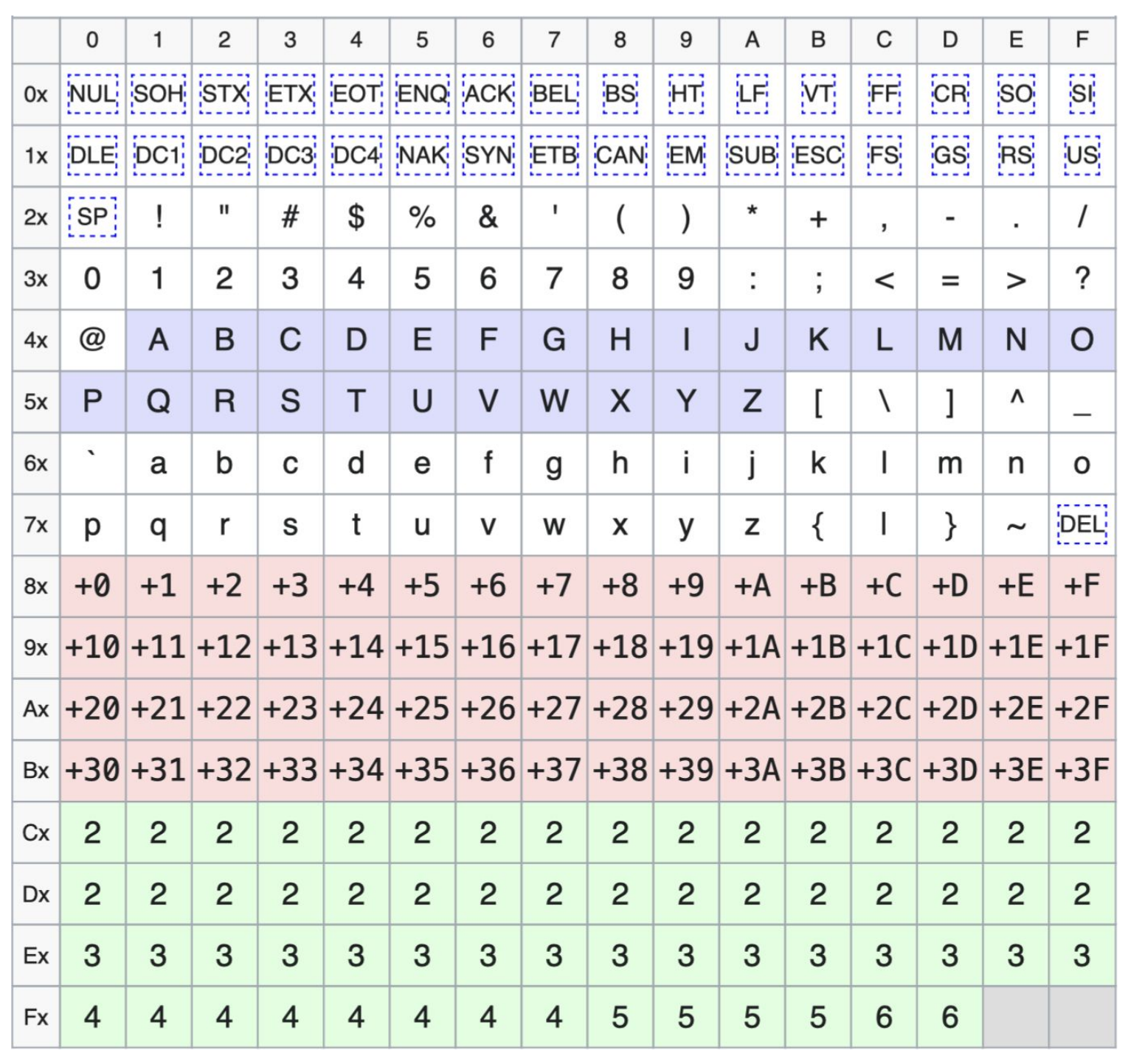
BYTE IS DEAD! LONG LIVE MYTE!

Tomasz Limisiewicz, Terra Blevins, Hila Gonen,
Orevaoghene Ahia, Luke Zettlemoyer

Morphological Bytes

We use morphological segmenter to redefine byte codespace.



