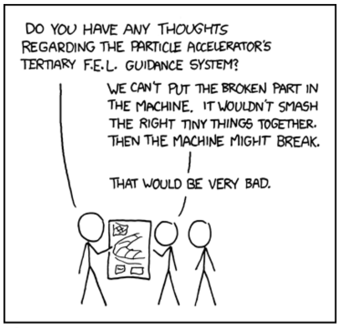

DO YOU HAVE ANY THOUGHTS REGARDING THE PARTICLE ACCELERATOR'S TERTIARY F.E.L. GUIDANCE SYSTEM?
WE CAN'T PUT THE BROKEN PART IN THE MACHINE. IT WOULDN'T SMASH THE RIGHT TINY THINGS TOGETHER. THEN THE MACHINE MIGHT BREAK.
THAT WOULD BE VERY BAD.

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

https://xkcd.com/547/

1

TEXT SIMPLIFICATION: IMPROVING INFORMATION ACCESSIBILITY

David Kauchak
Pomona College

Collaborators: Will Coster, Dan Feblowitz, Melissa Grueter, Colby Horn, Gondy Leroy, Katie Manduca and Max Schwarzer

2

Admin

Ethics discussion on Wednesday: read papers beforehand

Project status update due Wednesday

Next Monday, Quiz #4 (comprehensive)

Next Wednesday, presentations

3

Text simplification

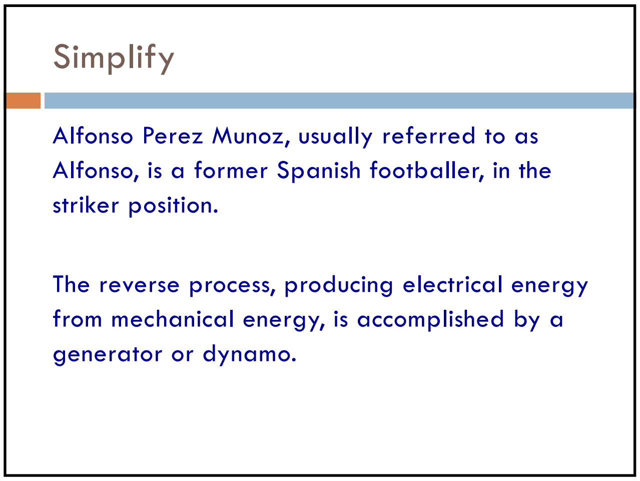
Any intelligent fool can make things bigger, more complex, and more violent. It takes a touch of genius and a lot of courage to move in the opposite direction.

- E. F. Schumacher

Goal:

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

4

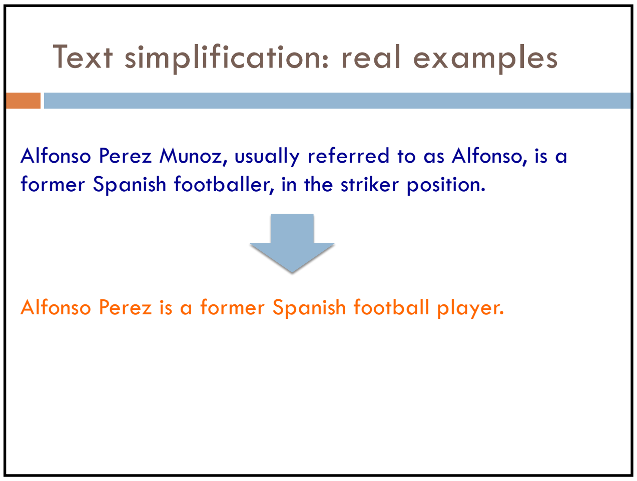


Simplify

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

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

7



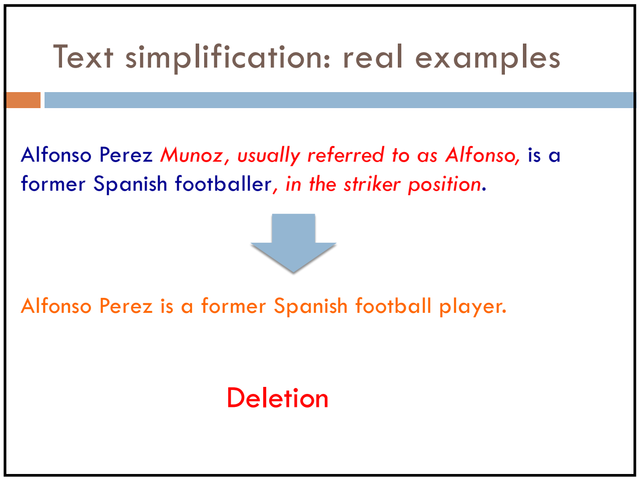
Text simplification: real examples

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.

8



Text simplification: real examples

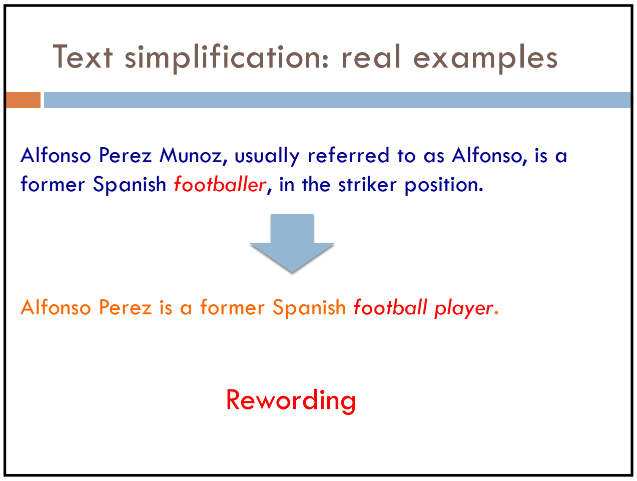
Alfonso Perez *Munoz, usually referred to as Alfonso,* is a former Spanish footballer*, in the striker position*.

↓

Alfonso Perez is a former Spanish football player.

Deletion

9



Text simplification: real examples

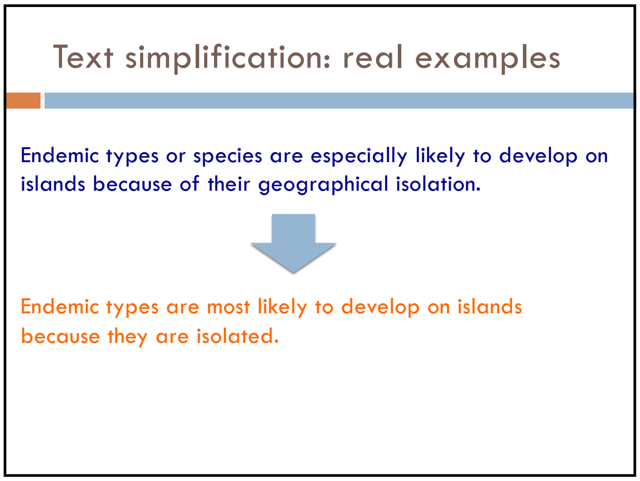
Alfonso Perez Munoz, usually referred to as Alfonso, is a former Spanish *footballer*, in the striker position.

↓

Alfonso Perez is a former Spanish *football player*.

