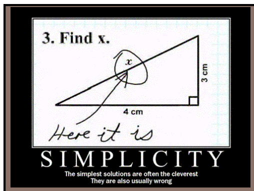

3. Find x.
x
3 cm
4 cm
Here it is
SIMPLICITY
The simplest solutions are often the cleverest
They are also usually wrong

TEXT SIMPLIFICATION

David Kauchak
CS159 – Fall 2014

Collaborators: Will Coster, Dan Feblowitz and Gondy Leroy

Admin

Paper draft due 5pm Wednesday
- Must be done with all of your experiments
- "Results" section is required

1 hr quiz on Tuesday

Review

Corpus analysis

Basic probability

Language modeling
- n-gram language models
- different smoothing techniques

Parsing
- CFG, PCFGs
- CKY algorithm
- improved models

Text and word similarity

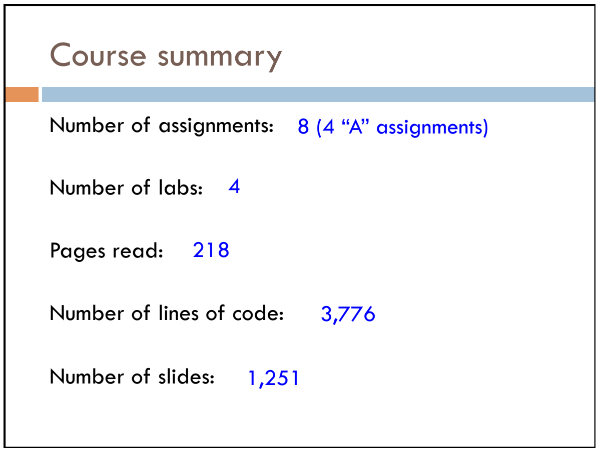
Review

Machine translation
- MT basics
- translation models
- word alignment

Machine learning
- ML basics
- NB (multinomial and Bernoulli)
- smoothing
- other models (*k*-NN, SVM)

NLP research topics
- text modeling
- text simplification

High-level themes
- Probabilistic modeling and data-driven modeling
- Evaluation



Course summary

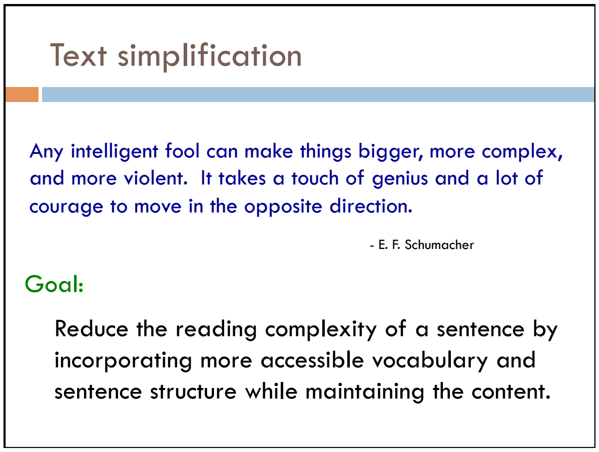
Number of assignments: 8 (4 "A" assignments)

Number of labs: 4

Pages read: 218

Number of lines of code: 3,776

Number of slides: 1,251



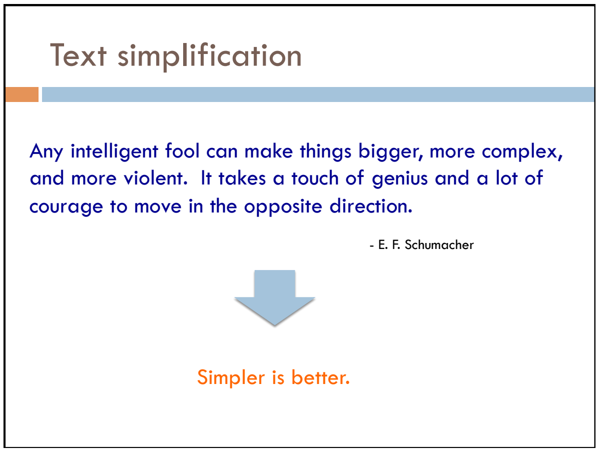
Text simplification

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.

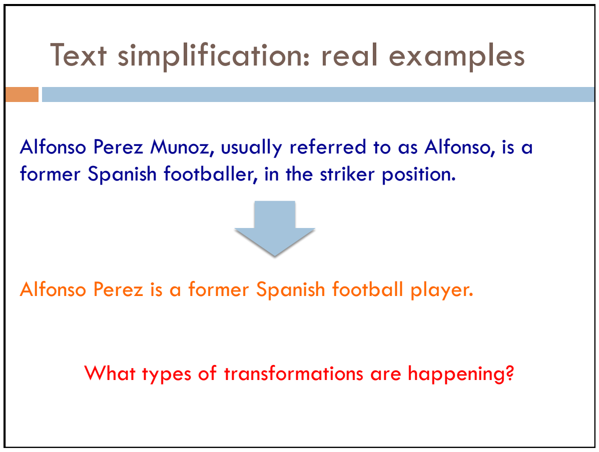


Text simplification

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

Simpler is better.

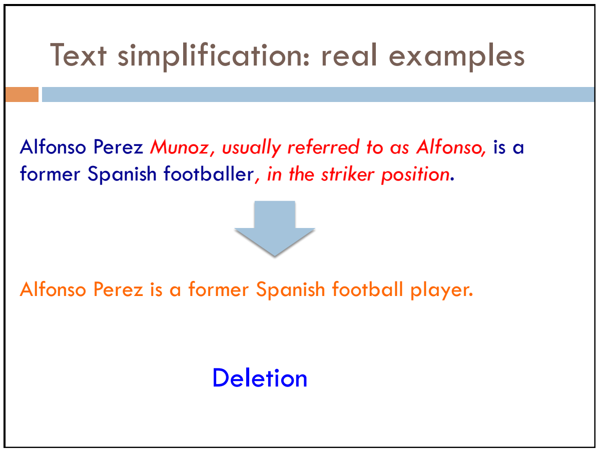


Text simplification: real examples

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?

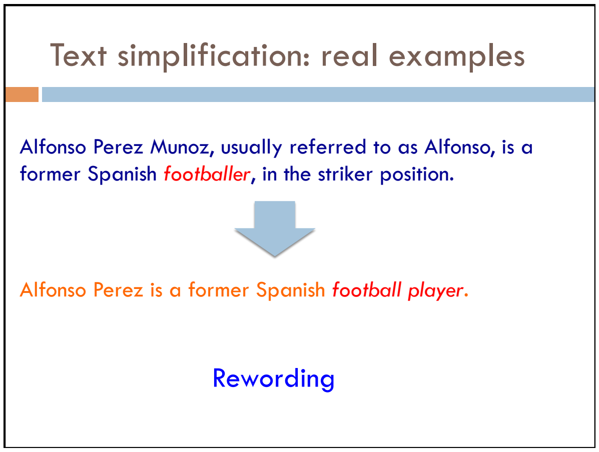


Text simplification: real examples

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

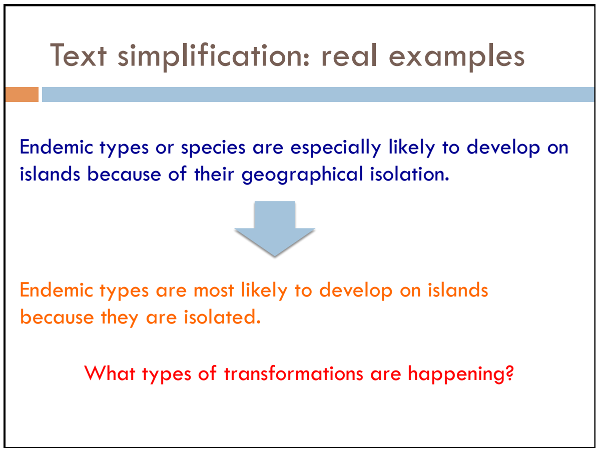


Text simplification: real examples

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

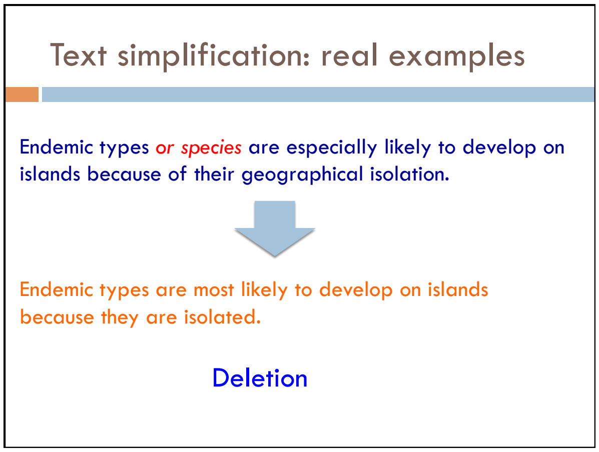


Text simplification: real examples

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?

