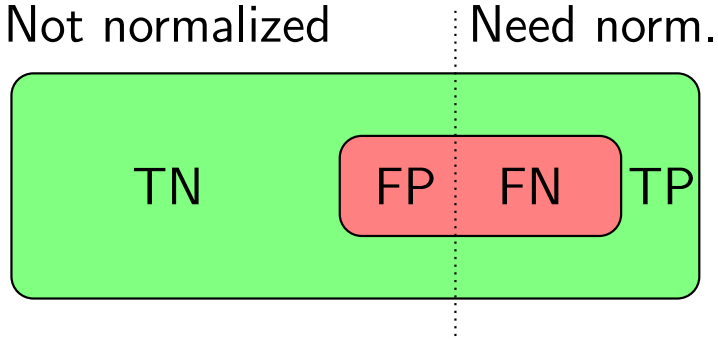

Not normalized
Need norm.
TN
FP
FN
TP

A Taxonomy for In-depth Evaluation of Normalization for User Generated Content

Rob van der Goot, Rik van Noord & Gertjan van Noord
University of Groningen
r.van.der.goot@rug.nl

26-01-2018

Lexical Normalization

- Pre-tokenized
- Word-word replacements

Lexical Normalization

RT	<USERNAME>	I	mis	bein	antisocial	:(
RT	<USERNAME>	I	miss	being	antisocial	:(

nt	havin	friens	was	the	bestest
not	having	friends	was	the	best

Taxonomy

Why?

- Find strengths/weaknesses normalization models
- Test effect different categories of normalization actions on other task
- Filter out undesirable categories

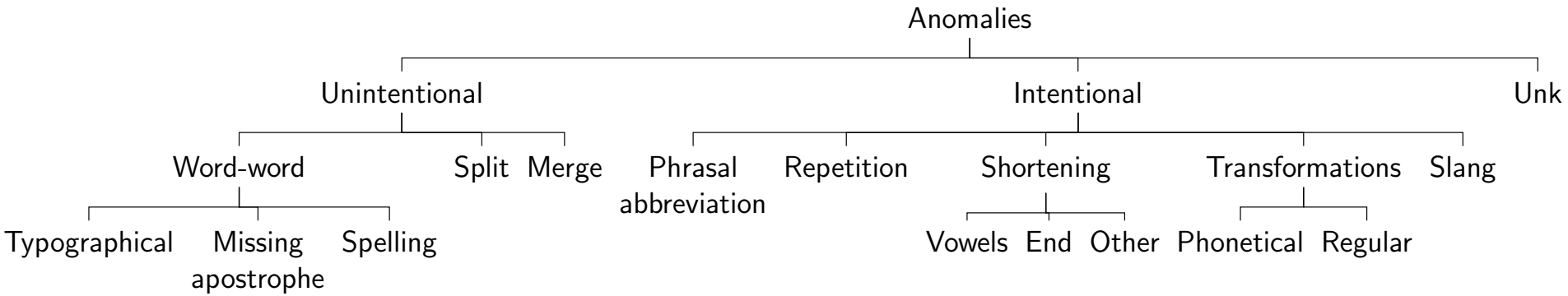
Taxonomy

Got it?

