Natural Language Processing

The Byte Pair Encoding Algorithm

Marco Kuhlmann
Department of Computer and Information Science

Target representations for tokenisation

- **Option 1:** Tokenise into words

  But: concept of "word" not universal; unknown words

- **Option 2:** Tokenise into individual characters

  But: may be too small a unit for learning

- **Option 3:** Tokenise into subwords

  Intuition: words are composed of morphemes

Byte Pair Encoding

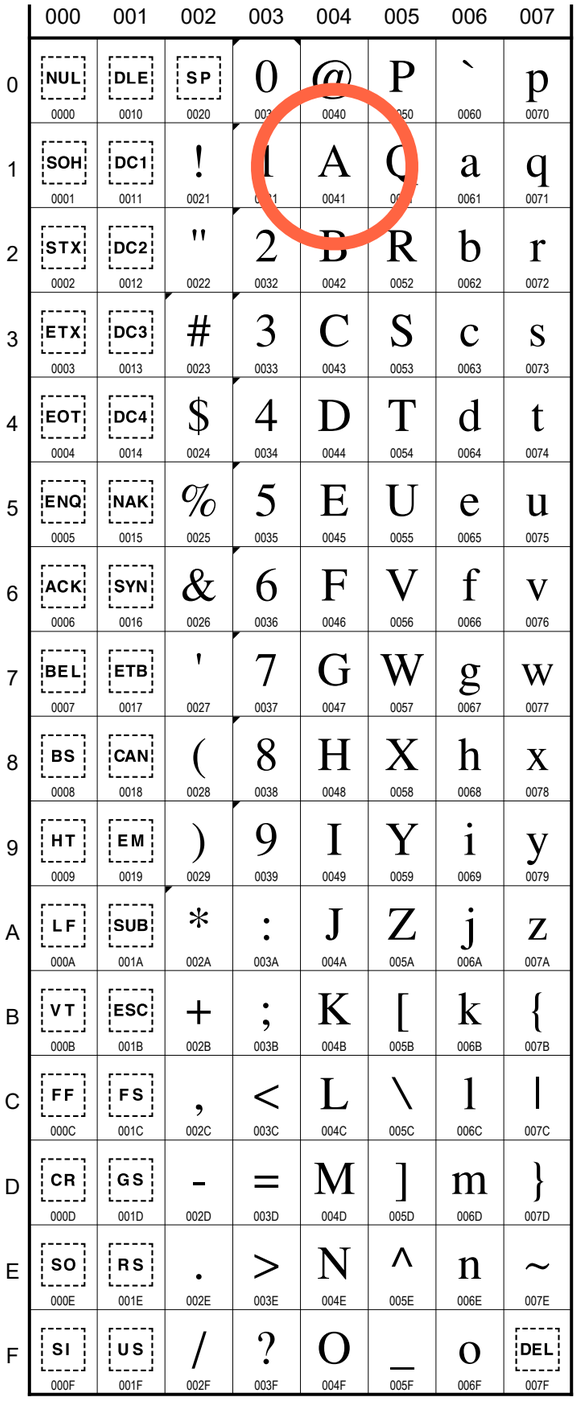
Byte Pair Encoding (BPE) is an algorithm for learning subword tokens from text.

- **Step 1:** Encode the text into a sequence of bytes. Initialise the token vocabulary with all single bytes.
- **Step 2:** Create a new token by merging the most frequent pair of consecutive tokens. Add the new token to the vocabulary.
- Repeat the previous step as long as the token vocabulary does not exceed a predefined maximum size.

Sennrich et al. (2016)

The Unicode Standard

- **Unicode** is a text encoding standard designed to support text from all the world's writing systems (that can be digitised).
- Version 16.0 supports 154,998 characters from 168 scripts.
- For backwards compability, the first 128 codepoints of Unicode are the same as ASCII.