Rewording

10

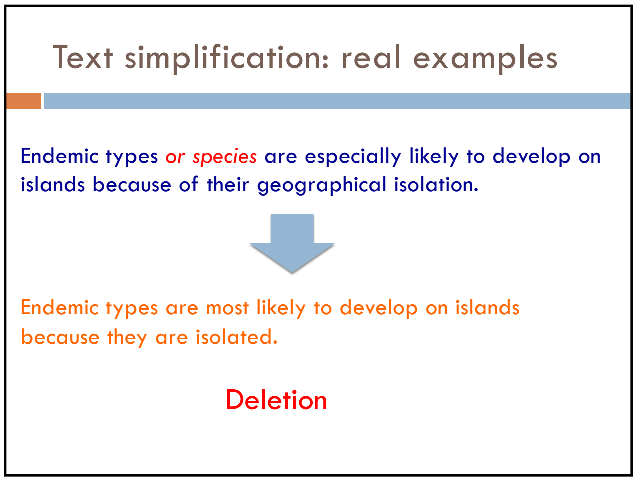


Text simplification: real examples

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.

11



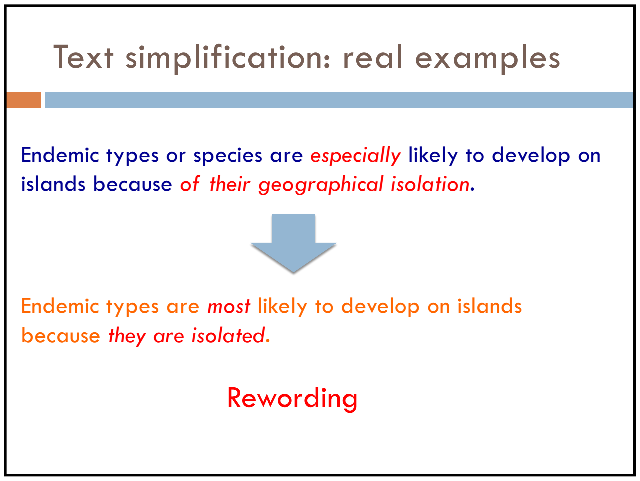
Text simplification: real examples

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

12



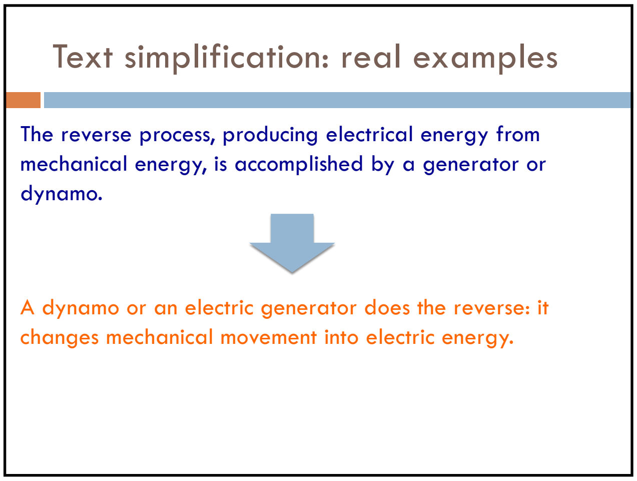
Text simplification: real examples

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

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.

14

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*.

15

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

16

Goals today

Introduce the text simplification problem ✓

Highlight why text simplification is important

Show some examples of text simplification approaches

Give one perspective on CS research

17

Why text simplification?

A lot of text data is available

WIKIPEDIA
The Free Encyclopedia

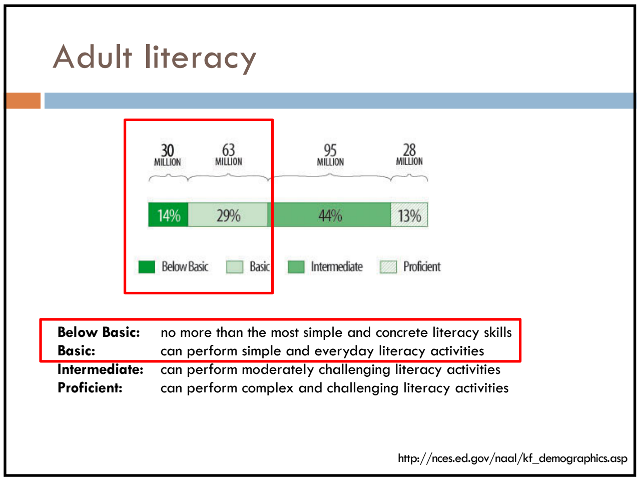
Google

acm

WebMD

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

18

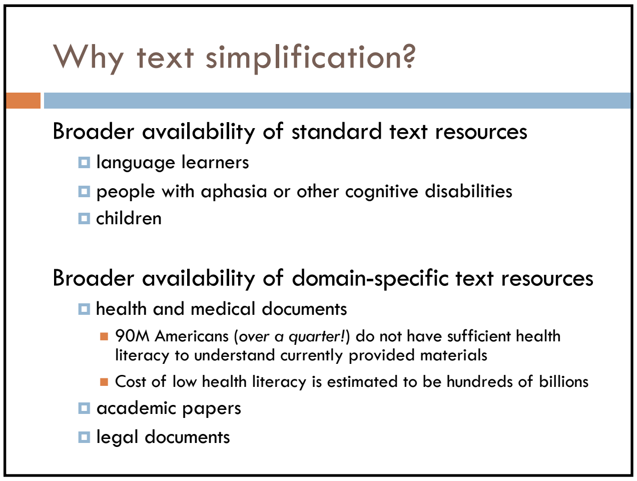


Adult literacy

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

19



Why text simplification?

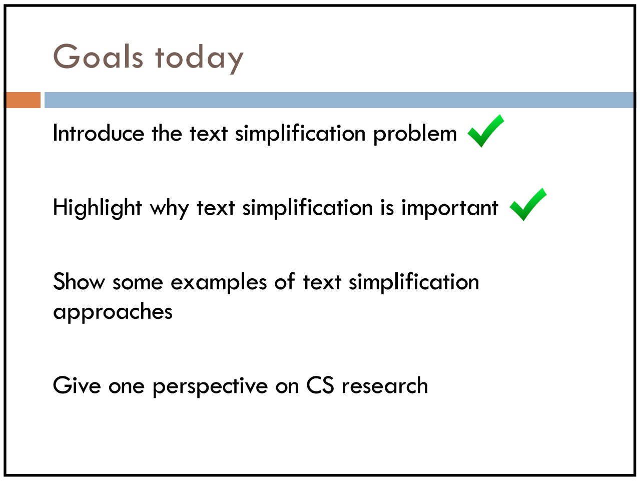
Broader availability of standard text resources

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

Broader availability of domain-specific text resources

- health and medical documents
  - 90M Americans (*over a quarter!*) 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

20



Goals today

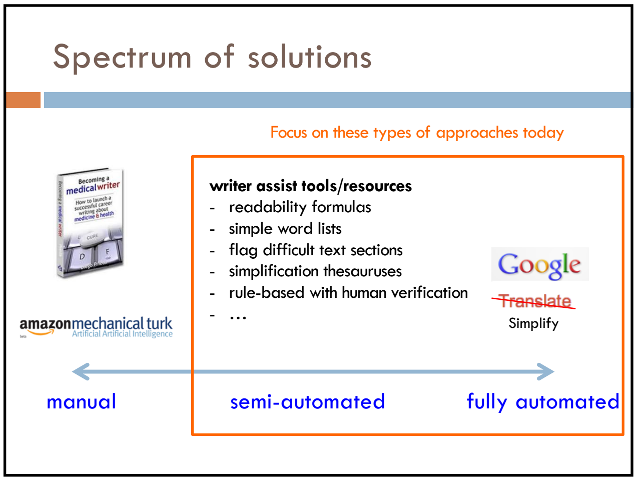
Introduce the text simplification problem ✓

