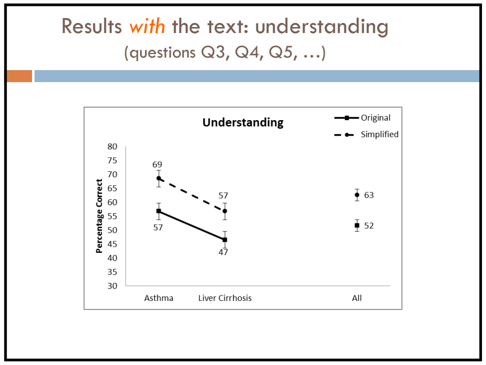
Q1
Q2
Q3
…

answer some questions related to the article topic

read one version of the article and answer some different questions *with* the text

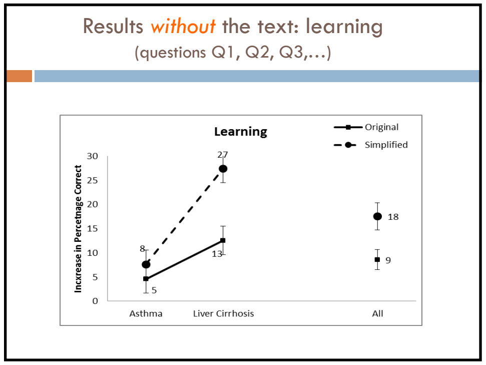
answer the same questions again!

46



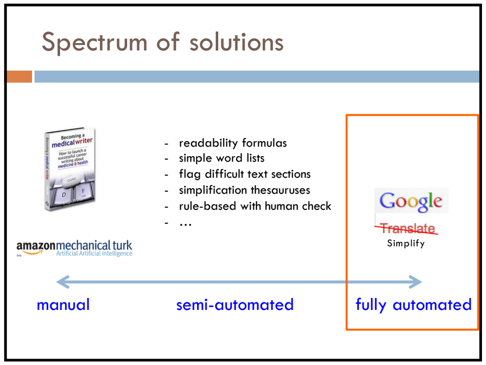
Results *with* the text: understanding
(questions Q3, Q4, Q5, …)

47



Results *without* the text: learning
(questions Q1, Q2, Q3,…)

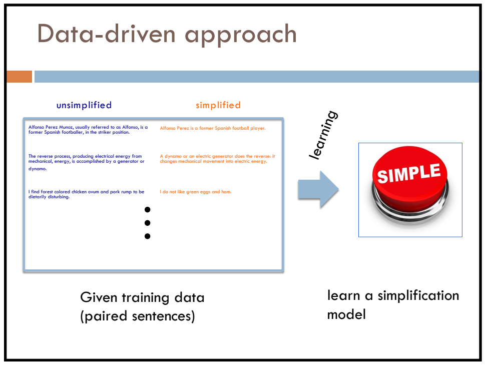
48



Spectrum of solutions

- readability formulas
- simple word lists
- flag difficult text sections
- simplification thesauruses
- rule-based with human check
- …

49



Data-driven approach

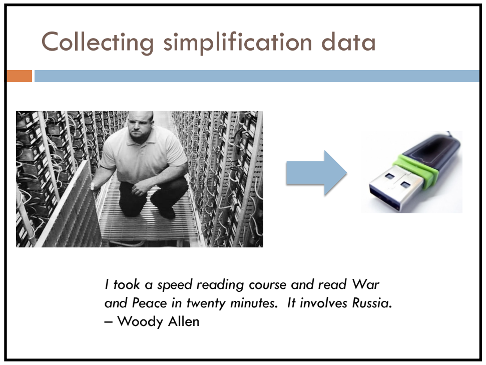
unsimplified	simplified
Alfonso Perez Munoz, usually referred to as Alfonso, is a former Spanish footballer, in the striker position.	Alfonso Perez is a former Spanish football player.
The reverse process, producing electrical energy from mechanical, energy, is accomplished by a generator or dynamo.	A dynamo or an electric generator does the reverse: it changes mechanical movement into electric energy.
I find forest colored chicken ovum and pork rump to be dietarily disturbing.	I do not like green eggs and ham.

