COMPARING IDENTIFIERS AND COMMENTS IN ENGINEERED AND NON-ENGINEERED CODE: A LARGE-SCALE EMPIRICAL STUDY

OTAVIO LEMOS (UNIFESP), MARCELO SUZUKI (UNIFESP),
ADRIANO DE PAULA (UNIFESP), AND CLAIRE LE GOES (CMU)

UNIFESP
UNIVERSIDADE FEDERAL DE SÃO PAULO
1933

Carnegie Mellon University

supported by: FAPESP

PROMISE

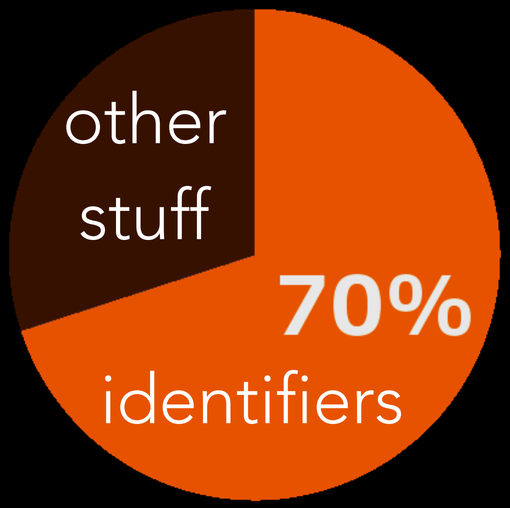
solid software engineering practices

⇩

identifiers and **comments**

CONTEXT

source-code



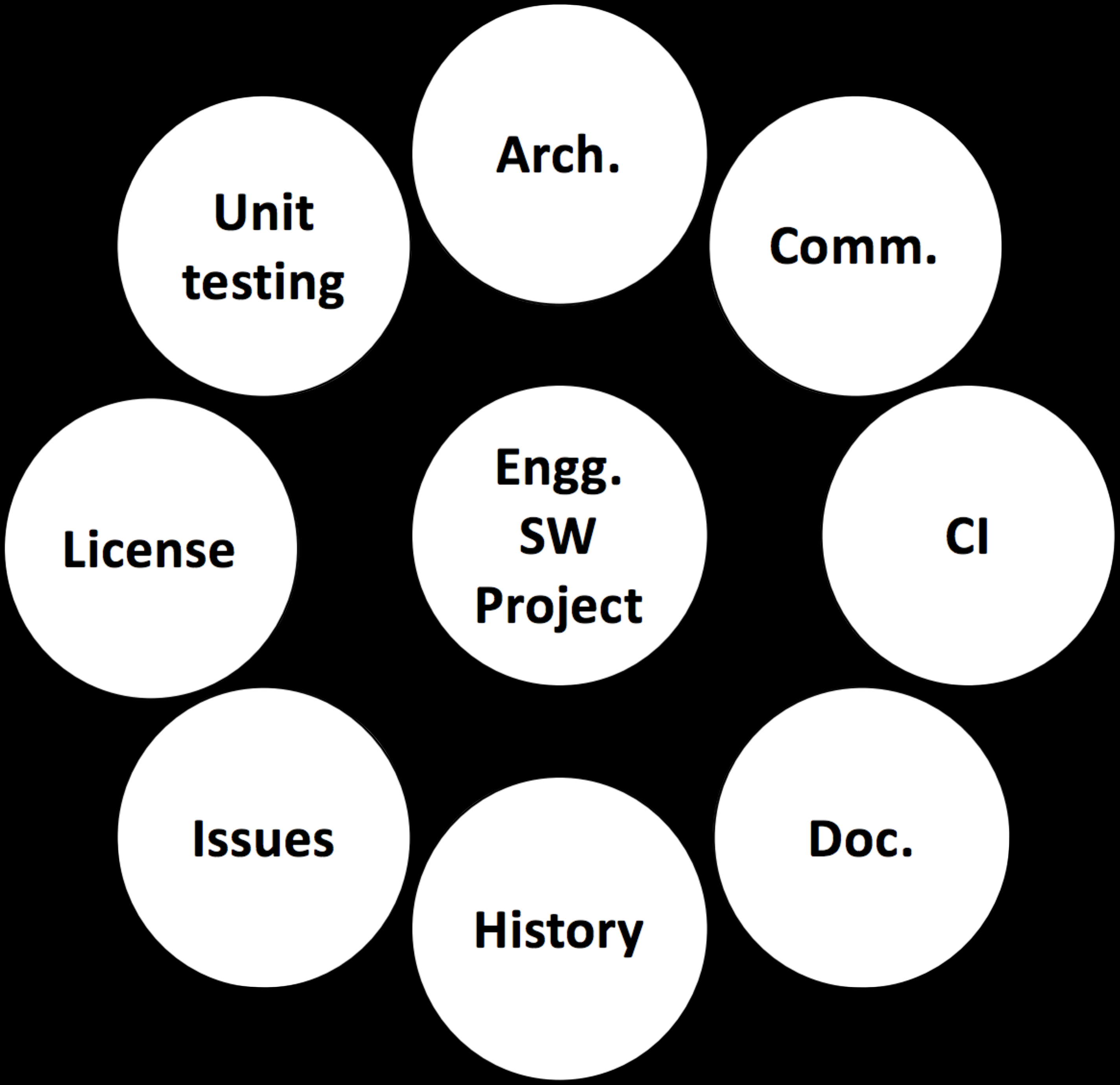
- after the code itself, **comments** → main source of documentation
- **engineered** vs. **non-engineered** code (*Munaiah* et al. 2017)