Highlight why text simplification is important ✓

Show some examples of text simplification approaches

Give one perspective on CS research

21



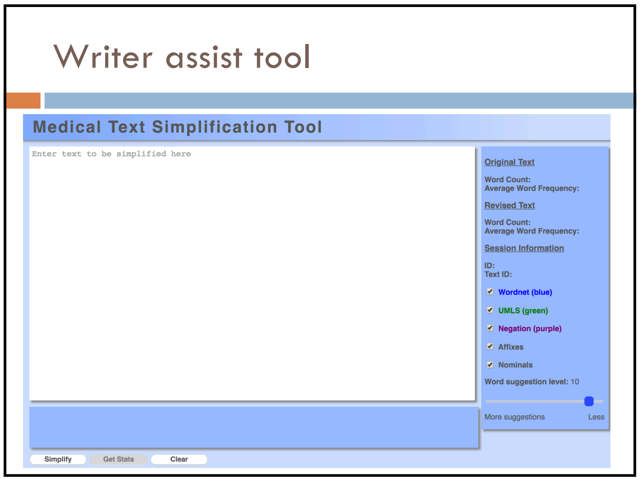
Spectrum of solutions

Focus on these types of approaches today

writer assist tools/resources

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

22



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)

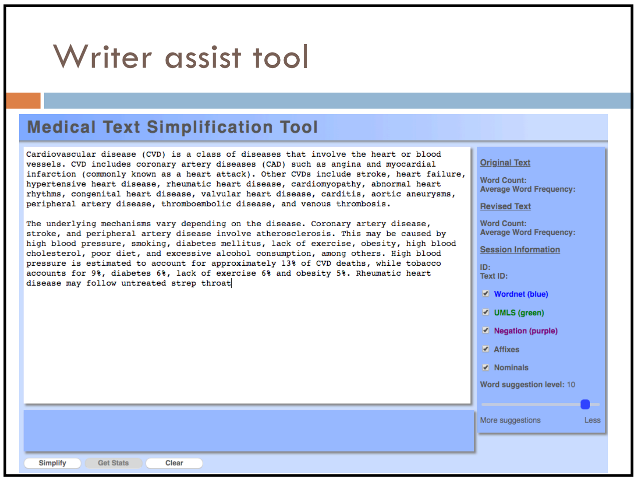
☑ Negation (purple)

☑ Affixes

☑ Nominals

Word suggestion level: 10

23



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)

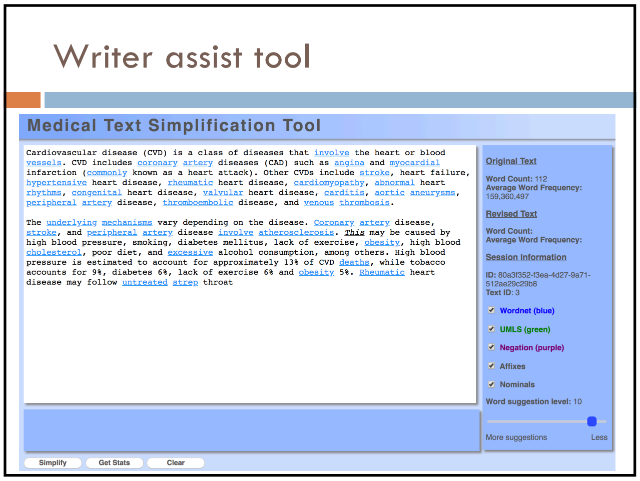
☑ Negation (purple)

☑ Affixes

☑ Nominals

Word suggestion level: 10

24



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)

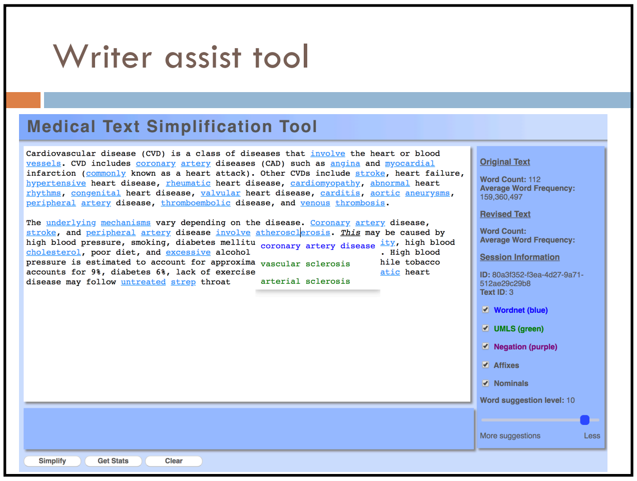
☑ Negation (purple)

☑ Affixes

☑ Nominals

Word suggestion level: 10

25



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)

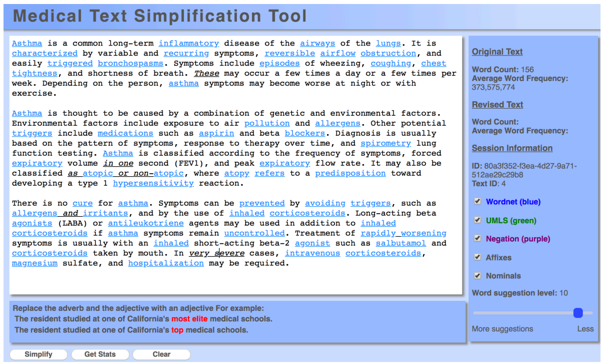
☑ Affixes

☑ Nominals

Word suggestion level: 10

26

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)

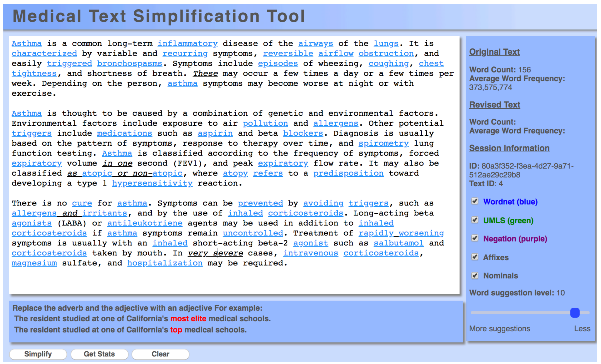
☑ Affixes

☑ Nominals

Word suggestion level: 10

27

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

28

Quantifying word difficulty

Hypothesis:

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

Proxy for "how often you see a word":

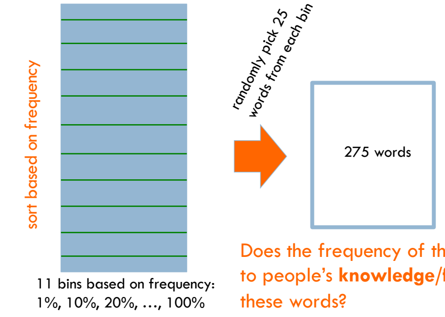
Frequency on the web!

Google bing Y!