⋮

Given training data (paired sentences)

learn a simplification model

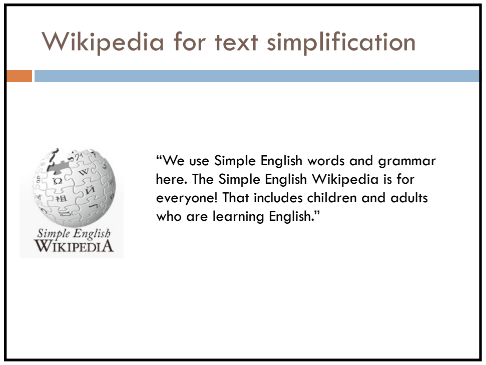
50



Collecting simplification data

I took a speed reading course and read War and Peace in twenty minutes. It involves Russia.
– Woody Allen

51



Wikipedia for text simplification

"We use Simple English words and grammar here. The Simple English Wikipedia is for everyone! That includes children and adults who are learning English."

Simple English
Wikipedia

52

Wikipedia for text simplification

Simple English WIKIPEDIA

"Simple does not mean little. Writing in Simple English means that simple words are used. It does not mean readers want simple information. Articles do not have to be short to be simple; expand articles, include a lot of information, but use basic vocabulary."

53

Wikipedia for text simplification

WIKIPEDIA The Free Encyclopedia

4.4M articles

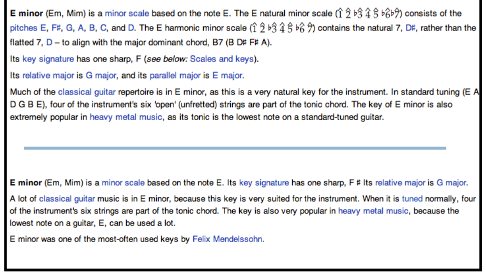
Simple English WIKIPEDIA

97K articles

unsimplified	simplified
Alfonso Perez Munoz, usually referred to as Alfonso, is a former Spanish footballer, in the striker position.	Alfonso Perez is a former Spanish football player.
The reverse process, producing electrical energy from mechanical, energy, is accomplished by a generator or dynamo.	A dynamo or an electric generator does the reverse: it changes mechanical movement into electric energy.
I find forest colored chicken ovum and pork rump to be dietarily disturbing.	I do not like green eggs and ham.

54

From aligned documents to aligned sentences



E minor (Em, Mim) is a minor scale based on the note E. The E natural minor scale $(\hat{1}\ \hat{2}\ \flat\hat{3}\ \hat{4}\ \hat{5}\ \flat\hat{6}\flat\hat{7})$ consists of the pitches E, F♯, G, A, B, C, and D. The E harmonic minor scale $(\hat{1}\ \hat{2}\ \flat\hat{3}\ \hat{4}\ \hat{5}\ \flat\hat{6}\ \hat{7})$ contains the natural 7, D♯, rather than the flatted 7, D – to align with the major dominant chord, B7 (B D♯ F♯ A).

Its key signature has one sharp, F (*see below:* Scales and keys).

Its relative major is G major, and its parallel major is E major.

Much of the classical guitar repertoire is in E minor, as this is a very natural key for the instrument. In standard tuning (E A D G B E), four of the instrument's six 'open' (unfretted) strings are part of the tonic chord. The key of E minor is also extremely popular in heavy metal music, as its tonic is the lowest note on a standard-tuned guitar.

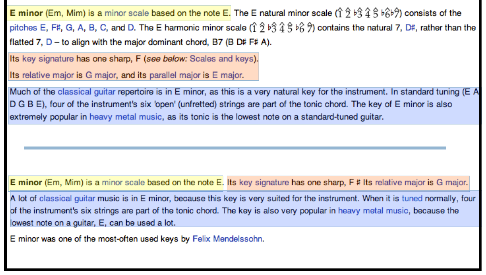
E minor (Em, Mim) is a minor scale based on the note E. Its key signature has one sharp, F ♯ Its relative major is G major.

A lot of classical guitar music is in E minor, because this key is very suited for the instrument. When it is tuned normally, four of the instrument's six strings are part of the tonic chord. The key is also very popular in heavy metal music, because the lowest note on a guitar, E, can be used a lot.

E minor was one of the most-often used keys by Felix Mendelssohn.

