Problems in Current Text Simplification Research

Wei Xu **UPenn**
Chris Callison-Burch UPenn
Courtney Napoles JHU

Follow @cocoweixu

TACL paper @ EMNLP Sep-20-2015

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 …)

Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” TACL (2015)

Goal of Text Simplification

grammaticality **meaning preservation** **simplicity**

		Human Evaluation		
INPUT	*Applesauce is a puree made of apples.*			
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)

Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” TACL (2015)

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)

Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” TACL (2015)

Brief History of Sentence Simplification

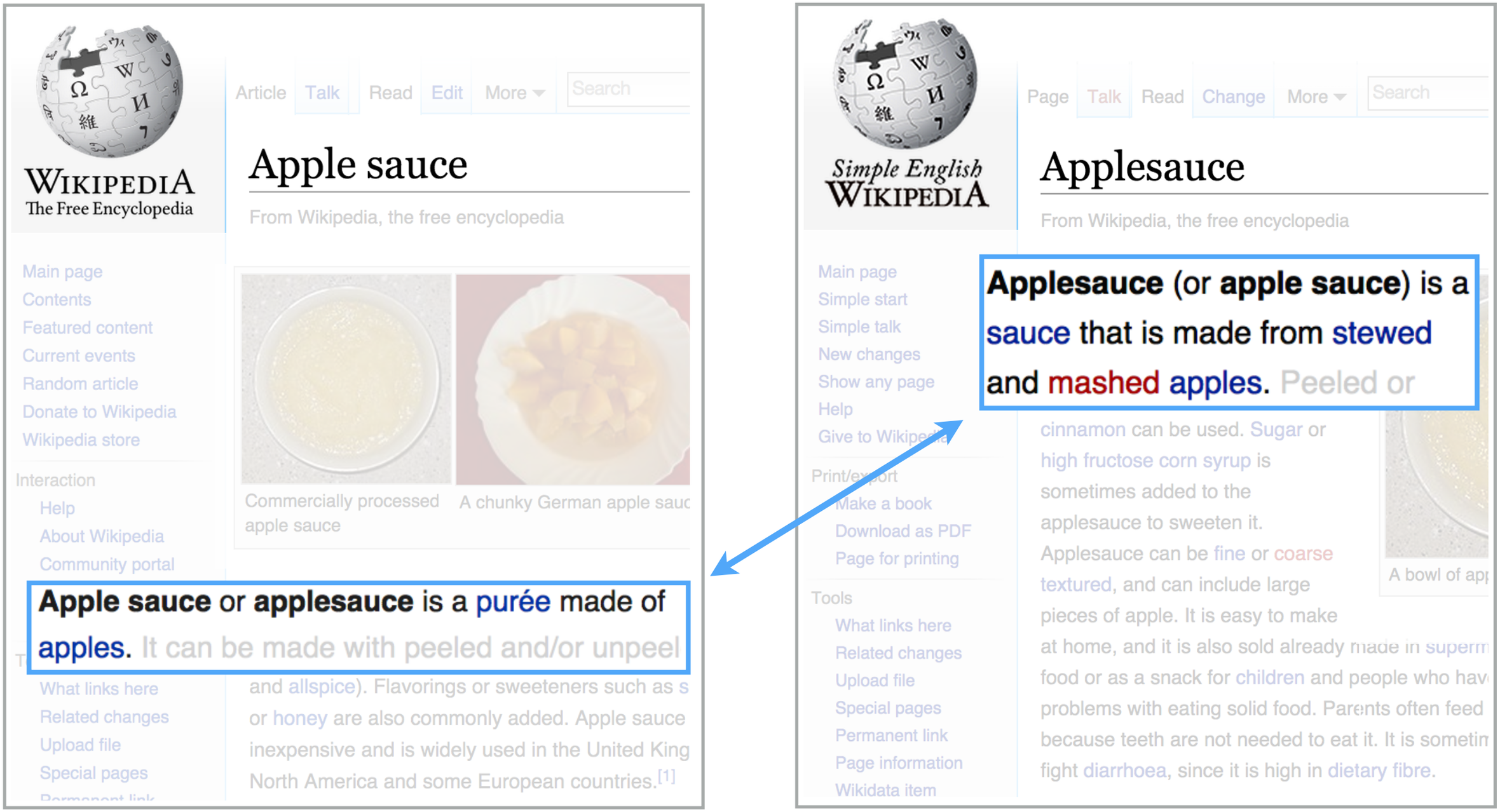
rule-based

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

Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” TACL (2015)

Parallel Wikipedia Corpus



Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” TACL (2015)

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

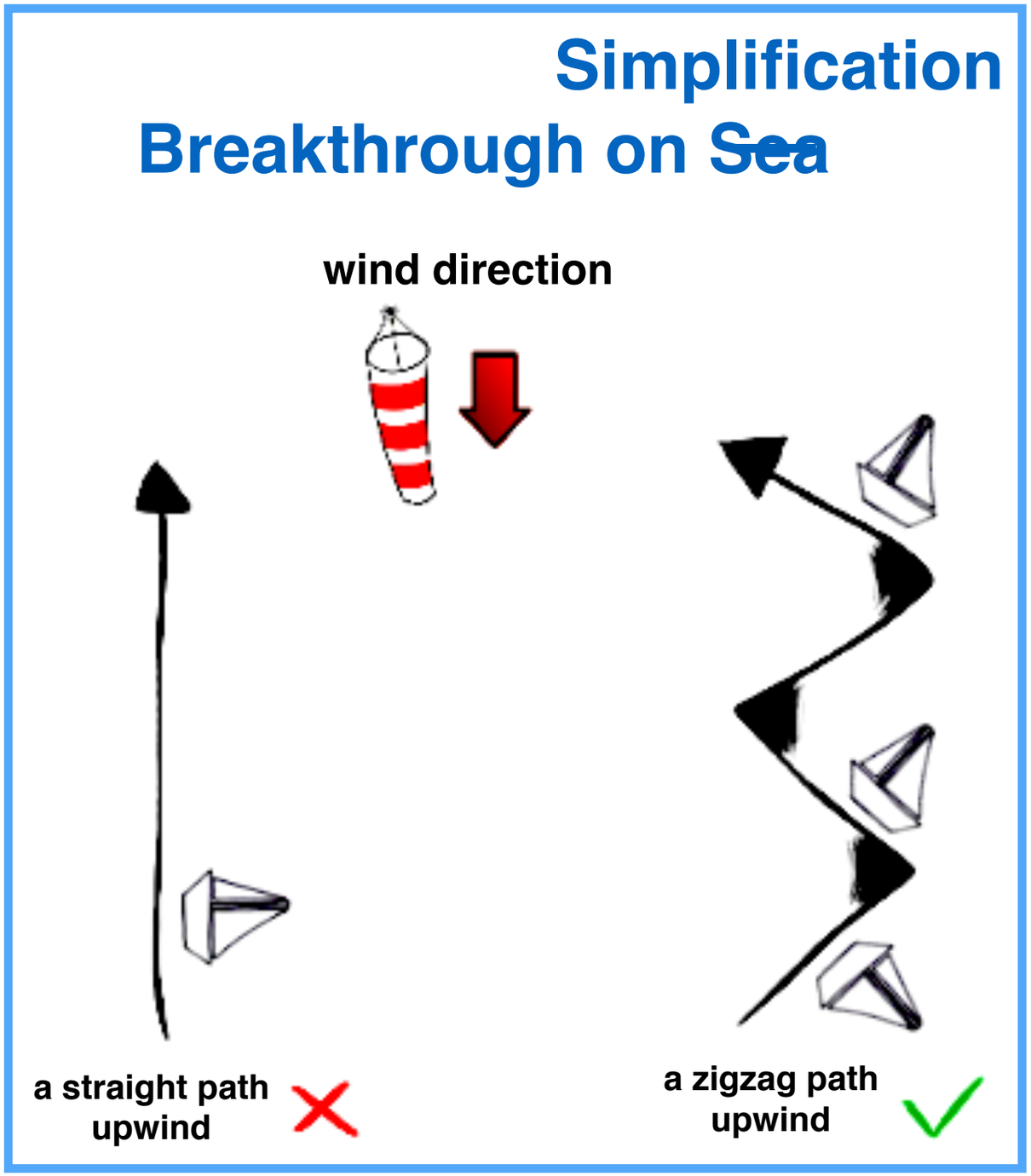
Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” TACL (2015)