	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0x	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1x	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2x	SP	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3x	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4x	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5x	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6x	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7x	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL
8x	+0	+1	+2	+3	+4	+5	+6	+7	+8	+9	+A	+B	+C	+D	+E	+F
9x	+10	+11	+12	+13	+14	+15	+16	+17	+18	+19	+1A	+1B	+1C	+1D	+1E	+1F
Ax	+20	+21	+22	+23	+24	+25	+26	+27	+28	+29	+2A	+2B	+2C	+2D	+2E	+2F
Bx	+30	+31	+32	+33	+34	+35	+36	+37	+38	+39	+3A	+3B	+3C	+3D	+3E	+3F
Cx	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Dx	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Ex	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Fx	4	4	4	4	4	4	4	4	5	5	5	5	6	6		

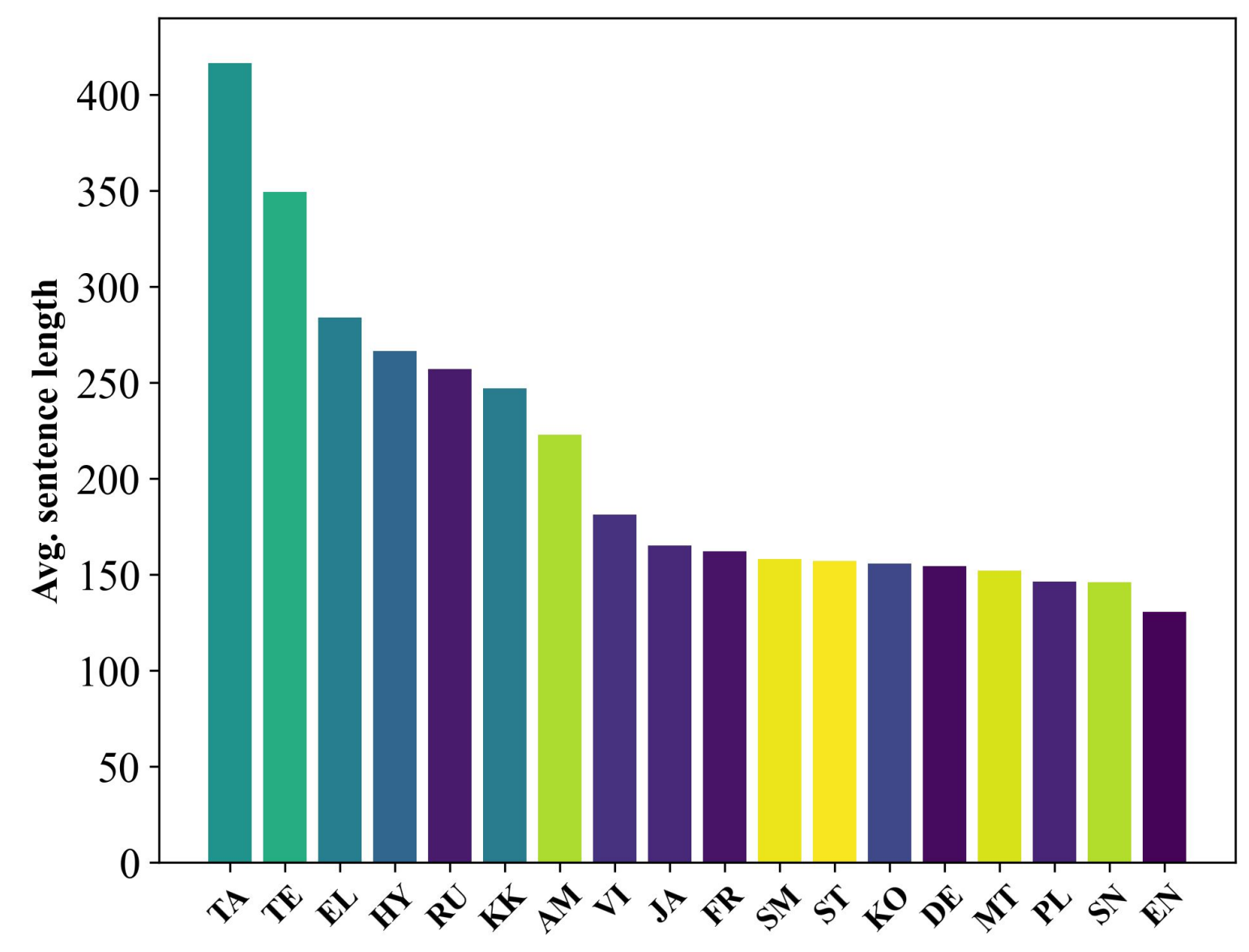
More Efficient

MYTE sequences are from 2 to 70% shorter than BYTEs ↘️

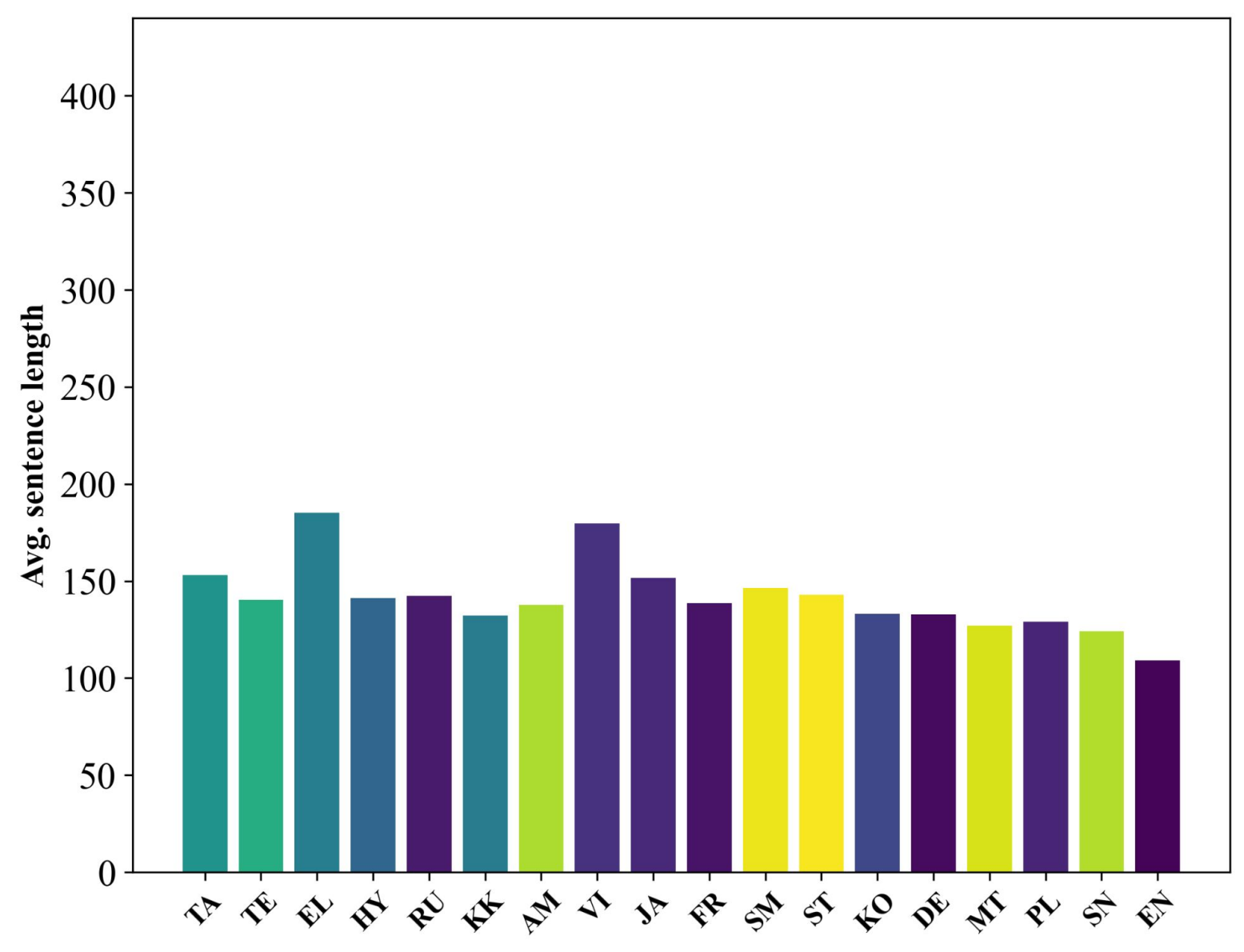
Fairer across Languages

Average length of parallel sentences FLORES 200 encoded in:

