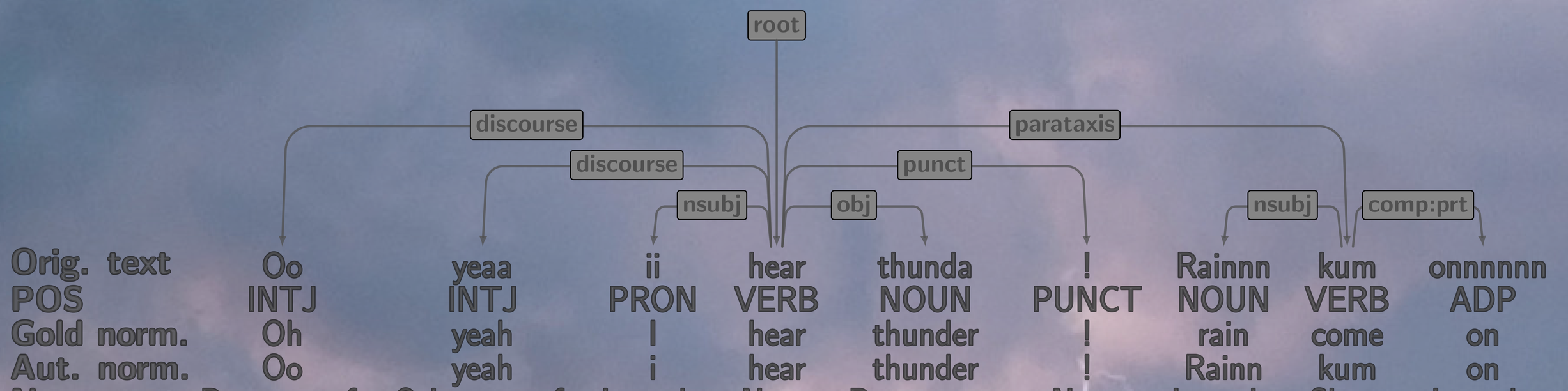
In-depth Analysis of the Effect of Lexical Normalization on the Dependency Parsing of Social Media