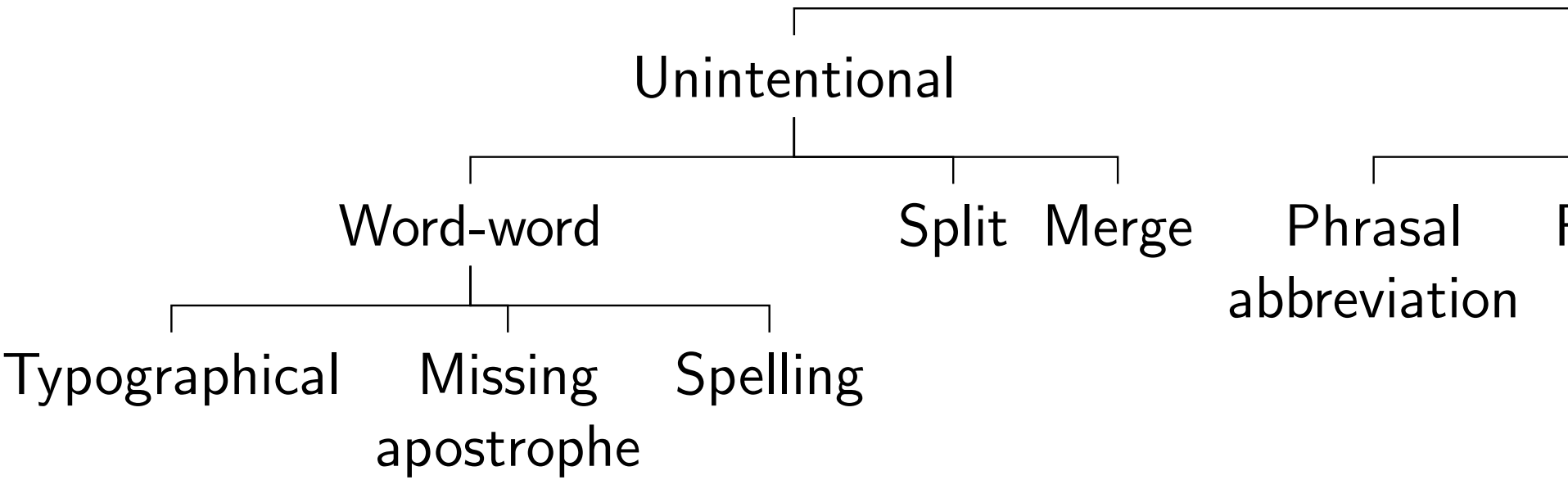
RT	<USERNAME>	I	mis	bein	antisocial	:(
RT	<USERNAME>	I	miss	being	antisocial	:(
		X	X			

nt	havin	friens	was	the	bestest
not	having	friends	was	the	best
X	X	X			X

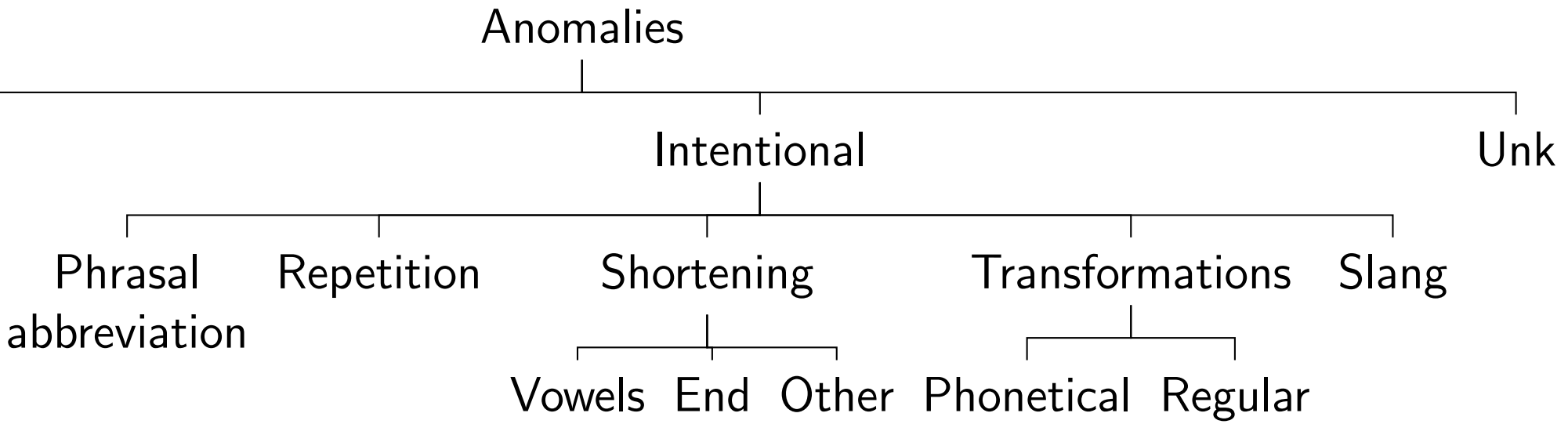
Taxonomy


Anomalies
Unintentional
Intentional
Unk
Word-word
Split
Merge
Phrasal abbreviation
Repetition
Shortening
Transformations
Slang
Typographical
Missing apostrophe
Spelling
Vowels
End
Other
Phonetical
Regular