Various signs

0900 ◌ँ DEVANAGARI SIGN INVERTED CANDRABINDU
= vaidika adhomukha candrabindu
0901 ◌ं DEVANAGARI SIGN CANDRABINDU
= anunasika
→ 0310 ◌̐ combining candrabindu
0902 ◌ं DEVANAGARI SIGN ANUSVARA
= bindu
0903 ◌ः DEVANAGARI SIGN VISARGA

Independent vowels

0904 अ DEVANAGARI LETTER SHORT A
• used for short e in Awadhi
• also used in Devanagari transliterations of some South Indian and Kashmiri languages by a publisher in Lucknow
0905 अ DEVANAGARI LETTER A
0906 आ DEVANAGARI LETTER AA
0907 इ DEVANAGARI LETTER I
0908 ई DEVANAGARI LETTER II
0909 उ DEVANAGARI LETTER U
090A ऊ DEVANAGARI LETTER UU
090B ऋ DEVANAGARI LETTER VOCALIC R
090C ऌ DEVANAGARI LETTER VOCALIC L
090D ऍ DEVANAGARI LETTER CANDRA E
090E ऎ DEVANAGARI LETTER SHORT E
• Kashmiri, Bihari languages
• also used for transcribing Dravidian short e
090F ए DEVANAGARI LETTER E
0910 ऐ DEVANAGARI LETTER AI
0911 ऑ DEVANAGARI LETTER CANDRA O
0912 ऒ DEVANAGARI LETTER SHORT O
• Kashmiri, Bihari languages
• also used for transcribing Dravidian short o
0913 ओ DEVANAGARI LETTER O
0914 औ DEVANAGARI LETTER AU

Consonants

0915 क DEVANAGARI LETTER KA
0916 ख DEVANAGARI LETTER KHA
0917 ग DEVANAGARI LETTER GA
0918 घ DEVANAGARI LETTER GHA
0919 ङ DEVANAGARI LETTER NGA
091A च DEVANAGARI LETTER CA
091B छ DEVANAGARI LETTER CHA
091C ज DEVANAGARI LETTER JA
091D झ DEVANAGARI LETTER JHA
091E ञ DEVANAGARI LETTER NYA
091F ट DEVANAGARI LETTER TTA
0920 ठ DEVANAGARI LETTER TTHA
0921 ड DEVANAGARI LETTER DDA
0922 ढ DEVANAGARI LETTER DDHA
0923 ण DEVANAGARI LETTER NNA
0924 त DEVANAGARI LETTER TA
0925 थ DEVANAGARI LETTER THA
0926 द DEVANAGARI LETTER DA
0927 ध DEVANAGARI LETTER DHA
0928 न DEVANAGARI LETTER NA
0929 ऩ DEVANAGARI LETTER NNNA
• for transcribing Dravidian alveolar n
≡ 0928 न 093C ़
092A प DEVANAGARI LETTER PA
092B फ DEVANAGARI LETTER PHA
092C ब DEVANAGARI LETTER BA
092D भ DEVANAGARI LETTER BHA
092E म DEVANAGARI LETTER MA
092F य DEVANAGARI LETTER YA
0930 र DEVANAGARI LETTER RA
0931 ऱ DEVANAGARI LETTER RRA
• for transcribing Dravidian alveolar r
• half form is represented as "Eyelash RA"
≡ 0930 र 093C ़
0932 ल DEVANAGARI LETTER LA
0933 ळ DEVANAGARI LETTER LLA
0934 ऴ DEVANAGARI LETTER LLLA
• for transcribing Dravidian l
≡ 0933 ळ 093C ़
0935 व DEVANAGARI LETTER VA
0936 श DEVANAGARI LETTER SHA
0937 ष DEVANAGARI LETTER SSA
0938 स DEVANAGARI LETTER SA
0939 ह DEVANAGARI LETTER HA

Dependent vowel signs

These dependent vowel signs are used in Kashmiri and in the Bihari languages (Bhojpuri, Magadhi, and Maithili).

093A ◌ॊ DEVANAGARI VOWEL SIGN OE
093B ◌ो DEVANAGARI VOWEL SIGN OOE

Various signs

093C ◌़ DEVANAGARI SIGN NUKTA
• for extending the alphabet to new letters
093D ऽ DEVANAGARI SIGN AVAGRAHA