RESEARCH QUESTIONS

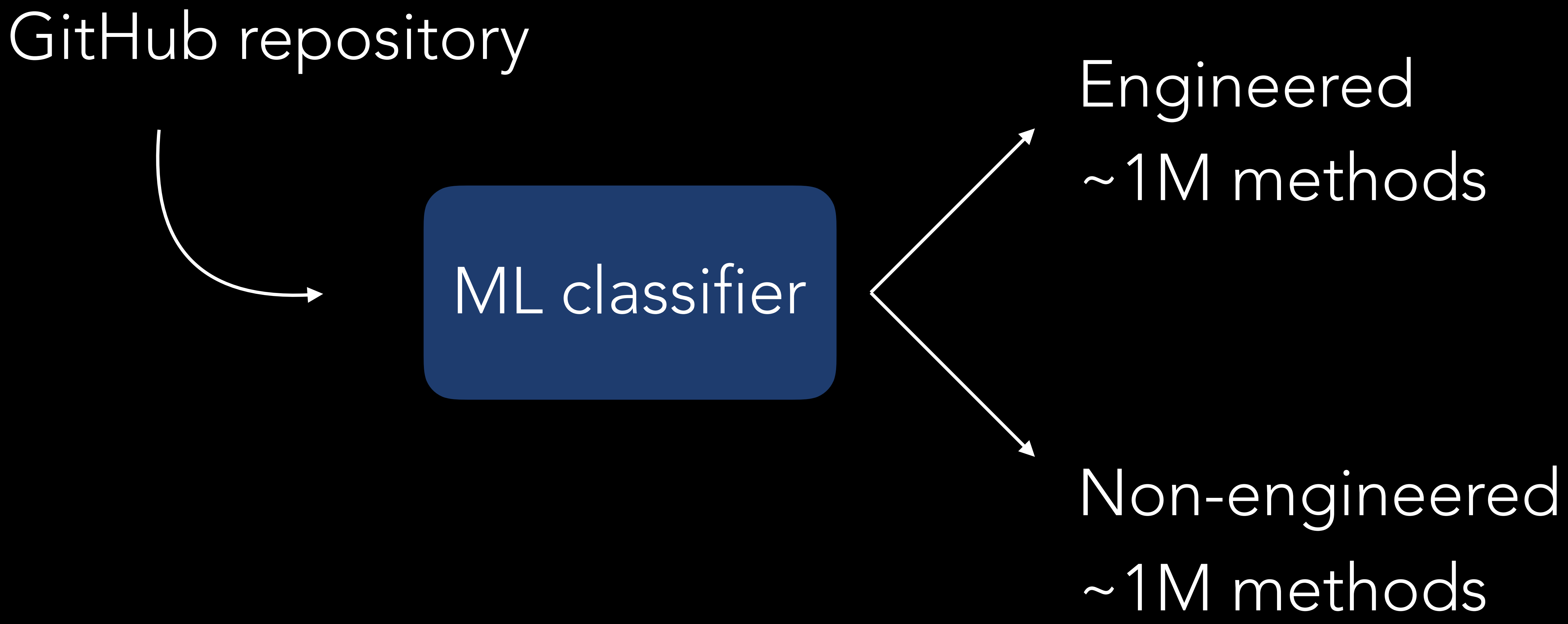
- ***RQ1***. Are identifiers significantly different in engineered versus non-engineered code? If so, what are these differences?

- ***RQ2***. Are comments significantly different in engineered versus non-engineered code? If so, what are these differences?

MUNAIAH ET AL.



MUNAIAH ET AL.