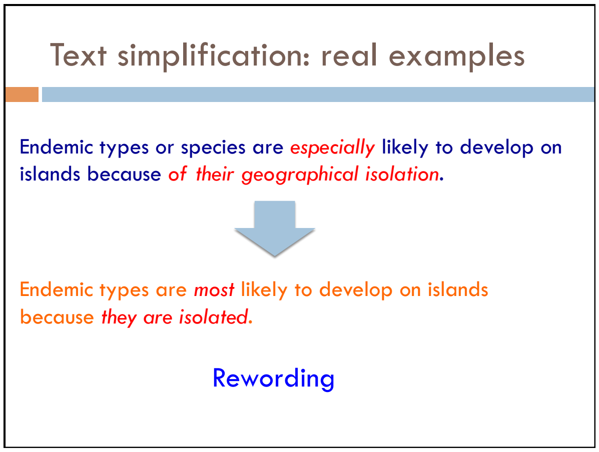


Text simplification: real examples

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

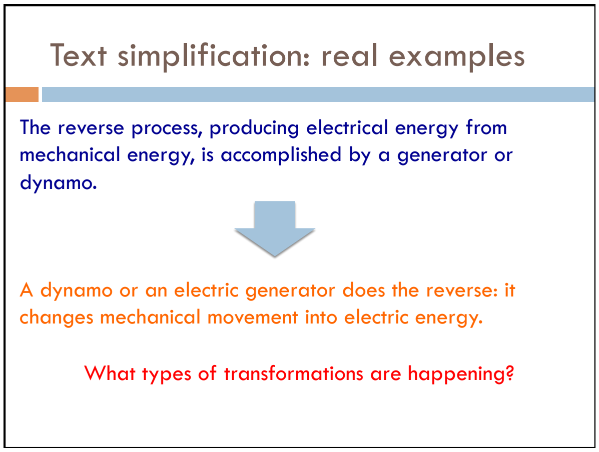


Text simplification: real examples

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

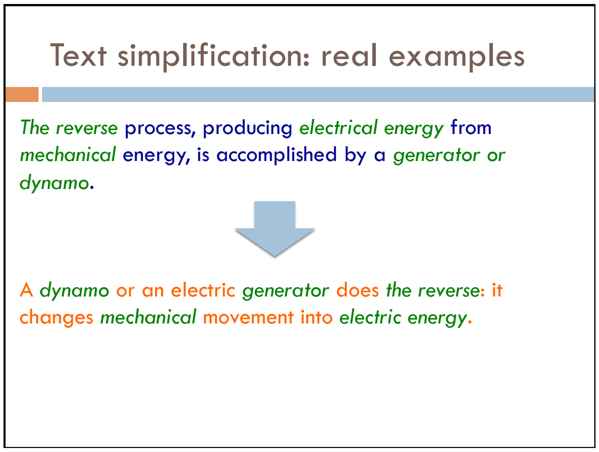


Text simplification: real examples

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?



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*.

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

Goals today

Introduce the text simplification problem ✓

Understand why it's important

Examine what makes text difficult/simple

Overview of approaches to text simplification

Why text simplification?

→

DO
NOT
PARK
HERE

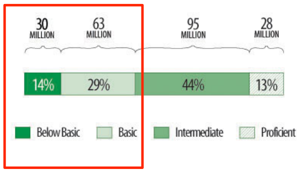
Why text simplification?

A lot of text data is available

Wikipedia The Free Encyclopedia · Google · acm · WebMD · Project Gutenberg

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

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

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

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.

What makes text difficult/simple?

?

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

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?

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

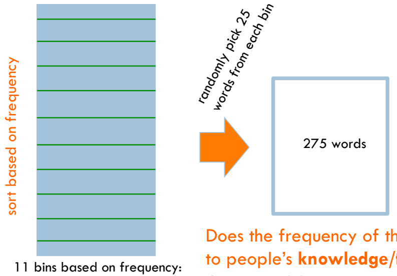
Proxy for "how often you see a word":

Frequency on the web!

Google bing Y!

Validating frequency hypothesis

Google unigrams: ~13M

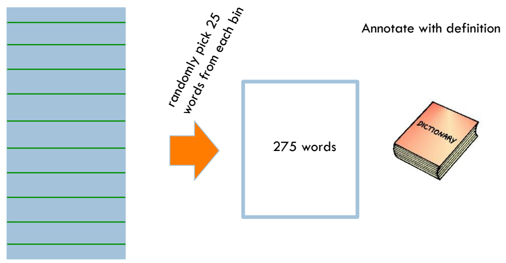


11 bins based on frequency: 1%, 10%, 20%, …, 100%

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

Validating frequency hypothesis

Google unigrams: ~13M



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

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

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

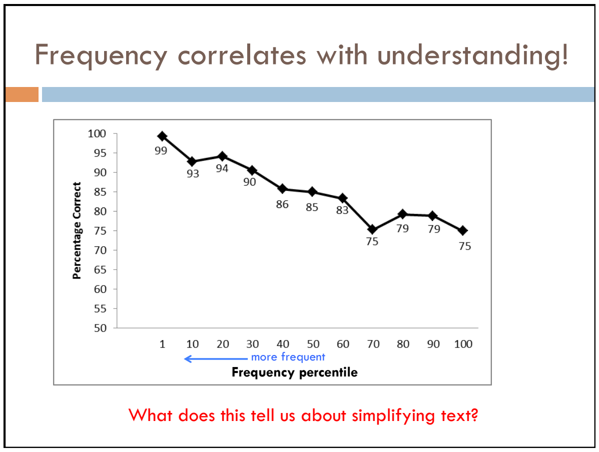
d) a woman policeman

random definitions from other words in data set

Study participants

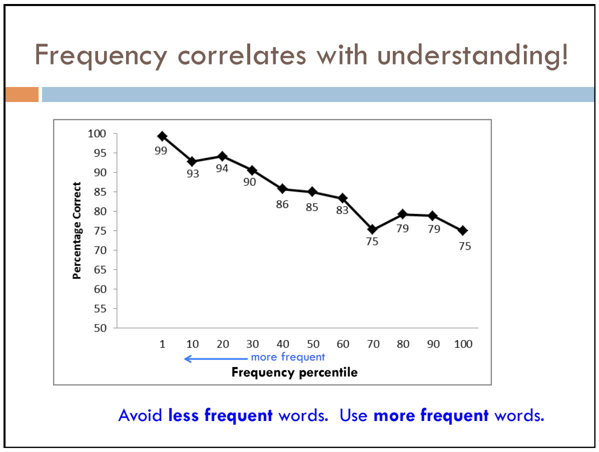
amazon
mechanical turk
beta

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



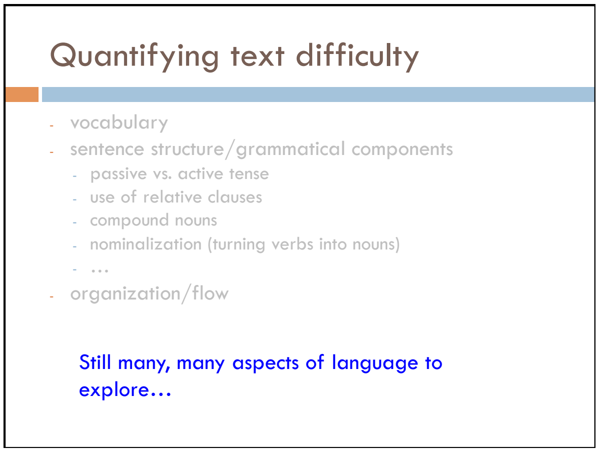
Frequency correlates with understanding!