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?



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

Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” TACL (2015)

“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?

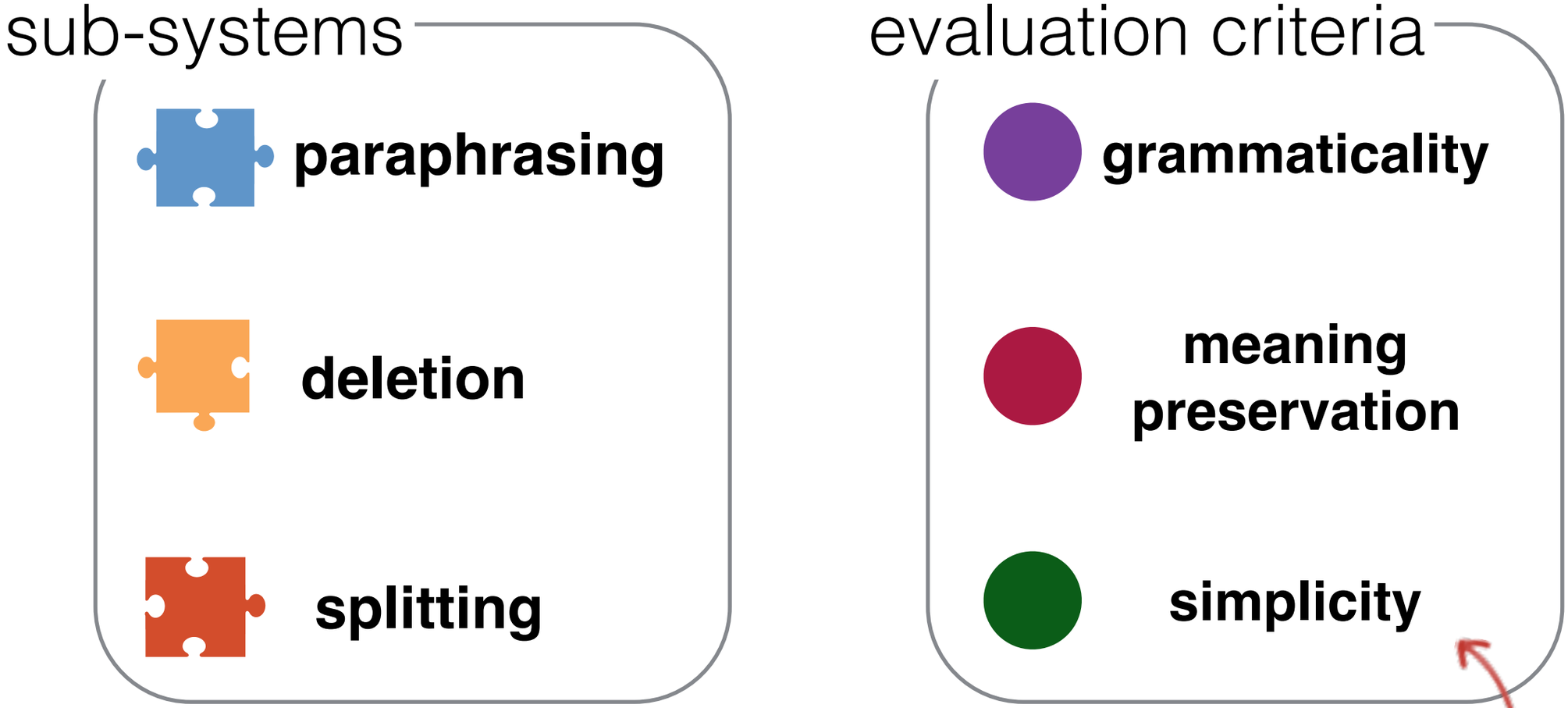
1 “state-of-the-art” competition 2 not easy to do

Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” TACL (2015)

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)

Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” TACL (2015)

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

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

Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” TACL (2015)

“Specific questions that need addressing are :

