NRW-FORSCHUNGSKOLLEG
ONLINE-PARTIZIPATION

When the Scale is Unclear

Analysis of the Interpretation of Rating Scales in Human Evaluation of Text Simplification

Regina Stodden
regina.stodden@hhu.de
Heinrich Heine University Düsseldorf, Germany
NRW-Forschungskolleg Online Partizipation

Research Questions

RQ1: How to carry out human text simplification evaluation?

RQ2: How do the raters interpret the rating scales?

Research Questions

RQ1: How to carry out human text simplification evaluation?

RQ2: How do the raters interpret the rating scales?

Text Simplification Evaluation

Manual Evaluation

- Main three rating dimensions (Alva-Manchego etal. (2020a))
  - Grammaticality (or fluency)
  - Simplicity
  - Meaning preservation (or adequacy)

- Comprehension studies (Lapan etal. (2021)):
  - Participants read either a original or a simplified text
  - Afterwards they answer questions related to the content of the text
  - Analysis if the questions were answered better with the original or simplified text

Simplicity

“Is the output simpler than the input?”

Sulem etal. (2018a): HSplit

absolute simplicity

Štajner etal. (2016): QATS

Structural simplicity

Sulem etal. (2018b): PWKP test

“Does the generated sentence(s) simplify the complex input?”

Narayan & Gardent (2014)

“The simplified sentence is easier to understand than the original sentence.”

Alva-Manchego etal. (2020b): ASSET;
Scialom etal. (2021): HL+SL

“How many successful lexical or syntactic paraphrases occurred in the simplification?”

Xu etal. (2016)

“How much simpler is sentence 2 than sentence 1?”

Schwarzer etal. (2021): Fusion

Simplicity - Scales

“Is the output simpler
than t

-2 - +2

Sulem etal. (2018

1 | 2 | 3
bad | ok | good

Štajner etal. (2016): QATS

Structural simplicity

Sulem etal. (2018b):
PWKP test

“Does the generated sentence(s) simplify the complex input?”

Narayan & Gardent (2014)

“The simplified sentence is easier
to understa

0 - 100
strongly disagree –
strongly agree

Alva-Manche
Scialom etal.

“How many successful lexical or syntactic paraphrases occurred in the simplification?”

Xu etal. (2016)

“How much
sentence 2 tha

-2 - +2
Much less simpler
– much simpler

Schwarzer etal. (2021): Fusion

Simplicity – Rating Groups

“Is the output simpler
than

3 experts

Sulem etal. (2018

ba

?

Štajner etal. (2016): QATS

Structural simplicity

Sulem etal. (2018b):
PWKP test

“Does the generated sentence(s) simplify the complex input?”

Narayan & Gardent (2014)

“The simplified sentence is easier
to understa

0

strong
strongly

12-35 crowd workers

Alva-Manche
Scialom etal.

“How many successful lexical or syntactic paraphrases occurred in the simplification?”

Xu etal. (2016)

“How much
sentence 2 tha

Much
– much si

3 crowd workers

Schwarzer etal. (2021): Fusion

Text Simplification Evaluation

RQ1: How to carry out human text simplification evaluation?

- The judgments are collected…
  - … on scales with different sizes, i.e., 3, 5 and 100,
  - … on scales with different point names, i.e., ”good” to ”bad” or ”strongly disagree” to ”strongly agree”
  - … by crowd workers or experts
  - … on different item types, i.e., questions or statements

Text Simplification Evaluation

RQ1: How to carry out human text simplification evaluation?

- The judgments are collected…
  - … on scales with different sizes, i.e., 3, 5 and 100,
  - … on scales with different point names, i.e., ”good” to ”bad” or ”strongly disagree” to ”strongly agree”
  - … by crowd worker
  - … on different item

Best practices of
human text simplification evaluation
are missing.

Research Questions

RQ1: How to carry out human text simplification evaluation?

RQ2: How do the raters interpret the rating scales?