29

Validating the frequency hypothesis

Google: ~13M unique "words"



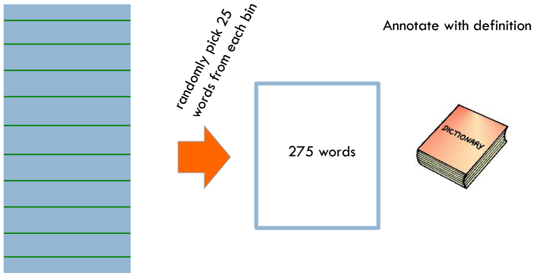
sort based on frequency
randomly pick 25 words from each bin
275 words
11 bins based on frequency:
1%, 10%, 20%, …, 100%

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

30

Validating the frequency hypothesis

Google: ~13M unique "words"



randomly pick 25 words from each bin

275 words

Annotate with definition

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

31

Validating the 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 color or shade

32

Validating the 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 color or shade

random definitions from other words in data set

33

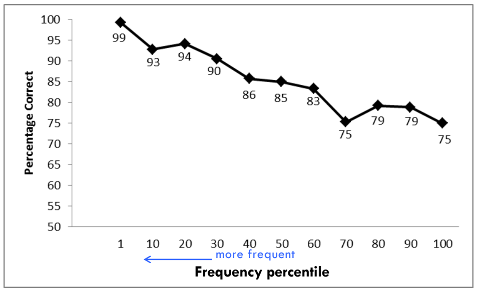
Study participants

amazon mechanical turk™
Artificial Artificial Intelligence

50 participants per word * 275 words = 13,750 total annotations!

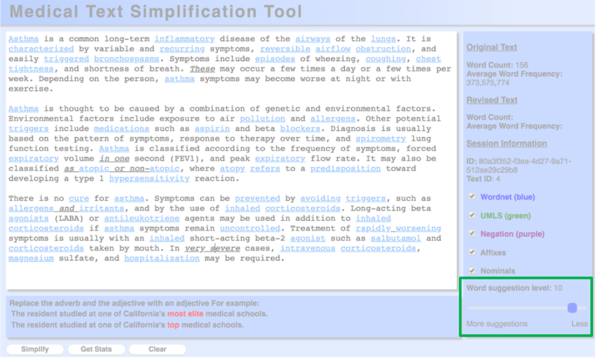
34

Frequency correlates with understanding!



35

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.

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

More suggestions — Less

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.

36

Spectrum of solutions

Becoming a medical writer
How to launch a successful career writing about medicine & health

amazonmechanicalturk
Artificial Artificial Intelligence

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

Google
~~Translate~~
Simplify

manual ↔ semi-automated ↔ fully automated

37

Most modern approaches are data-driven

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.
⋮	

learning →

Given data
(paired sentences)

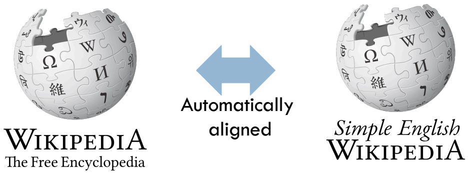
learn a simplification model

38

Simple English Wikipedia

39

Example data source: Wikipedia

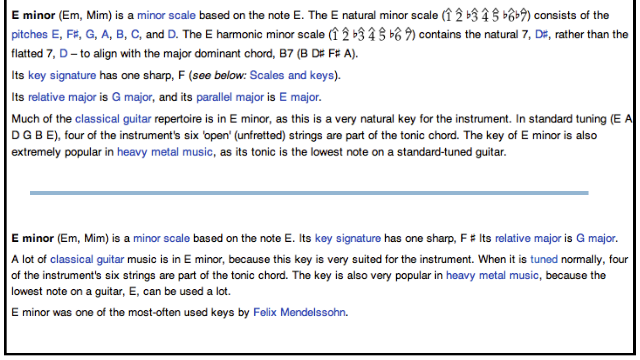


60K articles pairs

~200K aligned sentence pairs

40

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.

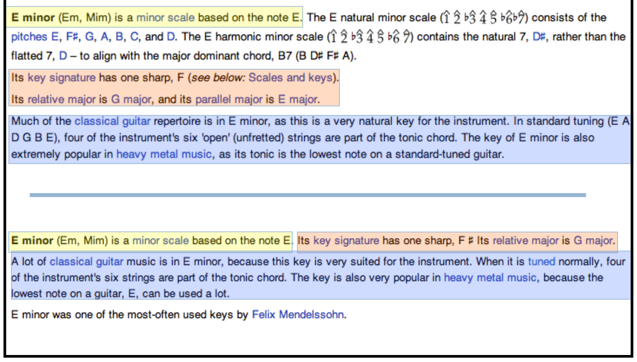
E minor (Em, Mim) is a minor scale based on the note E. Its key signature has one sharp, F ♯ Its relative major is G major.

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.

41

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.

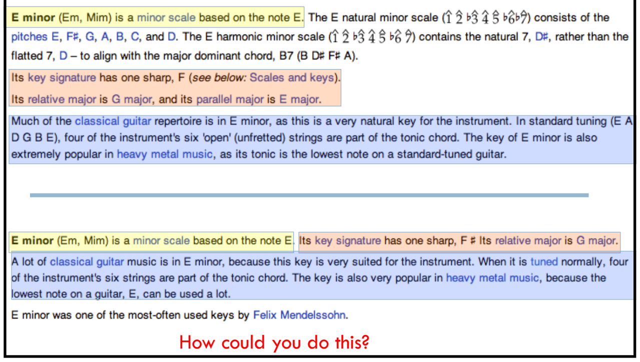
E minor (Em, Mim) is a minor scale based on the note E. Its key signature has one sharp, F ♯ Its relative major is G major.

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.

42

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.

How could you do this?

43

Simplification approaches

Many different data-driven approaches

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

44

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.

45

Lexical simplification

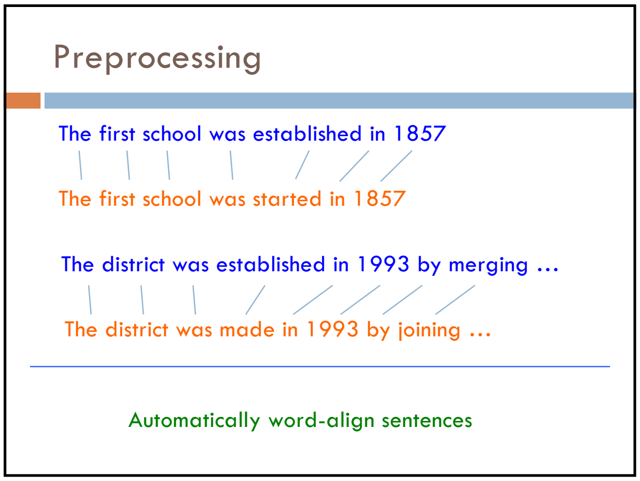
The ACL was established in 1962.

↓

The ACL was *started* in 1962.

How can we learn to do this from our data?