What does this tell us about simplifying text?



Frequency correlates with understanding!

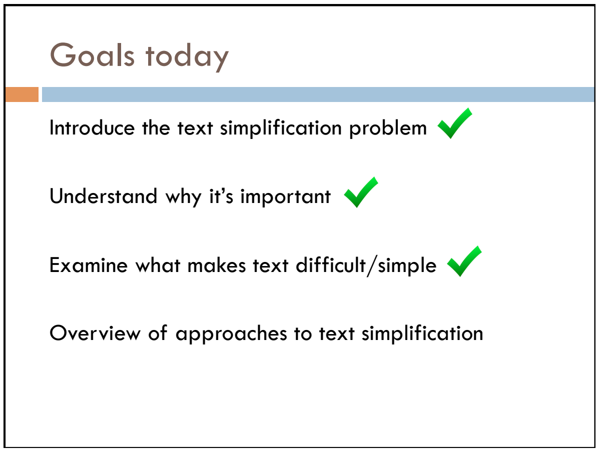
Avoid **less frequent** words. Use **more frequent** words.



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…



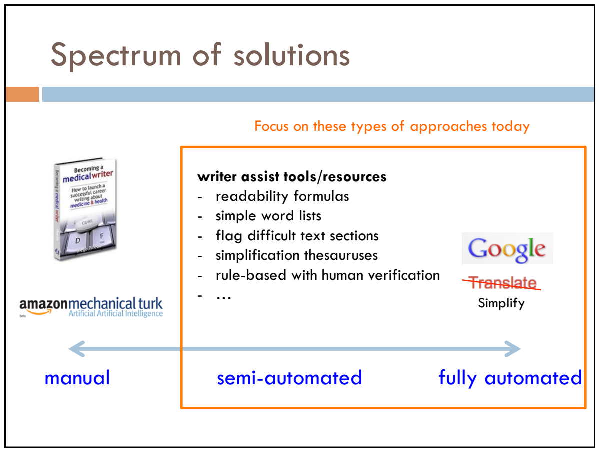
Goals today

Introduce the text simplification problem ✓

Understand why it's important ✓

Examine what makes text difficult/simple ✓

Overview of approaches to text simplification



Spectrum of solutions

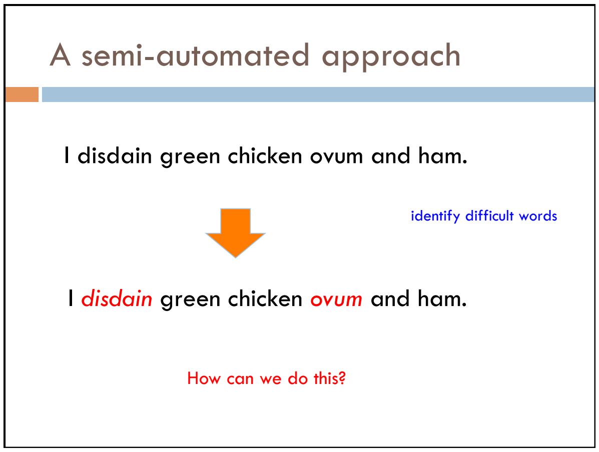
Focus on these types of approaches today

writer assist tools/resources

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

amazonmechanical turk
Artificial Artificial Intelligence
Google
~~Translate~~
Simplify

manual ↔ semi-automated ↔ fully automated



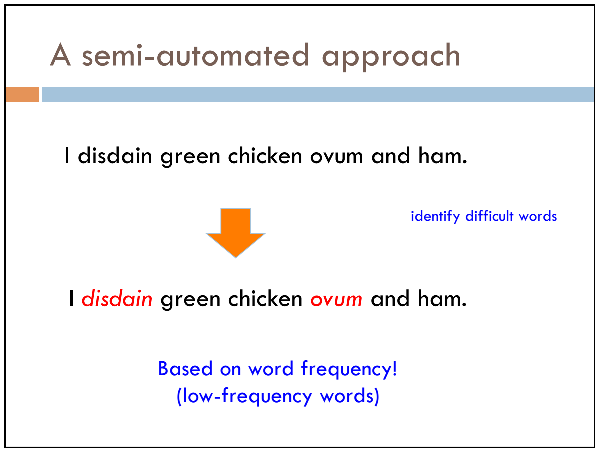
A semi-automated approach

I disdain green chicken ovum and ham.

identify difficult words

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

How can we do this?



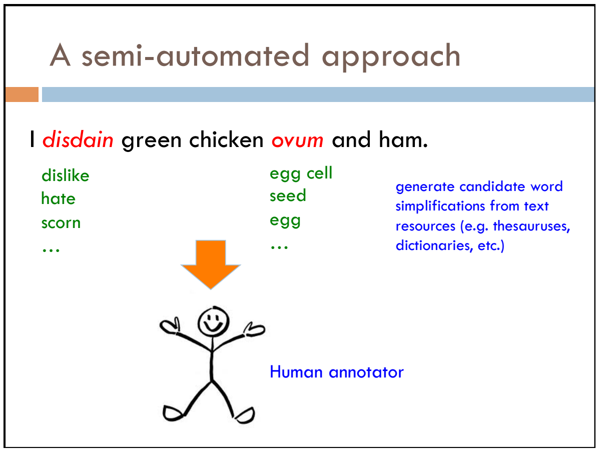
A semi-automated approach

I disdain green chicken ovum and ham.

identify difficult words

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

Based on word frequency!
(low-frequency words)



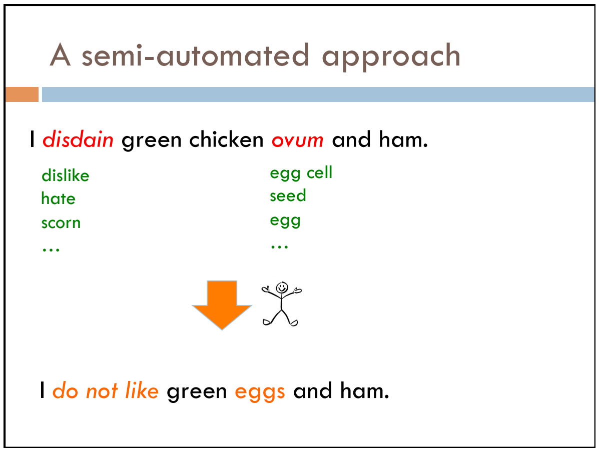
A semi-automated approach

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

disdain	ovum
dislike	egg cell
hate	seed
scorn	egg
…	…

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

Human annotator



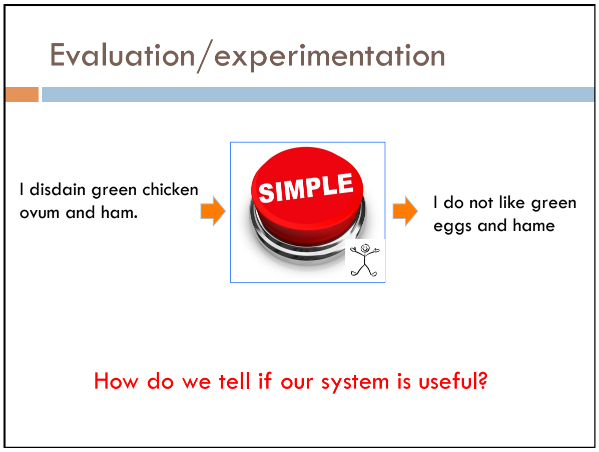
A semi-automated approach

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

dislike
hate
scorn
…

egg cell
seed
egg
…

I *do not like* green *eggs* and ham.