55

From aligned documents to aligned sentences



E minor (Em, Mim) is a minor scale based on the note E. The E natural minor scale $(\hat{1}\ \hat{2}\ \flat\hat{3}\ \hat{4}\ \hat{5}\ \flat\hat{6}\flat\hat{7})$ consists of the pitches E, F♯, G, A, B, C, and D. The E harmonic minor scale $(\hat{1}\ \hat{2}\ \flat\hat{3}\ \hat{4}\ \hat{5}\ \flat\hat{6}\ \hat{7})$ contains the natural 7, D♯, rather than the flatted 7, D – to align with the major dominant chord, B7 (B D♯ F♯ A).

Its key signature has one sharp, F (*see below:* Scales and keys).

Its relative major is G major, and its parallel major is E major.

Much of the classical guitar repertoire is in E minor, as this is a very natural key for the instrument. In standard tuning (E A D G B E), four of the instrument's six 'open' (unfretted) strings are part of the tonic chord. The key of E minor is also extremely popular in heavy metal music, as its tonic is the lowest note on a standard-tuned guitar.

E minor (Em, Mim) is a minor scale based on the note E. Its key signature has one sharp, F ♯ Its relative major is G major.

A lot of classical guitar music is in E minor, because this key is very suited for the instrument. When it is tuned normally, four of the instrument's six strings are part of the tonic chord. The key is also very popular in heavy metal music, because the lowest note on a guitar, E, can be used a lot.

E minor was one of the most-often used keys by Felix Mendelssohn.

56

Wikipedia for text simplification

Wikipedia
The Free Encyclopedia
4.4M articles

Simple English
Wikipedia
97K articles

unsimplified		simplified
Alfonso Perez Munoz, usually referred to as Alfonso, is a former Spanish footballer, in the striker position.	**167K aligned sentence pairs**	Alfonso Perez is a former Spanish football player.
The reverse process, producing electrical energy from mechanical, energy, is accomplished by a generator or dynamo.		A dynamo or an electric generator does the reverse: it changes mechanical movement into electric energy.
I find forest colored chicken ovum and pork rump to be dietarily disturbing.		I do not like green eggs and ham.

57

Simplification approaches

58

Simplification approaches

Many different data-driven approaches

- **Lexical (change a word at a time)**
- **Phrasal (change phrases)**
- Syntactic (use grammatical structure)
- Neural networks/LLMs

59

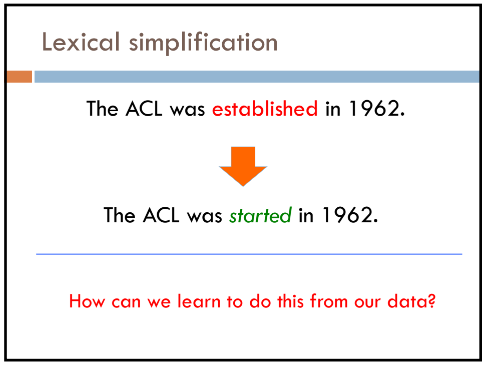
Lexical simplification

The ACL was established in 1962.

The ACL was *started* in 1962.

Simplification is accomplished by changing one word (or phrase) at a time.

60



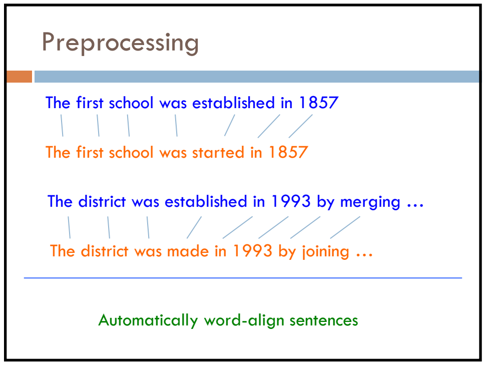
Lexical simplification

The ACL was established in 1962.

The ACL was *started* in 1962.

How can we learn to do this from our data?