GitHub repository

ML classifier

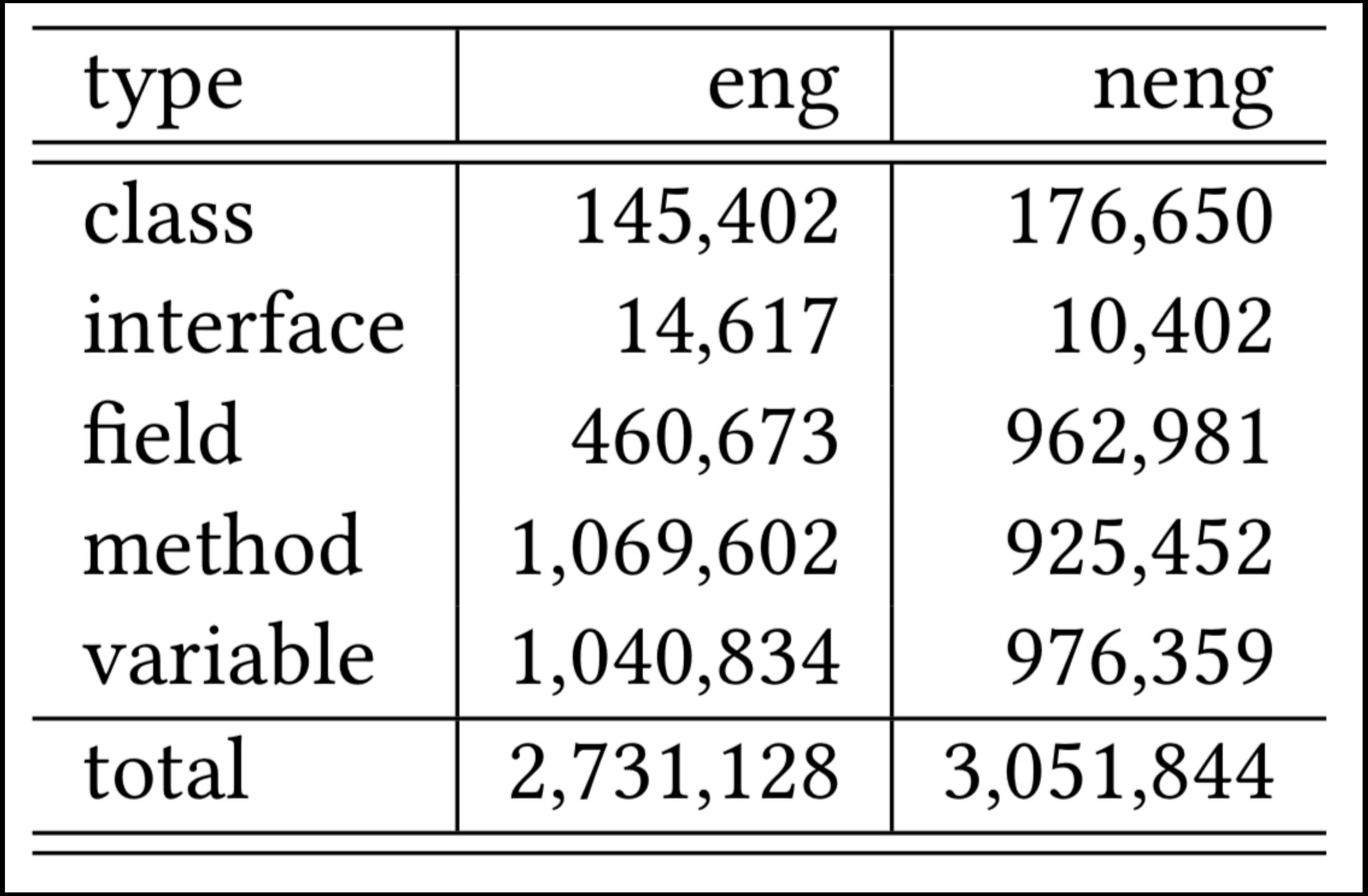
Engineered

~1M methods

Non-engineered

~1M methods

IDENTIFIER ANALYSIS

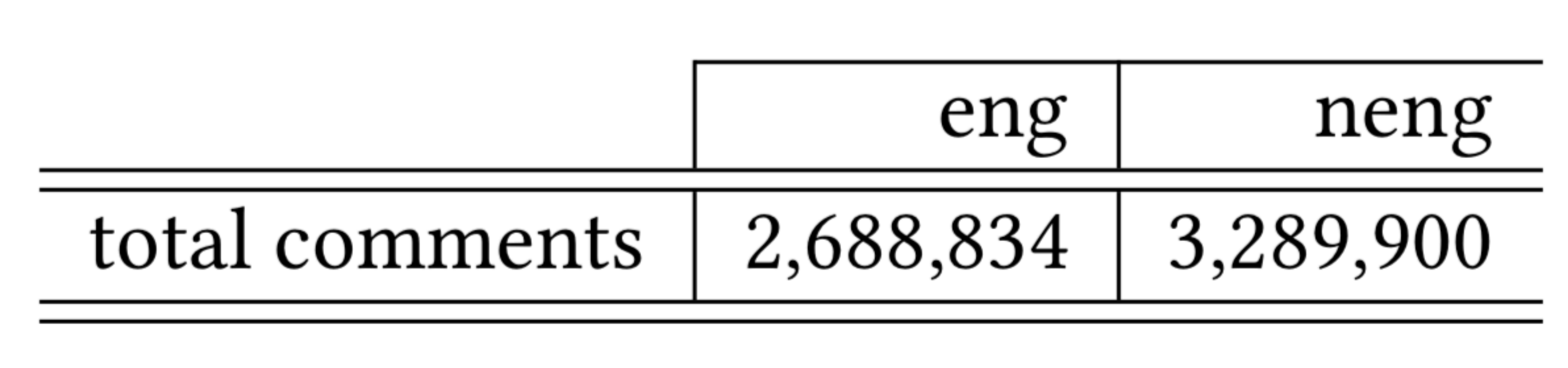


type	eng	neng
class	145,402	176,650
interface	14,617	10,402
field	460,673	962,981
method	1,069,602	925,452
variable	1,040,834	976,359
