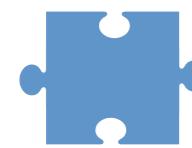




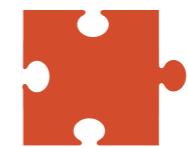
# Problems in Current Text Simplification Research

<b>Wei Xu</b>	UPenn
Chris Callison-Burch	UPenn
Courtney Napolis	JHU

# What is Text Simplification

 **paraphrasing**

 **deletion**

 **splitting**

**INPUT** *Applesauce is a puree made of apples.*

**OUT-1** *Applesauce is a soft paste.*

**OUT-2** *Applesauce is a paste. It is made of apples.*



- for children, disabled, non-native speakers ...
- for other NLP tasks (MT, summarization ...)

# Goal of Text Simplification



grammaticality



meaning  
preservation



simplicity

INPUT

*Applesauce is a puree made of apples.*

Human Evaluation

OUT-1

*Applesauce is a soft paste.*

5

4

5

OUT-2

*Applesauce is a paste. It is made of apples.*

5

5

4

(no reliable automatic evaluation yet)

# Brief History of Sentence Simplification

## rule-based

- 1997 Chandrasekar & Srinivas
- 1999 Dras (PhD thesis)
- 2000 Carroll, Minnen, Pearce, Canning, Devlin
- 2002 Canning (PhD thesis)
- 2004 Siddharthan (PhD thesis)

# Brief History of Sentence Simplification

rule-based

*Simple English*  
**WIKIPEDIA**



- 1997 Chandrasekar & Srinivas
- 1999 Dras (PhD thesis)
- 2000 Carroll, Minnen, Pearce, Canning, Devlin
- 2002 Canning (PhD thesis)
- 2004 Siddharthan (PhD thesis)
- 2010 Zhu, Bernhard, Gurevych**

# Parallel Wikipedia Corpus



The screenshot shows the English Wikipedia page for "Apple sauce". The title is "Apple sauce" and it is described as "From Wikipedia, the free encyclopedia". Below the title are two images: "Commercially processed apple sauce" (a smooth purée) and "A chunky German apple sauce" (pieces of apples). The main text starts with: "Apple sauce or applesauce is a purée made of apples. It can be made with peeled and/or unpeel". A blue arrow points from this text to the corresponding section on the Simple English Wikipedia page.

Article Talk Read Edit More Search

## Apple sauce

From Wikipedia, the free encyclopedia

Main page Contents Featured content Current events Random article Donate to Wikipedia Wikipedia store

Interaction Help About Wikipedia Community portal

**Apple sauce or applesauce is a purée made of apples. It can be made with peeled and/or unpeel**

What links here Related changes Upload file Special pages Permanent link



The screenshot shows the Simple English Wikipedia page for "Applesauce". The title is "Applesauce" and it is described as "From Wikipedia, the free encyclopedia". The main text starts with: "Applesauce (or apple sauce) is a sauce that is made from stewed and mashed apples. Peeled or". A blue box highlights this sentence. A blue arrow points from the corresponding section on the English Wikipedia page to this highlighted text. The page also includes a sidebar with links like Main page, Simple start, Simple talk, New changes, Show any page, Help, Give to Wikipedia, Print/export, Make a book, Download as PDF, Page for printing, Tools, What links here, Related changes, Upload file, Special pages, Permanent link, Page information, and Wikidata item.

Page Talk Read Change More Search

## Applesauce

From Wikipedia, the free encyclopedia

Main page Simple start Simple talk New changes Show any page Help Give to Wikipedia Print/export Make a book Download as PDF Page for printing

Tools What links here Related changes Upload file Special pages Permanent link Page information Wikidata item

**Applesauce (or apple sauce) is a sauce that is made from stewed and mashed apples. Peeled or**

cinnamon can be used. Sugar or high fructose corn syrup is sometimes added to the applesauce to sweeten it. Applesauce can be fine or coarse textured, and can include large pieces of apple. It is easy to make at home, and it is also sold already made in supermarket or as a snack for children and people who have problems with eating solid food. Parents often feed babies applesauce because teeth are not needed to eat it. It is sometimes used to help fight diarrhoea, since it is high in dietary fibre.

A bowl of applesauce

# Brief History of Sentence Simplification

rule-based

*Simple English*  
**WIKIPEDIA**



- 1997 Chandrasekar & Srinivas
- 1999 Dras (PhD thesis)
- 2000 Carroll, Minnen, Pearce, Canning, Devlin
- 2002 Canning (PhD thesis)
- 2004 Siddharthan (PhD thesis)
- 2010 Zhu, Bernhard, Gurevych**

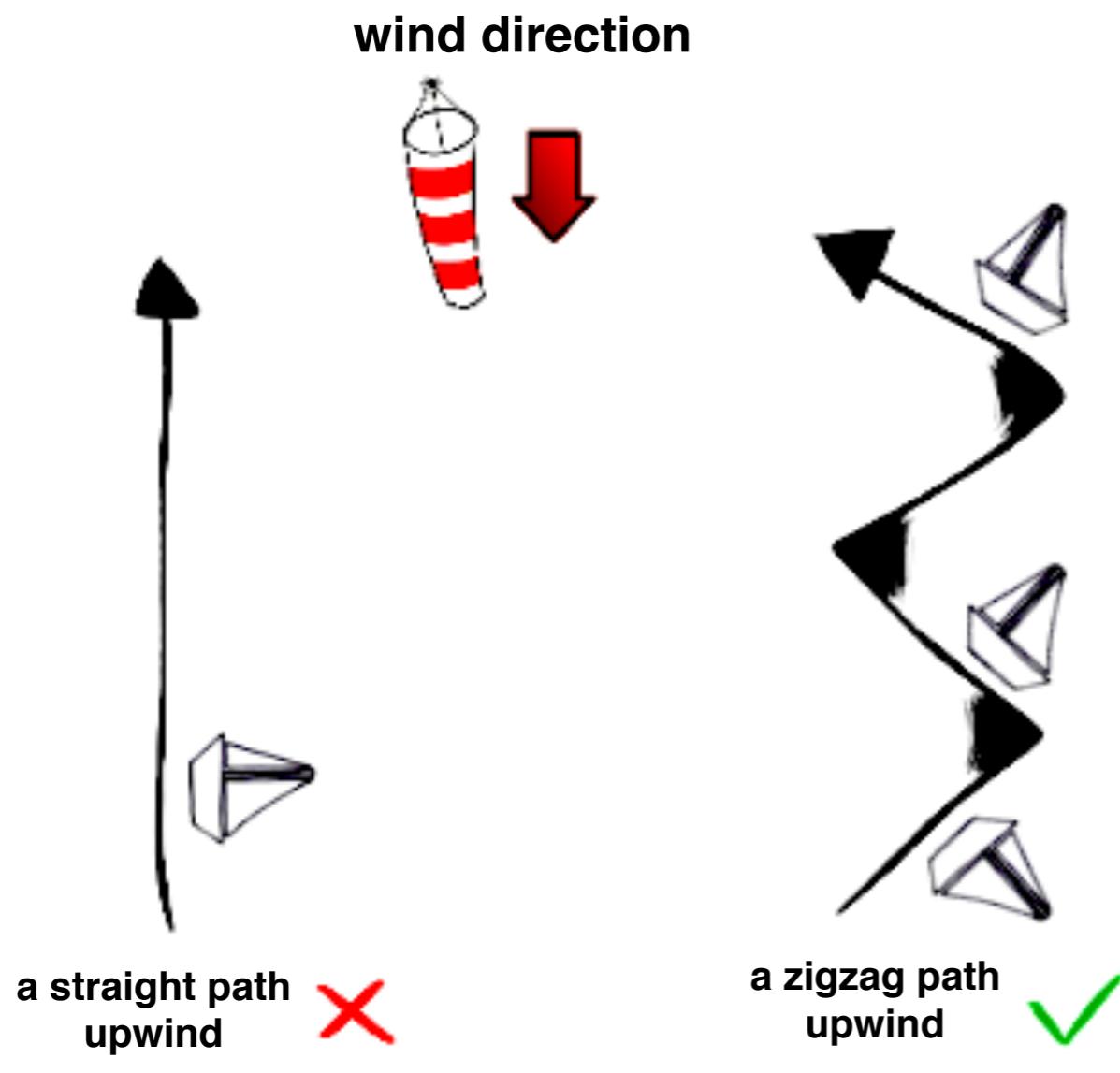
# Problems in Simplification Research

- State-of-the-art evaluation is suboptimal. But we have been doing this in the past 5 years\*.
- Simple Wikipedia data dominated in the past 5 years. But its quality was taken for granted. It limits the scope of research.