Rob van der Goot

university of groningen

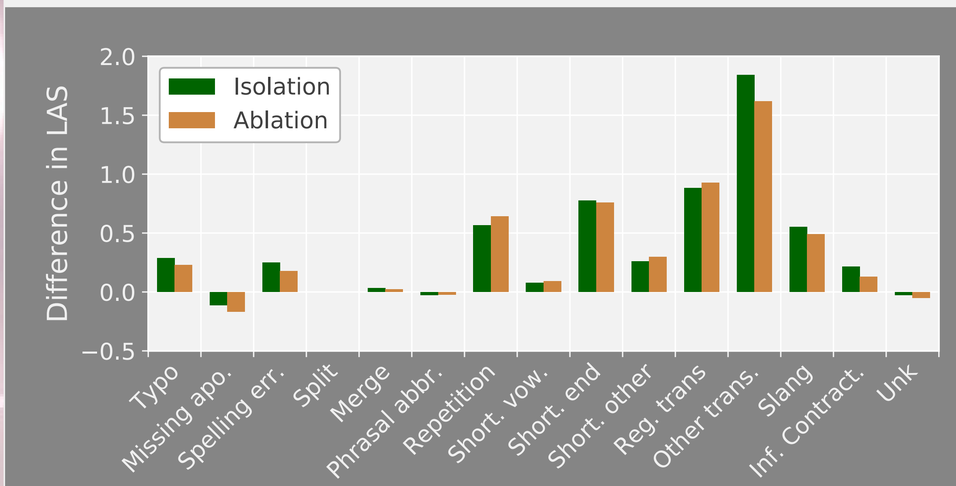
IT UNIVERSITY OF CPH



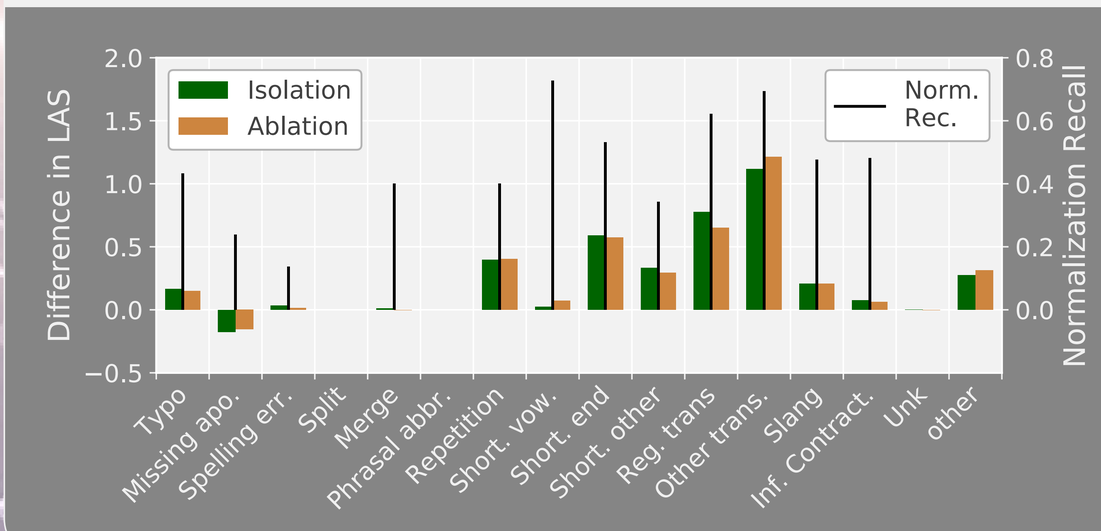
Taxonomy of Normalization Categories

	Category	Freq.	%
	None	3,743	81.76
Unintentional	Typo	30	0.66
	Missing apostrophe	176	3.84
	Spelling error	44	0.96
	Merge	10	0.22
Anomalies	Phrasal abbreviation	2	0.04
Intentional	Repetition	90	1.97
	Shortening. vowels	22	0.48
	Shortening end	64	1.40
	Shortening other	35	0.76
	Regular transformation	66	1.44
	Other transformation	186	4.06
	Slang	42	0.92
	Informal Contraction	56	1.22
	Unk	12	0.26

Results Gold Normalization

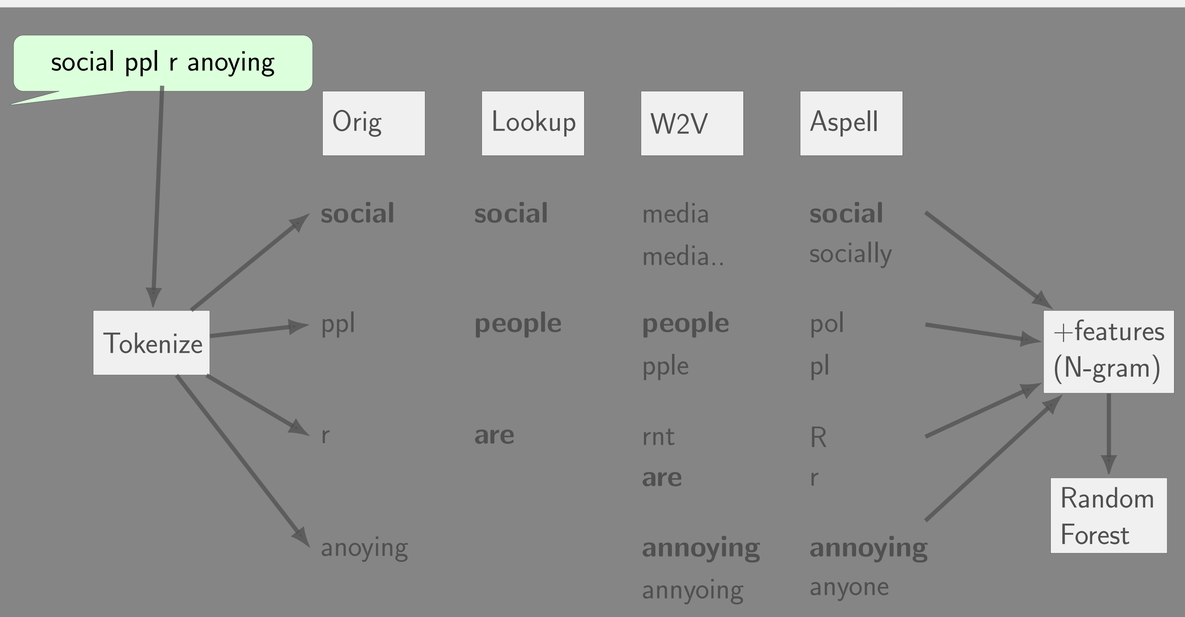


Results Automatic Normalization



MoNoise

TRY IT: www.robvandergoot.com/monoise



Conclusions

- Gold: the most important categories are ‘other transf.’, ‘regular transf.’ and ‘shortening end’
- Automatic: most potential for improvement for ‘other transf.’ and ‘slang’ categories
- Some categories are not beneficial for syntactic tasks
- Annotation guidelines matter!
- Novel dataset (dev+test) with all layers is released

Source & Data:
`https://bitbucket.org/robvanderg/taxeval/`