BYTES

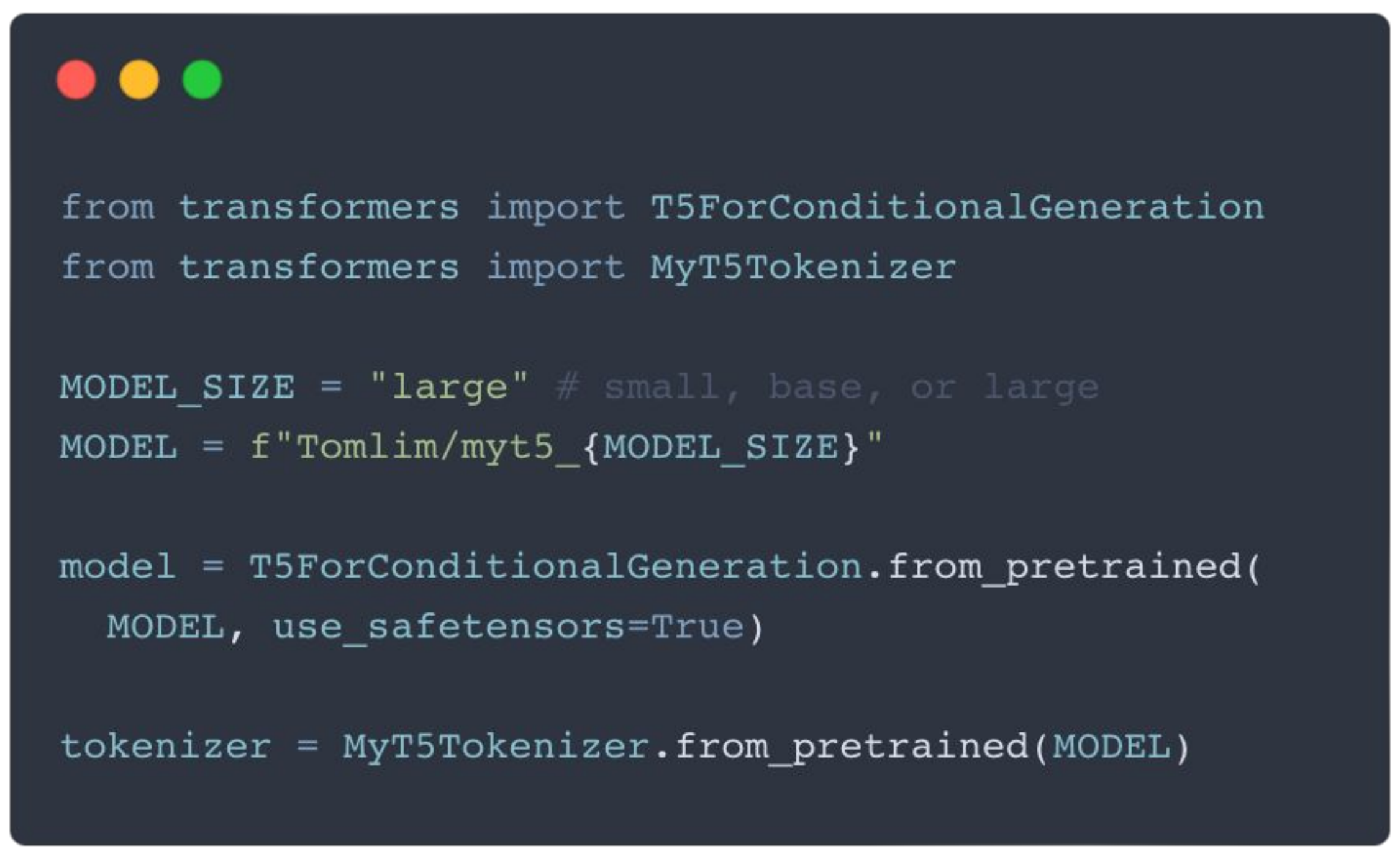


MYTES



MyT5 Models

We train T5 models in three sizes: small, base, and large and compare their performance with similar byte-level: ByT5 ➡️