Rating of Aligned Sentence Pairs

Original Sentence

"The collapsed Dome of the main church has been restored entirely."
(ASSET #287)

Simplified Sentence

"The Dome has been restored."
(ASSET #287)

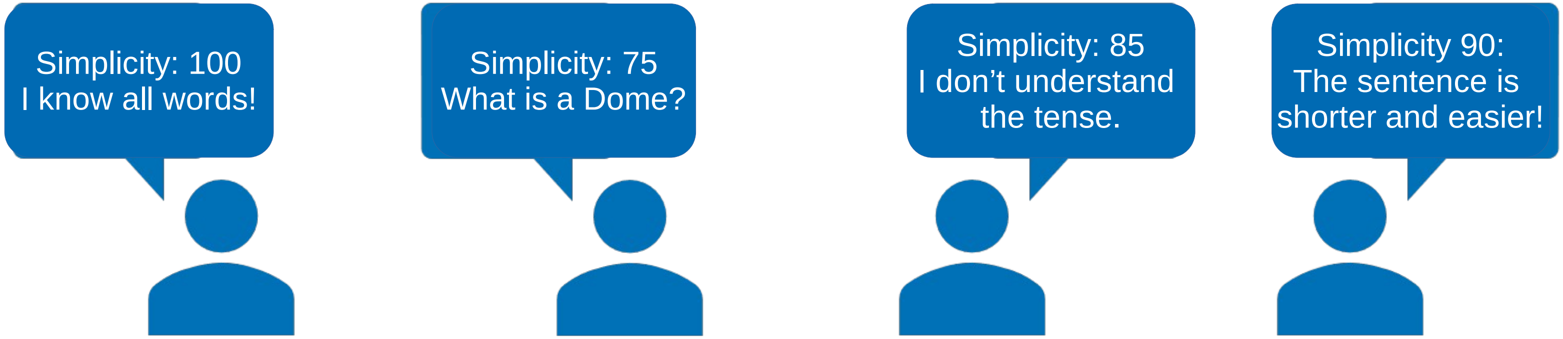
Rating of Aligned Sentence Pairs

Original Sentence

"The collapsed Dome of the main church has been restored entirely."
(ASSET #287)

Simplified Sentence

"The Dome has been restored."
(ASSET #287)



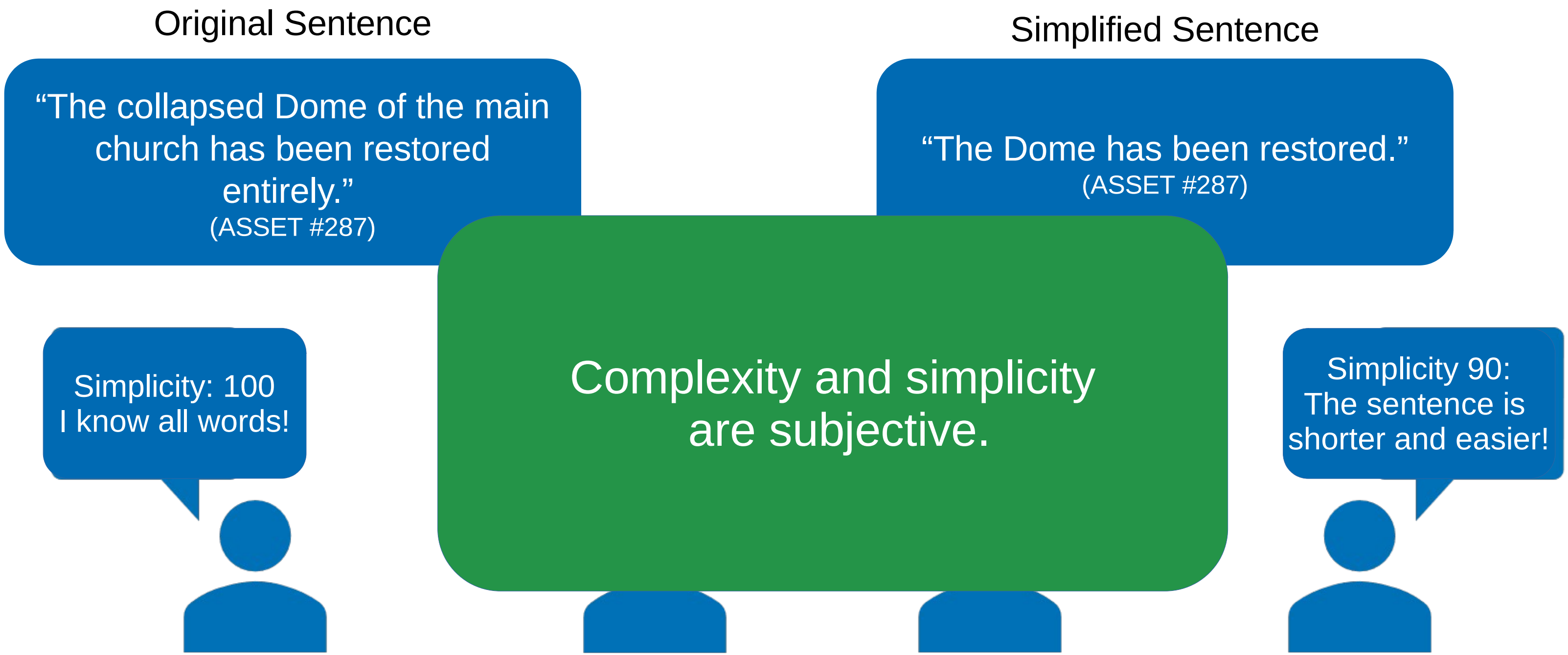
Simplicity: 100
I know all words!