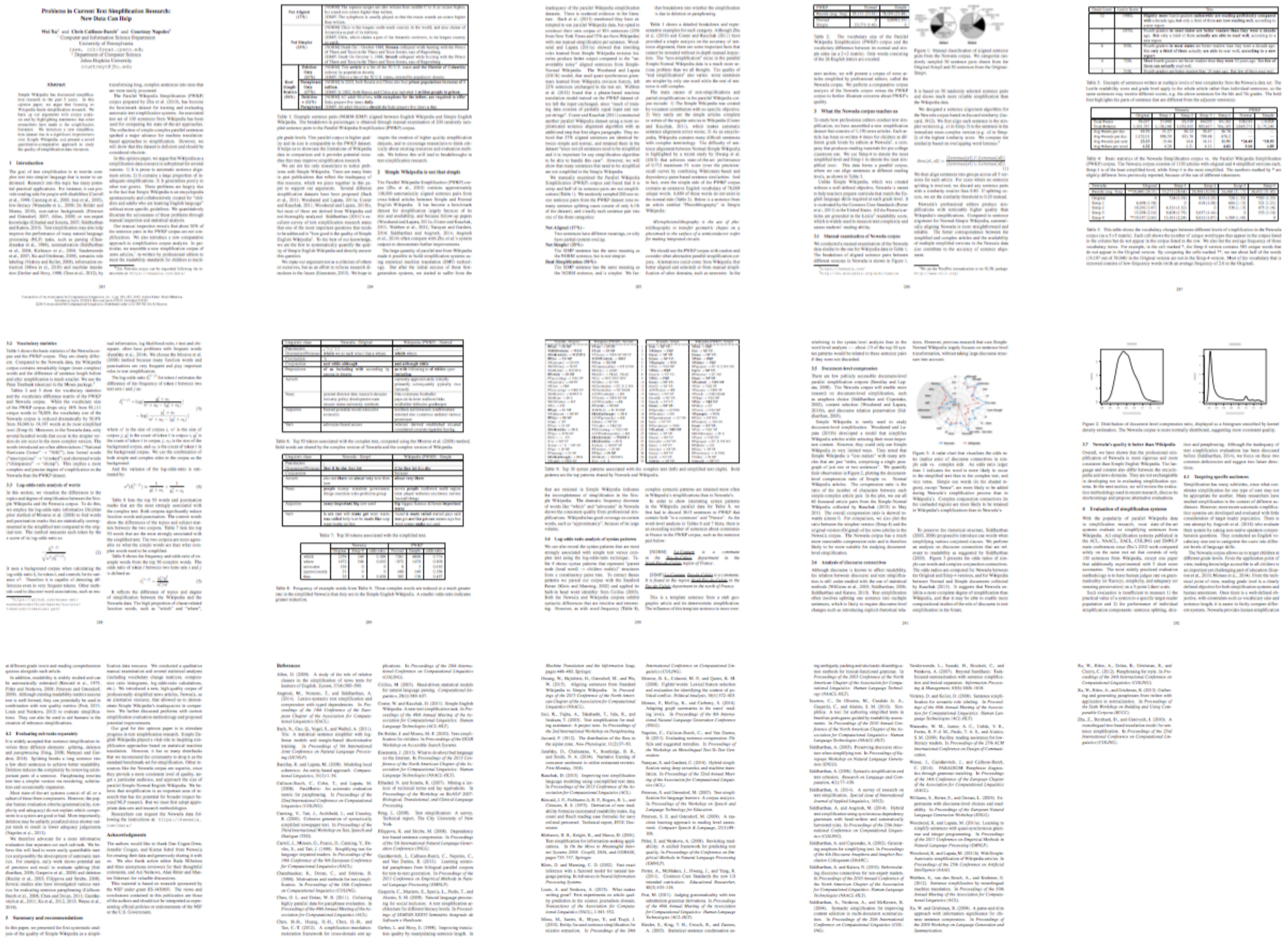
… we need to better understand the quality of Simple English Wikipedia, a resource that has been used to train many SMT based simplification systems…”

—— Advaith Siddharthan (2014 Survey)

WHAT’S NEW?

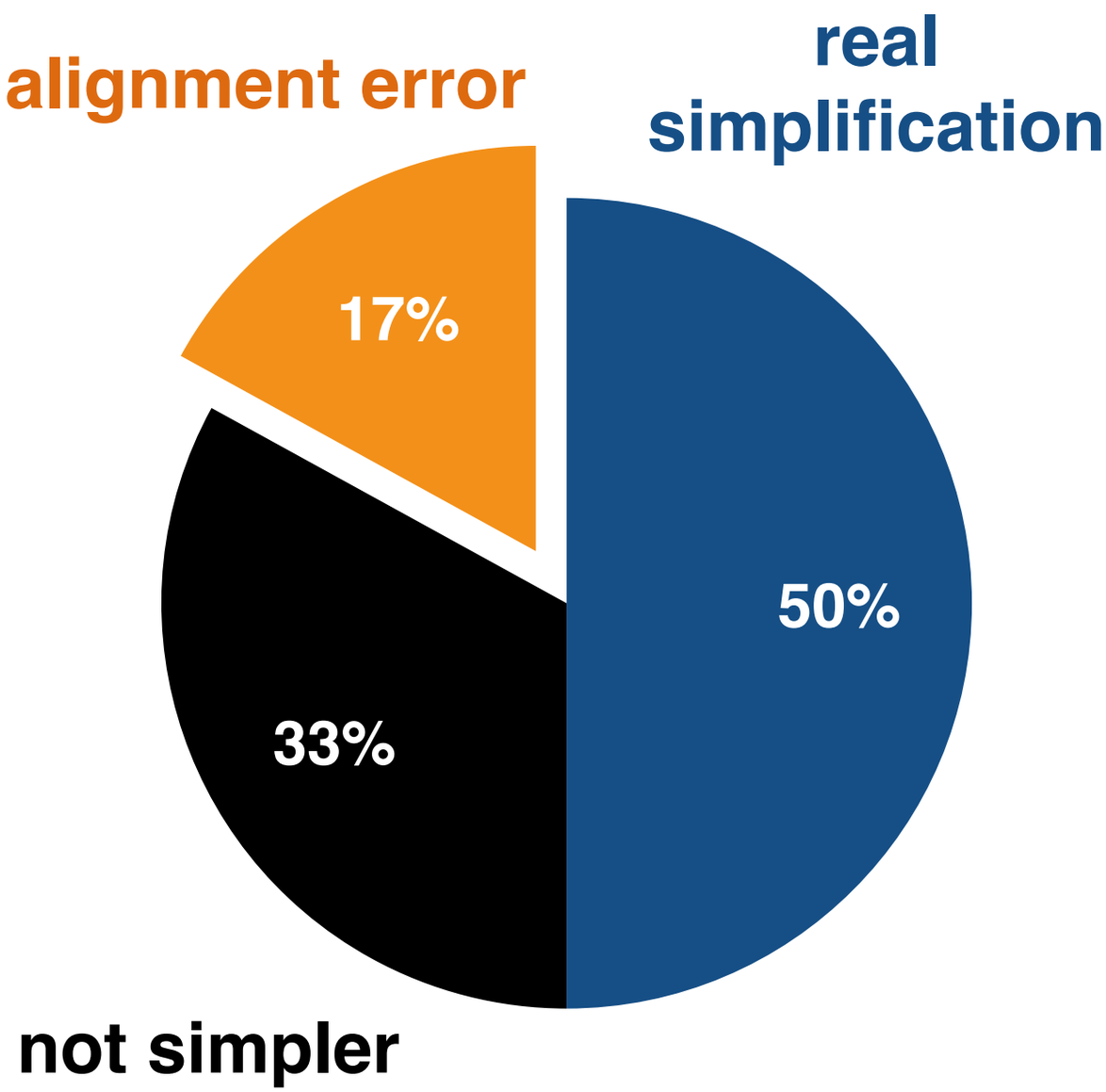
We quantitively and systematically answer this quest.

Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” TACL (2015)



Wei Xu, Chris Callison-Burch, Courtney Napoles. "Problems in Current Text Simplification Research: New Data Can Help" TACL (2015)

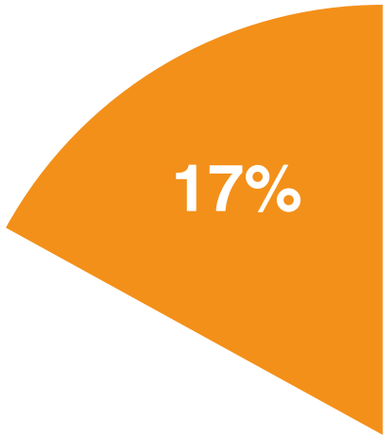
Quality of Parallel Wikipedia Corpus*



Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” TACL (2015)

Inaccuracy in Parallel Wikipedia Corpus*

alignment error (two sentences have different meaning)

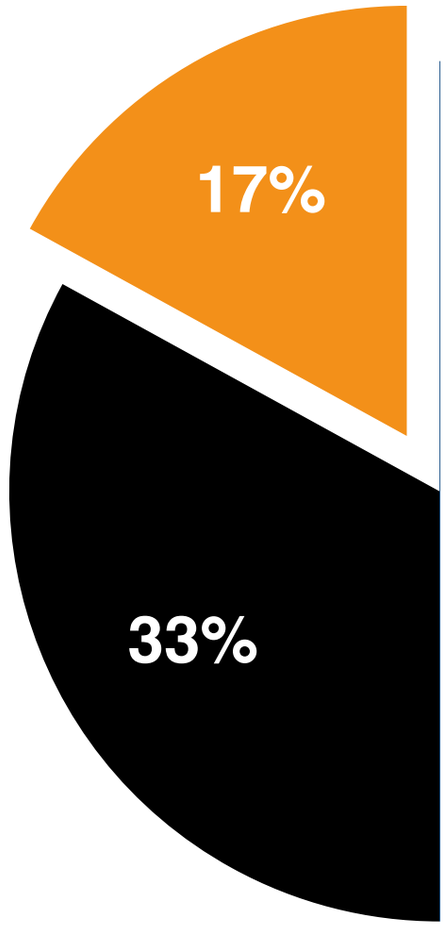


Best automatic sentence alignment gets about **0.7 F1** score (Hwang et al. 2015)

Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” TACL (2015)

Inadequacy in Parallel Wikipedia Corpus*

alignment error (two sentences have different meaning)



not simpler

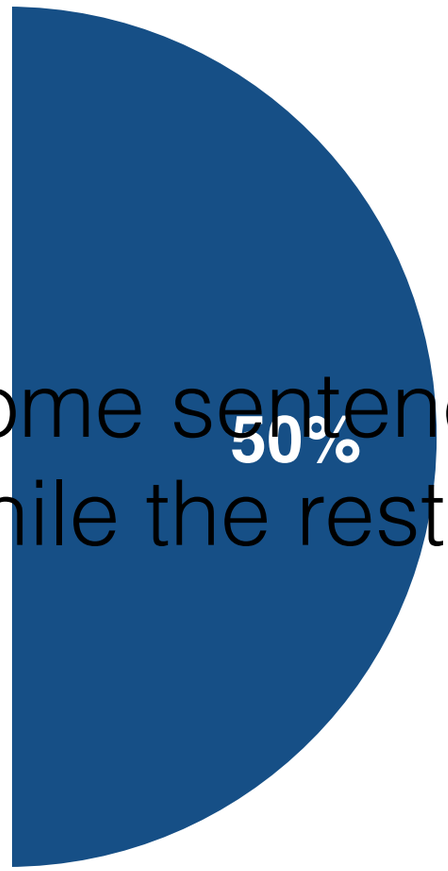
Best automatic sentence alignment gets about **0.7 F1** score (Hwang et al. 2015)

Sentences can have similar meaning but not simplification

Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” TACL (2015)

Inadequacy in Parallel Wikipedia Corpus*

real simplification ? (aligned and simpler)



Some sentences are simpler by only one word while the rest of sentence is still complex

Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” TACL (2015)

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

Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” TACL (2015)