total	2,731,128	3,051,844

IDENTIFIER ANALYSIS

- Length
- Format (_lead_underscore; UPPER_UNDERSCORE; lower_underscore; lowerCamel; UpperCamel; alllower; ALLUPPER)
- Number of words
- English

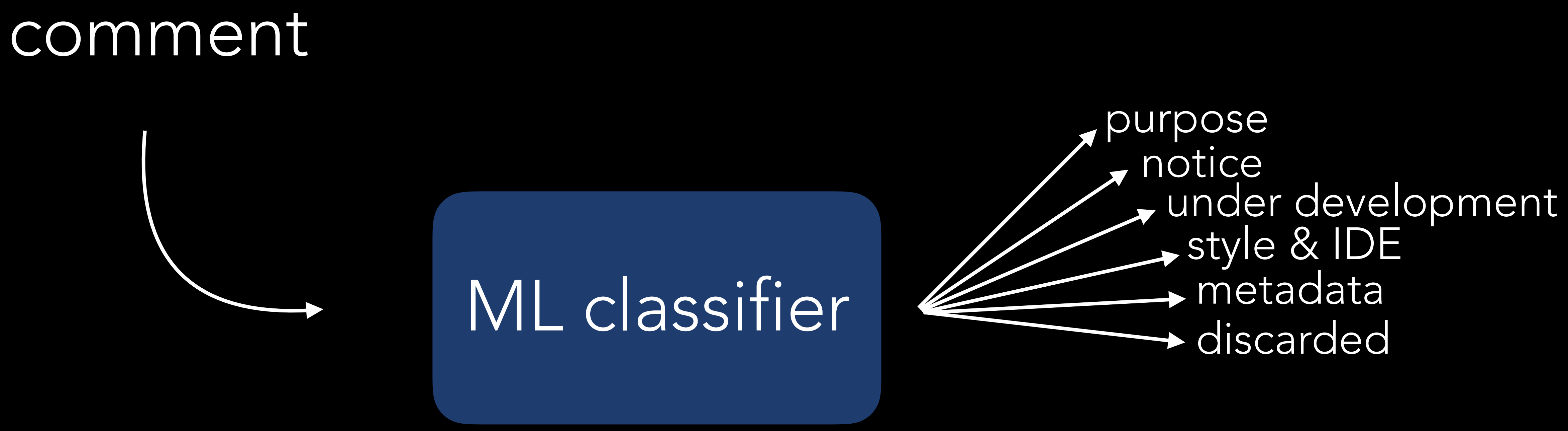
COMMENT ANALYSIS



	eng	neng
total comments	2,688,834	3,289,900

COMMENT ANALYSIS

(Pascarella & Bacchelli)