\* (Angrosh et al. 2014) tried comprehension quiz

# Why this is important?

## Simplification Breakthrough on Sea



- better understanding
- better **review**
- more diverse research
- better data and evaluation
- better model

“Recently, there have been several attempts at addressing the text simplification task as a monolingual translation problem ... However, they did not try to seek reasons for the success or the failure of their systems.”

— Štajner, Béchara, Saggion (2015)

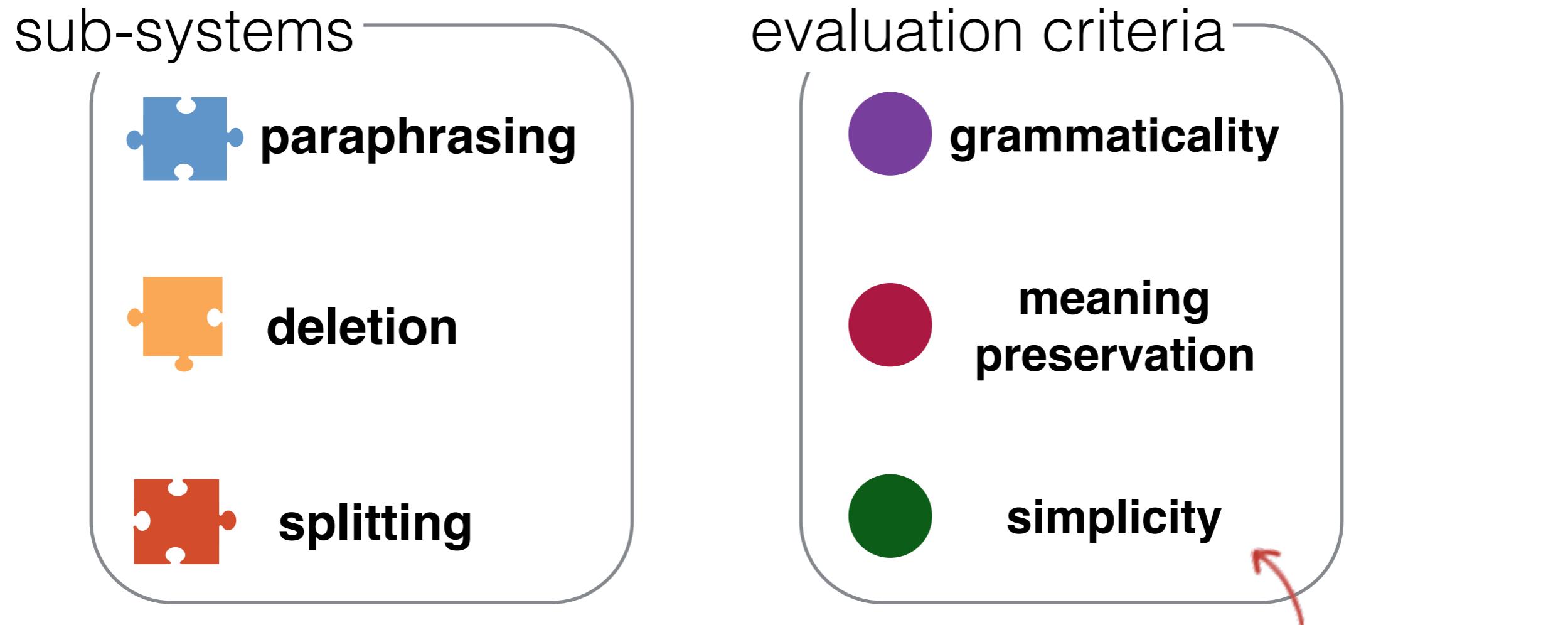
## WHY DID THIS HAPPEN?

- 1 “state-of-the-art” competition
- 2 not easy to do

# Opinion #1

Current evaluation **doesn't**  
tell us what's going on.

# System Comparability



We need more controlled evaluation:

- evaluate sub-tasks separately
- target specific audience (e.g. 10-12 year old)

**not easy to measure**

## Opinion #2

Simple Wikipedia is  
**not that simple**



*Simple English*  
**WIKIPEDIA**

Page

Talk

Read

Change

More ▾

Search

# Photolithography

From Wikipedia, the free encyclopedia

Main page

Photolithography is the combination of photography and lithography

**Microphotolithography** is the use of photolithography to transfer geometric shapes on a **photomask** to the surface of a **semiconductor wafer** for making **integrated circuits**.

“Specific questions that need addressing are :

... we need to better understand the quality of Simple English Wikipedia, a resource that has been used to train many SMT based simplification systems...”

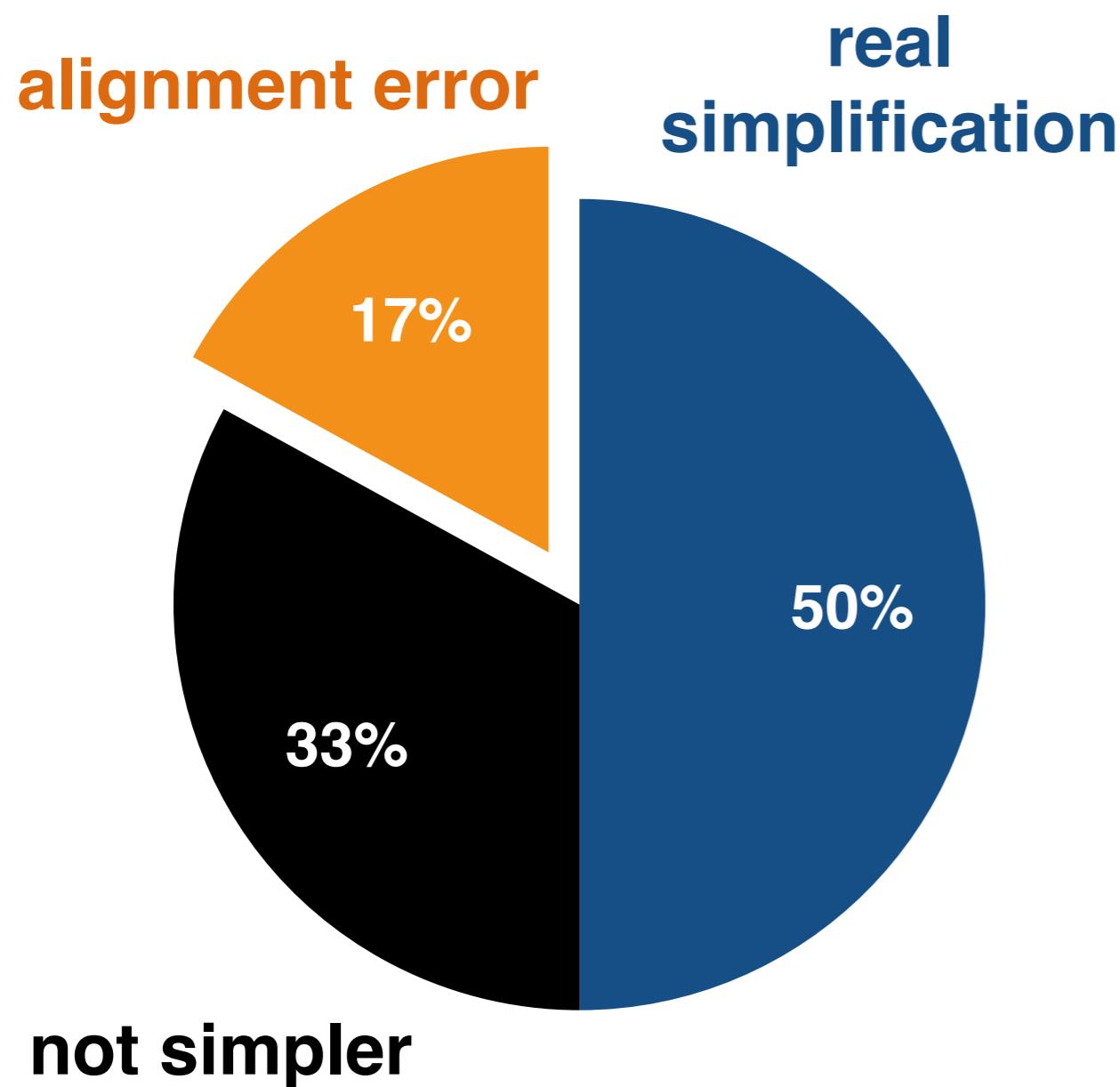
—— Advaith Siddharthan (2014 Survey)