Opinion #3

New data **can** help

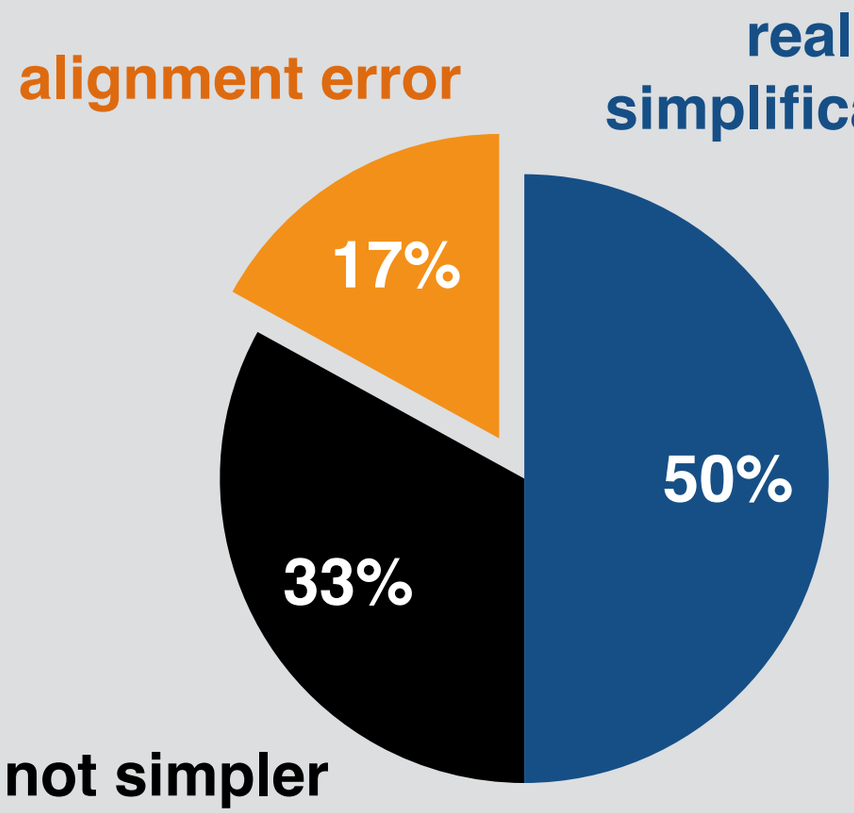
Newsela Dataset

every article at **5 levels** of simplification

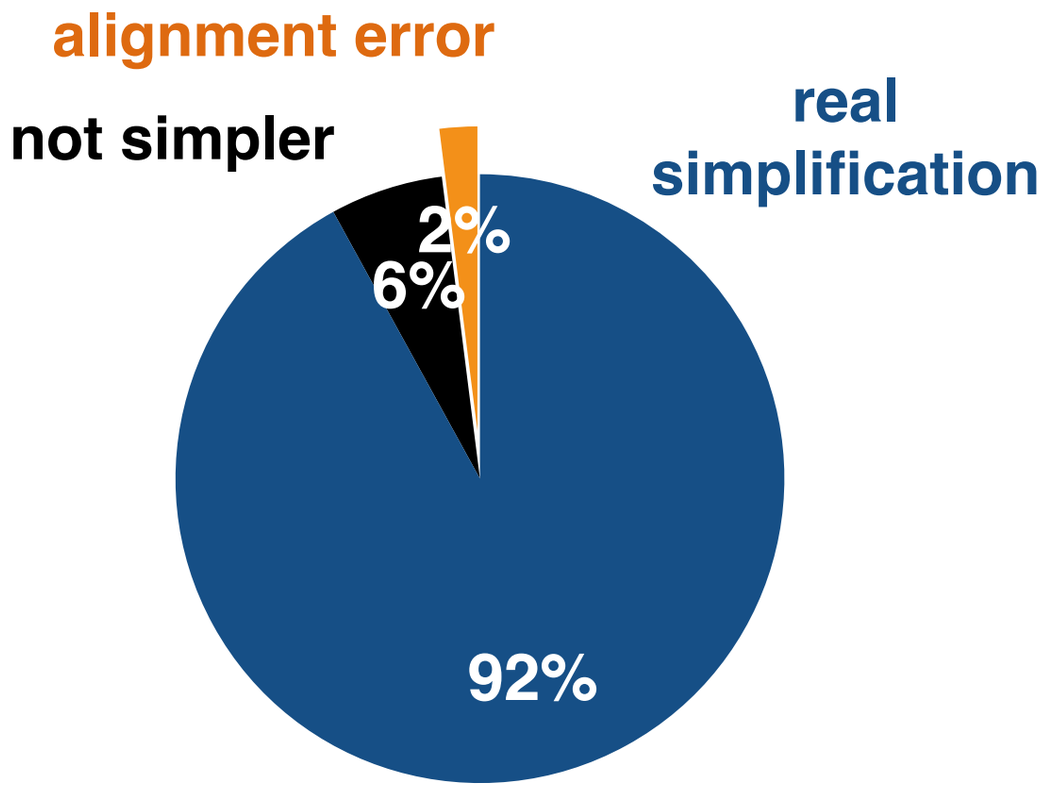
written by trained editors, comes with comprehension quizzes

Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” TACL (2015)

Wikipedia*



Newsela



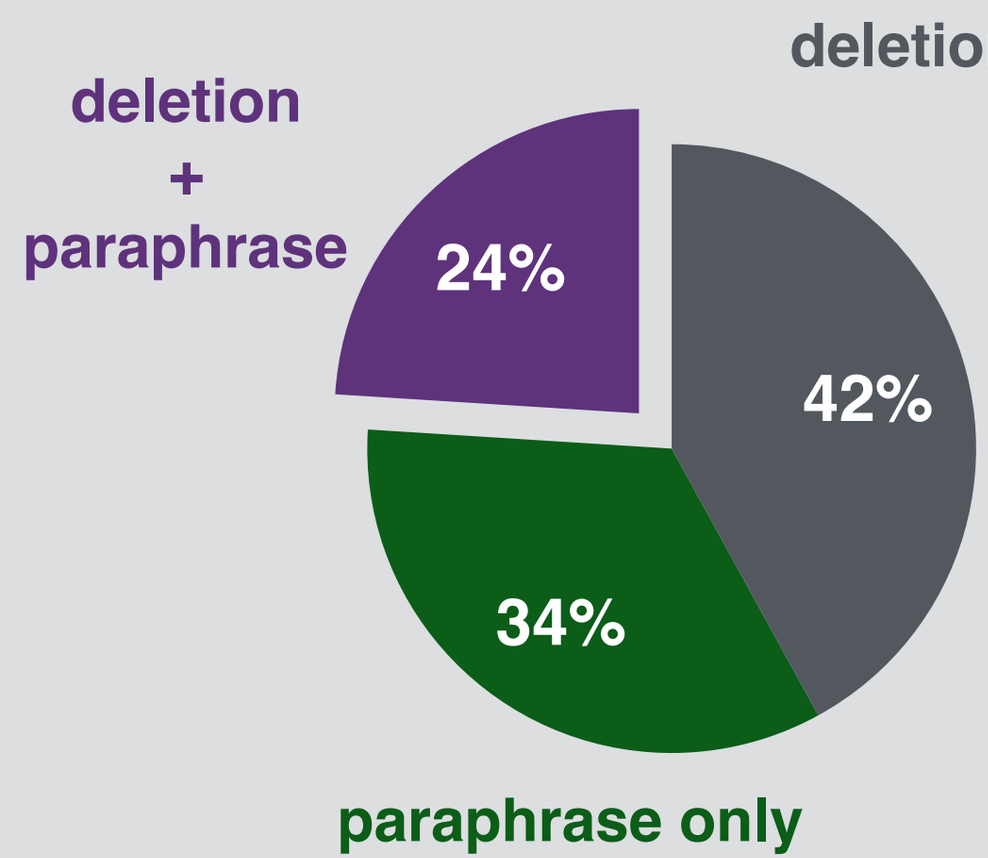
manual inspection of aligned sentence pairs

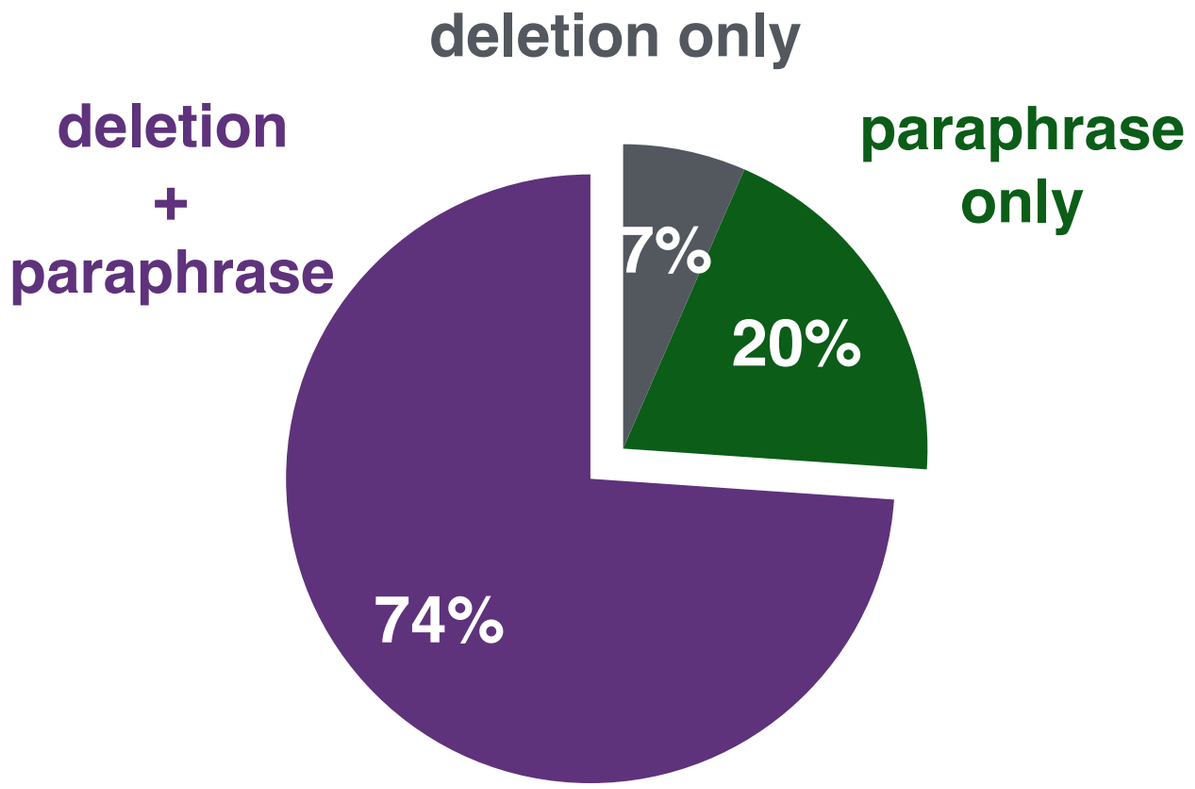
Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” TACL (2015)