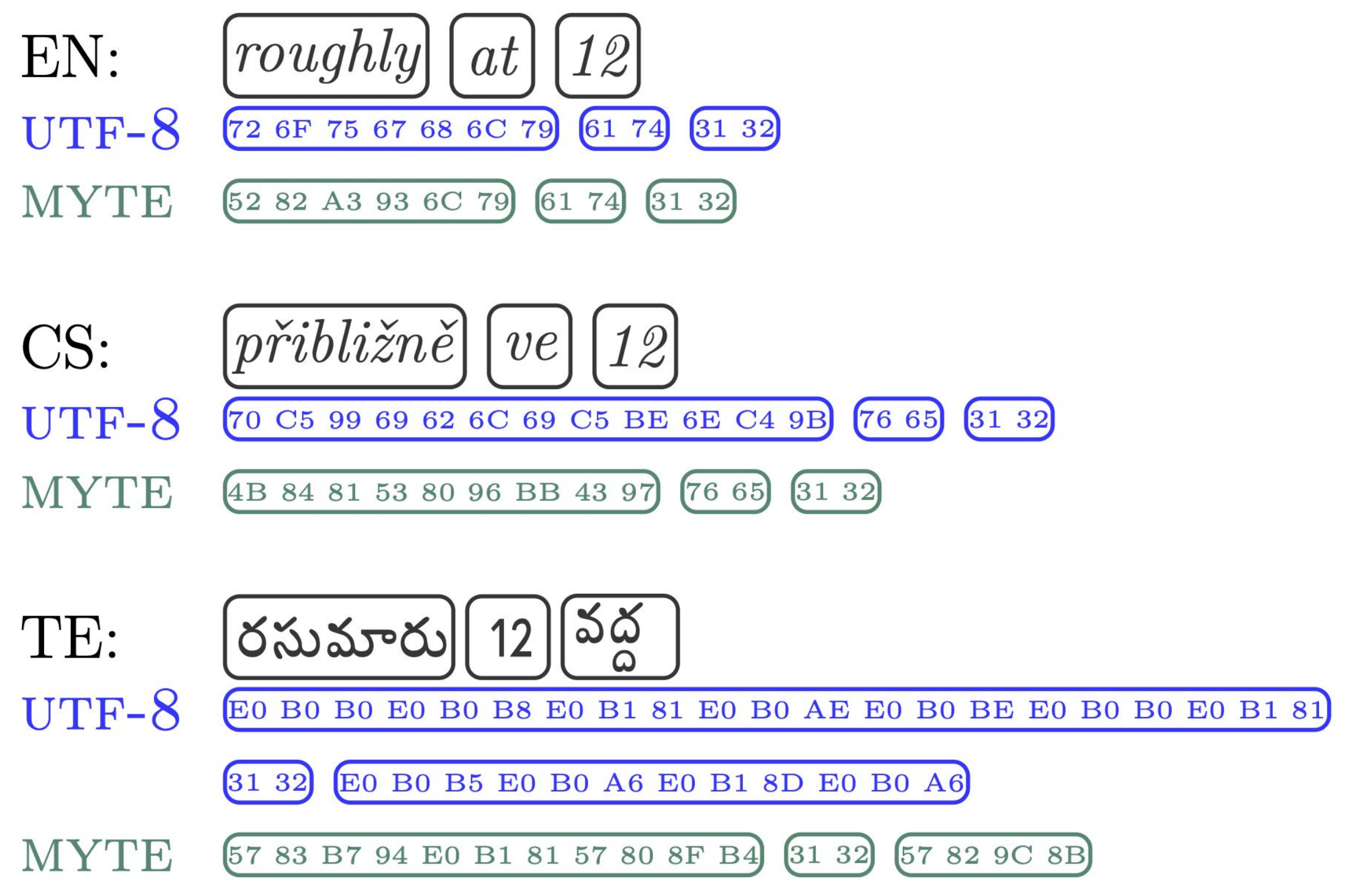


```
from transformers import T5ForConditionalGeneration
from transformers import MyT5Tokenizer

MODEL_SIZE = "large" # small, base, or large
MODEL = f"Tomlim/myt5_{MODEL_SIZE}"

model = T5ForConditionalGeneration.from_pretrained(
  MODEL, use_safetensors=True)

tokenizer = MyT5Tokenizer.from_pretrained(MODEL)
```