comment

ML classifier

purpose

notice

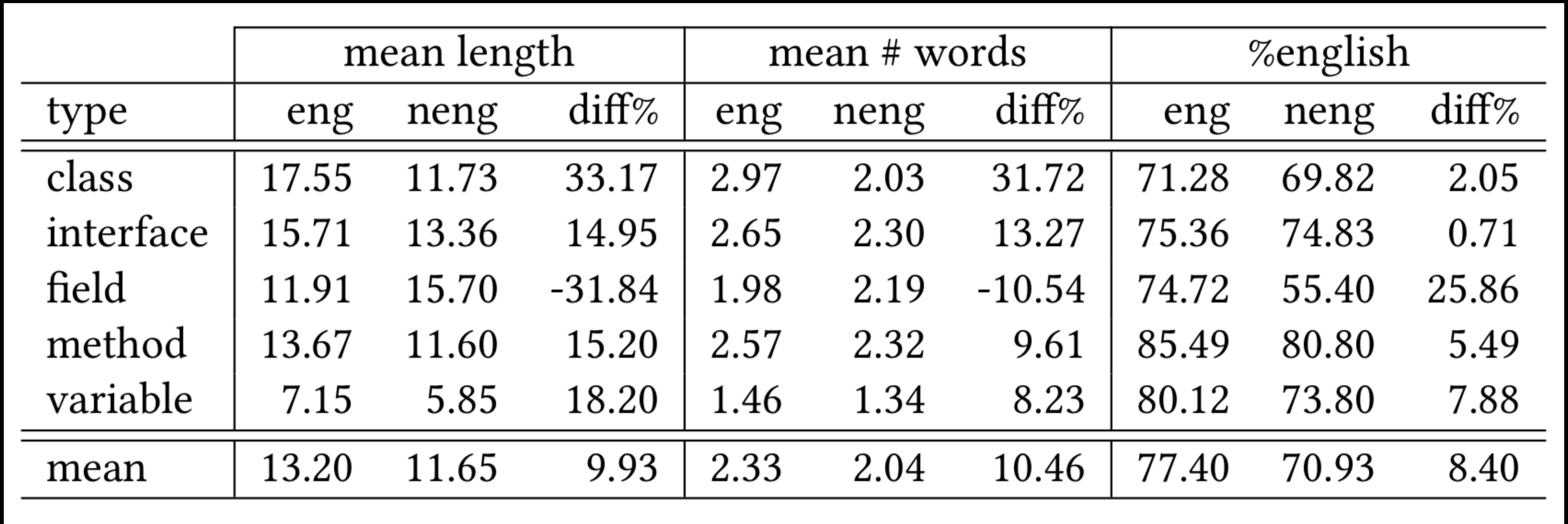
under development

style & IDE

metadata

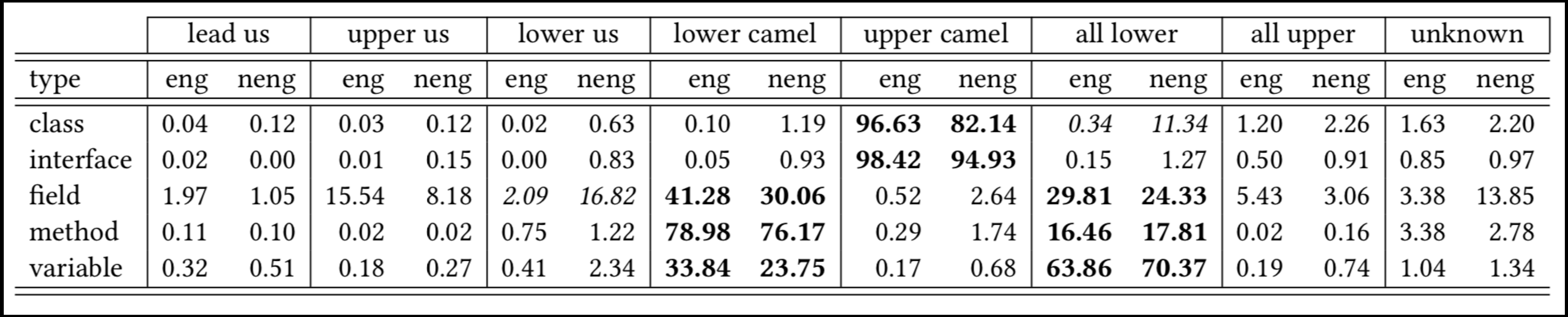
discarded

RESULTS: IDENTIFIERS



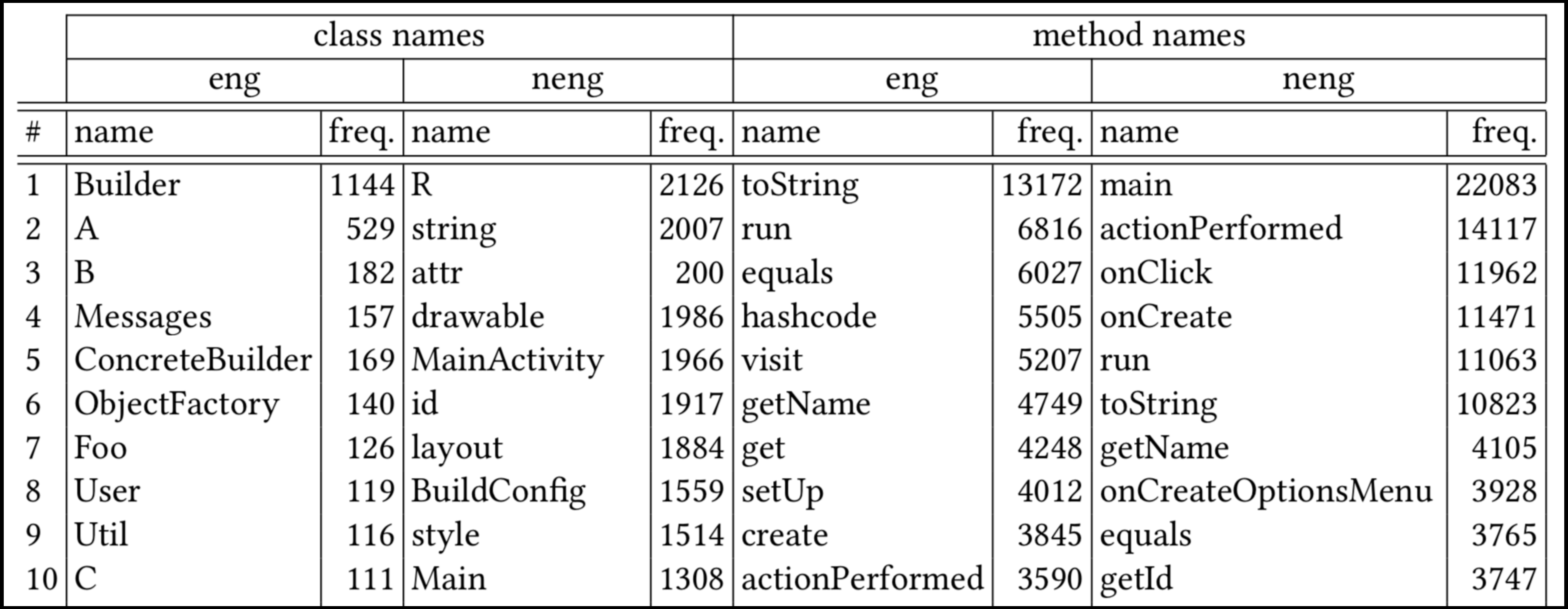
	mean length			mean # words			%english		
type	eng	neng	diff%	eng	neng	diff%	eng	neng	diff%
class	17.55	11.73	33.17	2.97	2.03	31.72	71.28	69.82	2.05
interface	15.71	13.36	14.95	2.65	2.30	13.27	75.36	74.83	0.71
field	11.91	15.70	-31.84	1.98	2.19	-10.54	74.72	55.40	25.86
method	13.67	11.60	15.20	2.57	2.32	9.61	85.49	80.80	5.49
variable	7.15	5.85	18.20	1.46	1.34	8.23	80.12	73.80	7.88
mean	13.20	11.65	9.93	2.33	2.04	10.46	77.40	70.93	8.40