## **WHAT'S NEW?**

We quantitatively and systematically answer this quest.

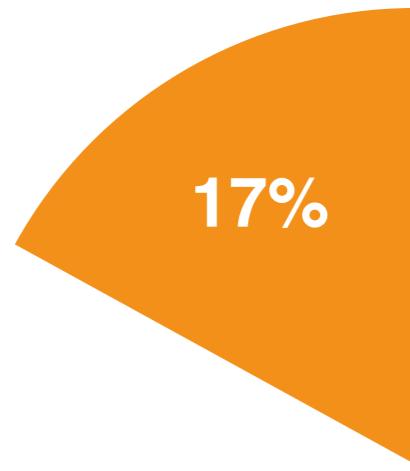


# Quality of Parallel Wikipedia Corpus\*



# Inaccuracy in Parallel Wikipedia Corpus\*

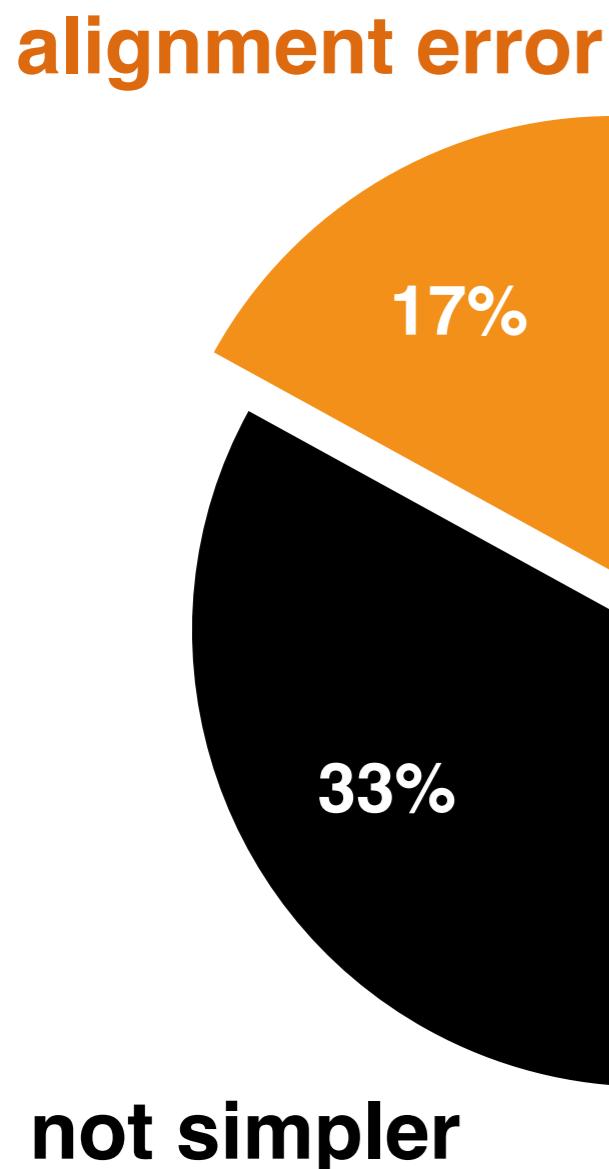
**alignment error**



(two sentences have different meaning)

Best automatic sentence alignment gets about **0.7 F1** score (Hwang et al. 2015)

# Inadequacy in Parallel Wikipedia Corpus\*

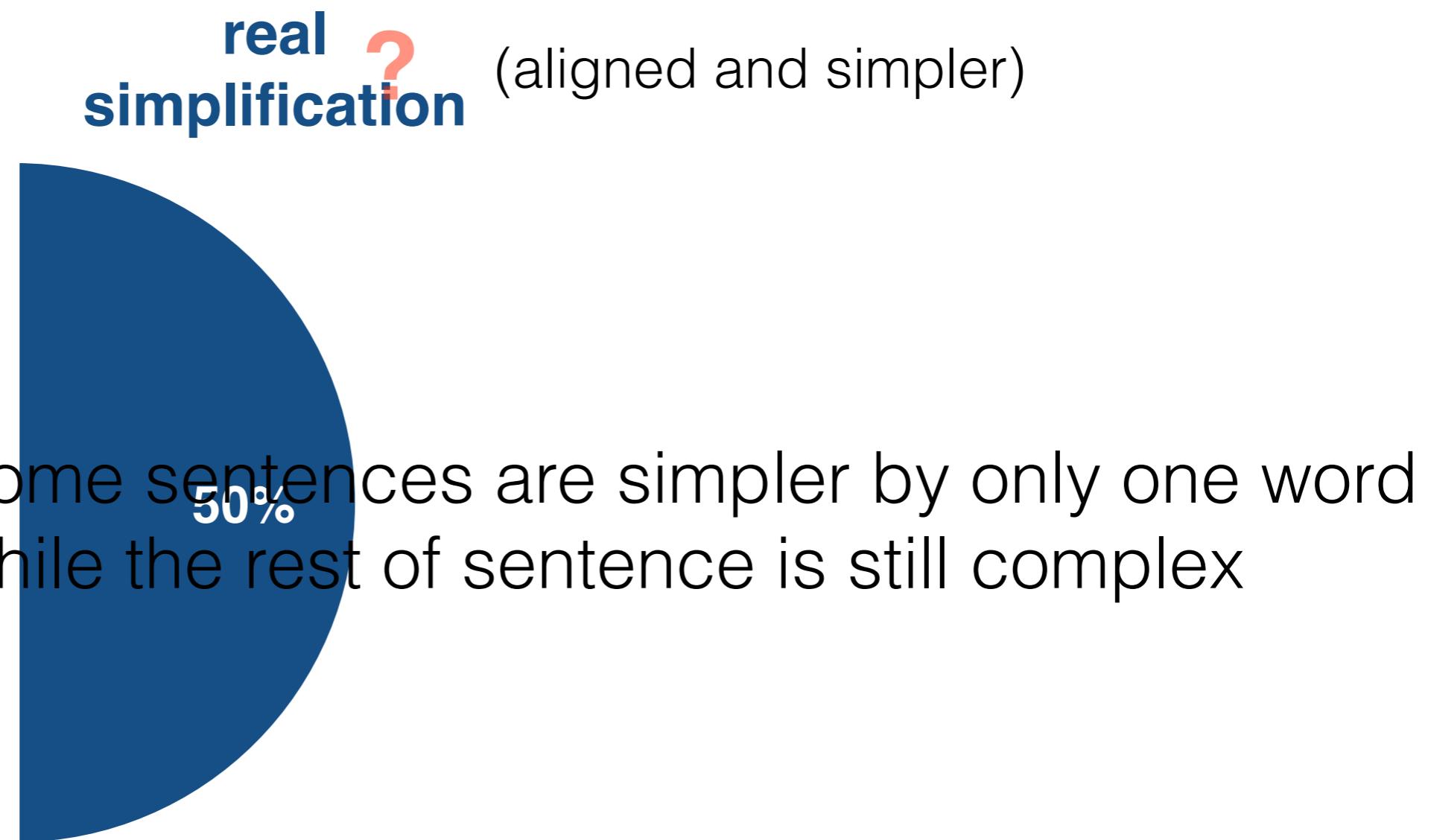


(two sentences have different meaning)