61



Preprocessing

The first school was established in 1857

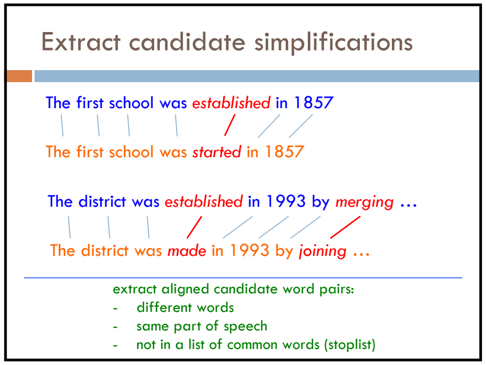
The first school was started in 1857

The district was established in 1993 by merging …

The district was made in 1993 by joining …

Automatically word-align sentences

62



Extract candidate simplifications

The first school was *established* in 1857

The first school was *started* in 1857

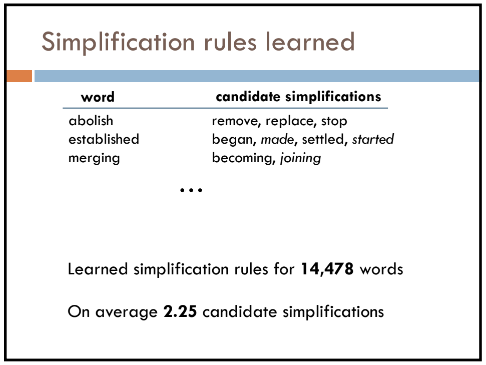
The district was *established* in 1993 by *merging* …

The district was *made* in 1993 by *joining* …

extract aligned candidate word pairs:

- different words
- same part of speech
- not in a list of common words (stoplist)

63



Simplification rules learned

word	candidate simplifications
abolish	remove, replace, stop
established	began, *made*, settled, *started*
merging	becoming, *joining*

…

Learned simplification rules for **14,478** words

On average **2.25** candidate simplifications

64

Not all rules apply in all contexts

The ACL was established in 1962.

The ACL was *began* in 1962. ✗

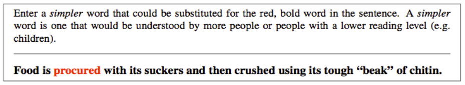
The ACL was *made* in 1962. ?

The ACL was *settled* in 1962. ✗

The ACL was *started* in 1962. ✓

65

Data for learning context



Enter a *simpler* word that could be substituted for the red, bold word in the sentence. A *simpler* word is one that would be understood by more people or people with a lower reading level (e.g. children).

Food is procured with its suckers and then crushed using its tough "beak" of chitin.

amazon mechanicalturk™
Artificial Artificial Intelligence

66

Collected data for 500 words

	simplification	# of people that suggested simplification (out of 50)
procured →	obtained	17
	gathered	9
	gotten	8
	grabbed	4
	acquired	2
	made	2
	…	

67

Learning to apply rules

500 examples

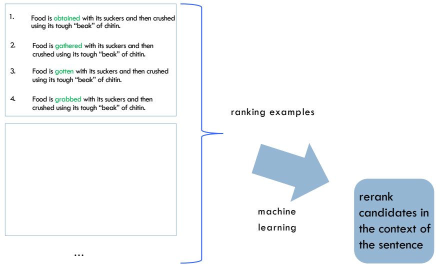
Food is procured with its suckers and then crushed using its tough "beak" of chitin.

1. Food is obtained with its suckers and then crushed using its tough "beak" of chitin.
2. Food is gathered with its suckers and then crushed using its tough "beak" of chitin.
3. Food is gotten with its suckers and then crushed using its tough "beak" of chitin.
4. Food is grabbed with its suckers and then crushed using its tough "beak" of chitin.

…

68

Learning a ranker



1. Food is obtained with its suckers and then crushed using its tough "beak" of chitin.
2. Food is gathered with its suckers and then crushed using its tough "beak" of chitin.
3. Food is gotten with its suckers and then crushed using its tough "beak" of chitin.
4. Food is grabbed with its suckers and then crushed using its tough "beak" of chitin.

…

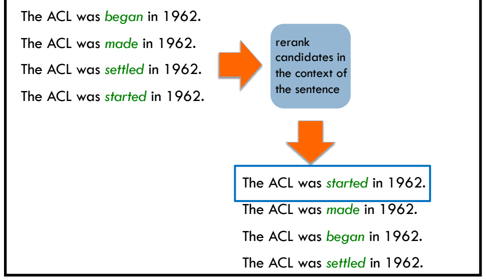
ranking examples

machine learning

rerank candidates in the context of the sentence

69

Applying the ranker



The ACL was *began* in 1962.

