[SIAN GOODING, EKATERINA KOCHMAR, ALAN BLACKWELL, ADVAITH SAKAR]

Comparative judgements are more consistent than binary classification for labelling word complexity

MOTIVATION

Lexical simplification (LS) systems replace complex words with simpler alternatives.

They aimed to **ameliorate** the situation.
↓
They aimed to **improve** the situation.

Complex word identification (CWI) is a sub-task of LS concerned with the detection of words in need of simplification. Current datasets for this task have low levels of annotator agreement:

Data	*IAA Statistic*	*Interpretation*
2012	κ = 0.386, 0.398	minimal agreement
2016	α = 0.244	inconclusive
2018	1% unanimous	idiosyncratic

Current drawbacks:
- Homogeneity of the annotator group is usually not controlled for
- Represented as a binary, not continuous task

HYPOTHESIS

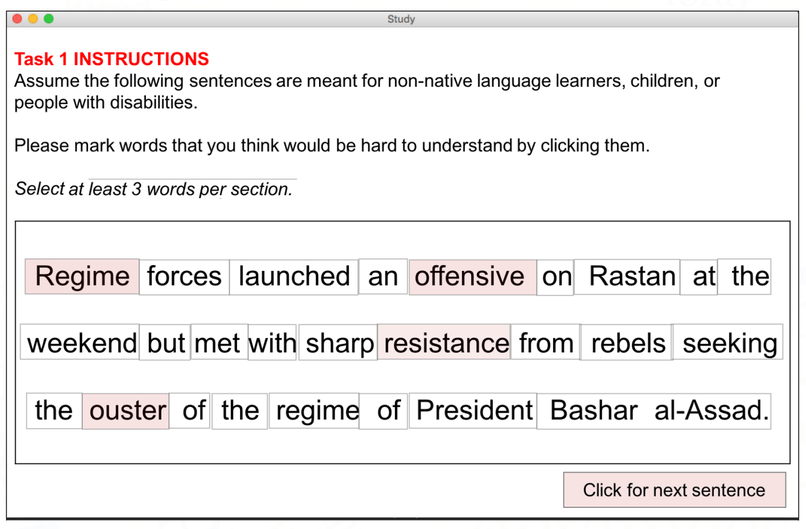
Do comparative judgments for CWI lead to higher inter-annotator agreement and higher quality labelled data than binary?

Additional Questions

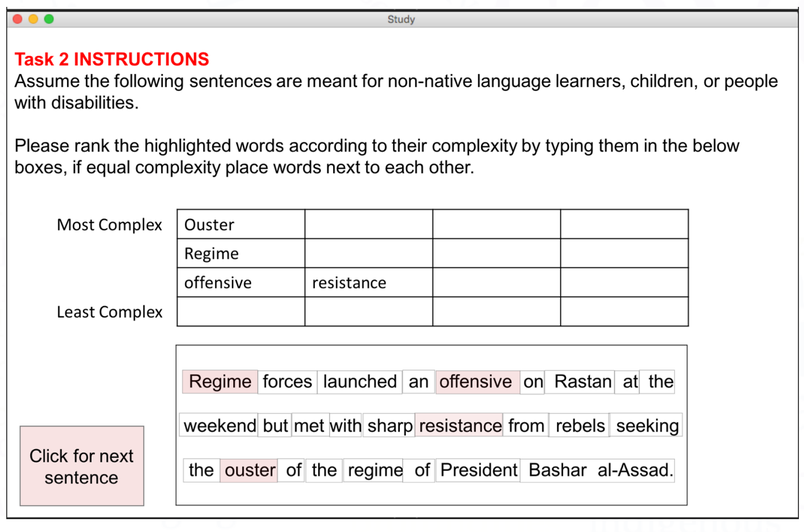
- Does controlling for the homogeneity of the group of annotators contribute to higher agreement?
- Can comparative judgments be made in a significantly shorter period of time than binary judgments for word complexity?

INTERFACE

Within-subjects design used; annotators were asked to label word complexity in continuous vs binary fashion:



Binary Judgement Interface



Continuous Judgement Interface

STUDY

Annotator group:

- 30 annotators
- 20 sentences
- ~25 minutes per participant

- Same first language (English)
- Same level of educational background
- Similar age range (21-30)

Annotators were presented with professionally written news sentences from Yimam et al. (2017) dataset:

- 10 sentences presented per interface
- Chosen to contain a range of word complexities based on the number of annotations from Yimam et al. (2017):

hard ∈ [10, 20]	*politicizing* (14)
medium ∈ [6, 9]	*warily* (9)
low ∈ [1, 5]	*trip* (2)

RESULTS

	Comparative Judgement	*Binary Judgement*
Kappa Coefficient	0.6775 *(moderate)*	0.3937 *(minimal)*
Alpha Coefficient	0.6821	0.4960
Avg Time (s)	28.77	38.69

- According to Cohen (1968), our Kappa results indicate *moderate* agreement for comparative judgements and *minimal* for the binary annotation task supporting our hypothesis

Binary judgement :
- 62 distinct words from 10 sentences marked as complex by annotators
- Higher agreement than previously reported studies: α = 0.496 vs α = 0.244 in Paetzold and Specia (2016)

Comparative judgement:
- Higher agreement than previously reported studies: κ = 0.6775 vs κ = 0.398 in Specia et al. (2012)
- 9.92s less time per sentence on average than binary judgements

CONCLUSIONS

- This study demonstrates the advantage of annotating datasets using *comparative judgments rather than binary classifications*, both for *efficiency* and *accuracy*.
- Our results also show *higher agreement coefficients* for both binary and relative judgment tasks when compared to previously collected datasets.
- Our work supports the case that the concept of word complexity, and thus the level of agreement, is *aligned between individuals that share a common background.*

Future steps for this research include:
1. more thorough investigation of effects of annotator group homogeneity on the inter-annotator agreement
2. more detailed larger study of the efficiency of the comparative judgments as opposed to binary judgments

- Our results are applicable to other natural language tasks where specific user experiences like simplicity can be modelled as an ordering so that they may be optimized or personalized.

CONTACT INFORMATION

Sian Gooding, Ekaterina Kochmar, Alan Blackwell, Advaith Sakar [shg36, ek358, afb21] @cam.ac.uk, advaith@microsoft.com