Simplicity: 75
What is a Dome?

Simplicity: 85
I don't understand the tense.

Simplicity 90:
The sentence is shorter and easier!

Rating of Aligned Sentence Pairs



Original Sentence

“The collapsed Dome of the main church has been restored entirely.”
(ASSET #287)

Simplified Sentence

“The Dome has been restored.”
(ASSET #287)

Simplicity: 100
I know all words!

Complexity and simplicity are subjective.

Simplicity 90:
The sentence is shorter and easier!

Rating of no-change Pairs

Original Sentence

“Their eyes are quite small, and their visual Acuity is poor.”
(ASSET #90)

Simplified Sentence

“Their eyes are quite small, and their visual Acuity is poor.”
(ASSET #90)

Absolute simplicity ratings of the simplified sentence are still subjective.

But,
relative simplicity ratings of no-change pairs should be the same.

Rating of no-change Pairs

Original Sentence

“Their eyes are quite small, and their visual Acuity is poor.”
(ASSET #90)

Simplified Sentence

“Their eyes are quite small, and their visual Acuity is poor.”
(ASSET #90)

- Simplicity Rating:
  - As simple as before, but could be worse
  - Not easier to understand than before , worst case

- Rate with neutral scale element?
- Rate with lowest scale element?

It depends on the rating scale and its definition.

Data

	QATS Štajner etal. (2016)	**HSplit** Sulem etal. (2018a)	**PWKP test** Sulem etal. (2018b)	**ASSET** Alva-Manchego etal. (2020b)	**system-likert** Scialom etal. (2021)	**human-likert** Scialom etal. (2021)	**Fusion** Schwarzer etal. (2021)
# sentence pairs	631	1960	500	100	151	108	2920
# no-change sentence pairs	107	346	20	5	2	3	338
# no-change in %	16.96	17.65	4	5	0.97	1.35	11.58
# no-change pair raters		3	5	23	19	30	3
# no-change annotation records	321	1384	80	225	90	126	2028