The ACL was *made* in 1962.

The ACL was *settled* in 1962.

The ACL was *started* in 1962.

rerank candidates in the context of the sentence

The ACL was *started* in 1962.

The ACL was *made* in 1962.

The ACL was *began* in 1962.

The ACL was *settled* in 1962.

70

Results

Previous approach:

- Coverage: 85% (of the words that could be changed)
- Accuracy: 54% (of the suggestions are correct)

Our approach:

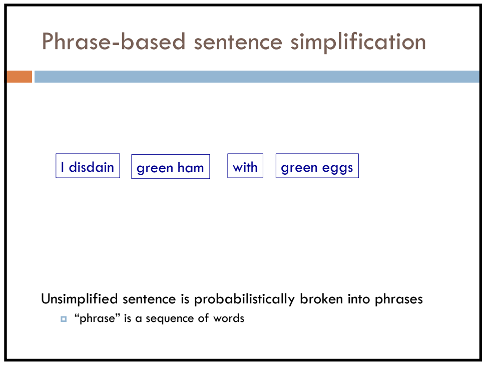
- Coverage: 86%
- Accuracy: **76%**

71

Phrase-based sentence simplification

I disdain green ham with green eggs

72



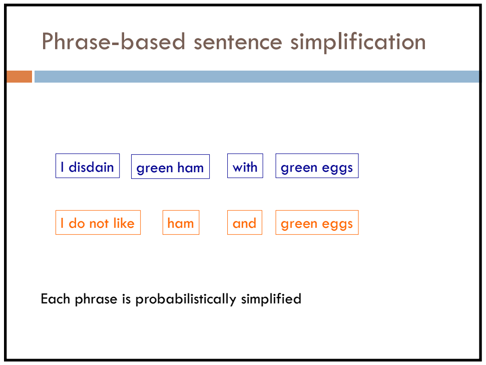
Phrase-based sentence simplification

I disdain
green ham
with
green eggs

Unsimplified sentence is probabilistically broken into phrases

- "phrase" is a sequence of words

73

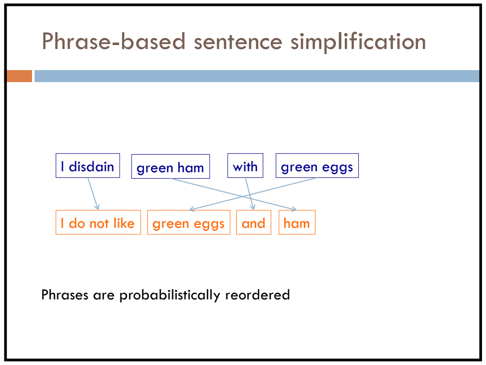


Phrase-based sentence simplification

I disdain
green ham
with
green eggs
I do not like
ham
and
green eggs

Each phrase is probabilistically simplified

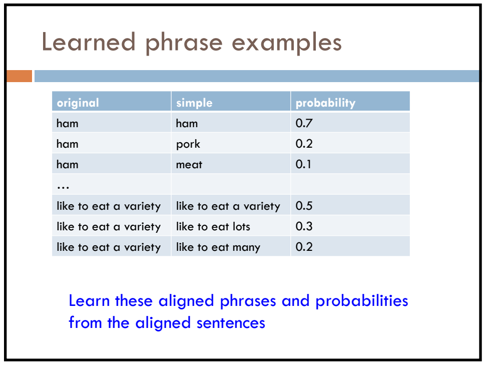
74



Phrase-based sentence simplification

Phrases are probabilistically reordered

75

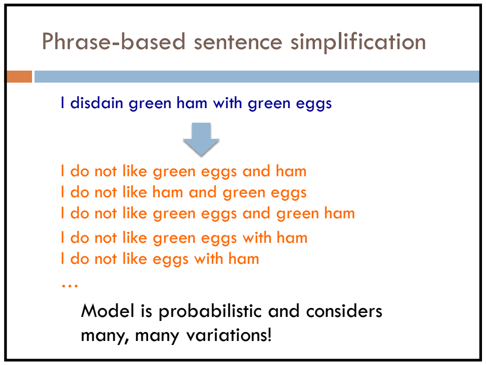


Learned phrase examples

original	simple	probability
ham	ham	0.7
ham	pork	0.2
ham	meat	0.1
…		
like to eat a variety	like to eat a variety	0.5
like to eat a variety	like to eat lots	0.3
like to eat a variety	like to eat many	0.2