Dependent vowel signs

093E ◌ा DEVANAGARI VOWEL SIGN AA
093F ◌ि DEVANAGARI VOWEL SIGN I
• stands to the left of the consonant
0940 ◌ी DEVANAGARI VOWEL SIGN II
0941 ◌ु DEVANAGARI VOWEL SIGN U
0942 ◌ू DEVANAGARI VOWEL SIGN UU
0943 ◌ृ DEVANAGARI VOWEL SIGN VOCALIC R
0944 ◌ॄ DEVANAGARI VOWEL SIGN VOCALIC RR
0945 ◌ॅ DEVANAGARI VOWEL SIGN CANDRA E
= candra
0946 ◌ॆ DEVANAGARI VOWEL SIGN SHORT E
• Kashmiri, Bihari languages
• also used for transcribing Dravidian short e
0947 ◌े DEVANAGARI VOWEL SIGN E
0948 ◌ै DEVANAGARI VOWEL SIGN AI
0949 ◌ॉ DEVANAGARI VOWEL SIGN CANDRA O
094A ◌ॊ DEVANAGARI VOWEL SIGN SHORT O
• Kashmiri, Bihari languages
• also used for transcribing Dravidian short o
094B ◌ो DEVANAGARI VOWEL SIGN O
094C ◌ौ DEVANAGARI VOWEL SIGN AU

Virama

094D ◌् DEVANAGARI SIGN VIRAMA
= halant (the preferred Hindi name)
• suppresses inherent vowel

Dependent vowel signs

094E ◌ॎ DEVANAGARI VOWEL SIGN PRISHTHAMATRA E
• character has historic use only
• combines with E to form AI, with AA to form O, and with O to form AU

Encoding text into bytes

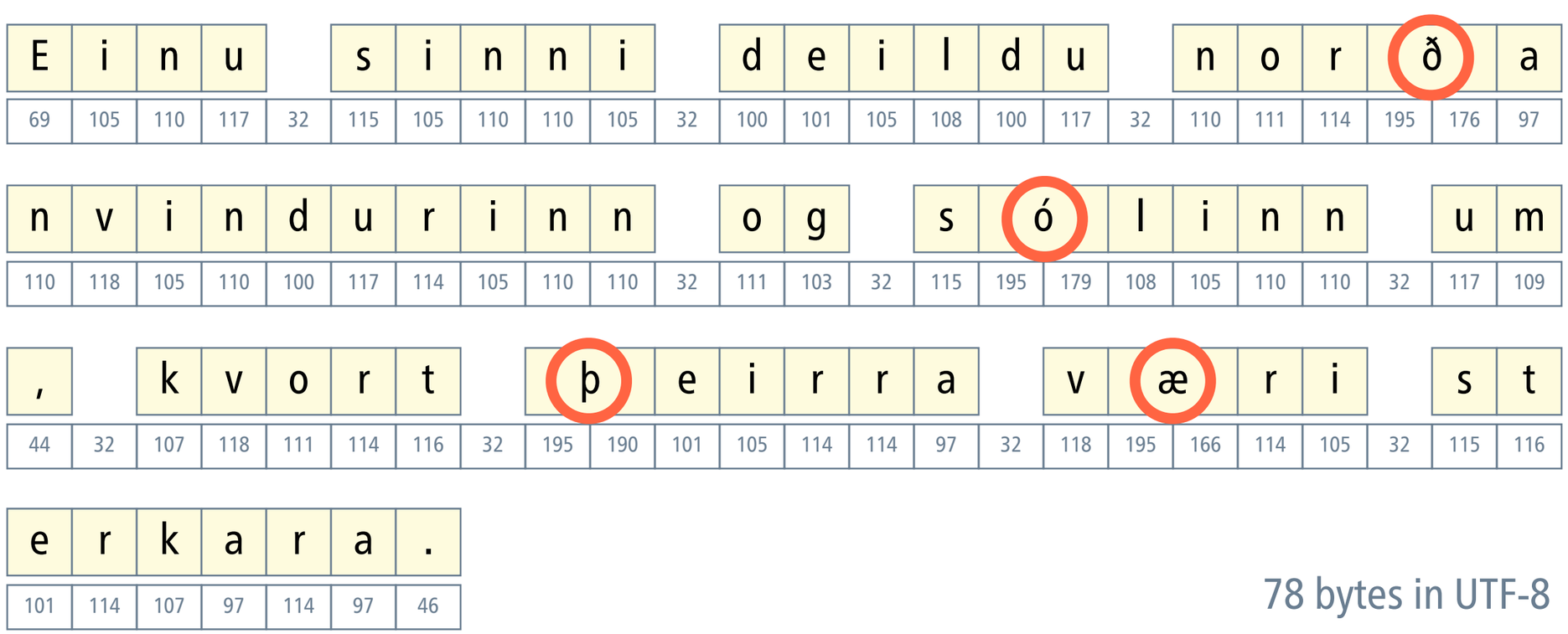
- Encoding all (more than 1 million) Unicode characters into bytes requires more than one byte per character.
- **UTF-8 (8-bit Unicode Transformation Format)** is the most widely used encoding scheme for Unicode.
- It uses a variable-width encoding of 1–4 bytes per character.