46



Preprocessing

The first school was established in 1857

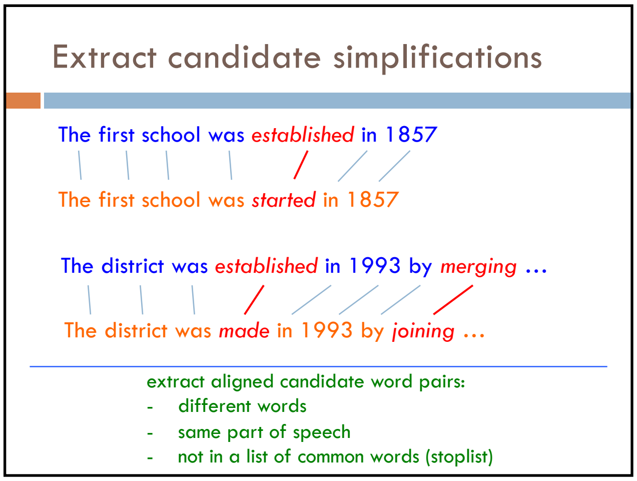
The first school was started in 1857

The district was established in 1993 by merging …

The district was made in 1993 by joining …

Automatically word-align sentences

47



Extract candidate simplifications

The first school was *established* in 1857

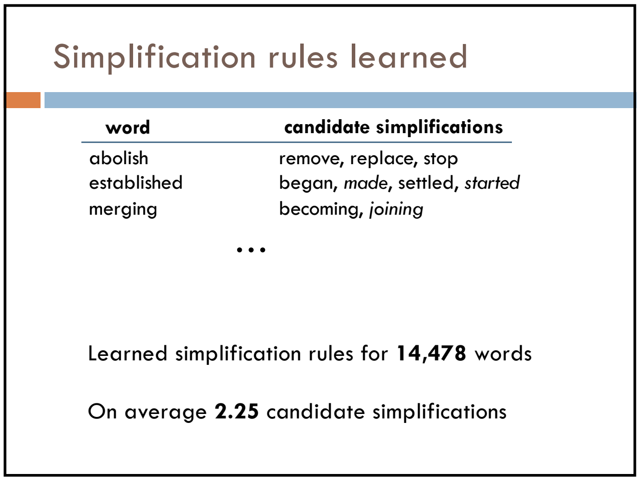
The first school was *started* in 1857

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)

48



Simplification rules learned

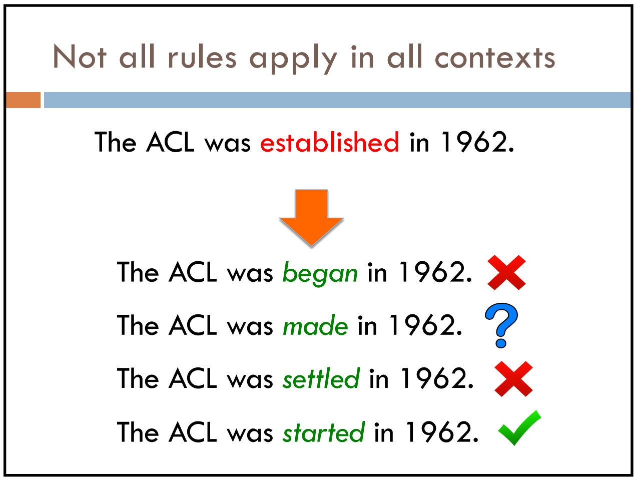
word	candidate simplifications
abolish	remove, replace, stop
established	began, *made*, settled, *started*
merging	becoming, *joining*

…

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

On average **2.25** candidate simplifications

49



Not all rules apply in all contexts

The ACL was established in 1962.

The ACL was *began* in 1962. ✗

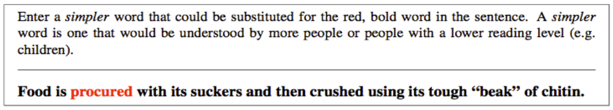
The ACL was *made* in 1962. ?

The ACL was *settled* in 1962. ✗

The ACL was *started* in 1962. ✓

50

Data for learning context



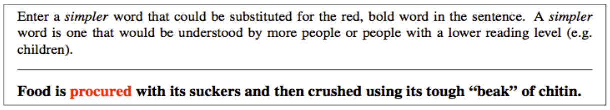
Enter a *simpler* word that could be substituted for the red, bold word in the sentence. A *simpler* word is one that would be understood by more people or people with a lower reading level (e.g. children).

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

amazon mechanical turk™
Artificial Artificial Intelligence

51

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 mechanical turk™
Artificial Artificial Intelligence

52

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
	…	

53

Learning to apply rules

500 examples

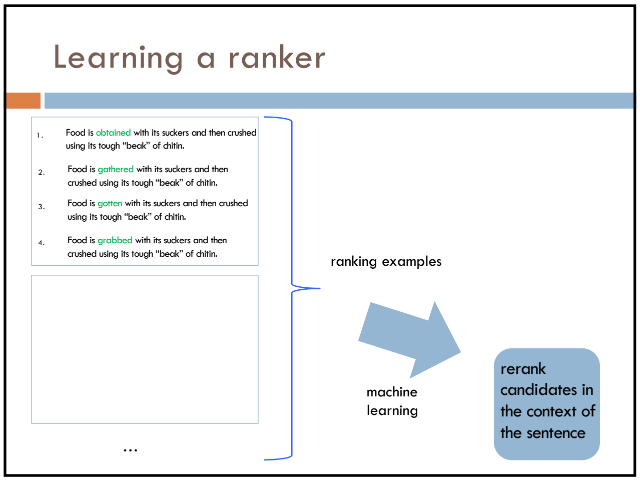
Food is procured with its suckers and then crushed using its tough "beak" of chitin.

↓

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.

…

54



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.

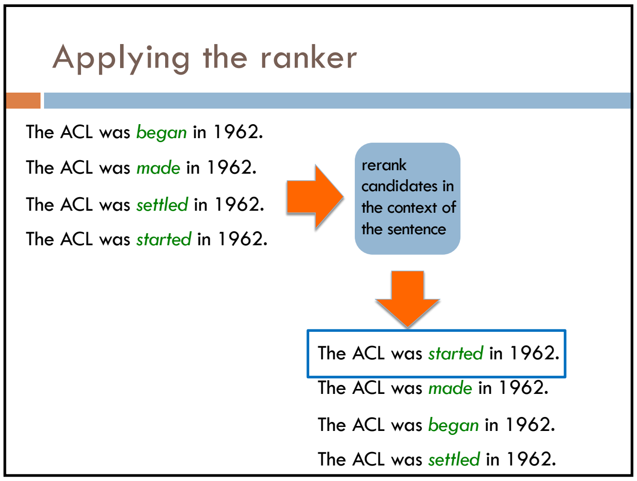
…

ranking examples

machine learning

rerank candidates in the context of the sentence

55



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

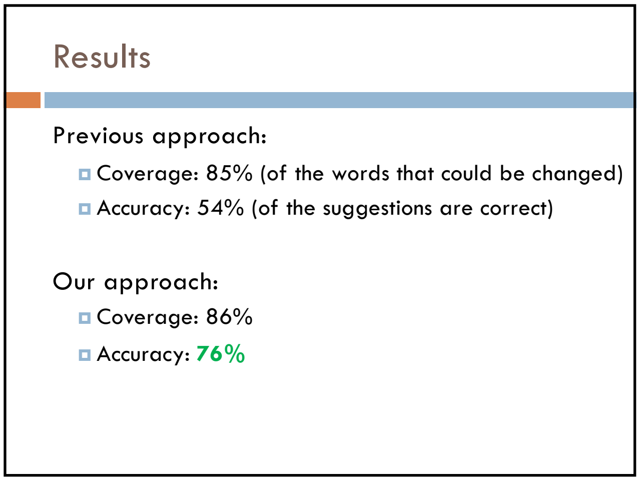
The ACL was *started* in 1962.

The ACL was *made* in 1962.