Learn these aligned phrases and probabilities from the aligned sentences

76



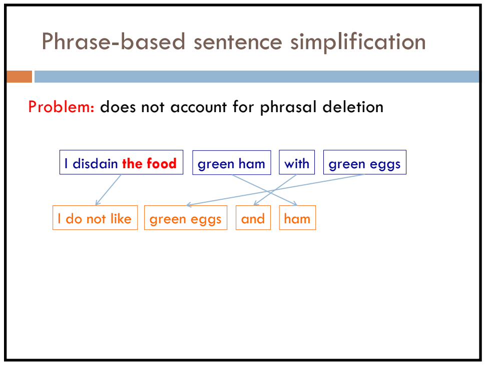
Phrase-based sentence simplification

I disdain green ham with green eggs

I do not like green eggs and ham
I do not like ham and green eggs
I do not like green eggs and green ham
I do not like green eggs with ham
I do not like eggs with ham
…

Model is probabilistic and considers many, many variations!

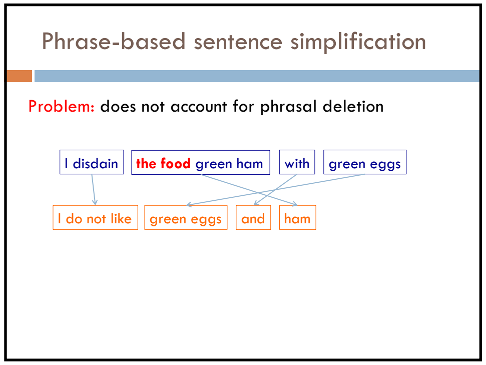
77



Phrase-based sentence simplification

Problem: does not account for phrasal deletion

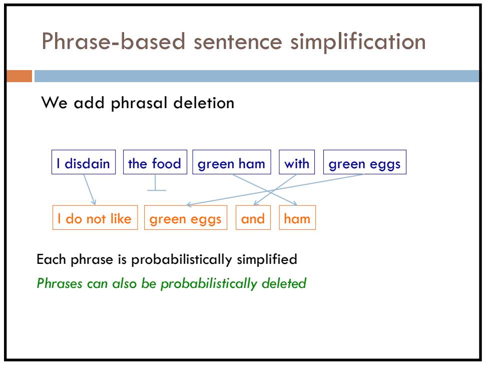
78



Phrase-based sentence simplification

Problem: does not account for phrasal deletion

79



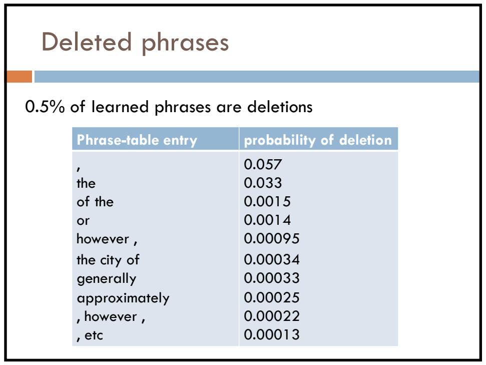
Phrase-based sentence simplification

We add phrasal deletion

Each phrase is probabilistically simplified

Phrases can also be probabilistically deleted

80

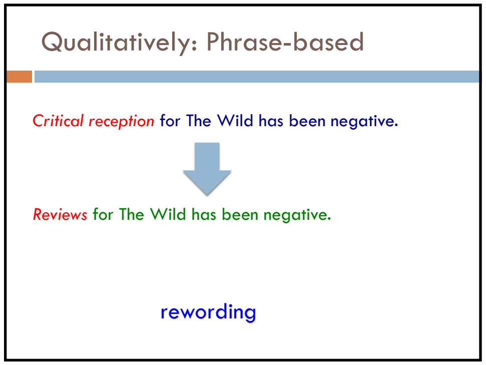


Deleted phrases

0.5% of learned phrases are deletions

Phrase-table entry	probability of deletion
,	0.057
the	0.033
of the	0.0015
or	0.0014
however ,	0.00095
the city of	0.00034
generally	0.00033
approximately	0.00025
, however ,	0.00022
, etc	0.00013