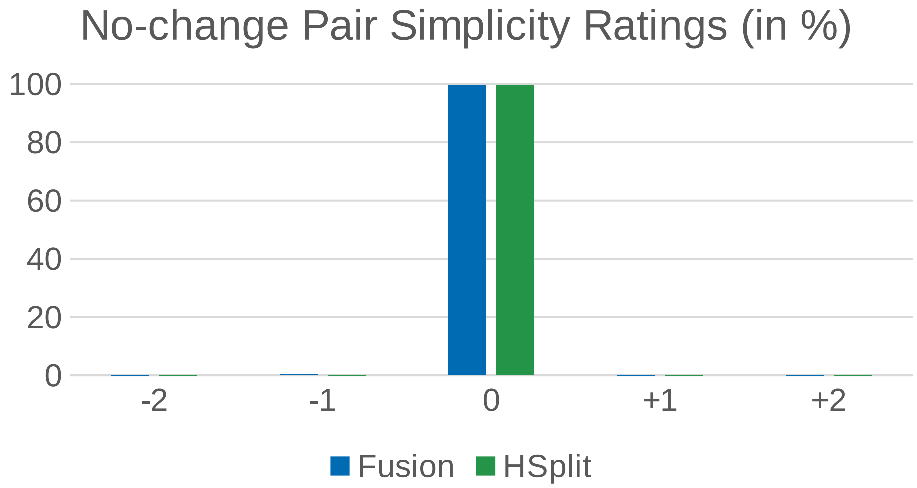
Data

	QATS Štajner etal. (2016)	HSplit Sulem etal. (2018a)	PWKP test Sulem etal. (2018b)	ASSET Alva-Manchego etal. (2020b)	system-likert Scialom etal. (2021)	human-likert Scialom etal. (2021)	Fusion Schwarzer etal. (2021)
# sentence pairs	631	1960	500	100	151	108	2920
# no-change sentence pairs	107	346	20	5	2	3	338
# no-change in %	16.96	17.65	4	5	0.97	1.35	11.58
# no-change pair raters		3	5	23	19	30	3
# no-change annotation records	321	1384	80	225	90	126	2028

Hypotheses

Hypothesis 1

In HSplit and Fusion, the **simplicity** rating of no-change pairs are equal to the **neutral element, i.e., 0**.

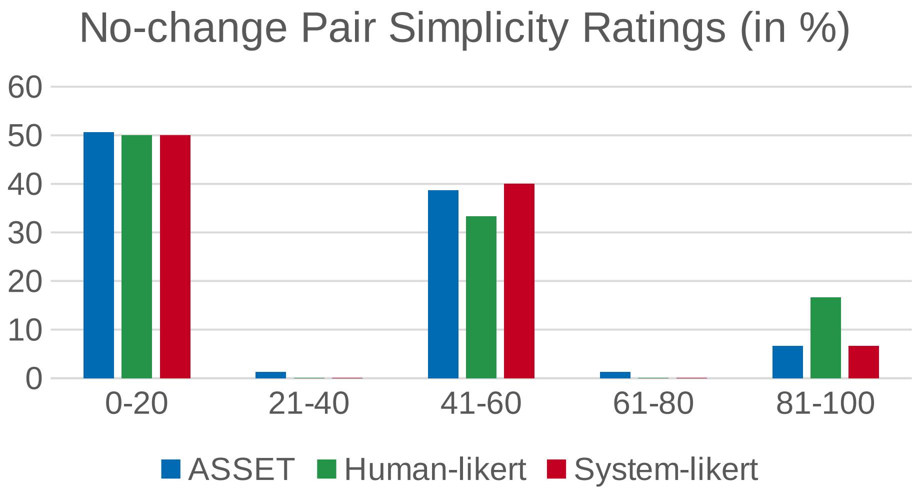


✓

Hypotheses

Hypothesis 2

In ASSET, human-likert, and system-likert, the **simplicity** ratings of no-change pairs are equal to the **lowest element of the scale, i.e., 0**, as it indicates the worst simplification.