Wikipedia*

Newsela

Good simplification needs more paraphrasing.





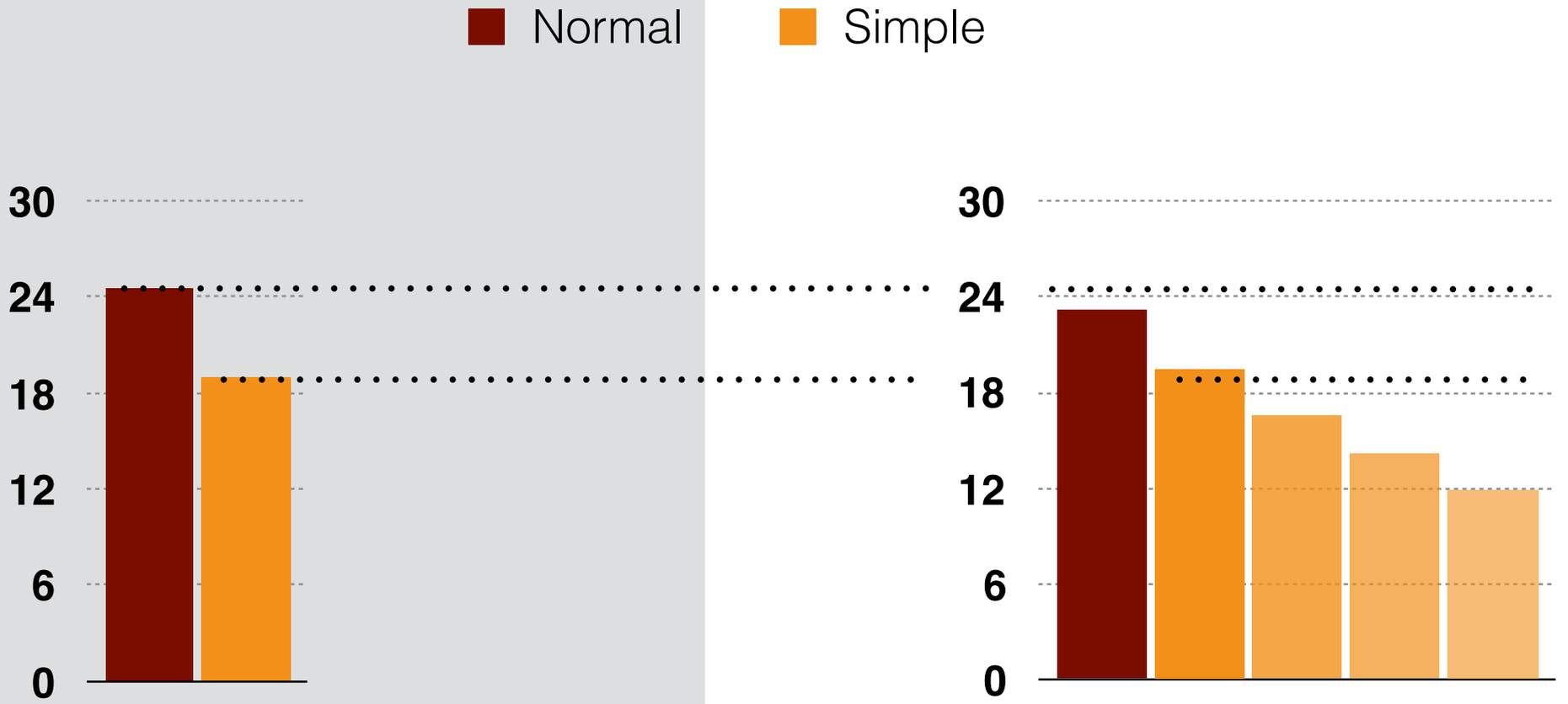
degree of paraphrasing

Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” TACL (2015)

Wikipedia*

Newsela

Good simplification could be much shorter.



sentence length (#words)