The ACL was *began* in 1962.

The ACL was *settled* in 1962.

56



Results

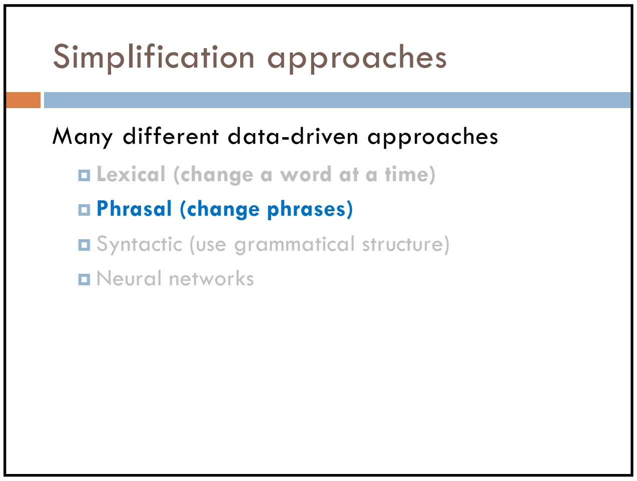
Previous approach:

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

Our approach:

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

57

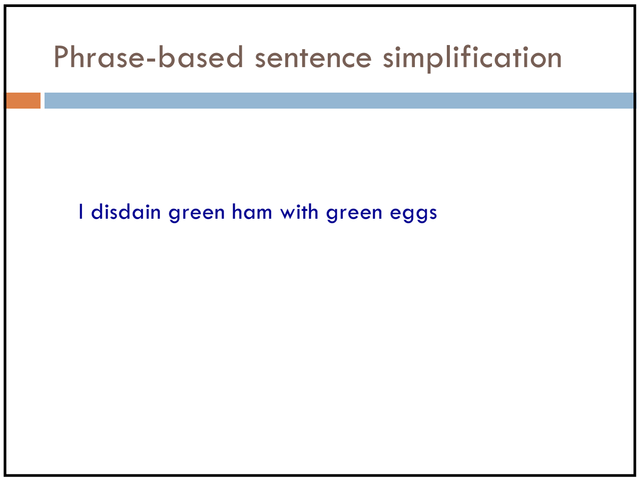


Simplification approaches

Many different data-driven approaches

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

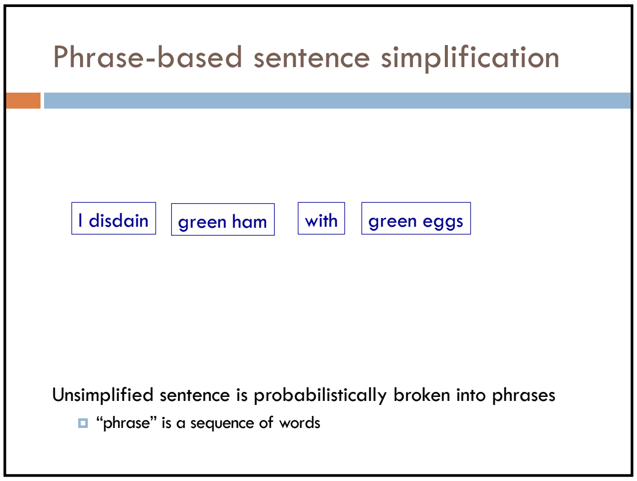
58



Phrase-based sentence simplification

I disdain green ham with green eggs

59



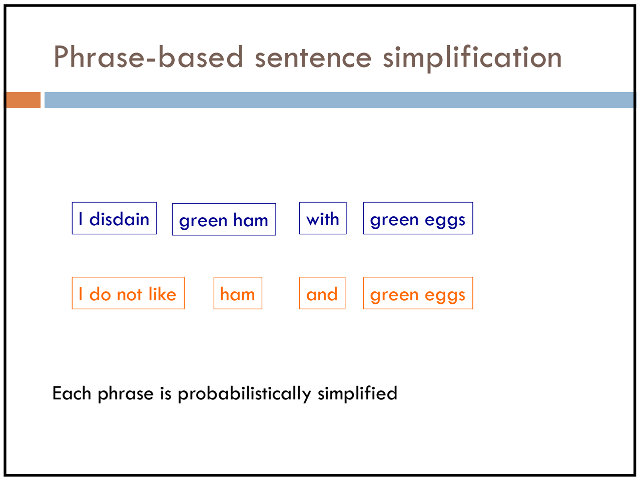
Phrase-based sentence simplification

I disdain | green ham | with | green eggs

Unsimplified sentence is probabilistically broken into phrases

- “phrase” is a sequence of words

60



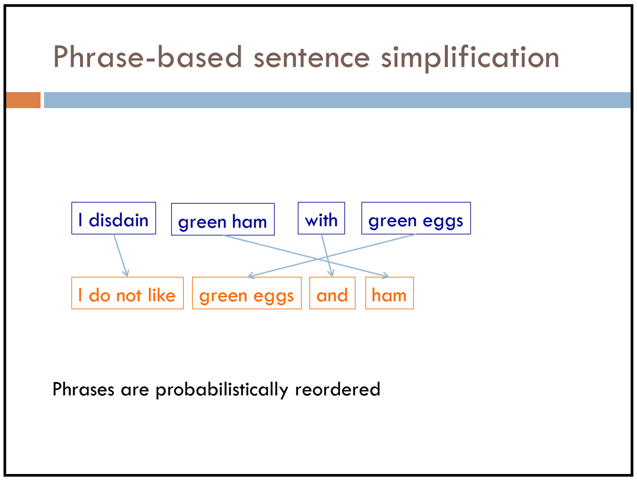
Phrase-based sentence simplification

I disdain | green ham | with | green eggs

I do not like | ham | and | green eggs

Each phrase is probabilistically simplified

61



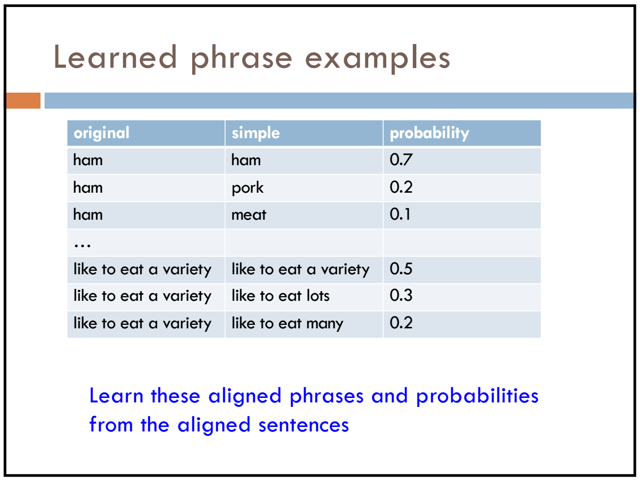
Phrase-based sentence simplification

I disdain | green ham | with | green eggs

I do not like | green eggs | and | ham

Phrases are probabilistically reordered

62

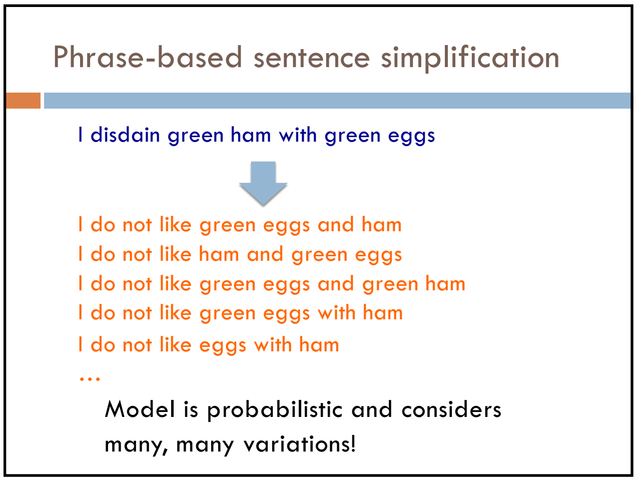


Learned phrase examples

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

63



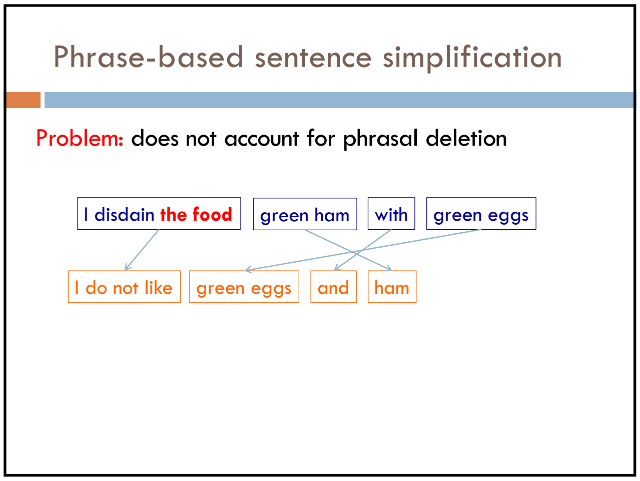
Phrase-based sentence simplification