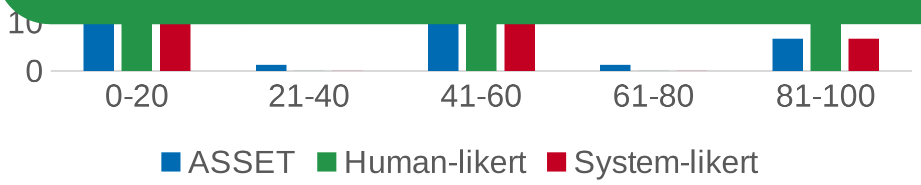
✕

Hypotheses

Hypothesis 2

In ASSET, human-likert, and system-likert, the **simplicity** ratings of no-change pairs are equal to the **lowest element of the scale, i.e., 0**, as it indicates the worst simplification.

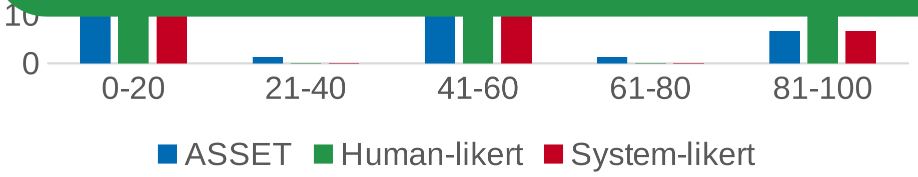
Different scale interpretations?



Hypotheses

Hypothesis 2

In ASSET, human-likert, and system-likert, the **simplicity** ratings of no-change pairs are equal to the **lowest element of the scale, i.e., 0**, as it indicates the worst simplification.



Different scale interpretations?
Either...

- ...0 for higher complexity and <u>**0**</u> for same complexity, or
- ...0 for higher complexity and <u>**50**</u> for same complexity

Hypotheses

Hypothesis 4

If **different interpretations** of the scales exist, the rater groups’ ratings significantly **differ** for sentence pairs in which the original and the simplified sentences are **not identical**.

Hypotheses

Hypothesis 4

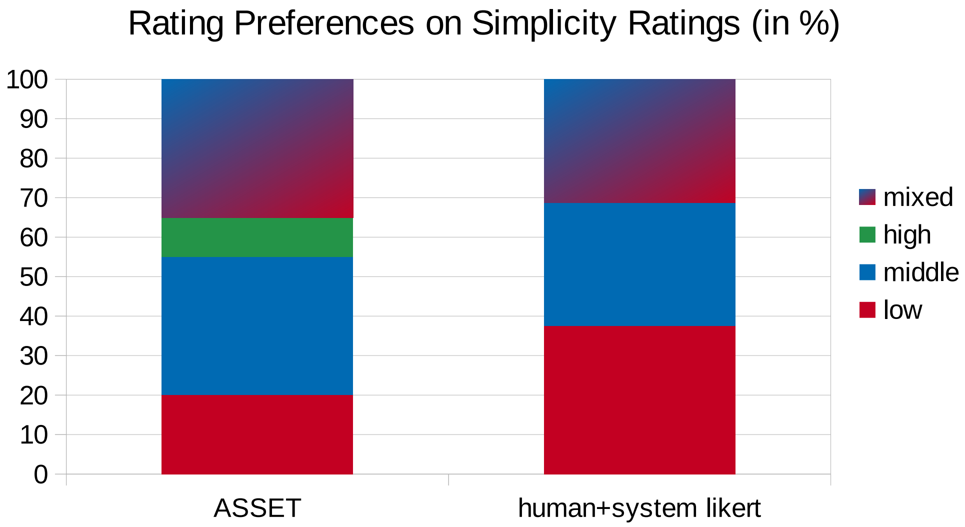
If **different interpretations** of the scales exist, the rater groups’ ratings significantly **differ** for sentence pairs in which the original and the simplified sentences are **not identical**.

- We compare ratings of annotators, who rated more than one no-change pair of ASSET or human- and system-likert.
  - ASSET: 20 rater
  - human+system likert: 16 rater
- Does annotators prefered a rating over all their no-change pairs?

Hypotheses

Hypothesis 4

If **different interpretations** of the scales exist, the rater groups’ ratings significantly **differ** for sentence pairs in which the original and the simplified sentences are **not identical**.