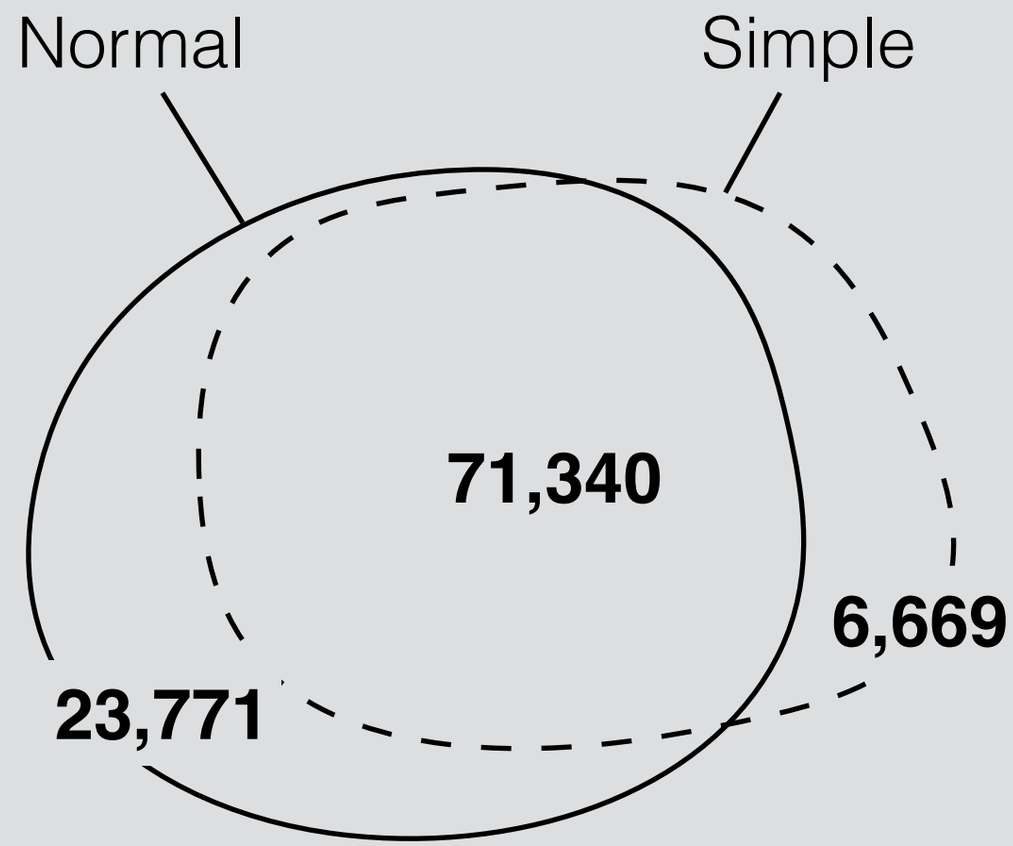
see syntax analysis in the paper

Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” TACL (2015)

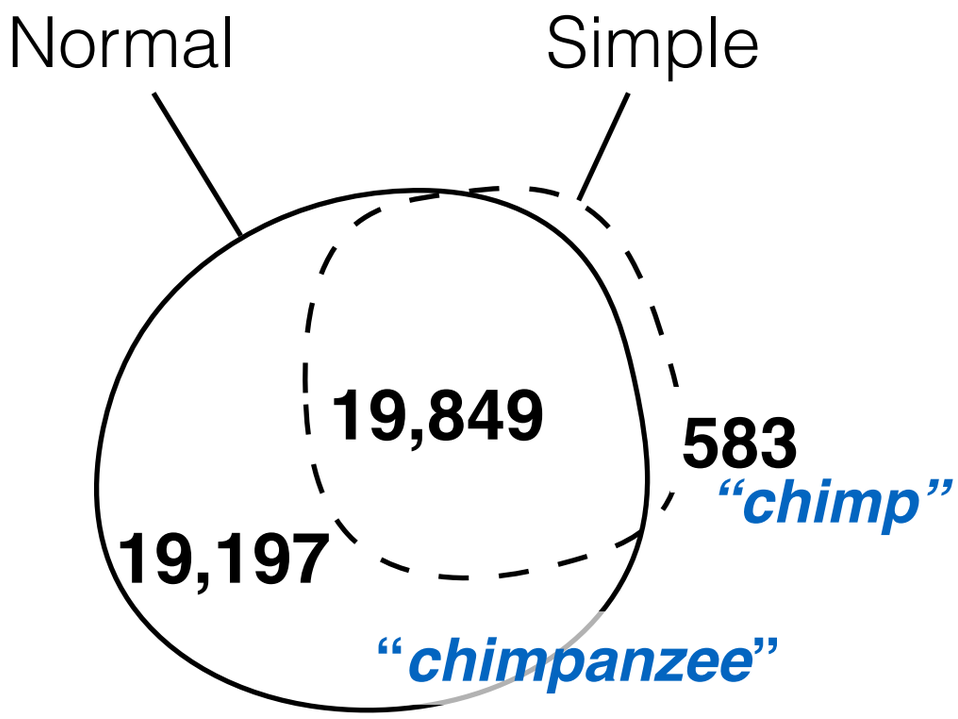
Wikipedia*
(total 2.6 million tokens)

Newsela
(total 1.3 million tokens)

Good simplification uses a much smaller vocabulary.



18% reduction



48% reduction

vocabulary size (#unique words)

Wei Xu, Chris Callison-Burch, Courtney Napoles. "Problems in Current Text Simplification Research: New Data Can Help" TACL (2015)

Wikipedia*

Newsela

Good simplification reduces certain function word usage.

commune	***,***
,	***and***
as	***"***
and	***—***
northern	***of***
northwestern	***which***
film	***as***
;	***percent***
southwestern	***including***
footballer	***director***

most significantly reduced words

(weighted log-odds-ratio analysis w/ informative Dirichlet prior)

Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” TACL (2015)