RESULTS: IDENTIFIERS



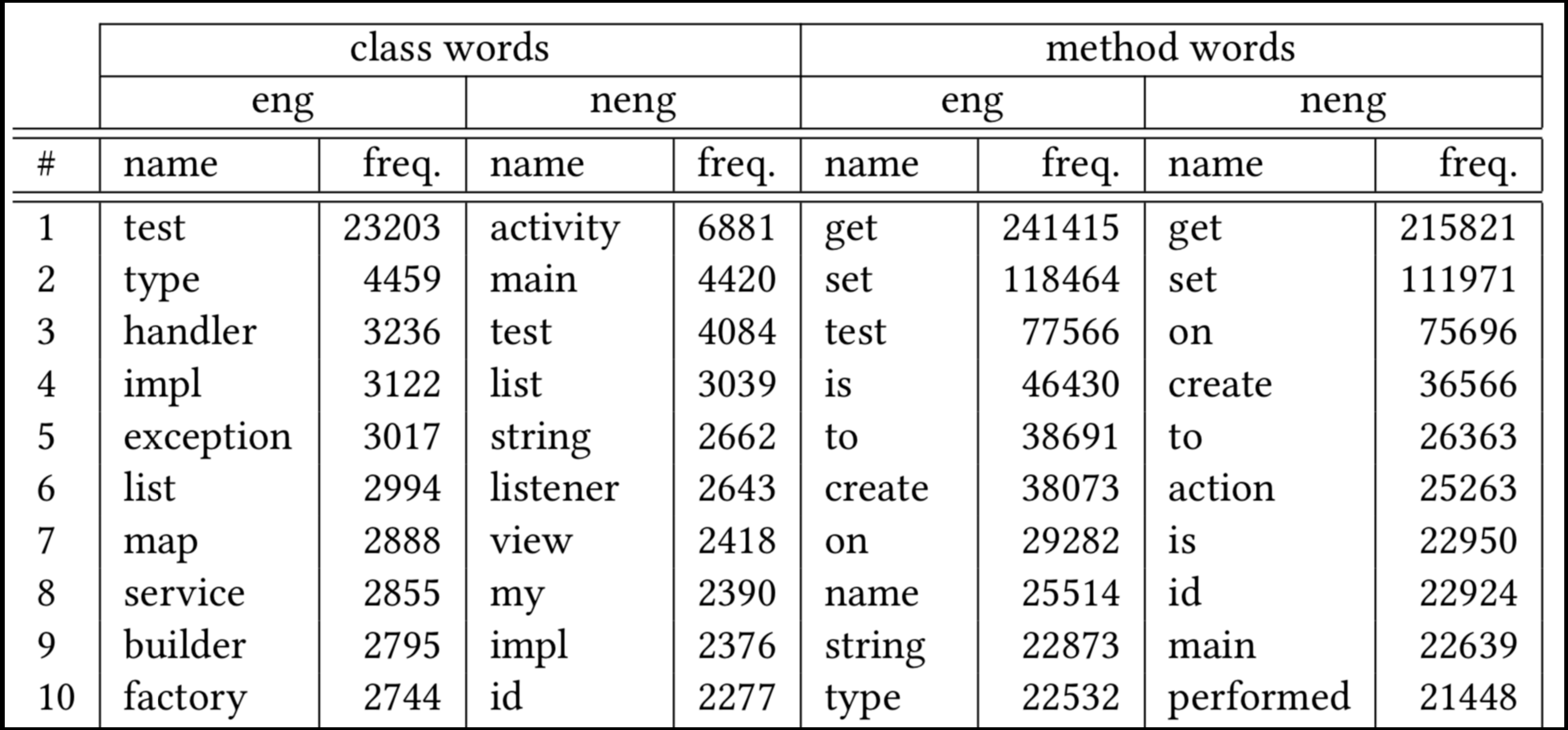
	lead us		upper us		lower us		lower camel		upper camel		all lower		all upper		unknown	
type	eng	neng	eng	neng	eng	neng	eng	neng	eng	neng	eng	neng	eng	neng	eng	neng
class	0.04	0.12	0.03	0.12	0.02	0.63	0.10	1.19	**96.63**	**82.14**	*0.34*	*11.34*	1.20	2.26	1.63	2.20
interface	0.02	0.00	0.01	0.15	0.00	0.83	0.05	0.93	**98.42**	**94.93**	0.15	1.27	0.50	0.91	0.85	0.97
field	1.97	1.05	15.54	8.18	*2.09*	*16.82*	**41.28**	**30.06**	0.52	2.64	**29.81**	**24.33**	5.43	3.06	3.38	13.85
method	0.11	0.10	0.02	0.02	0.75	1.22	**78.98**	**76.17**	0.29	1.74	**16.46**	**17.81**	0.02	0.16	3.38	2.78
variable	0.32	0.51	0.18	0.27	0.41	2.34	**33.84**	**23.75**	0.17	0.68	**63.86**	**70.37**	0.19	0.74	1.04	1.34

RESULTS: IDENTIFIERS



	class names				method names			
	eng		neng		eng		neng	
#	name	freq.	name	freq.	name	freq.	name	freq.
1	Builder	1144	R	2126	toString	13172	main	22083
2	A	529	string	2007	run	6816	actionPerformed	14117
3	B	182	attr	200	equals	6027	onClick	11962
4	Messages	157	drawable	1986	hashcode	5505	onCreate	11471
5	ConcreteBuilder	169	MainActivity	1966	visit	5207	run	11063
6	ObjectFactory	140	id	1917	getName	4749	toString	10823
7	Foo	126	layout	1884	get	4248	getName	4105
8	User	119	BuildConfig	1559	setUp	4012	onCreateOptionsMenu	3928
9	Util	116	style	1514	create	3845	equals	3765
10	C	111	Main	1308	actionPerformed	3590	getId	3747