81



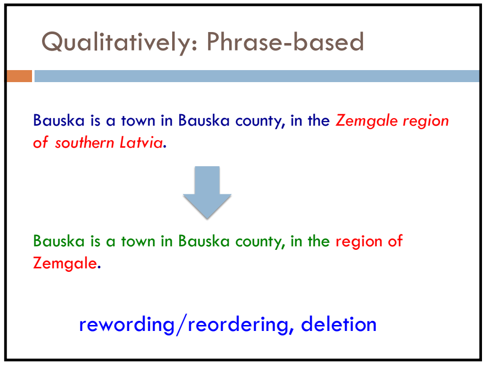
Qualitatively: Phrase-based

Critical reception for The Wild has been negative.

Reviews for The Wild has been negative.

rewording

82



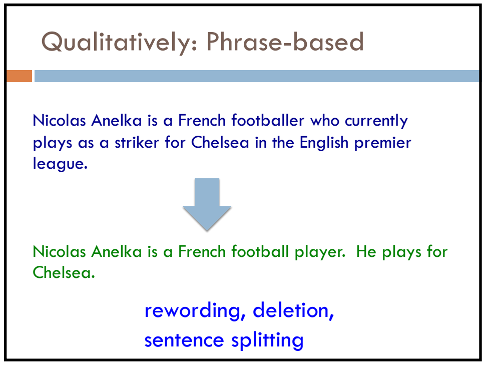
Qualitatively: Phrase-based

Bauska is a town in Bauska county, in the *Zemgale region of southern Latvia*.

Bauska is a town in Bauska county, in the region of Zemgale.

rewording/reordering, deletion

83



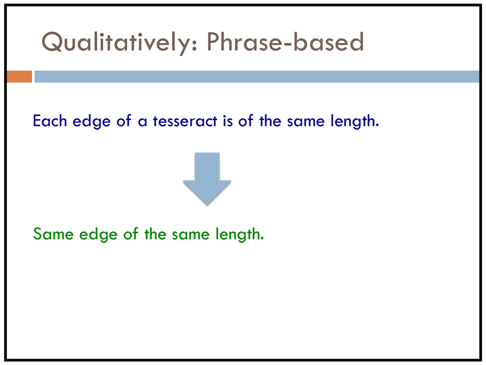
Qualitatively: Phrase-based

Nicolas Anelka is a French footballer who currently plays as a striker for Chelsea in the English premier league.

Nicolas Anelka is a French football player. He plays for Chelsea.

rewording, deletion, sentence splitting

84

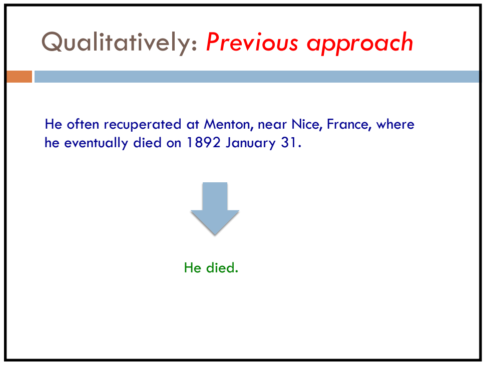


Qualitatively: Phrase-based

Each edge of a tesseract is of the same length.

Same edge of the same length.

85

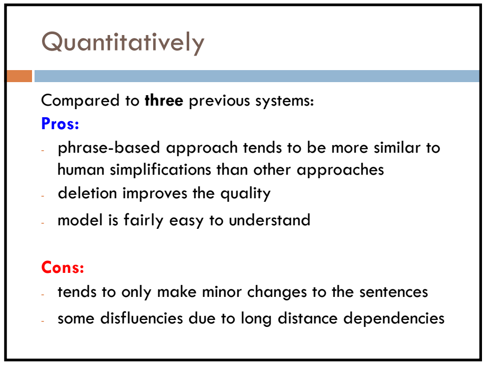


Qualitatively: *Previous approach*

He often recuperated at Menton, near Nice, France, where he eventually died on 1892 January 31.

He died.