Wikipedia*

Newsela

- Normal
- Simple

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

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

most significantly reduced words

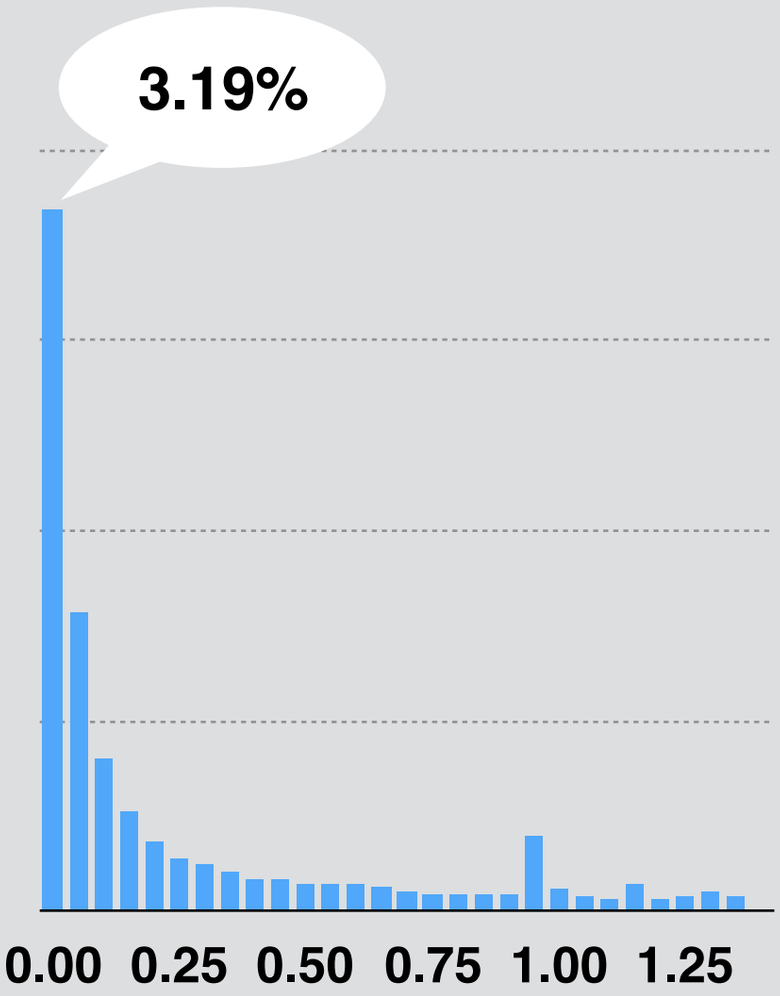
see syntax analysis in the paper

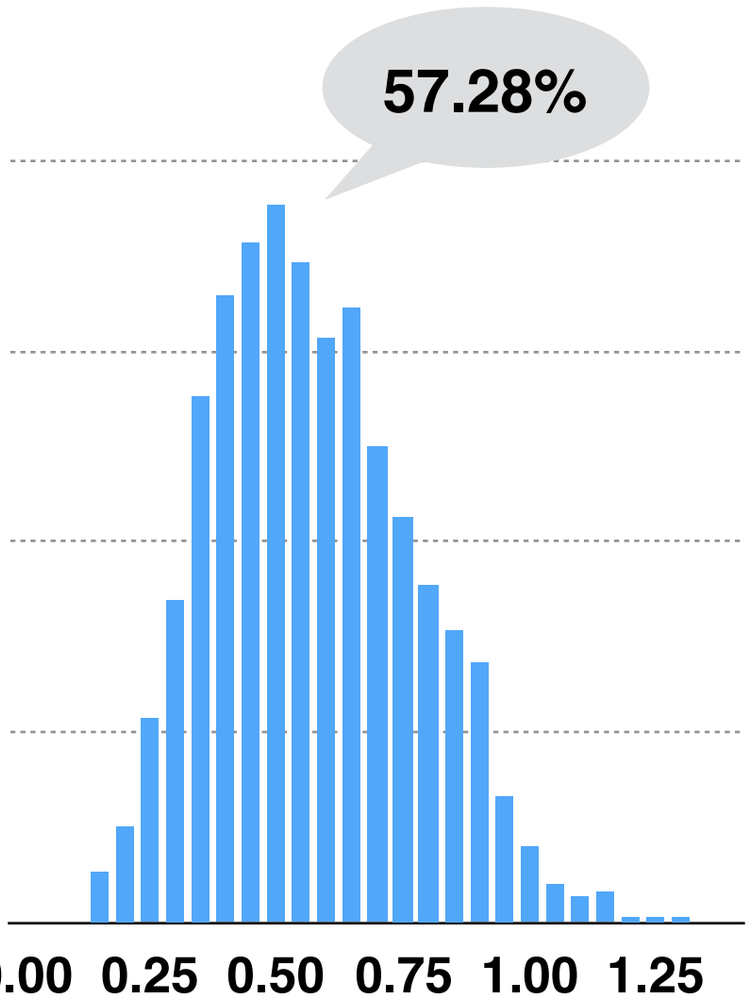
Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” TACL (2015)

Wikipedia*

Newsela

Wikipedia is <u>not</u> suitable for full-document simplification.





document compression ratio (simple/normal)

see discourse analysis in the paper

Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” TACL (2015)

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

Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” TACL (2015)

Thank you

Questions?

Opinions?

Sponsor: NSF
Newsela data are available at https://newsela.com/data/

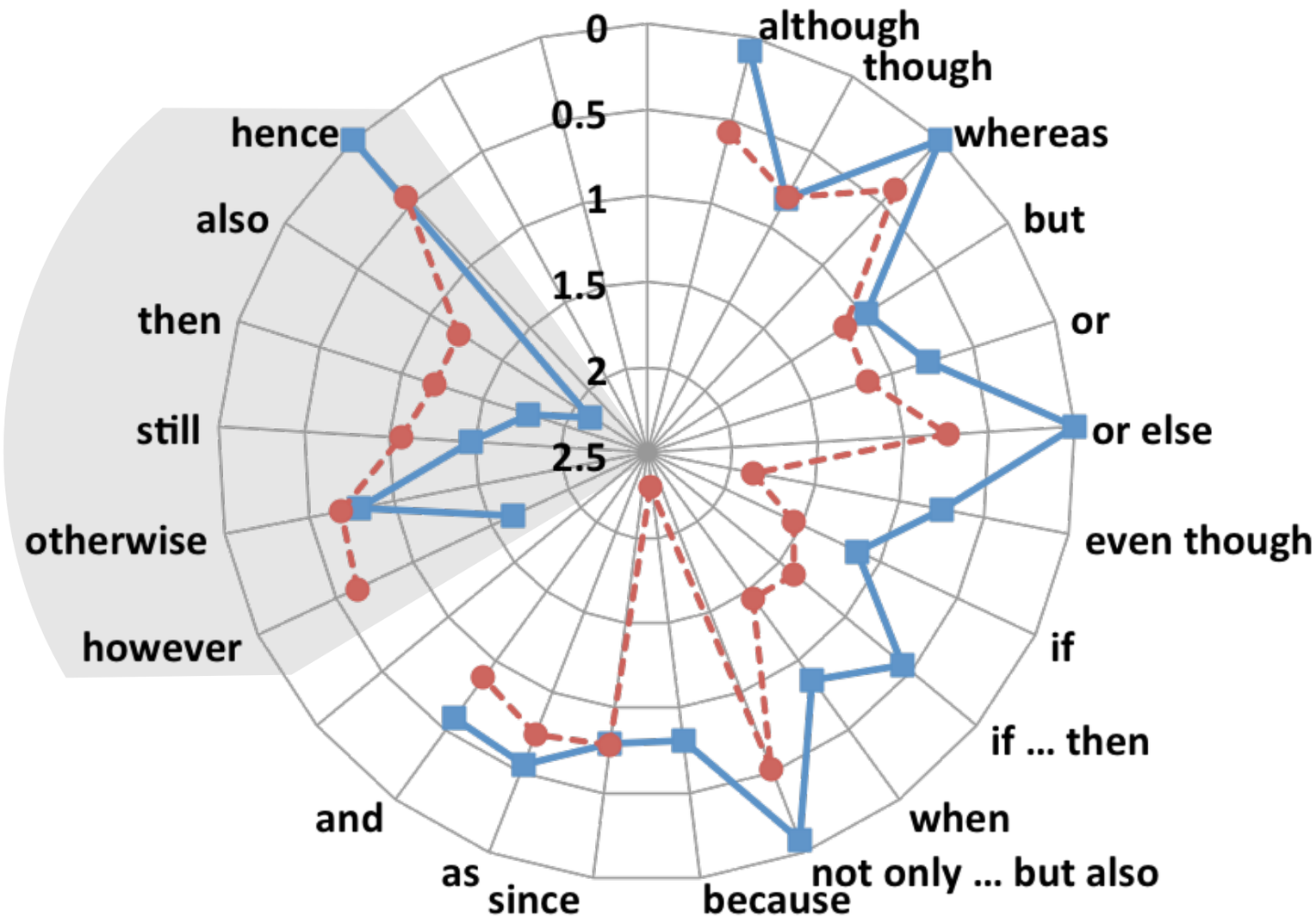
Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” TACL (2015)

Back Up

Wikipedia*

Newsela

simple cue words



complex conjunctions

change of discourse connectives (odds-ratio)

Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” TACL (2015)

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 difficulty 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