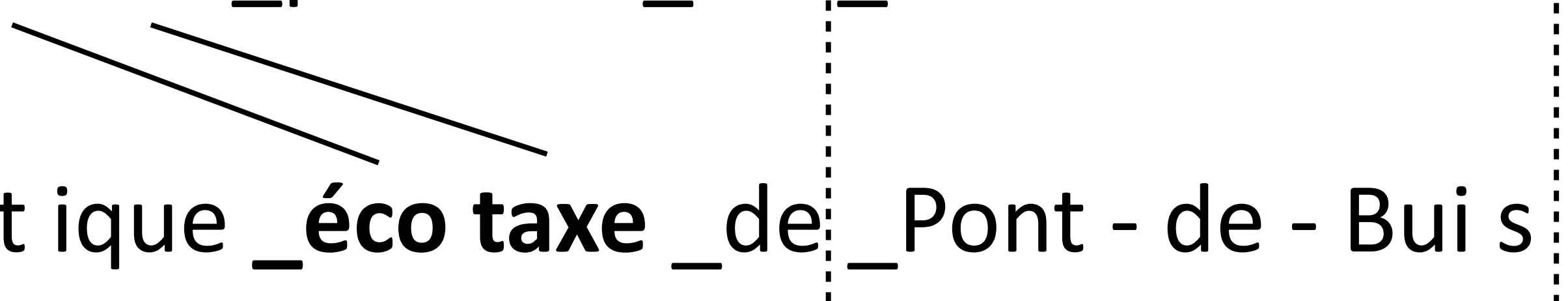


# Handling Rare Words

- ▶ Words are a difficult unit to work with: copying can be cumbersome, word vocabularies get very large
- ▶ Character-level models don't work well
- ▶ Compromise solution: use subword tokens, which may be full words but may also be parts of words

Input: \_the \_**eco tax** \_port i co \_in \_Po nt - de - Bu is...

Output: \_le \_port ique \_**éco taxe** \_de \_Pont - de - Bui s



- ▶ Can achieve transliteration with this, subword structure makes some translations easier to achieve

# Byte Pair Encoding (BPE)

- ▶ Start with every individual byte (character) as its own symbol

```
for i in range(num_merges):  
    pairs = get_stats(vocab)  
    best = max(pairs, key=pairs.get)  
    vocab = merge_vocab(best, vocab)
```

- ▶ Count bigram character cooccurrences in dictionary
- ▶ Merge the most frequent pair of adjacent characters

- ▶ Vocabulary stats are weighted over a large corpus
- ▶ Doing 30k merges => vocabulary of 30000 word pieces. Includes many whole words:

*and there were no re\_fueling stations anywhere*

*one of the city's more un\_princi\_pled real estate agents*

# Word Pieces

- ▶ Alternative to BPE

while voc size < target voc size:

- Build a language model over your corpus

- Merge pieces that lead to highest improvement in language model perplexity

- ▶ Issues: what LM to use? How to make this tractable?

- ▶ SentencePiece library from Google: unigram LM

# Comparison

(a)	<b>Original:</b>	furiously		(b)	<b>Original:</b>	tricycles			
	<b>BPE:</b>	_fur	iously		<b>BPE:</b>	_t	ric	y	cles
	<b>Unigram LM:</b>	_fur	ious   ly		<b>Unigram LM:</b>	_tri	cycle	s	
(c)	<b>Original:</b>	Completely preposterous suggestions							
	<b>BPE:</b>	_Comple	t	ely	_prep	ost	erous	_suggest	ions
	<b>Unigram LM:</b>	_Complete	ly	_pre	post	er	ous	_suggestion	s

- ▶ BPE produces less linguistically plausible units than word pieces (unigram LM)
- ▶ Some evidence that unigram LM works better in pre-trained transformer models
- ▶ Other work explores ensembling across multiple tokenizations