✓

Hypotheses

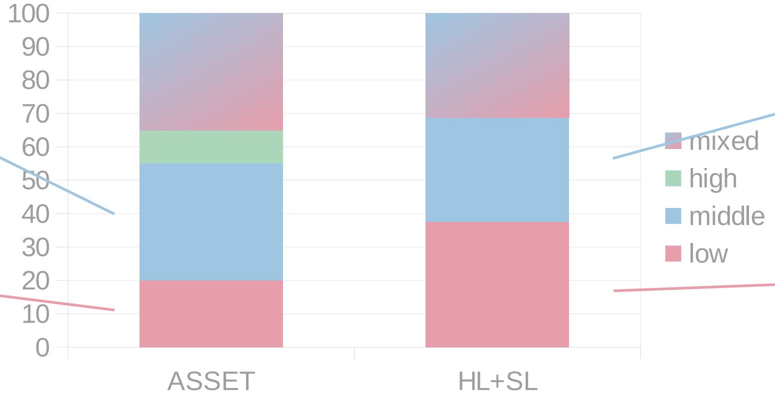
Hypothesis 4

If **different interpretations** of the scales exist, the rater groups' ratings significantly **differ** for sentence pairs in which the original and the simplified sentences are **not identical**.

- Average of rating scores of non-identical sentence pairs by annotators, who annotated more than one no-change pair.

n_{raters}: 6
$n_{ratings}$: 292
Avg.: 44.43
STD: 33.22

n_{raters}: 7
$n_{ratings}$: 571
Avg.: 35.58
STD: 37.24



n_{raters}: 5
$n_{ratings}$: 634
Avg.: 63.77
STD: 33.88

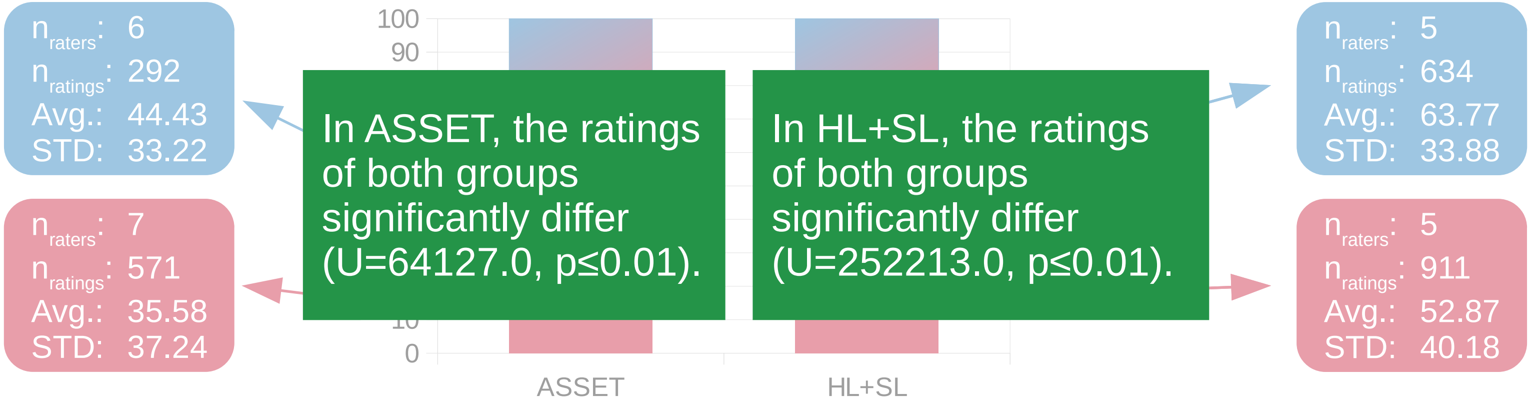
n_{raters}: 5
$n_{ratings}$: 911
Avg.: 52.87
STD: 40.18

Hypotheses

Hypothesis 4

If **different interpretations** of the scales exist, the rater groups’ ratings significantly **differ** for sentence pairs in which the original and the simplified sentences are **not identical**.