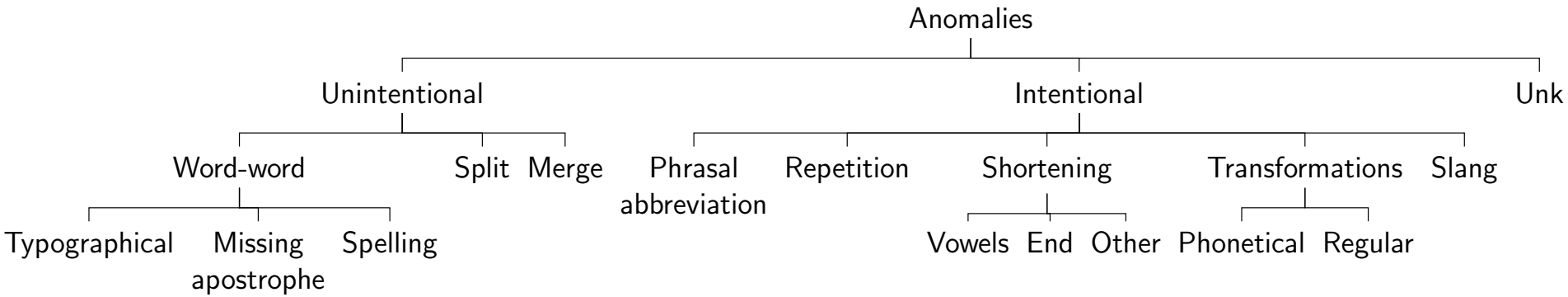
Taxonomy


Unintentional
Word-word
Split
Merge
Phrasal abbreviation
Typographical
Missing apostrophe
Spelling

Taxonomy


Anomalies
Intentional
Unk
Phrasal abbreviation
Repetition
Shortening
Transformations
Slang
Vowels
End
Other
Phonetical
Regular

Taxonomy


Anomalies
Unintentional
Intentional
Unk
Word-word
Split
Merge
Phrasal abbreviation
Repetition
Shortening
Transformations
Slang
Typographical
Missing apostrophe
Spelling
Vowels
End
Other
Phonetical
Regular

Taxonomy

1. Typographical error
   spirite↦spirit, complaing↦complaining, throwg↦throw
2. Missing apostrophe
   im↦i'm, yall↦y'all, microsofts↦microsoft's
3. Spelling error
   favourite↦favorite, dieing↦dying, theirselves↦themselves
4. Split
   pre order↦preorder, screen shot↦screenshot
5. Merge
   alot↦a lot, nomore↦no more, appstore↦app store
6. Phrasal abbreviation
   lol↦laughing out loud, pmsl↦pissing myself laughing
7. Repetition
   soooo↦so, weiiiiird↦weird

Taxonomy

8. Shortening vowels
   pls↦please, wrked↦worked, rmx↦remix
9. Shortening end
   gon↦gonna, congrats↦congratulations, g↦girl
10. Shortening other
    cause↦because, smth↦something, tl↦timeline,
11. Phonetic transformation
    hackd↦hacked, gentille↦gentle, rizky↦risky
12. Regular transformation
    foolin↦fooling, wateva↦whatever, droppin↦dropping
13. Slang
    cuz↦because, fina↦going to, plz↦please
14. Unknown
    skepta↦sunglasses, putos↦photos

Annotation

Dataset:

- LexNorm2015
- Shared Task 2nd Workshop on Noisy user-generated Text (W-NUT)
- Train: 2,950 Tweets / 44,385 words
- 8.9% normalized

Annotation

Guidelines

- On unique replacement pairs
- First annotator: all train data (1,204 pairs)
- Second annotator: 150 replacement pairs
- One category per pair