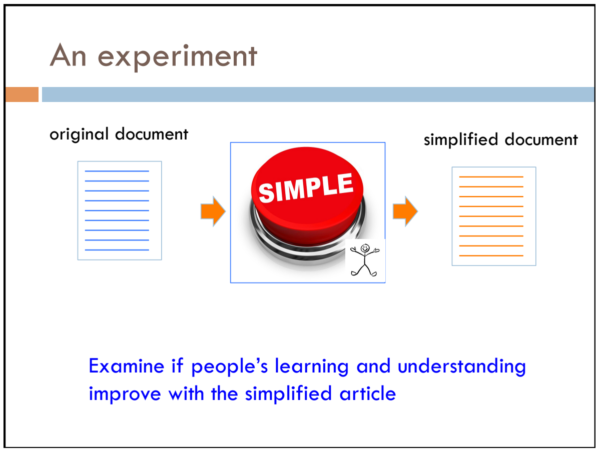


Evaluation/experimentation

I disdain green chicken ovum and ham.

I do not like green eggs and hame

How do we tell if our system is useful?

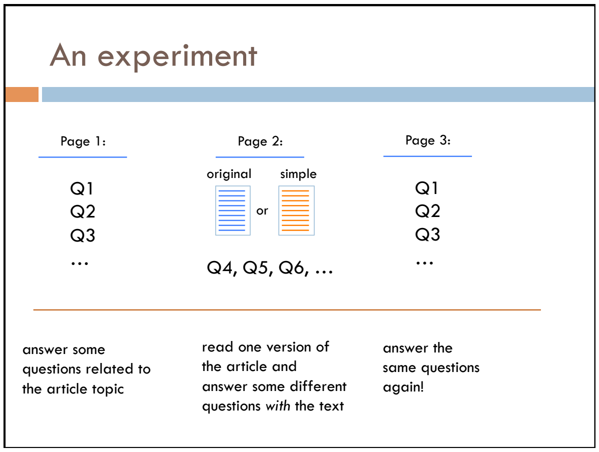


An experiment

original document

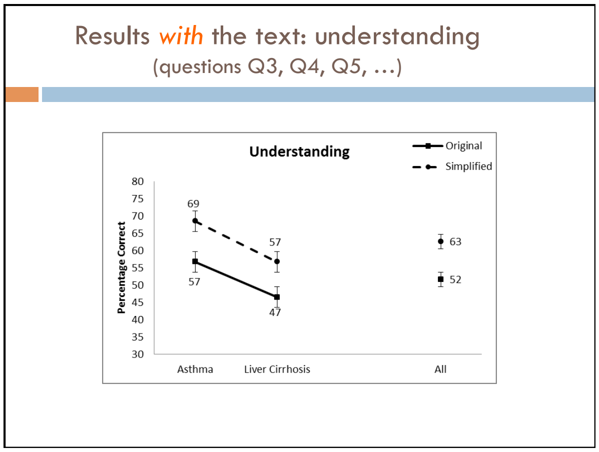
simplified document

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

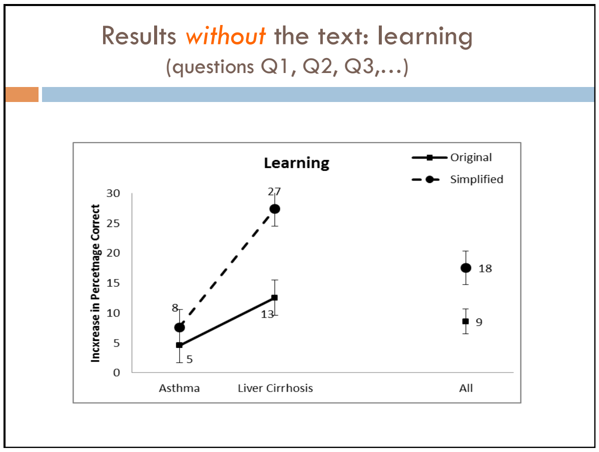


An experiment

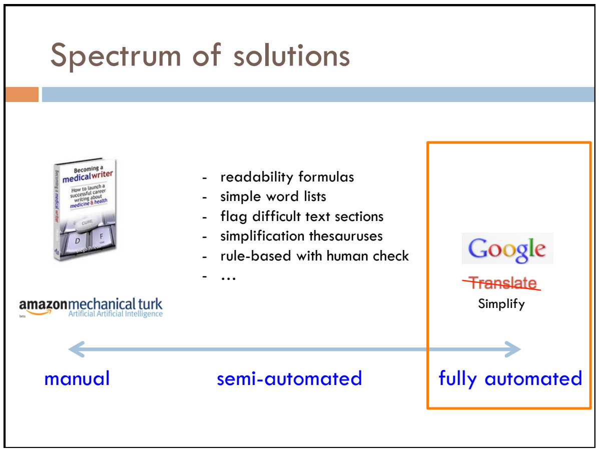
Page 1:	Page 2:	Page 3:
Q1 Q2 Q3 …	original or simple Q4, Q5, Q6, …	Q1 Q2 Q3 …
answer some questions related to the article topic	read one version of the article and answer some different questions *with* the text	answer the same questions again!



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

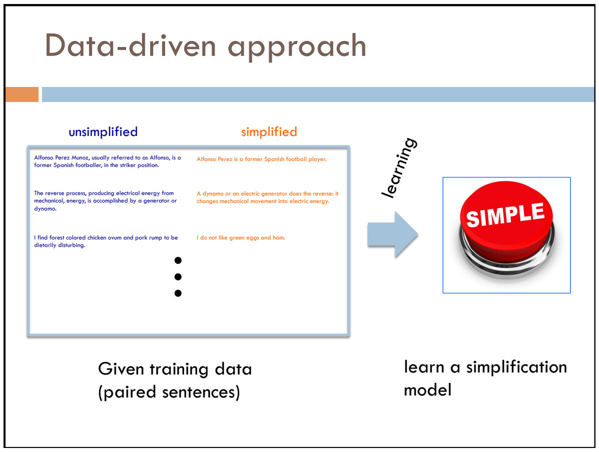


Results *without* the text: learning
(questions Q1, Q2, Q3,…)



Spectrum of solutions

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



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

Collecting simplification data

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

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."

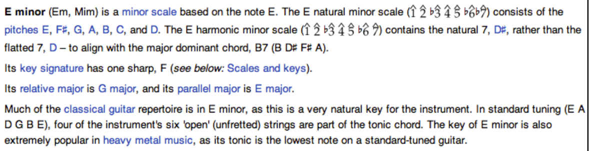
Wikipedia for text simplification

"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."

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.

From aligned documents to aligned sentences

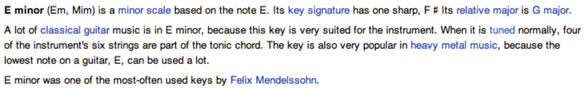


E minor (Em, Mim) is a minor scale based on the note E. The E natural minor scale ($\hat{1}\ \hat{2}\ \flat\hat{3}\ \hat{4}\ \hat{5}\ \flat\hat{6}\flat\hat{7}$) consists of the pitches E, F♯, G, A, B, C, and D. The E harmonic minor scale ($\hat{1}\ \hat{2}\ \flat\hat{3}\ \hat{4}\ \hat{5}\ \flat\hat{6}\ \hat{7}$) 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.

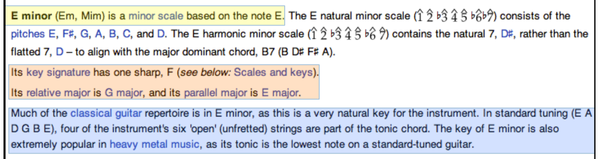


E minor (Em, Mim) is a minor scale based on the note E. Its key signature has one sharp, F ♯ Its relative major is G major.

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.