Problem: Byte encoding long and suboptimal for many languages

We propose Morphological BYTES to improve over UTF-8 encoding in:

- **Fairness**: comparable sequence length for the same information
- **Efficiency**: shorter sequence length



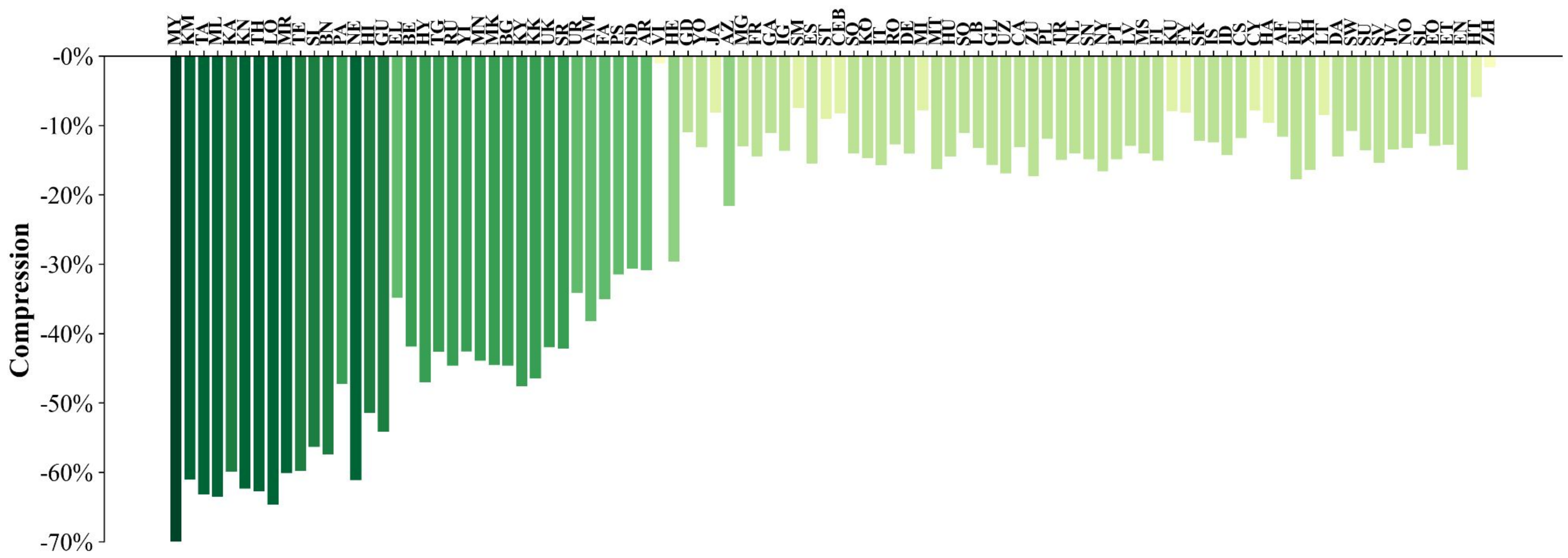
EN: roughly | at | 12
UTF-8 72 6F 75 67 68 6C 79 | 61 74 | 31 32
MYTE 52 82 A3 93 6C 79 | 61 74 | 31 32

CS: přibližně | ve | 12
UTF-8 70 C5 99 69 62 6C 69 C5 BE 6E C4 9B | 76 65 | 31 32
MYTE 4B 84 81 53 80 96 BB 43 97 | 76 65 | 31 32