Annotation

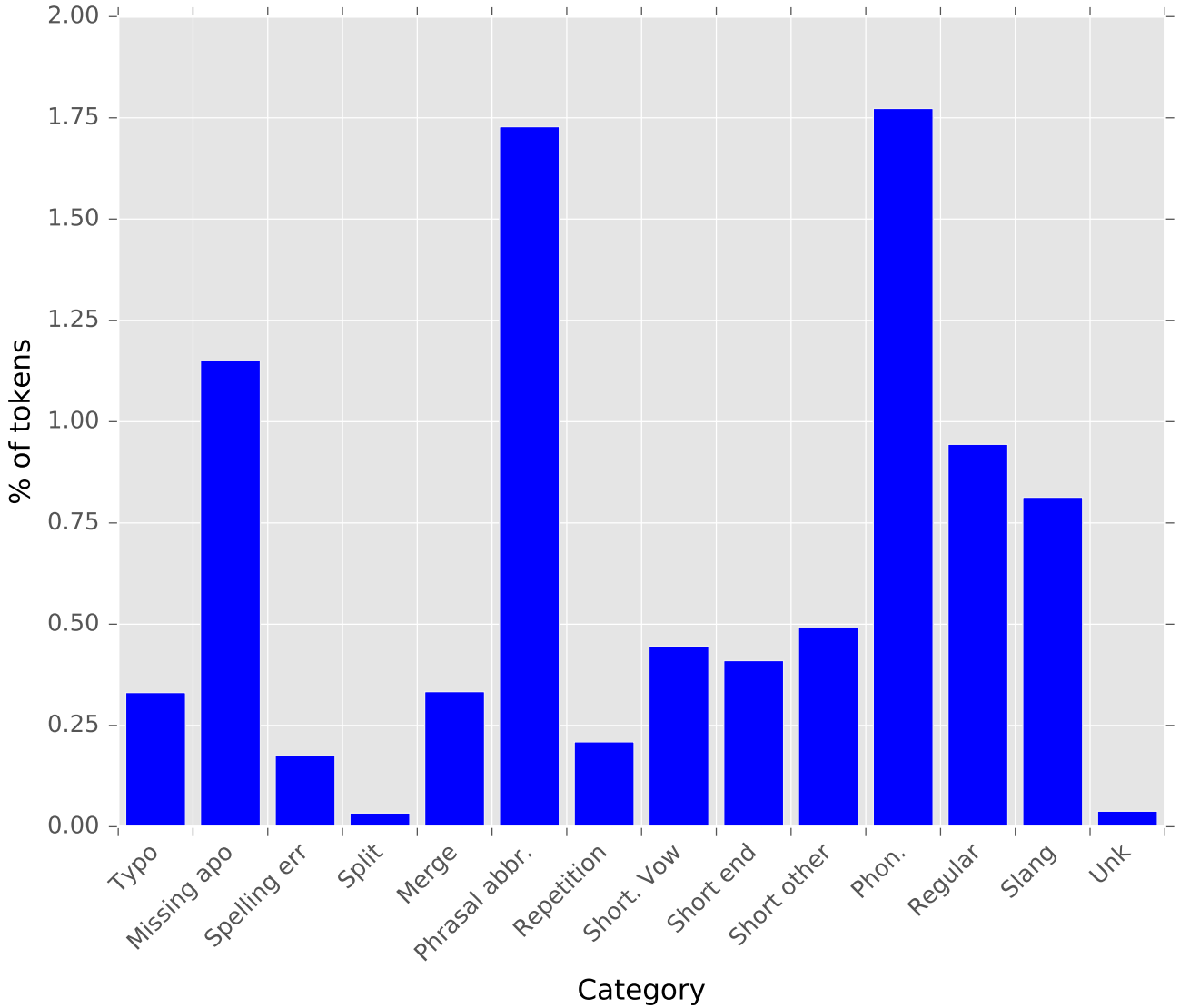
Disagreements: ($\kappa = 0.807$)

9	7	diffffff	different
8	9	t	to
12	8	talkn	talking
9	11	custa	custand
1	3	shat	shit
6	7	tunee	tune
1	8	downgrding	downgrading
12	8	thx	thank
11	8	yur	your
12	8	yor	your
1	7	wearr	wear
10	6	gf	girlfriend