From aligned documents to aligned sentences

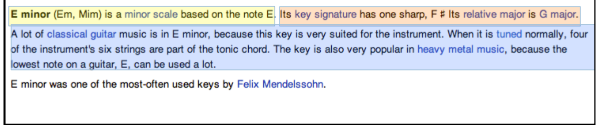


E minor (Em, Mim) is a minor scale based on the note E. The E natural minor scale ($\hat{1}\ \hat{2}\ \flat\hat{3}\ \hat{4}\ \hat{5}\ \flat\hat{6}\flat\hat{7}$) consists of the pitches E, F♯, G, A, B, C, and D. The E harmonic minor scale ($\hat{1}\ \hat{2}\ \flat\hat{3}\ \hat{4}\ \hat{5}\ \flat\hat{6}\ \hat{7}$) 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.

Wikipedia for text simplification

4.4M articles

97K articles

unsimplified

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.

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

167K aligned sentence pairs

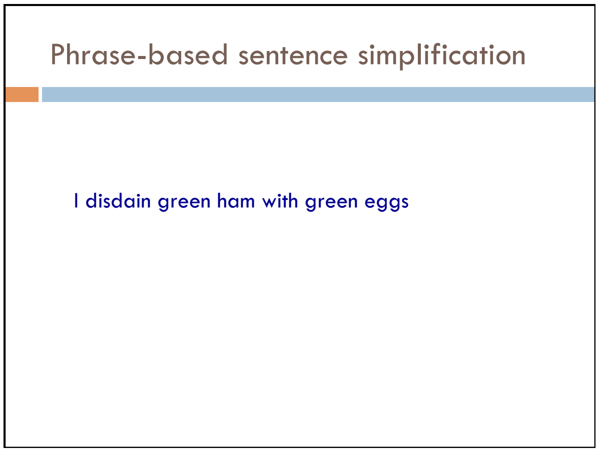
simplified

Alfonso Perez is a former Spanish football player.

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

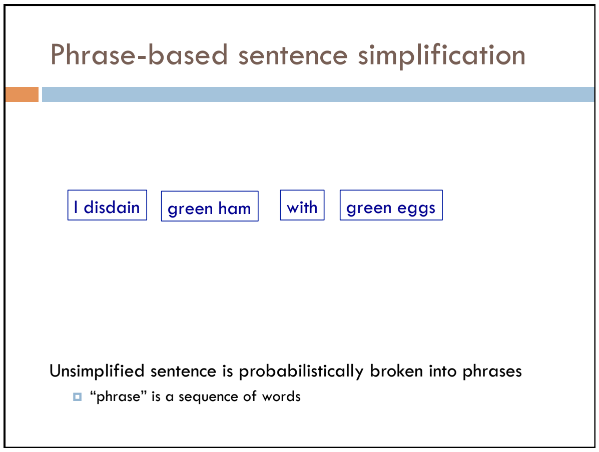
I do not like green eggs and ham.

Simplification approaches



Phrase-based sentence simplification

I disdain green ham with green eggs

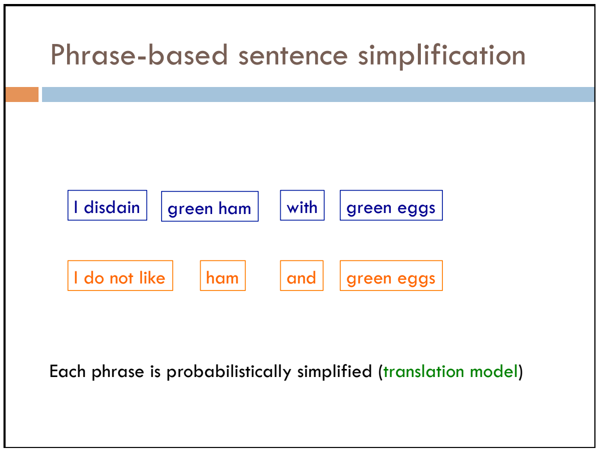


Phrase-based sentence simplification

I disdain | green ham | with | green eggs

Unsimplified sentence is probabilistically broken into phrases

- "phrase" is a sequence of words

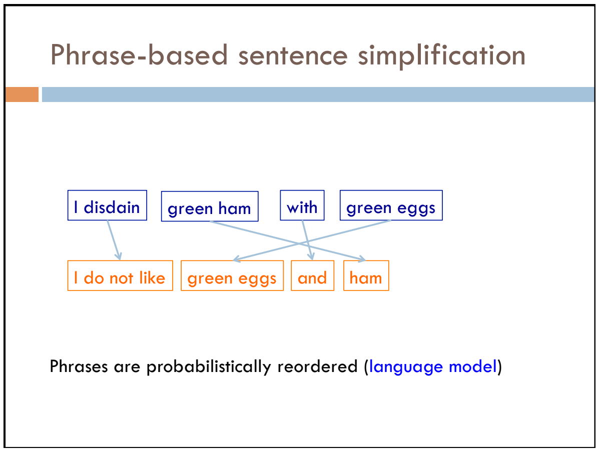


Phrase-based sentence simplification

I disdain | green ham | with | green eggs

I do not like | ham | and | green eggs

Each phrase is probabilistically simplified (translation model)

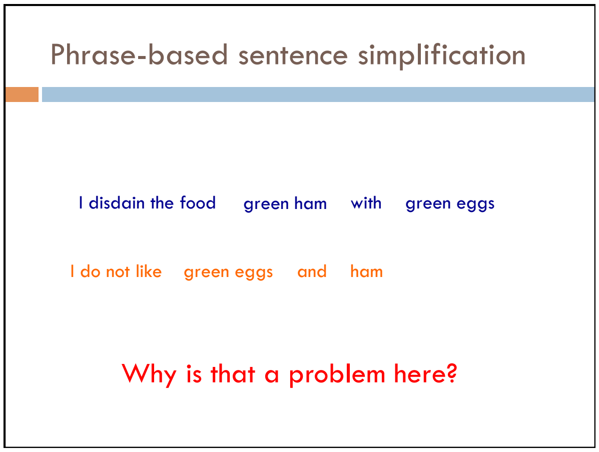


Phrase-based sentence simplification

I disdain | green ham | with | green eggs

I do not like | green eggs | and | ham

Phrases are probabilistically reordered (language model)

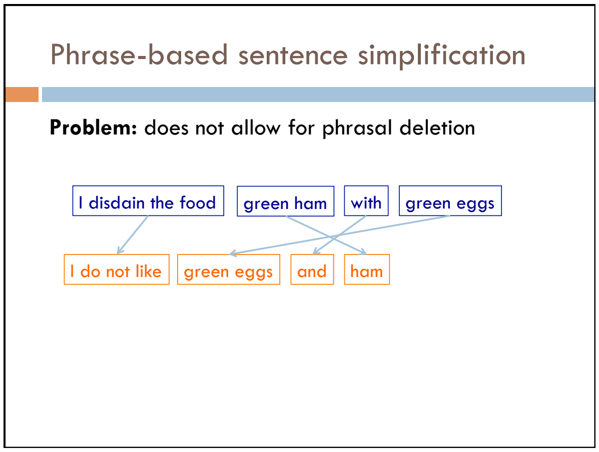


Phrase-based sentence simplification

I disdain the food green ham with green eggs

I do not like green eggs and ham

Why is that a problem here?