Best automatic sentence alignment gets about **0.7 F1** score (Hwang et al. 2015)

Sentences can have similar meaning but not simplification

# Inadequacy in Parallel Wikipedia Corpus\*



# Issues with Parallel Wikipedia Corpus

- suboptimal for estimating “translation” probabilities
- suboptimal for developing automatic metrics
- suboptimal for tuning MT system
- unsuitable for document-level simplification

# Opinion #3

New data **can** help

# Newsela Dataset

**NEWSELA**

WAR & PEACE SCIENCE KIDS MONEY HEALTH

SCIENCE 1738

## Archaeologist may have found remains of ancient Egyptian Queen Nefertiti

By Robert Gebelhoff, Washington Post. 08.17.15



The 3,330-year-old bust of Nefertiti sits in an exhibition in the Kulturforum in Berlin, Germany, March 1, 2005. Photo: AP/Herbert Knosowski

Nefertiti — she's an ancient Egyptian queen and the source of a fantastic mystery regarding the iconic remnants of long-lost royalty.

For decades, archaeologists have speculated on the location of the queen's remains, the last royal mummy missing from the dynasty of the famous King Tutankhamun, better known as King Tut. But now, an archaeologist claims that he has found her

**MAX**  
1140L  
960L  
720L  
420L  
 WRITE  
 QUIZ

**NEWSELA**

WAR & PEACE SCIENCE KIDS MONEY LAW HEALTH

SCIENCE 1738

## Mystery of ancient Egypt solved? Tomb of queen may be hidden near King Tut

By Washington Post, adapted by Newsela staff 08.17.15



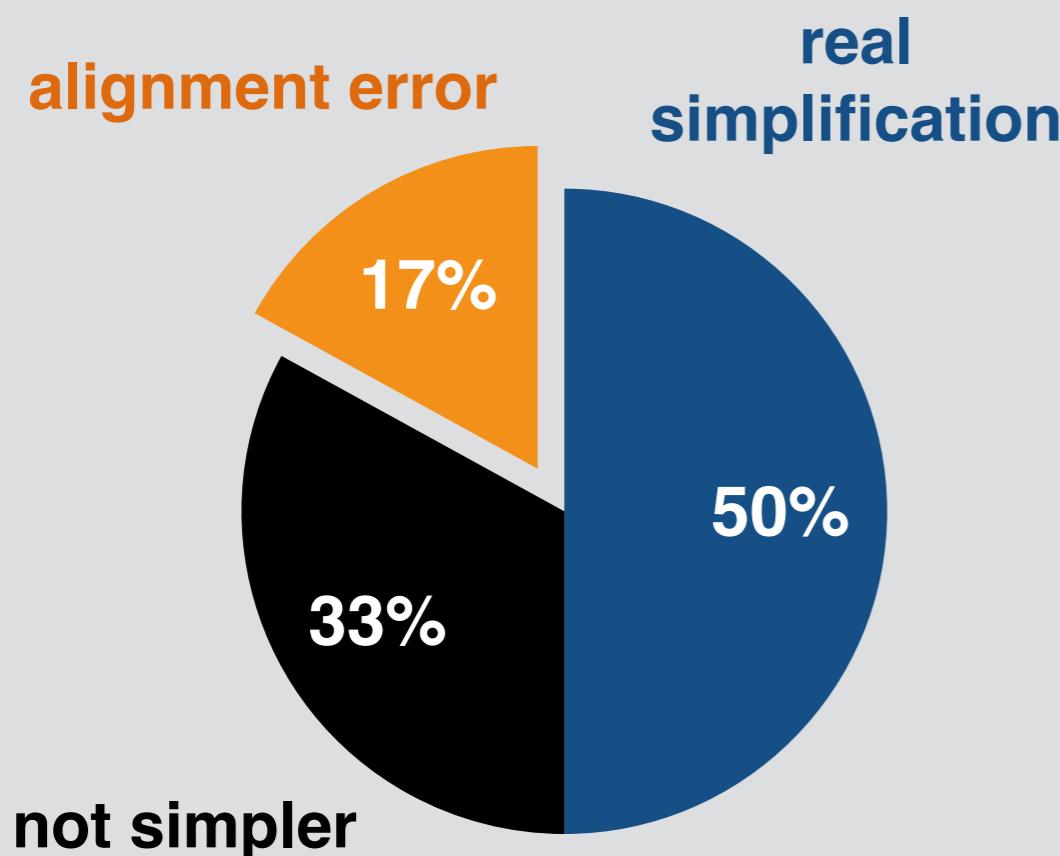
The 3,330-year-old bust of Nefertiti sits in an exhibition in the Kulturforum in Berlin, Germany, March 1, 2005. Photo: AP/Herbert Knosowski

The ancient Egyptian Queen Nefertiti has long been at the center of a mystery. For years, archaeologists have wondered where her tomb might be hidden. Nefertiti belonged to the family line of the famous King Tutankhamun, better known as King Tut. Indeed, some believe she was Tut's mother. While the other royals in her line are

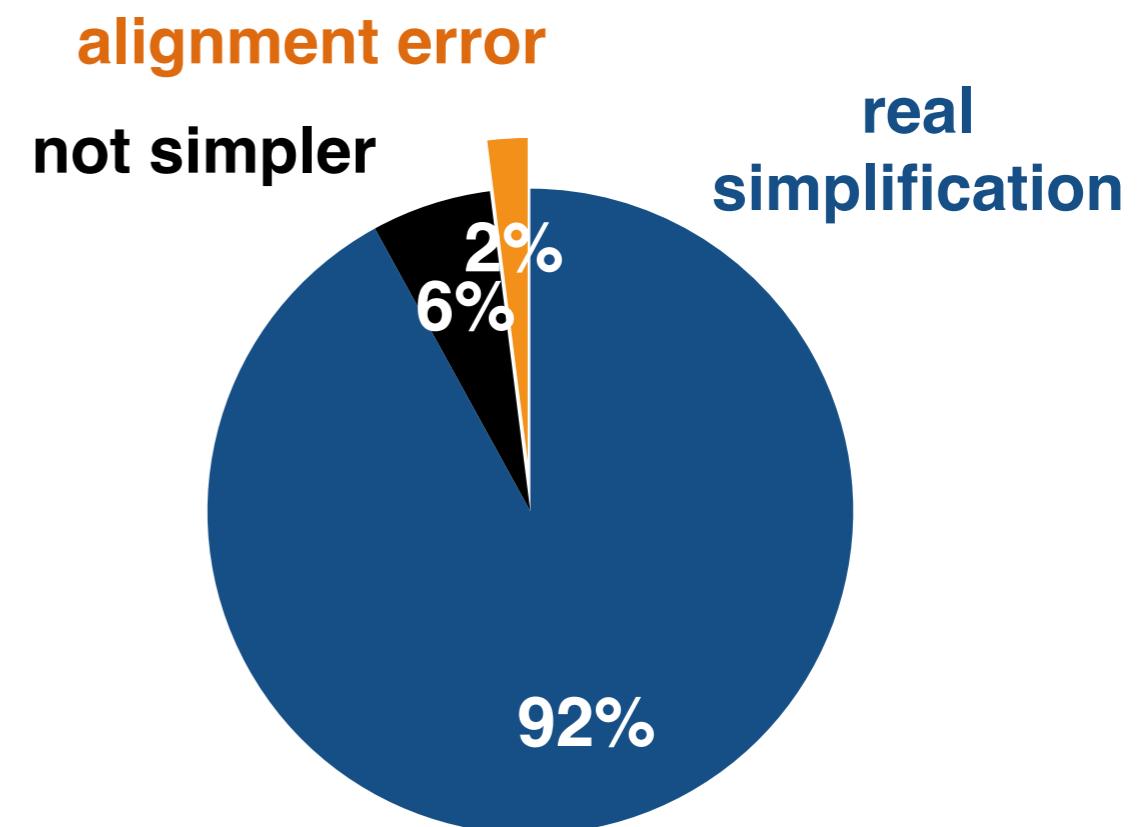
1140L  
960L  
720L  
420L  
 WRITE  
 QUIZ