- Average of rating scores of non-identical sentence pairs by annotators, who annotated more than one no-change pair.



n_{raters}: 6
$n_{ratings}$: 292
Avg.: 44.43
STD: 33.22

n_{raters}: 7
$n_{ratings}$: 571
Avg.: 35.58
STD: 37.24

In ASSET, the ratings of both groups significantly differ (U=64127.0, p≤0.01).

In HL+SL, the ratings of both groups significantly differ (U=252213.0, p≤0.01).

n_{raters}: 5
$n_{ratings}$: 634
Avg.: 63.77
STD: 33.88

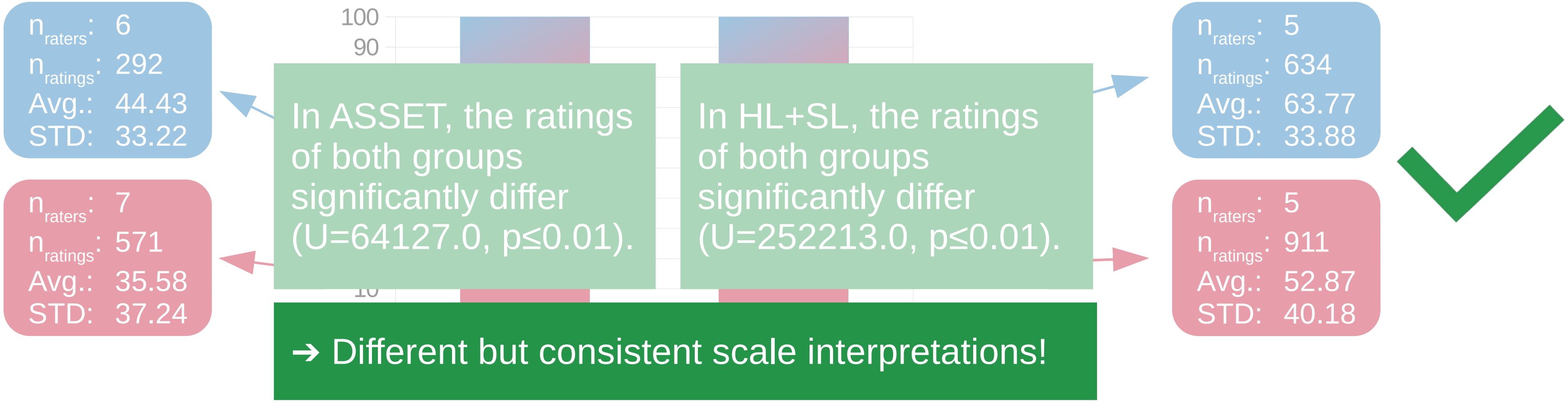
n_{raters}: 5
$n_{ratings}$: 911
Avg.: 52.87
STD: 40.18

Hypotheses

Hypothesis 4

If **different interpretations** of the scales exist, the rater groups' ratings significantly **differ** for sentence pairs in which the original and the simplified sentences are **not identical**.

- Average of rating scores of non-identical sentence pairs by annotators, who annotated more than one no-change pair.



n_{raters}: 6
$n_{ratings}$: 292
Avg.: 44.43
STD: 33.22

n_{raters}: 7
$n_{ratings}$: 571
Avg.: 35.58
STD: 37.24

In ASSET, the ratings of both groups significantly differ (U=64127.0, p≤0.01).

In HL+SL, the ratings of both groups significantly differ (U=252213.0, p≤0.01).

n_{raters}: 5
$n_{ratings}$: 634
Avg.: 63.77
STD: 33.88

n_{raters}: 5
$n_{ratings}$: 911
Avg.: 52.87
STD: 40.18

✓

➔ Different but consistent scale interpretations!