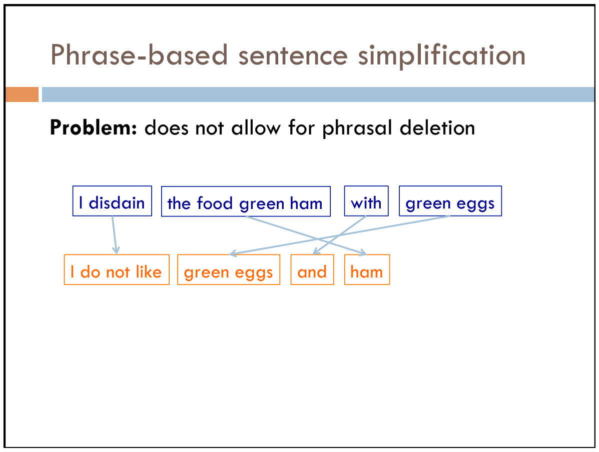


Phrase-based sentence simplification

Problem: does not allow for phrasal deletion

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

I do not like | green eggs | and | ham

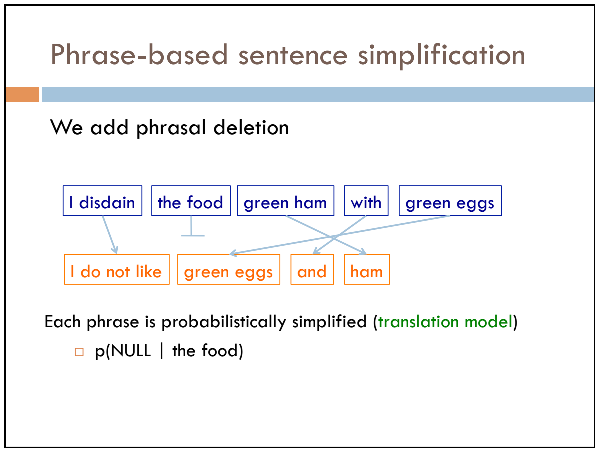


Phrase-based sentence simplification

Problem: does not allow for phrasal deletion

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

I do not like | green eggs | and | ham



Phrase-based sentence simplification

We add phrasal deletion

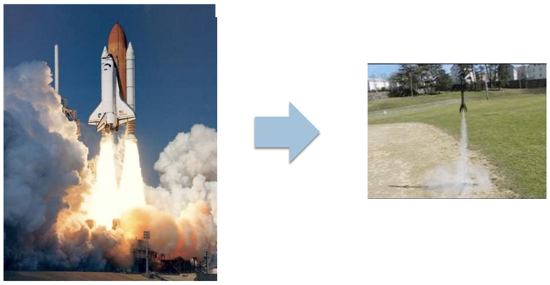
I disdain | the food | green ham | with | green eggs

I do not like | green eggs | and | ham

Each phrase is probabilistically simplified (translation model)

- □ p(NULL | the food)

Phrase-based performance



Experiments

5 approaches

- **none** – output the unsimplified sentence
- **K&M** – noisy channel sentence compression with PCFGs
  - Only allows for deletion
  - Uses syntactic information
- **T3** – Cohn and Lapata (2009)
  - All transformation operations
  - Uses syntactic information
  - Only been previously employed for sentence compression
- **Moses** – noisy channel, phrase-based without deletion
- **Moses+Del** – with delection

Evaluation

3 measures

- BLEU (0-1.0)
  - weighted mean of n-gram precisions
  - brevity penalty to avoid overly short results
- word-F1 (0-1.0)
  - F1 measure of system word occurrences
  - F1 combines precision and recall into one measure
- Simple String Accuracy - SSA (0-1.0)
  - length normalized edit distance

machine translation

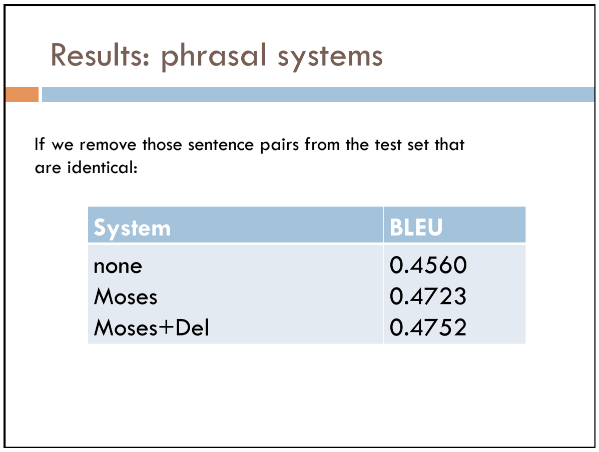
sentence compression

Results

System	BLEU	word-F1	SSA
none	0.5937	0.5967	0.6179
K&M	0.4352	0.4352	0.4871
T3*	0.2437	0.2190	0.3651
Moses	0.5987	0.6076	0.6224
Moses+Del	0.6046	0.6149	0.6259

All results are significantly different at the p=0.01 level

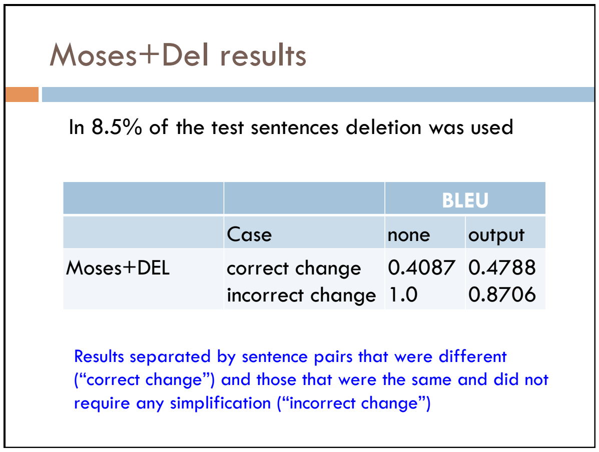
* T3 was only trained on 30K sentence pairs



Results: phrasal systems

If we remove those sentence pairs from the test set that are identical:

System	BLEU
none	0.4560
Moses	0.4723
Moses+Del	0.4752

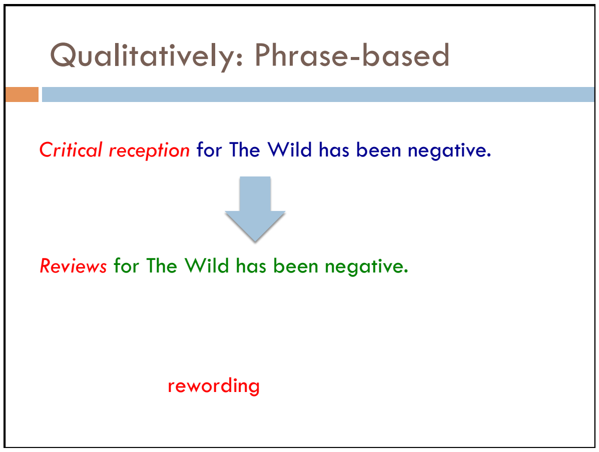


Moses+Del results

In 8.5% of the test sentences deletion was used

		BLEU	
	Case	none	output
Moses+DEL	correct change	0.4087	0.4788
	incorrect change	1.0	0.8706

Results separated by sentence pairs that were different ("correct change") and those that were the same and did not require any simplification ("incorrect change")



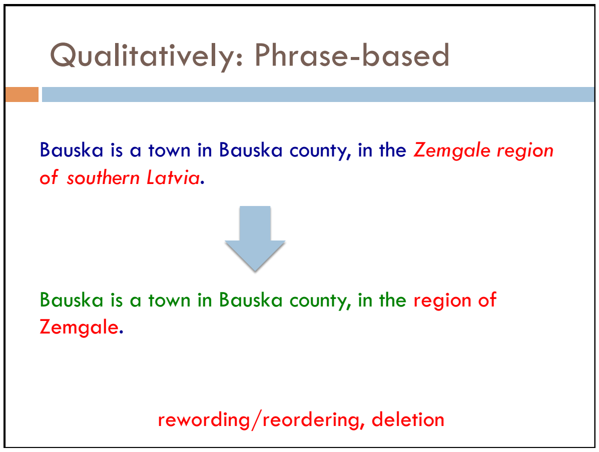
Qualitatively: Phrase-based

Critical reception for The Wild has been negative.