every article at **5 levels** of simplification  
written by trained editors, comes with comprehension quizzes

# Wikipedia\*



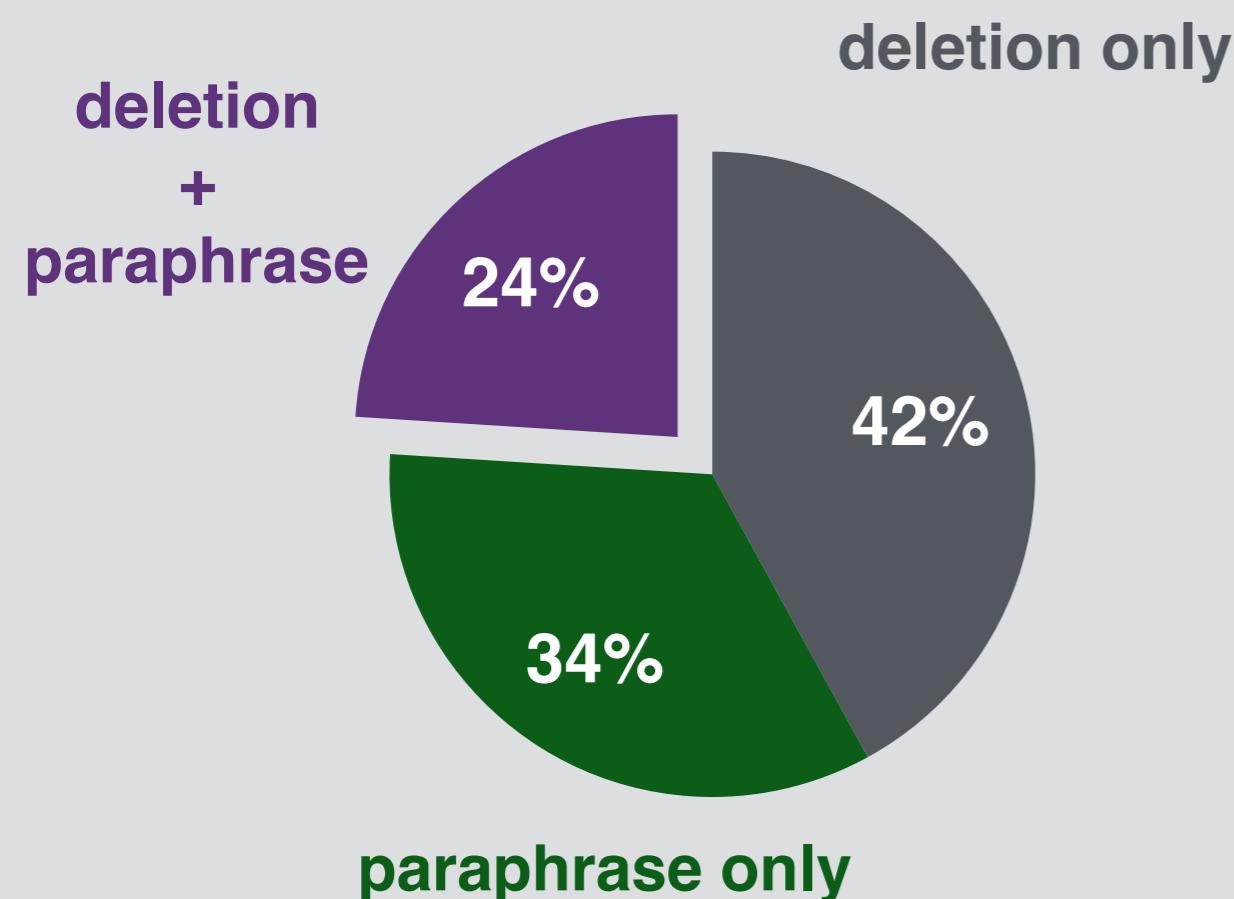
# Newsela



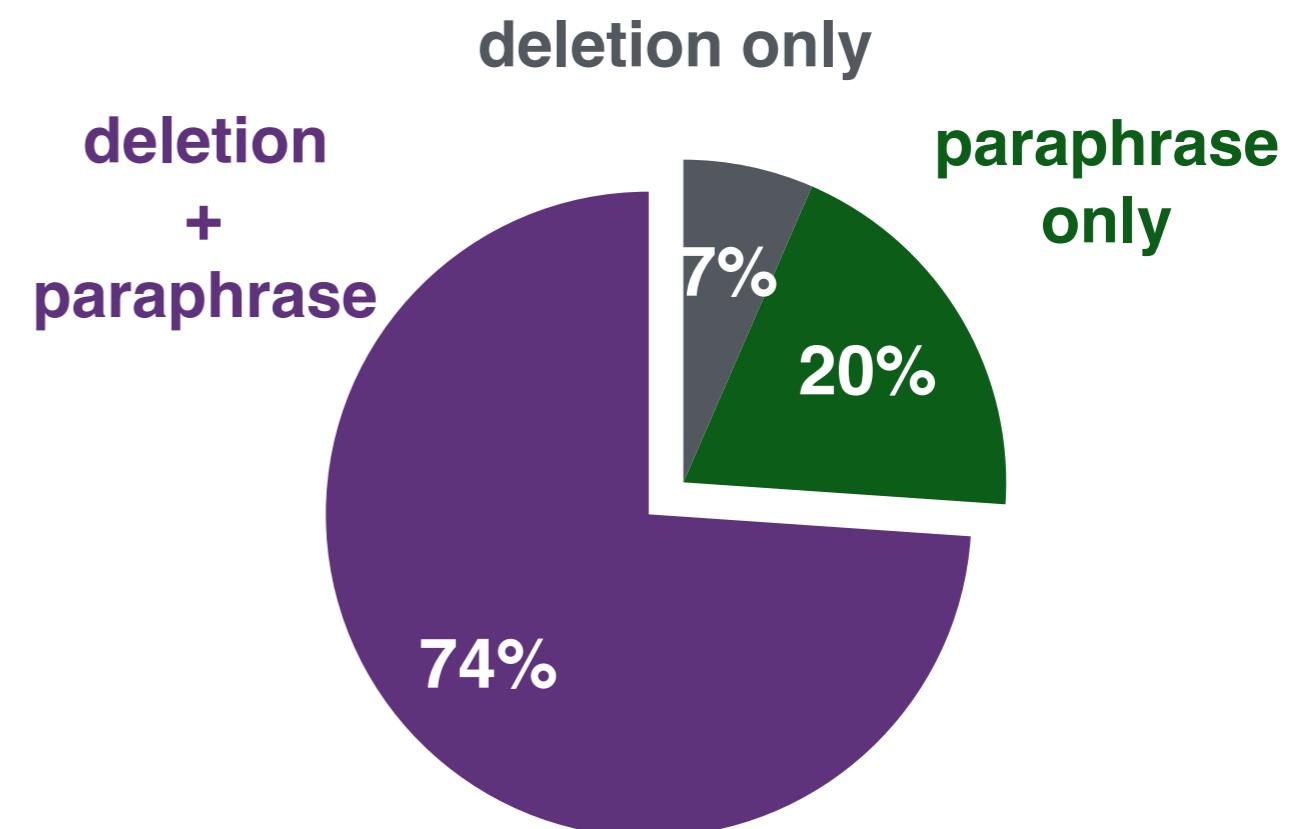
manual inspection of aligned sentence pairs