86



Quantitatively

Compared to **three** previous systems:

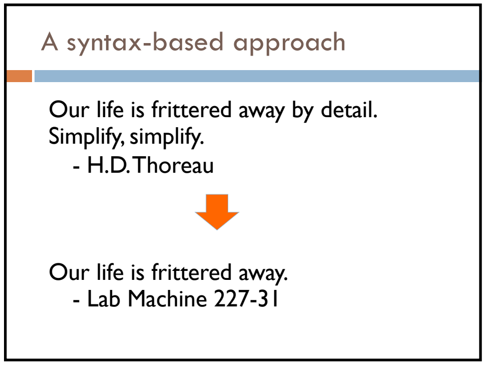
Pros:

- phrase-based approach tends to be more similar to human simplifications than other approaches
- deletion improves the quality
- model is fairly easy to understand

Cons:

- tends to only make minor changes to the sentences
- some disfluencies due to long distance dependencies

87

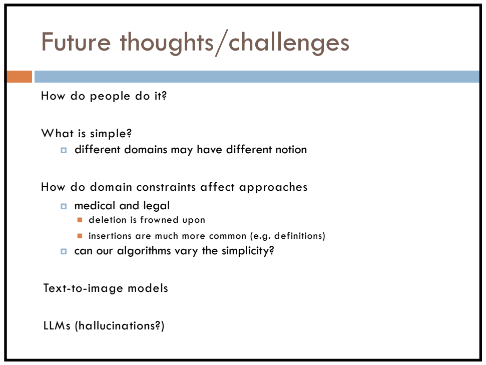


A syntax-based approach

Our life is frittered away by detail.
Simplify, simplify.
- H.D.Thoreau

Our life is frittered away.
- Lab Machine 227-31

88



Future thoughts/challenges

How do people do it?

What is simple?

- different domains may have different notion

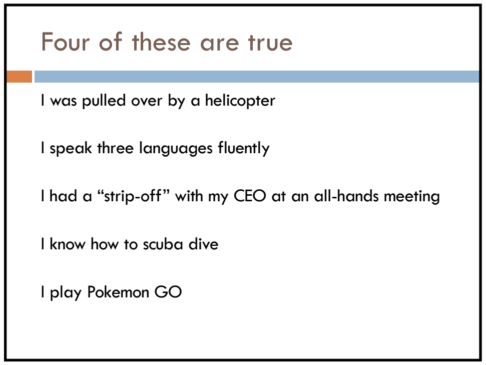
How do domain constraints affect approaches

- medical and legal
  - deletion is frowned upon
  - insertions are much more common (e.g. definitions)
- can our algorithms vary the simplicity?

Text-to-image models

LLMs (hallucinations?)

89



Four of these are true

I was pulled over by a helicopter

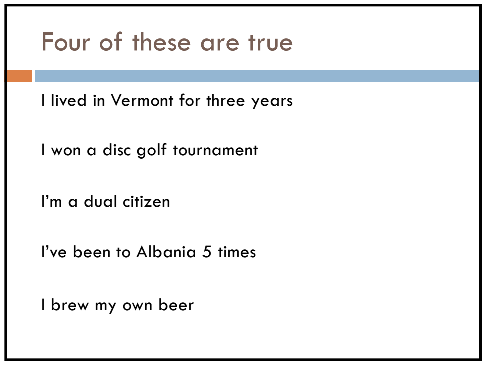
I speak three languages fluently

I had a "strip-off" with my CEO at an all-hands meeting

I know how to scuba dive

I play Pokemon GO

90



Four of these are true

I lived in Vermont for three years

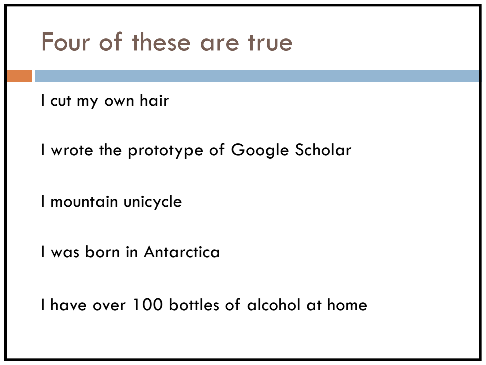
I won a disc golf tournament

I'm a dual citizen

I've been to Albania 5 times

I brew my own beer

91



Four of these are true

I cut my own hair

I wrote the prototype of Google Scholar

I mountain unicycle

I was born in Antarctica

I have over 100 bottles of alcohol at home

92