↓

Reviews for The Wild has been negative.

rewording



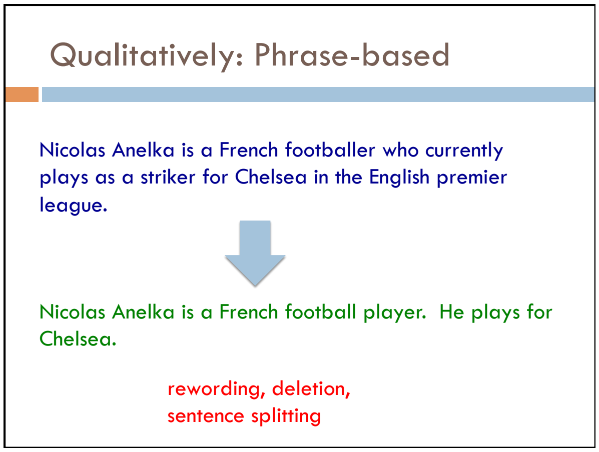
Qualitatively: Phrase-based

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

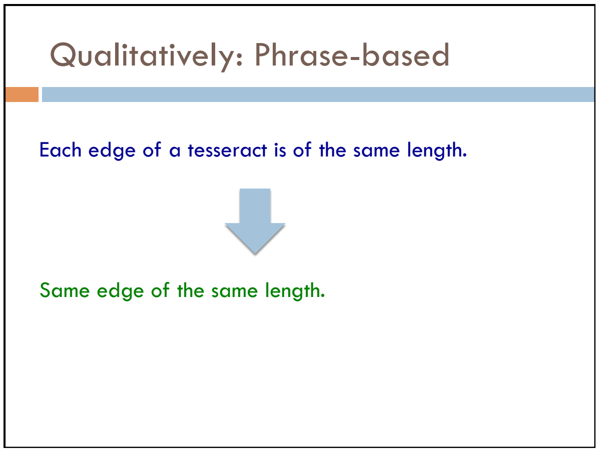


Qualitatively: Phrase-based

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.

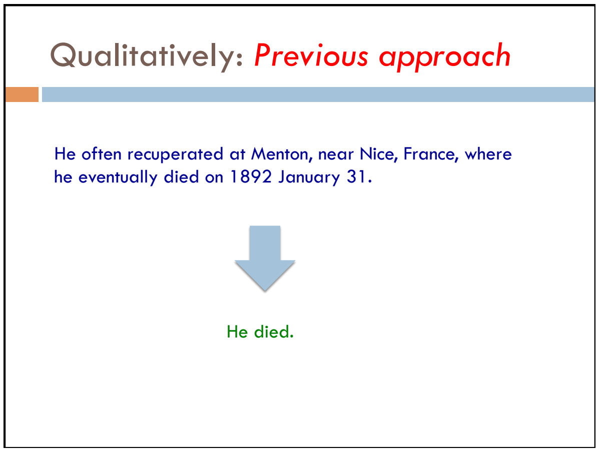
rewording, deletion,
sentence splitting



Qualitatively: Phrase-based

Each edge of a tesseract is of the same length.

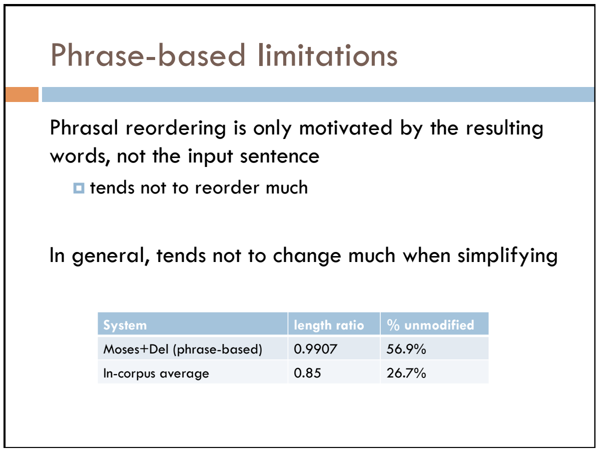
Same edge of the same length.



Qualitatively: *Previous approach*

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

He died.



Phrase-based limitations

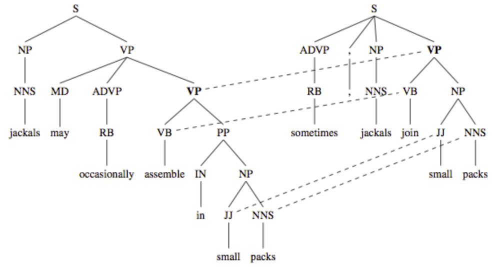
Phrasal reordering is only motivated by the resulting words, not the input sentence

- tends not to reorder much

In general, tends not to change much when simplifying

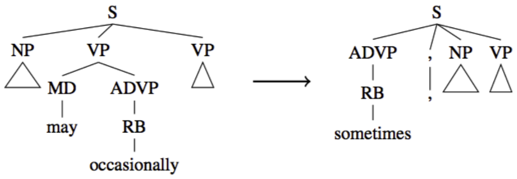
System	length ratio	% unmodified
Moses+Del (phrase-based)	0.9907	56.9%
In-corpus average	0.85	26.7%

Syntax-based approach



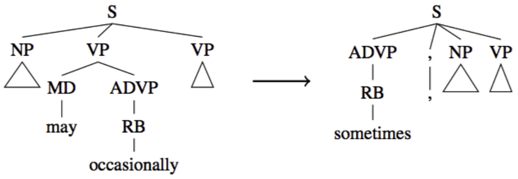
Rather than operating on phrases, operate on grammar trees

Learn probabilistic, syntax-based rules



They may occasionally eat ➡ sometimes, they eat

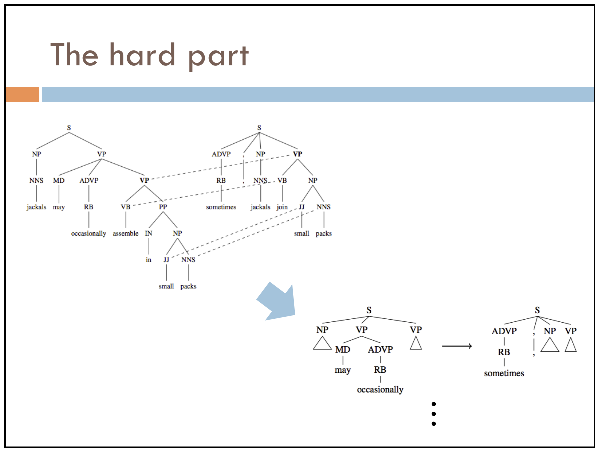
Learn probabilistic, syntax-based rules



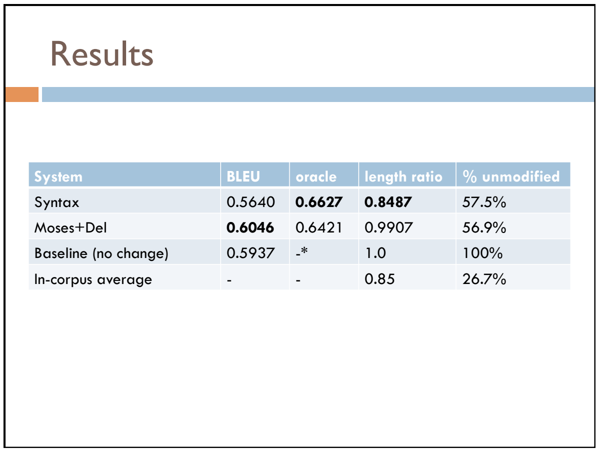
The scary cats from the park may occasionally walk around on two legs ➡ sometimes, the scary cats from the park walk around on two legs