# Wikipedia\*

**Good simplification needs more paraphrasing.**

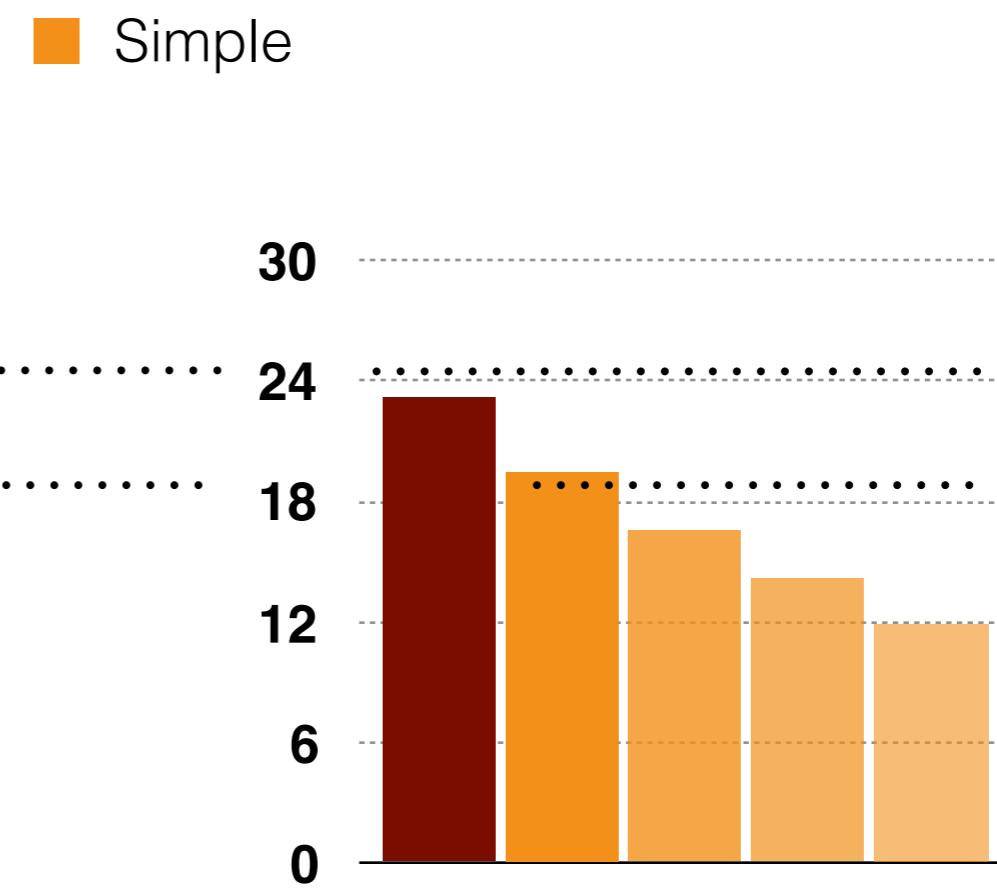
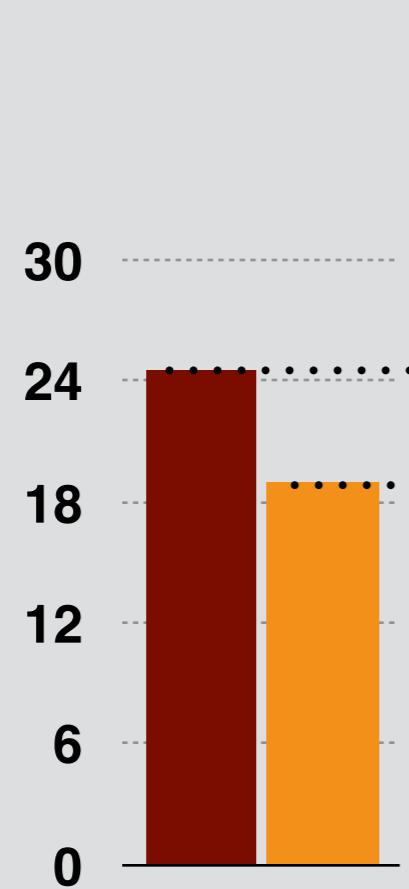


# Newsela



# Wikipedia\*

**Good simplification could be much shorter.**

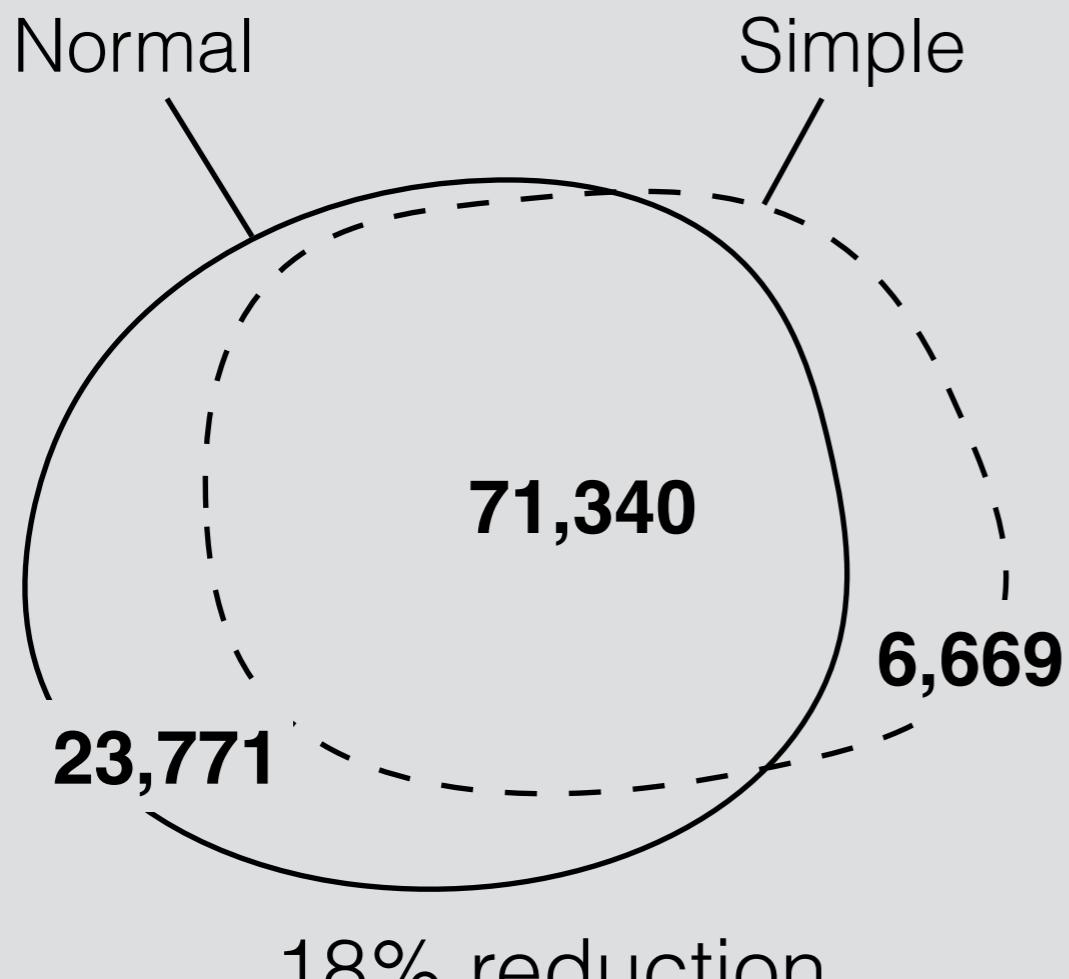


sentence length (#words)  
**see syntax analysis in the paper**

# Wikipedia\*

(total 2.6 million tokens)