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!

64



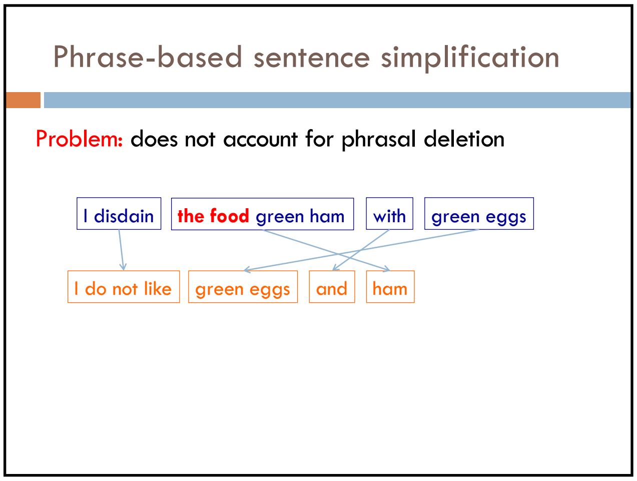
Phrase-based sentence simplification

Problem: does not account for phrasal deletion

I disdain **the food** | green ham | with | green eggs

I do not like | green eggs | and | ham

65



Phrase-based sentence simplification

Problem: does not account for phrasal deletion

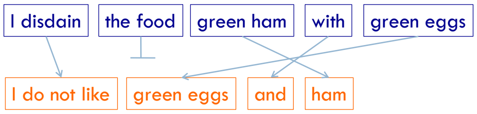
I disdain | **the food** green ham | with | green eggs

I do not like | green eggs | and | ham

66

Phrase-based sentence simplification

We add phrasal deletion



Each phrase is probabilistically simplified

Phrases can also be probabilistically deleted

67

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

68

Qualitatively: Phrase-based

Critical reception for The Wild has been negative.

Reviews for The Wild has been negative.

rewording

69

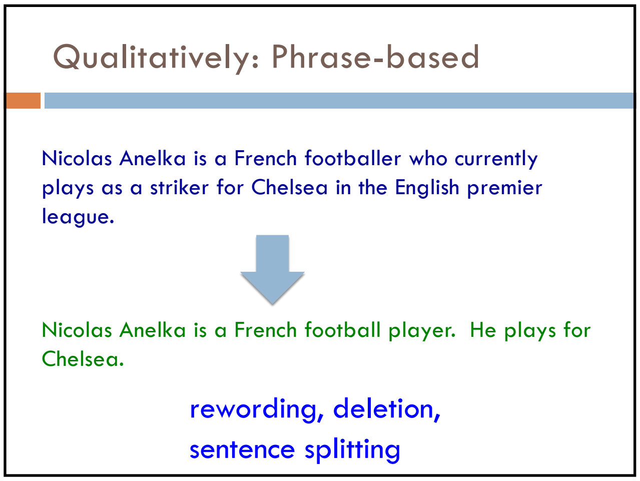
Qualitatively: Phrase-based

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

70



Qualitatively: Phrase-based

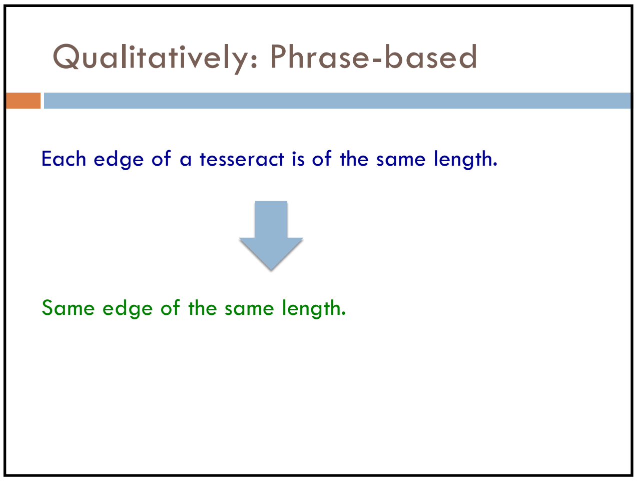
Nicolas Anelka is a French footballer who currently plays as a striker for Chelsea in the English premier league.

↓

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

rewording, deletion,
sentence splitting

71



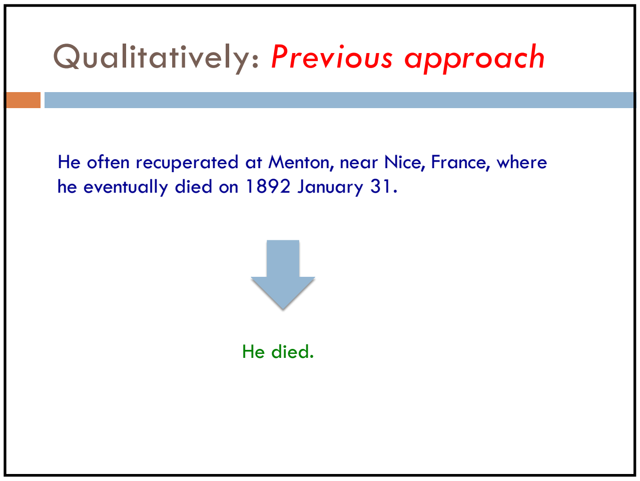
Qualitatively: Phrase-based

Each edge of a tesseract is of the same length.

↓

Same edge of the same length.