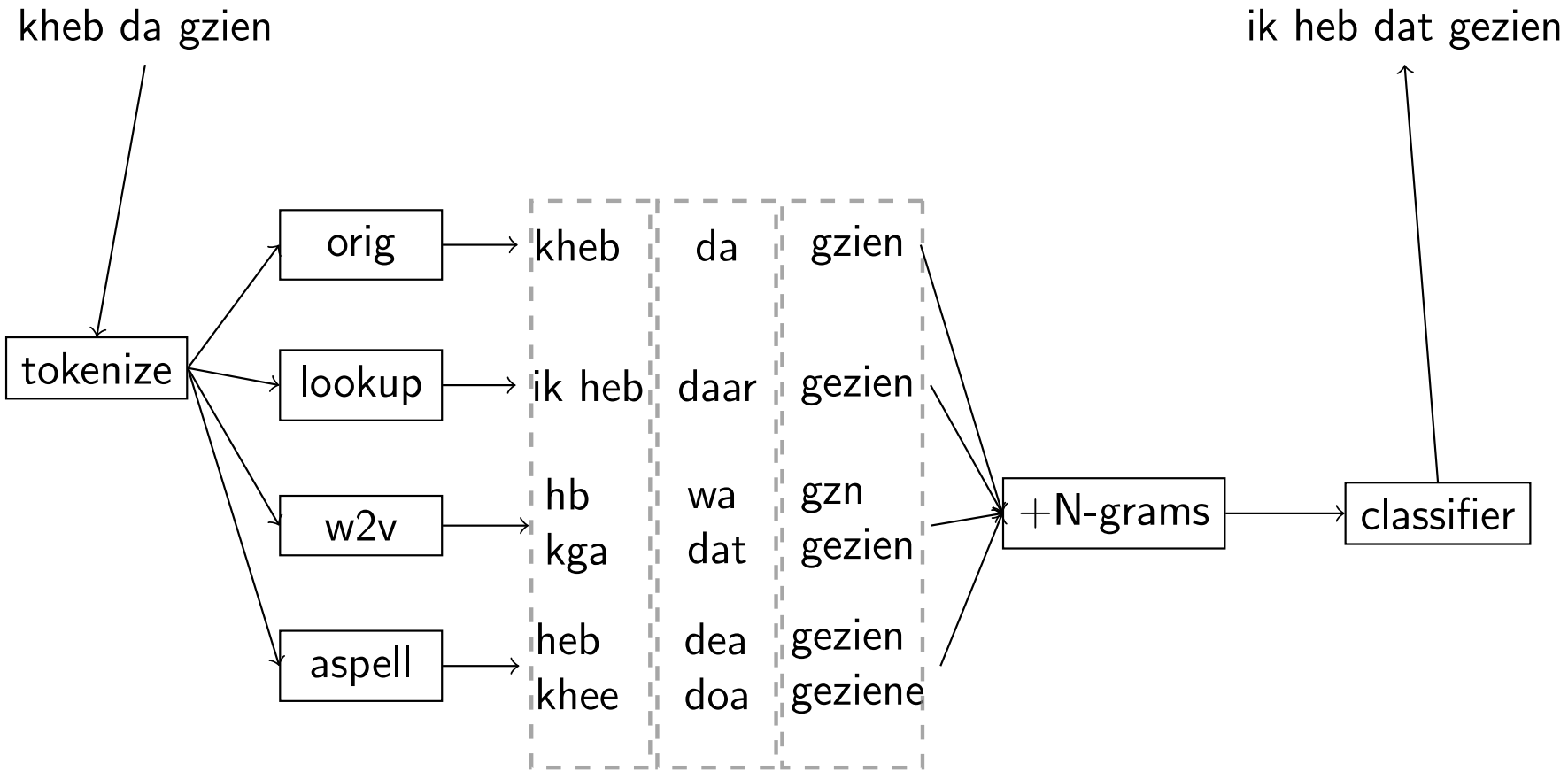
Annotation

Most	social	pple	r	troublesome	..	lol
most	social	people	are	troublesome	..	laughing out loud
0	0	short vow	phon.	0	..	phrasal abbrev.