  The first byte indicates how many additional bytes are part of the character.

Encoding text into bytes

Einu sinni deildu norðanvindurinn og sólin um, kvort þeirra væri sterkara.

74 Unicode characters



78 bytes in UTF-8

The North Wind and the Sun were disputing which was the stronger, when a traveler came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveler take his cloak off should be considered stronger than the other.

Vocabulary (without single bytes)

Token ID	Token
256	
257	
258	
259	
260	

Pair counts

Token pair	Count
e + SPACE	11
SPACE + t	10
h + e	9
t + h	9
d + SPACE	7

The North Wind and the Sun were disputing which was the stronger, when a traveler came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveler take his cloak off should be considered stronger than the other.

Vocabulary (without single bytes)

Token ID	Token
256	[e]
257	
258	
259	
260	

Pair counts

Token pair	Count
e + SPACE	11
SPACE + t	10
h + e	9
t + h	9
d + SPACE	7

Th[e]North Wind and th[e]Sun wer[e]disputing which was th[e]stronger, when a traveler cam[e]along wrapped in a warm cloak. They agreed that th[e]on[e]who first succeeded in making th[e]traveler tak[e]his cloak off should b[e]considered stronger than th[e]other.

Vocabulary (without single bytes)

Token ID	Token
256	[e]
257	
258	
259	
260	

Pair counts

Token pair	Count
t + h	9
SPACE + t	9
d + SPACE	7
e + r	7
h + [e]	6

Th[e]North Wind and th[e]Sun wer[e]disputing which was th[e]stronger, when a traveler cam[e]along wrapped in a warm cloak. They agreed that th[e]on[e]who first succeeded in making th[e]traveler tak[e]his cloak off should b[e]considered stronger than th[e]other.

Vocabulary (without single bytes)

Token ID	Token
256	[e]
257	[th]
258	
259	
260	

Pair counts

Token pair	Count
t + h	9
SPACE + t	9
d + SPACE	7
e + r	7
h + [e]	6

Th[e]Nor[th] Wind and [th][e]Sun wer[e]disputing which was [th][e]stronger, when a traveler cam[e]along wrapped in a warm cloak. They agreed [th]at [th][e]on[e]who first succeeded in making [th][e]traveler tak[e]his cloak off should b[e]considered stronger [th]an [th][e]o[th]er.

Vocabulary (without single bytes)

Token ID	Token
256	[e]
257	[th]
258	
259	
260	

Pair counts

Token pair	Count
d + SPACE	7
SPACE + [th]	7
e + r	7
SPACE + w	6
i + n	5

Th[e]Nor[th] Wind and [th][e]Sun wer[e]disputing which was [th][e]stronger, when a traveler cam[e]along wrapped in a warm cloak. They agreed [th]at [th][e]on[e]who first succeeded in making [th][e]traveler tak[e]his cloak off should b[e]considered stronger [th]an [th][e]o[th]er.

Vocabulary (without single bytes)

Token ID	Token
256	[e]
257	[th]
258	[d]
259	
260	

Pair counts

Token pair	Count
d + SPACE	7
SPACE + [th]	7
e + r	7
SPACE + w	6
i + n	5

Th[e]Nor[th] Win[d]an[d][th][e]Sun wer[e]disputing which was [th][e]stronger, when a traveler cam[e]along wrappe[d]in a warm cloak. They agree[d][th]at [th][e]on[e]who first succeede[d]in making [th][e]traveler tak[e]his cloak off shoul[d]b[e]considere[d]stronger [th]an [th][e]o[th]er.