Conclusion

- Simplicity ratings on a rating scale from -2 to +2 rated by crowd workers and experts are consistent.
- Simplicity ratings on a rating scale from 0 to 100 with crowd workers are not consistent.
  - ➔ The scales are unclear.
  - ➔ Neutral element might be helpful to make them more clear.
  - ➔ Different scale interpretations exist.
- The rater can be grouped by their different scale interpretations.
  - – The ratings are consistent per group.
  - – The ratings of the groups differ significantly.
  - – The complete ratings of the corpora, might be split regarding the raters‘ scale interpretation.

Conclusion II

- We should be aware of different scale interpretations and inconsistency in the human judgements.
- The unclear scales might influence human and automatic text simplification evaluation.
- The analysis of no-change pairs can be used a sanity check, if the ratings scales are interpreted as expected.
- The meaning preservation scales on all corpora have more consistent ratings.
    - ➔ The meaning preservation scales seem easier to understand than the simplicity scales.

Discussion & Future Work

- More analysis on (best) ratings scales for text simplification is required.
- Recommendations or best practices on how to evaluate TS are required.
- Some open points for the best practices:
    - Which scale statement should we recommend?
    - Which scale size should we recommend?
    - Should we recommend a scale with or without neutral element?
    - Should we recommend to ask experts or crowd workers?
    - Should we recommend to validate the ratings and the scale interpretation of the annotators by analyzing the no-change pairs as kind of a sanity check?

References

- F. Alva-Manchego, C. Scarton, L. Specia, *Data-driven sentence simplification: Survey and benchmark*, Computational Linguistics 46 (2020a) 135–187. URL: https://aclanthology.org/2020.cl-1.4.
- F. Alva-Manchego, L. Martin, A. Bordes, C. Scarton, B. Sagot, L. Specia, *ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations*, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020b, pp. 4668–4679. URL: https://aclanthology.org/2020.acl-main.424.
- P. Laban, T. Schnabel, P. Bennett, M. A. Hearst (2021). *Keep it Simple: Unsupervised Simplification of Multi-Paragraph Text*, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2021, pp. 6365–6378. URL: https://aclanthology.org/2021.acl-long.498/
- M. Schwarzer, T. Tanprasert, D. Kauchak, *Improving human text simplification with sentence fusion*, in: Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15), Association for Computational Linguistics, Mexico City, Mexico, 2021, pp. 106–114. URL: https://aclanthology.org/2021.textgraphs-1.10.
- T. Scialom, L. Martin, J. Staiano, Éric Villemonte de la Clergerie, B. Sagot, *Rethinking automatic evaluation in sentence simplification*, 2021. arXiv:2104.07560.
- S. Štajner, M. Popović, H. Saggion, L. Specia, M. Fishel, *Shared task on quality assessment for text simplification*, in: Proceedings of the Workshop on Quality Assessment for Text Simplification (QATS), Association for Computational Linguistics, Portorož, Slovenia, 2016, pp. 22–37. URL: http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-QATS_Proceedings.pdf#page=28.
- E. Sulem, O. Abend, A. Rappoport, *Simple and effective text simplification using semantic and neural methods*, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018a, pp. 162–173. URL: https://aclanthology.org/P18-1016.
- E. Sulem, O. Abend, A. Rappoport, *Semantic structural evaluation for text simplification*, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018b, pp. 685–696. URL: https://aclanthology.org/N18-1063.
- W. Xu, C. Napoles, E. Pavlick, Q. Chen, C. Callison-Burch, *Optimizing statistical machine translation for text simplification*, Transactions of the Association for Computational Linguistics 4 (2016) 401–415. URL: https://aclanthology.org/Q16-1029.
- All icons are copied from Font Awesome. URL: https://fontawesome.com/license

When the Scale is Unclear

Analysis of the Interpretation of Rating Scales in Human Evaluation of Text Simplification

Regina Stodden
regina.stodden@hhu.de
Heinrich Heine University Düsseldorf, Germany
NRW-Forschungskolleg Online Partizipation

The code of the analysis will be available on github soon:
https://github.com/rstodden/TS-scale-interpretation