Annotation


2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
% of tokens
Typo
Missing apo
Spelling err
Split
Merge
Phrasal abbr.
Repetition
Short. Vow
Short end
Short other
Phon.
Regular
Slang
Unk
Category

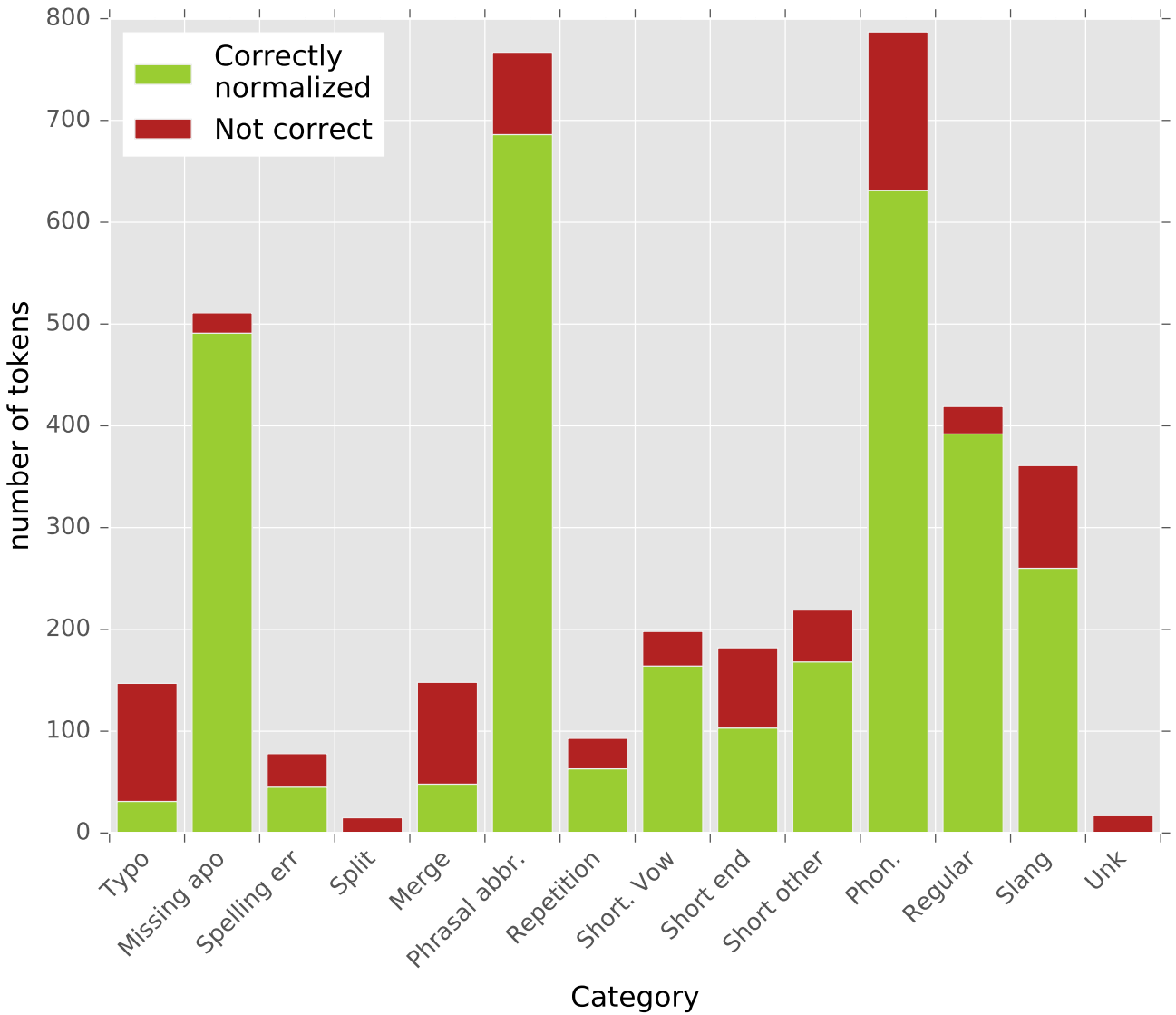
Evaluation MoNoise



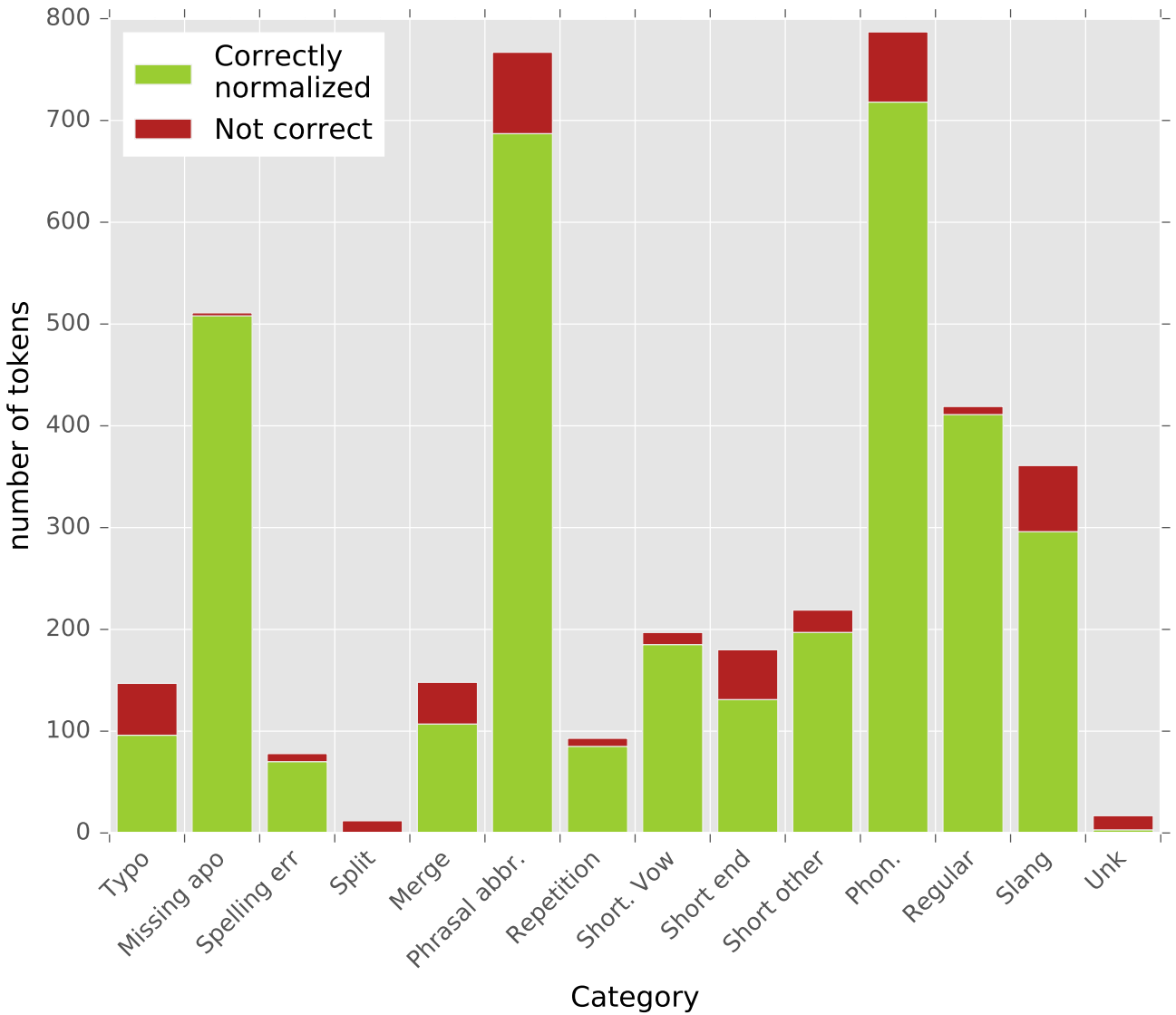
Evaluation MoNoise

`www.let.rug.nl/rob/monoise`

Evaluation MoNoise



Evaluation MoNoise



Evaluation MoNoise

`https://bitbucket.org/robvanderg/normtax`
`https://bitbucket.org/robvanderg/monoise`

Extensions

- Effect of different normalization modules on different categories
- Classify other corpora automatically?
- Test effect of each category on end-task (automatic vs gold)