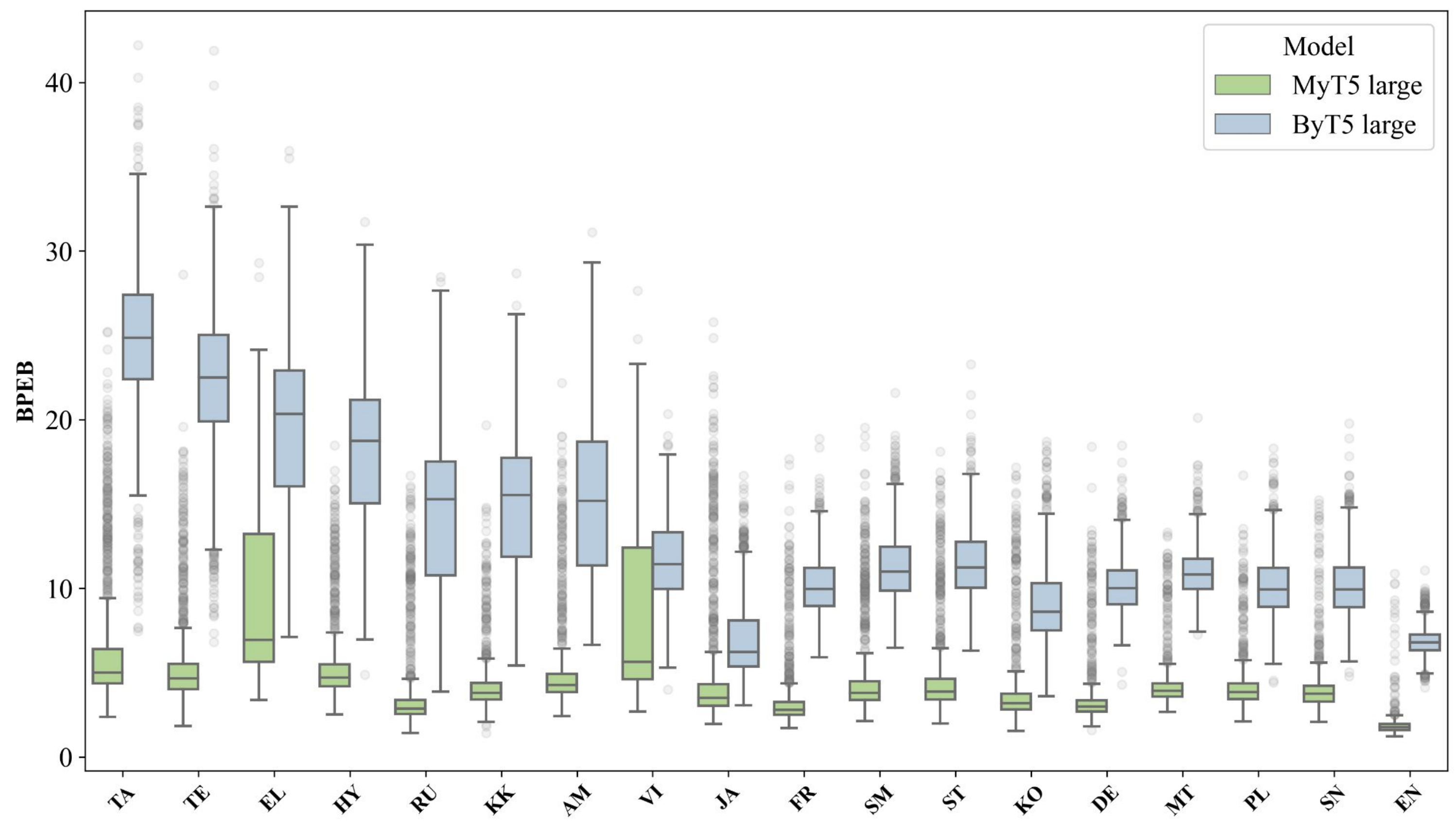
TE: రసుమారు | 12 | వద్ద
UTF-8 E0 B0 B0 E0 B0 B8 E0 B1 81 E0 B0 AE E0 B0 BE E0 B0 B0 E0 B1 81 | 31 32 | E0 B0 B5 E0 B0 A6 E0 B1 8D E0 B0 A6
MYTE 57 83 B7 94 E0 B1 81 57 80 8F B4 | 31 32 | 57 82 9C 8B

- **MYTE outperforms UTF-8 Bytes, efficiently representing texts in diverse languages, especially in non-Latin scripts.**
- **All 104 tested languages are encoded in less MYTEs than BYTEs**
- **MYTE speeds up and improves multilingual language modeling**
- **MYTE is more efficient for end-tasks with comparable results**

Compression: Length in Mytes vs. UTF-8 Bytes



LM Results: Perplexity of Parallel Sentences



End-tasks

	ByT5		MyT5	
	score	time	score	time
QA	73.2	36.2	**75.3**	**35.6**
NER	**81.5**	13.8	80.8	**12.6**
SemP	**25.1**	13.2	19.6	**12.4**
MT	20.1	15.9	**20.4**	**12.6**