An aside

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

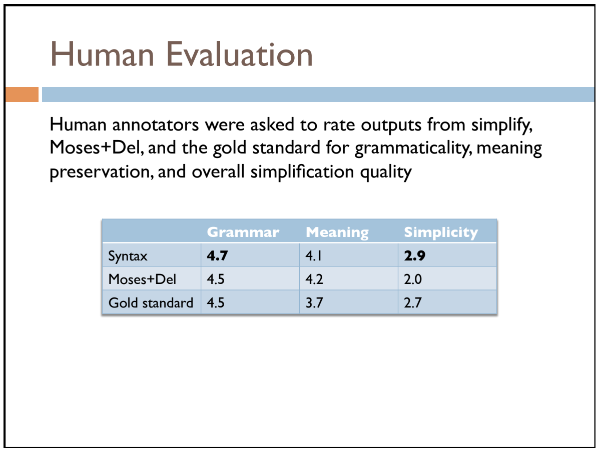


The hard part



Results

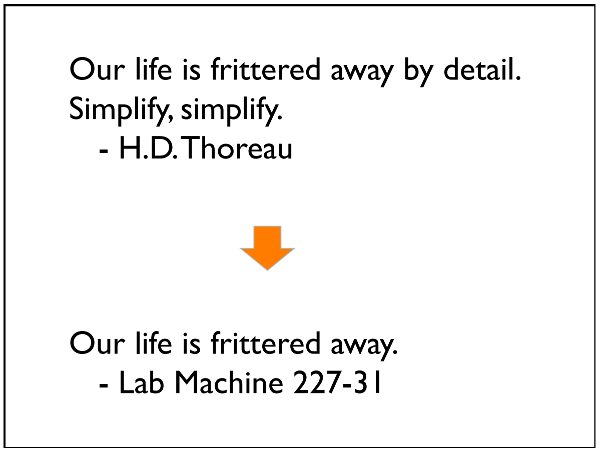
System	BLEU	oracle	length ratio	% unmodified
Syntax	0.5640	**0.6627**	**0.8487**	57.5%
Moses+Del	**0.6046**	0.6421	0.9907	56.9%
Baseline (no change)	0.5937	-*	1.0	100%
In-corpus average	-	-	0.85	26.7%



Human Evaluation

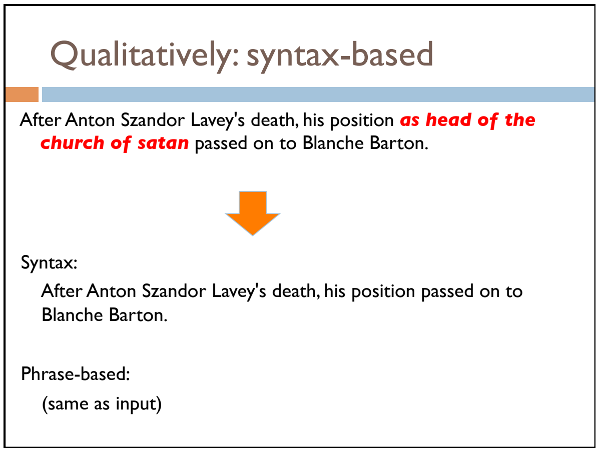
Human annotators were asked to rate outputs from simplify, Moses+Del, and the gold standard for grammaticality, meaning preservation, and overall simplification quality

	Grammar	Meaning	Simplicity
Syntax	**4.7**	4.1	**2.9**
Moses+Del	4.5	4.2	2.0
Gold standard	4.5	3.7	2.7



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

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



Qualitatively: syntax-based

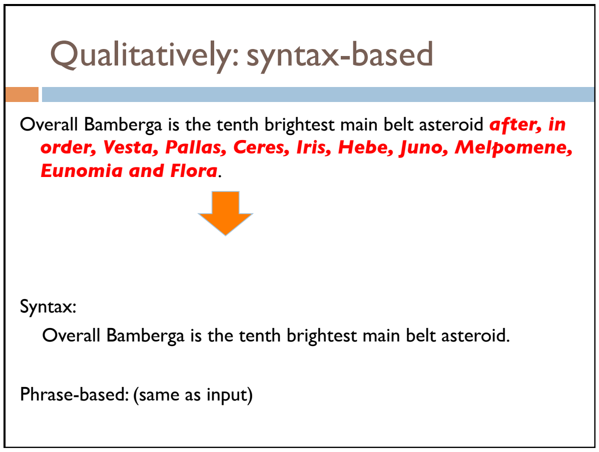
After Anton Szandor Lavey's death, his position ***as head of the church of satan*** passed on to Blanche Barton.

Syntax:

After Anton Szandor Lavey's death, his position passed on to Blanche Barton.

Phrase-based:

(same as input)



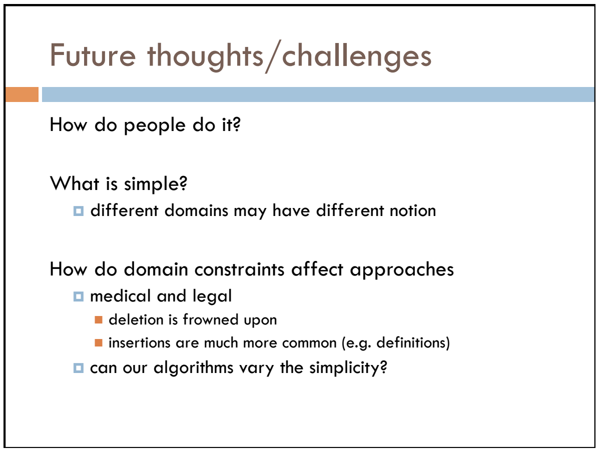
Qualitatively: syntax-based

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)



Future thoughts/challenges

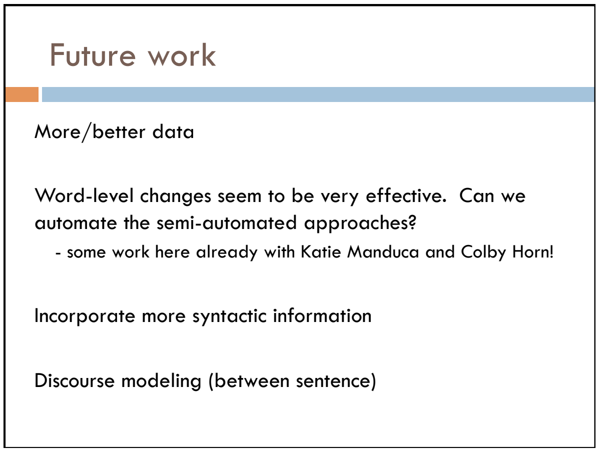
How do people do it?

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?



Future work

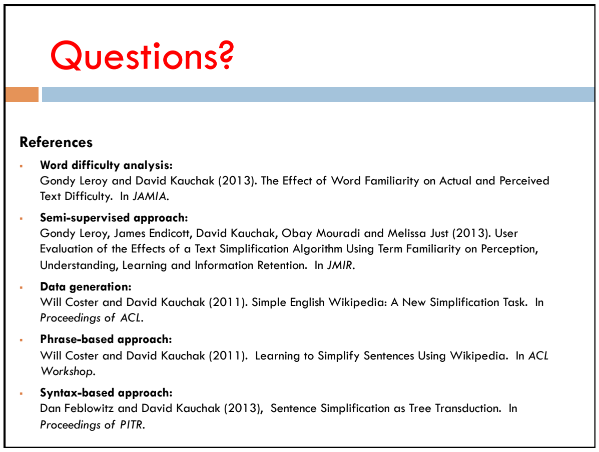
More/better data

Word-level changes seem to be very effective. Can we automate the semi-automated approaches?

- some work here already with Katie Manduca and Colby Horn!

Incorporate more syntactic information

Discourse modeling (between sentence)



Questions?

References

- **Word difficulty analysis:**
  Gondy Leroy and David Kauchak (2013). The Effect of Word Familiarity on Actual and Perceived Text Difficulty. In *JAMIA*.
- **Semi-supervised approach:**
  Gondy Leroy, James Endicott, David Kauchak, Obay Mouradi and Melissa Just (2013). User Evaluation of the Effects of a Text Simplification Algorithm Using Term Familiarity on Perception, Understanding, Learning and Information Retention. In *JMIR*.
- **Data generation:**
  Will Coster and David Kauchak (2011). Simple English Wikipedia: A New Simplification Task. In *Proceedings of ACL*.
- **Phrase-based approach:**
  Will Coster and David Kauchak (2011). Learning to Simplify Sentences Using Wikipedia. In *ACL Workshop*.
- **Syntax-based approach:**
  Dan Feblowitz and David Kauchak (2013), Sentence Simplification as Tree Transduction. In *Proceedings of PITR*.