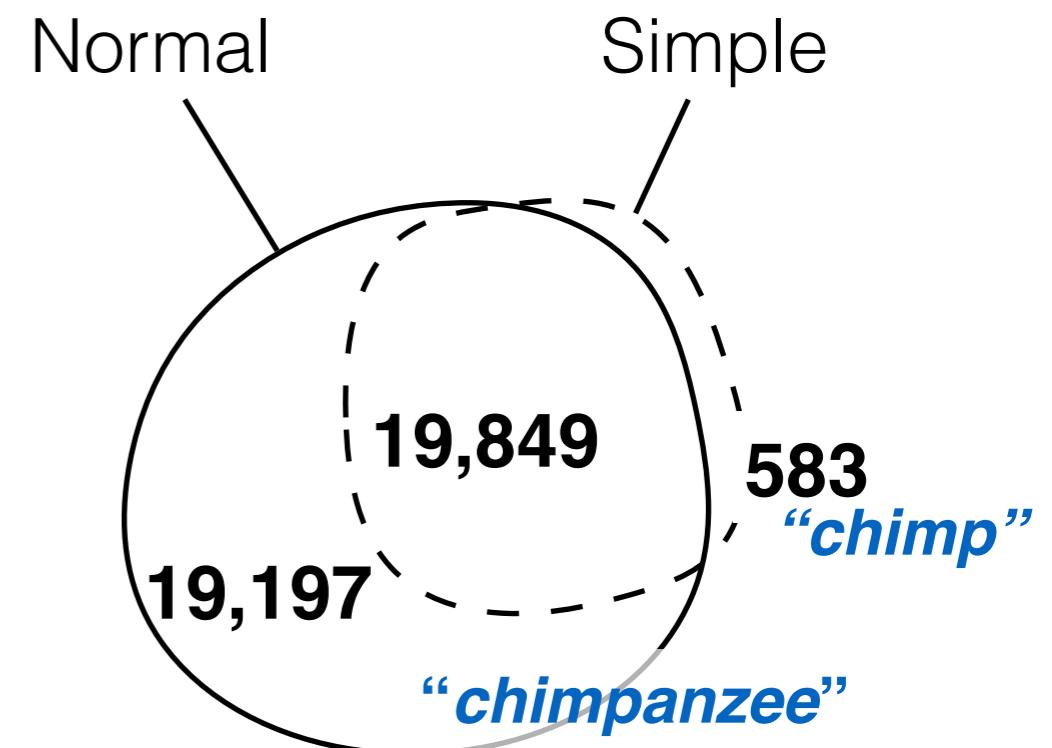
**Good simplification uses a much smaller vocabulary.**



vocabulary size (#unique words)

# Newsela

(total 1.3 million tokens)



# Wikipedia\*

Good simplification reduces certain function word usage.

*commune*  
,

*as*

*and*

*northern*

*northwestern*

*film*

;

*southwestern*

*footballer*

# Newsela

,

*and*

"

—

*of*

*which*

*as*

*percent*

*including*

*director*

most significantly reduced words  
(weighted log-odds-ratio analysis w/ informative Dirichlet prior)