Vocabulary (without single bytes)

Token ID	Token
256	[e]
257	[th]
258	[d]
259	
260	

Pair counts

Token pair	Count
e + r	7
SPACE + w	6
i + n	5
[th] + [e]	5
n + SPACE	5

Th[e]Nor[th] Win[d]an[d][th][e]Sun wer[e]disputing which was [th][e]stronger, when a traveler cam[e]along wrappe[d]in a warm cloak. They agree[d][th]at [th][e]on[e]who first succeede[d]in making [th][e]traveler tak[e]his cloak off shoul[d]b[e]considere[d]stronger [th]an [th][e]o[th]er.

Vocabulary (without single bytes)

Token ID	Token
256	[e]
257	[th]
258	[d]
259	[er]
260	

Pair counts

Token pair	Count
e + r	7
SPACE + w	6
i + n	5
[th] + [e]	5
n + SPACE	5

Th[e]Nor[th] Win[d]an[d][th][e]Sun w[er][e]disputing which was [th][e]strong[er], when a travel[er] cam[e]along wrappe[d]in a warm cloak. They agree[d][th]at [th][e]on[e]who first succeede[d]in making [th][e]travel[er] tak[e]his cloak off shoul[d]b[e]consid[er]e[d]strong[er] [th]an [th][e]o[th][er].

Vocabulary (without single bytes)

Token ID	Token
256	[e]
257	[th]
258	[d]
259	[er]
260	

Pair counts

Token pair	Count
SPACE + w	6
i + n	5
[th] + [e]	5
n + SPACE	5
n + g	5

Th[e]Nor[th] Win[d]an[d][th][e]Sun w[er][e]disputing which was [th][e]strong[er], when a travel[er] cam[e]along wrappe[d]in a warm cloak. They agree[d][th]at [th][e]on[e]who first succeede[d]in making [th][e]travel[er] tak[e]his cloak off shoul[d]b[e]consid[er]e[d]strong[er] [th]an [th][e]o[th][er].

Vocabulary (without single bytes)

Token ID	Token
256	[e]
257	[th]
258	[d]
259	[er]
260	[w]

Pair counts

Token pair	Count
SPACE + w	6
i + n	5
[th] + [e]	5
n + SPACE	5
n + g	5

Th[e]Nor[th] Win[d]an[d][th][e]Sun[w][er][e]disputing
[w]hich[w]as [th][e]strong[er],[w]hen a travel[er] cam[e]along
[w]rappe[d]in a[w]arm cloak. They agree[d][th]at [th]
[e]on[e]who first succeede[d]in making [th][e]travel[er]
tak[e]his cloak off shoul[d]b[e]consid[er]e[d]strong[er] [th]an
[th][e]o[th][er].

Vocabulary (without single bytes)

Token ID	Token
256	[e]
257	[th]
258	[d]
259	[er]
260	[w]

Some comments on BPE

- The tokens obtained using BPE match varying spans of source text, from single characters to whole words and beyond.
- The tokens are not guaranteed to have any apparent linguistic meaning, but often resemble words or morphemes.

  BPE = “poor man’s morphology”

- BPE solves the problem with unknown words: Every text can be tokenised; in the worst case, it is tokenised as bytes.

Tokenisation in language models

Model	Release year	Tokenisation method	Vocabulary size
BERT	2018	WordPiece	30 K
GPT-2	2019	BPE	50 K
GPT-3.5	2022	BPE	100 K
GPT-40	2024	BPE	200 K
Llama 3	2024	BPE	128 K

Tiktokenizer

o200k_base

The North Wind and the Sun were disputing which was the stronger, when a traveler came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveler take his cloak off should be considered stronger than the other.

Token count
49

The North Wind and the Sun were disputing which was t he stronger, when a traveler came along wrapped in a warm cloak. They agreed that the one who first succee ded in making the traveler take his cloak off should be considered stronger than the other.

976, 7180, 28551, 326, 290, 11628, 1504, 28301, 289, 1118, 673, 290, 26929, 11, 1261, 261, 72819, 5831, 42 51, 31831, 306, 261, 9144, 152842, 13, 3164, 12863, 4 84, 290, 1001, 1218, 1577, 53434, 306, 4137, 290, 728 19, 2304, 1232, 152842, 1277, 1757, 413, 9474, 26929, 1572, 290, 1273, 13

Show whitespace