72



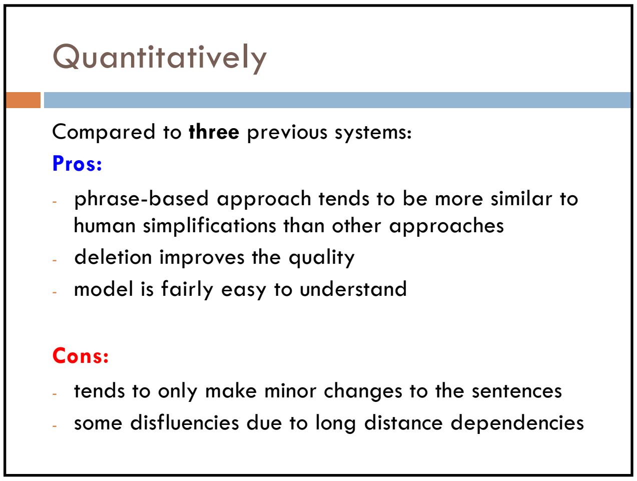
Qualitatively: *Previous approach*

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

↓

He died.

73



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

74

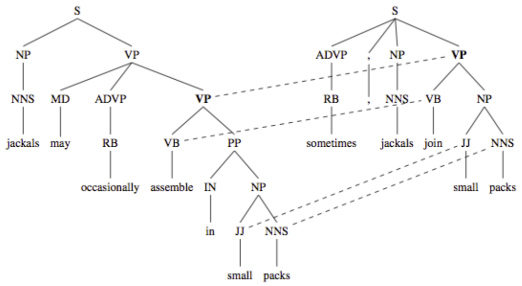
Simplification approaches

Many different data-driven approaches

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

75

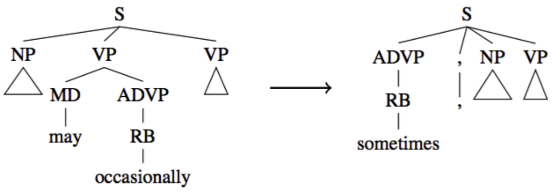
Syntax-based approach



Rather than operating on phrases, operate on grammar trees

76

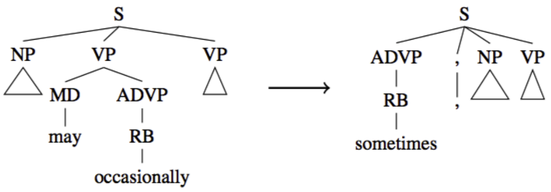
Learn probabilistic, syntax-based rules



They may occasionally eat → sometimes, they eat

77

Learn probabilistic, syntax-based rules



The scary cats from the park may occasionally walk around on two legs → sometimes, the scary cats from the park walk around on two legs

78

An aside

sometimes, the scary cats from the park
walk around on two legs

79

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

80

Qualitatively: syntax-based

Overall Bamberga is the tenth brightest main belt asteroid ***after, in order, Vesta, Pallas, Ceres, Iris, Hebe, Juno, Melpomene, Eunomia and Flora***.

Syntax:
Overall Bamberga is the tenth brightest main belt asteroid.

Phrase-based: (same as input)

81

Quantitatively

Compared to **phrase-based**:

Pros:

- Much more significant simplifications
- More grammatical
- Simpler

Cons:

- Was sometimes aggressive about removing content

82

Goals today

Introduce the text simplification problem ✓

Highlight why text simplification is important ✓

Show some examples of text simplification approaches ✓

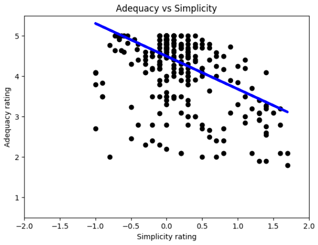
Give one perspective on CS research ✓

83

Future thoughts/challenges

What is simple?

- different domains may have different notion
- how do we measure/evaluate simplicity



84

Collaborators!

Will Coster (Pomona)
Dan Feblowitz (Pomona)
Melissa Grueter (Pomona)
Colby Horn (Middlebury)
Katie Manduca (Middlebury)
Max Schwarzer (Pomona)
Mui Tanprasert (Pomona)
Gondy Leroy (University of Arizona)

85

Course recap

Corpus analysis

Regex

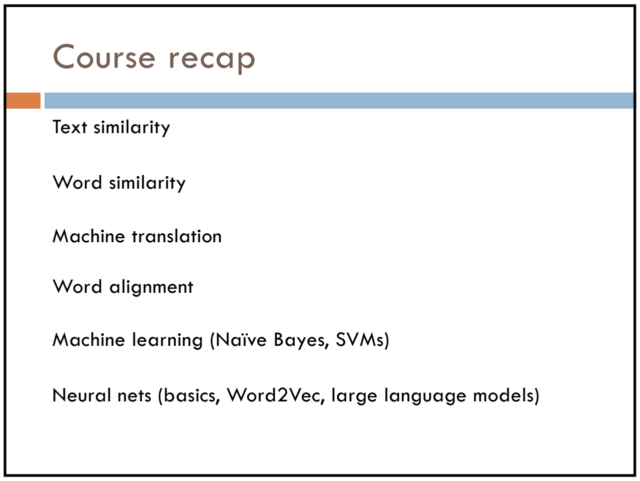
Language modeling (smoothing!)

Linguistics basics

Learning grammars PCFGs

Parsing

86



Course recap

Text similarity

Word similarity

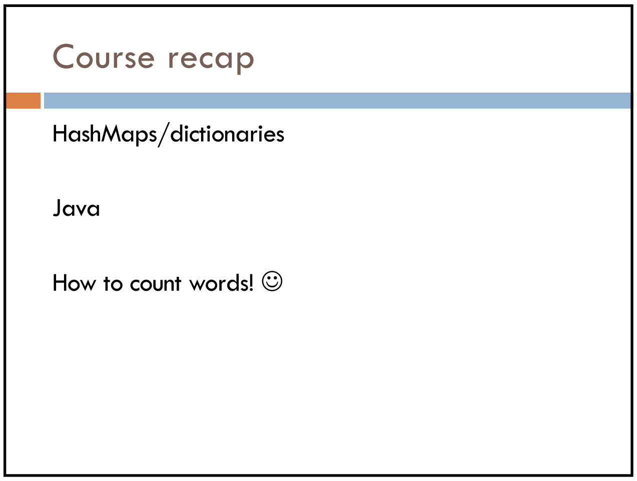
Machine translation

Word alignment

Machine learning (Naïve Bayes, SVMs)

Neural nets (basics, Word2Vec, large language models)

87



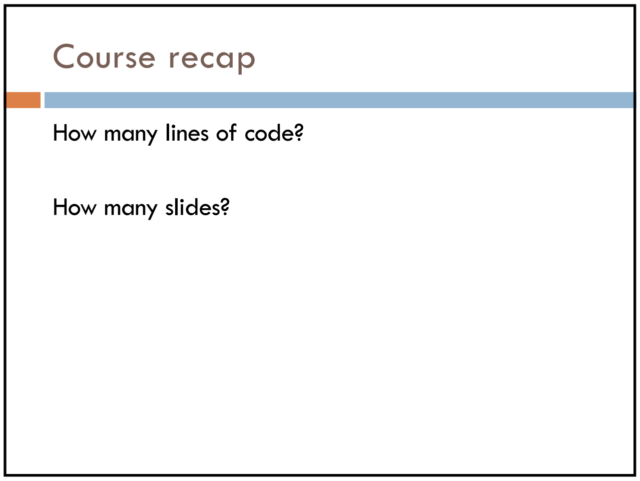
Course recap

HashMaps/dictionaries

Java

How to count words! ☺

88



Course recap

How many lines of code?

How many slides?

89