# Wikipedia\*

- █ Normal
- █ Simple

*Postal officials recently tried to ... ... , **which** could ... ....*

*Postal officials recently tried to ... .... **That** could ... ....*

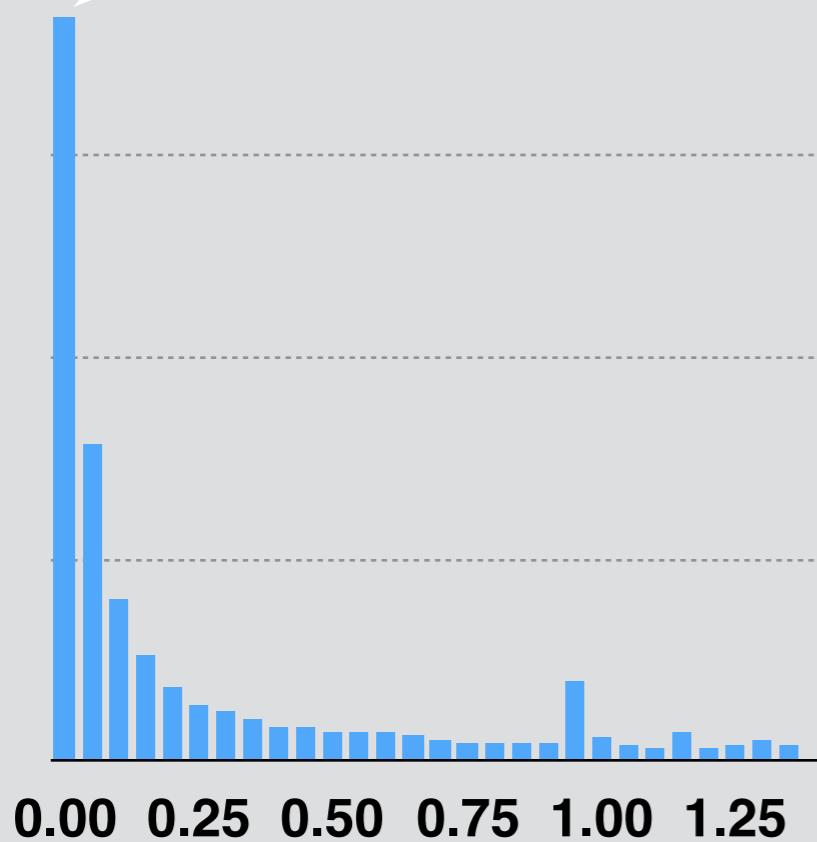
# Newsela

most significantly reduced words  
**see syntax analysis in the paper**

# Wikipedia\*

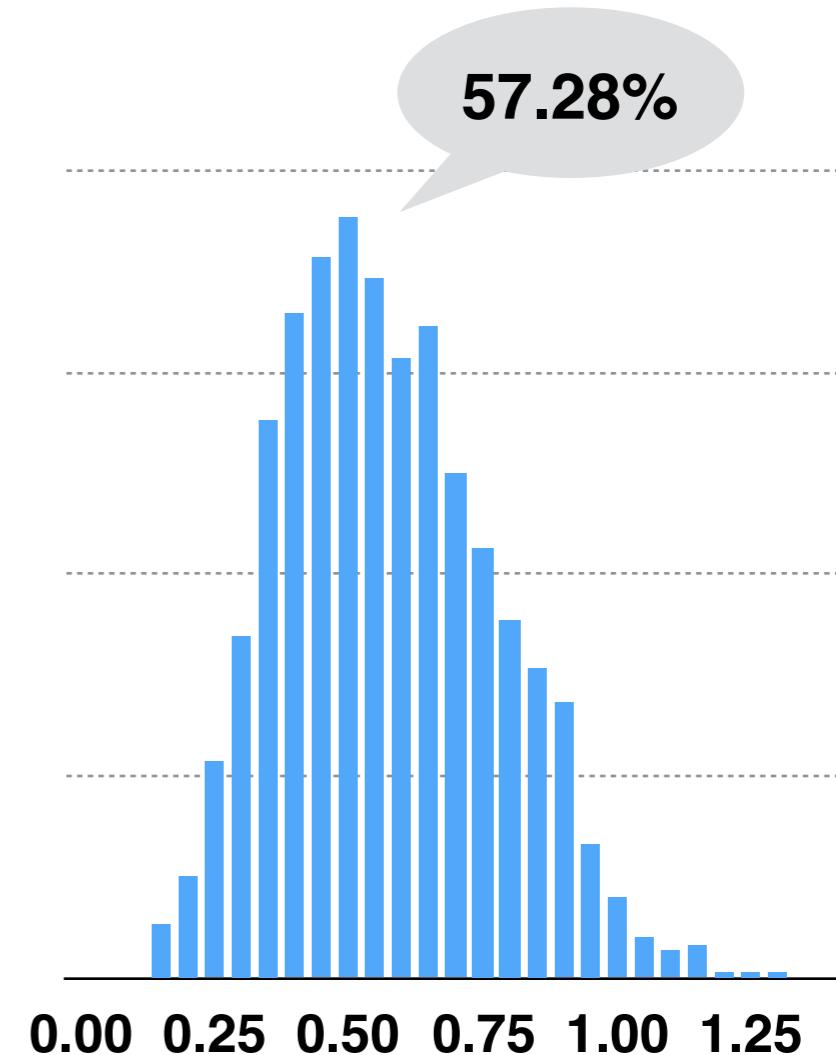
**Wikipedia is not suitable for full-document simplification.**

3.19%



# Newsela

57.28%



document compression ratio (simple/normal)  
**see discourse analysis in the paper**

**Opinion #1**

**Current evaluation  
doesn't tell us what's going on.**

**Opinion #2**

**Simple Wikipedia is **not** that simple.**

**Opinion #3**

**New data **can** help.**

# My Suggestions

- **to reviewers:**
  - be open-minded to papers that may not follow previous evaluation setup, may not outperform the “state-of-the-art” on Wikipedia
  - be sympathetic towards papers specially on data construction\*, data analysis\* and automatic evaluation metrics
  - read our paper

\* (Pellow & Maxine, 2014 HCOMP; Marcelo & Specia, 2014 PITR)

# My Suggestions

- **to researchers:**
  - consider working on text simplification (“pre-BLEU age”)
  - improve evaluation
  - make your system replicable
  - read our paper

# Thank you

Questions?  
Opinions?

**Sponsor: NSF**  
**Newsela data are available at [https://newsela.com/  
data/](https://newsela.com/data/)**

# Back Up

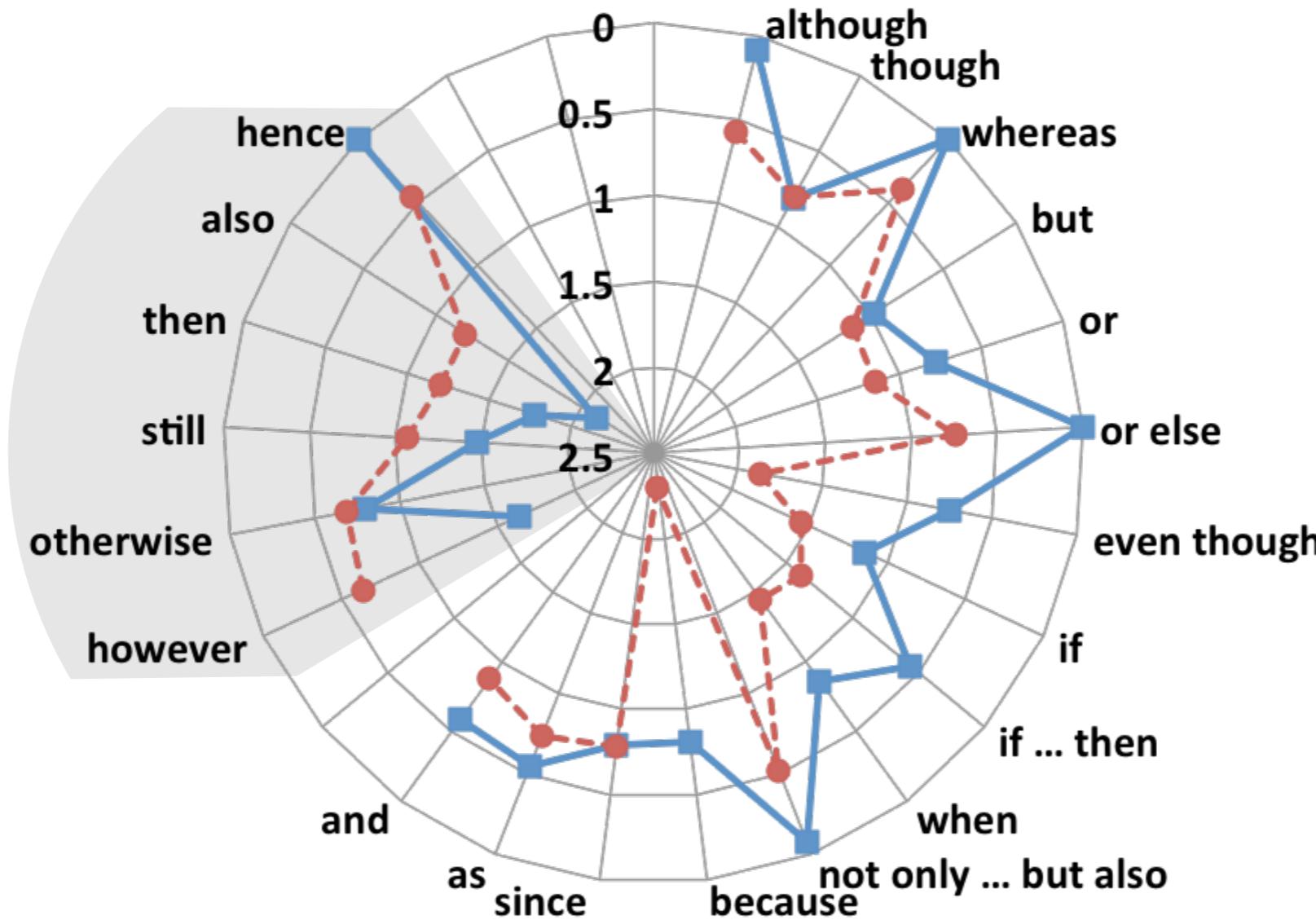
# Wikipedia\*



# Newsela



*simple  
cue  
words*



*complex  
conjunctions*

change of discourse connectives (odds-ratio)

# Reasons of Quality Issues

## in Parallel Wikipedia Corpus

- The Simple Wikipedia was created by volunteers with no specific objective;
- Articles in Simple Wikipedia do not necessarily map Normal Wikipedia;
- As an encyclopedia, Wikipedia contains extremely difficult words and sentences.

# Newsela Dataset

Original

Slightly more fourth-graders nationwide are reading proficiently compared with a decade ago, but only a third of them are now reading well, according to a new report.

Simple-1

Fourth-graders in most states are better readers than they were a decade ago. But only a third of them actually are able to read well, according to a new report.

Simple-2

Fourth-graders in most states are better readers than they were a decade ago. But only a third of them actually are able to read well, according to a new report.

Simple-3

Most fourth-graders are better readers than they were 10 years ago. But few of them can actually read well.

Simple-4

Fourth-graders are better readers than 10 years ago. But few of them read well.

# Newsela Dataset

1,130 news articles

Time: 2013 January ~ 2015 March

Source: Chicago Tribune, Seattle Times,  
LA Times, The Baltimore Sun

Original: 56k sentences — Simple: 64k sentences