RESULTS: IDENTIFIERS

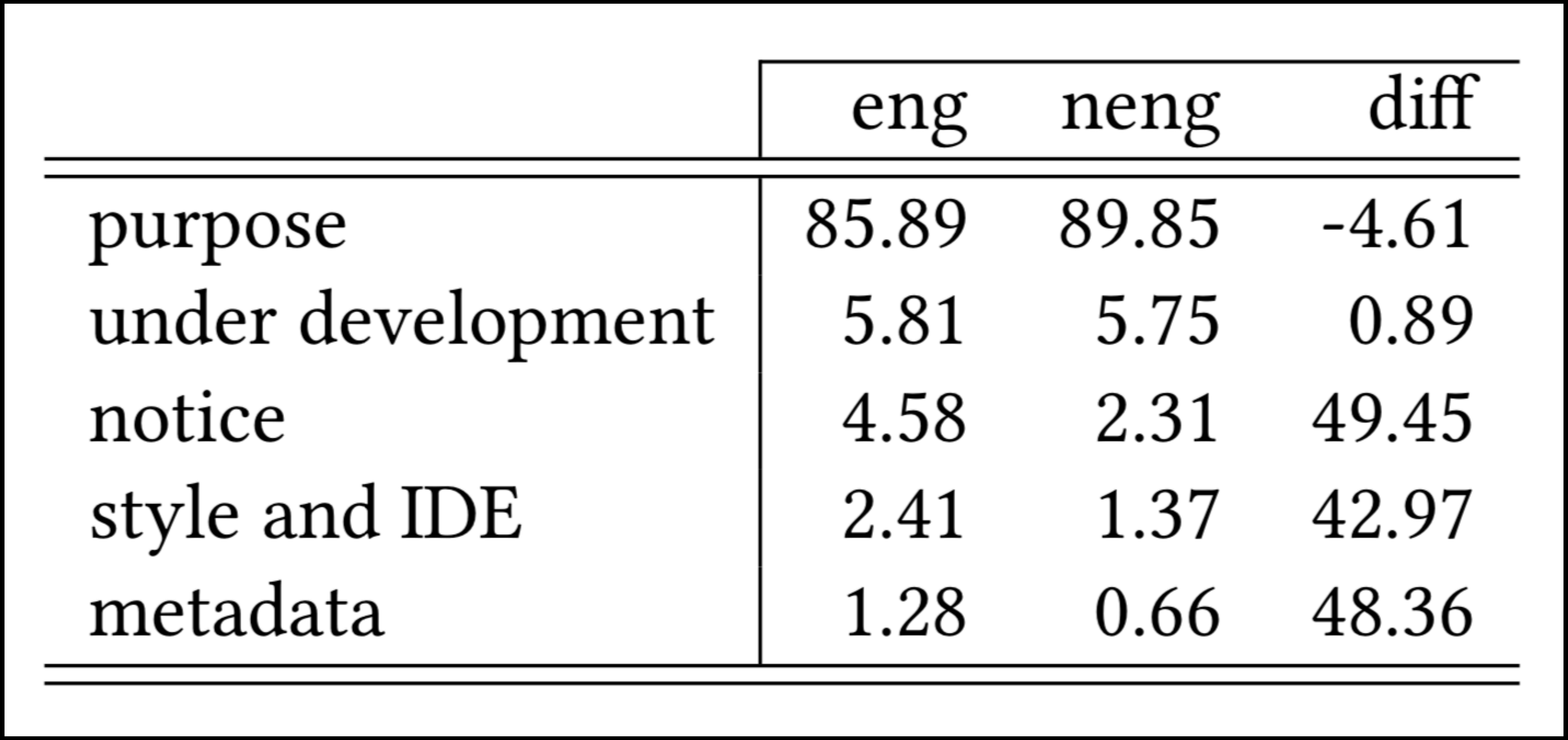


	class words				method words			
	eng		neng		eng		neng	
#	name	freq.	name	freq.	name	freq.	name	freq.
1	test	23203	activity	6881	get	241415	get	215821
2	type	4459	main	4420	set	118464	set	111971
3	handler	3236	test	4084	test	77566	on	75696
4	impl	3122	list	3039	is	46430	create	36566
5	exception	3017	string	2662	to	38691	to	26363
6	list	2994	listener	2643	create	38073	action	25263
7	map	2888	view	2418	on	29282	is	22950
8	service	2855	my	2390	name	25514	id	22924
9	builder	2795	impl	2376	string	22873	main	22639
10	factory	2744	id	2277	type	22532	performed	21448

RESULTS: COMMENTS

- mean length: 94.92 eng — 181.95 non-eng

RESULTS: COMMENTS



	eng	neng	diff
purpose	85.89	89.85	-4.61
under development	5.81	5.75	0.89
notice	4.58	2.31	49.45
style and IDE	2.41	1.37	42.97
metadata	1.28	0.66	48.36

CONCLUSIONS

- SE practices ***do impact*** on **identifiers** and **comments**
- ids longer in eng; comments longer in neng
- ids with formats non-conformant with Java conventions — much more common in neng code
- sophisticated comments — eng