بابەتی 4: لەتکردنی تۆکن

خولی سەرهێڵی زمانەوانیی کۆمپیوتەری

سینا ئەحمەدی

sinaahmadi.github.io/KurdishCL

ناوەڕۆک

لەتکردنی تۆکن

چۆن دەق لە سەر کۆمپیوتەر نیشان دەدەین؟

لە زۆر بابەتی پڕۆسەس کردنی زماندا پێویستمان بە داشکاندنی دەق (segmentation) بۆ دەرهێنانی پێکهاتە زمانەوانییەکان یان ئاسانکاریی تەکنیکی هەیە.

- هەنگاوێکی پێش-پڕۆسەس کردنی زۆرینەی ئەرکەکانی پڕۆسەس کردنی زمانە
- لەتەکان یەکەی جیاوازی دەقن (پیت، وشە، هتد)
- ئاستی زبر وەک داشکاندنی دەق بە سەر پاراگراف و ڕستەکاندا

مافەکانی مرۆڤ کۆمەڵێک بنچینە یان ڕێسای ڕەوشتین کە لەسەر پێوانەی ئاکاریی مرۆڤەکان ساز کراون و بەگوێرەی یاسای ناوخۆیی و نێودەوڵەتی پارێزگارییان لێ دەکرێت. بەگشتی دەتوانین ئەم بنەمایانە بەم شێوەیە ناو بنین کە مافی ڕەوای هەموو مرۆڤێکن و شیاوی گۆڕین و دەستکاریی نین و بەیەکسانی و بەهۆی مرۆڤبوونی کەسەکەوە پێی ڕەوا بینراوە. بێ گوێدانە تەمەن، نەتەوە، ئایین، زمان، ڕەنگ و، ڕەگەز هەمووان وەکو مرۆڤ هاوتان. ئەم مافانە لە هەموو شوێنێک و لە هەموو کاتێکدا جێبەجێ دەبن بەسەر مرۆڤەکاندا، لە هەر کوێیەک بن لە جیهاندا، بێ جیاوازی. لە سایەی سەروەریی یاسا و سۆزی مرۆڤانەوە چاوەڕوان دەکرێت کە هەموو پێڕەوی ڕێزگرتنی مافەکانی مرۆڤی هەبێت بۆ مرۆڤەکانی تر، هیچ کات ئەو مافانە لە مرۆڤ ناترازێن، مافی مرۆڤ بۆ هەموو کەسێک مافێکی ڕەسەنە و پێوەی لکاوە، مەگەر لە دۆخێکی یاساییی تایبەتی و بە هۆکاری سەپاندنی لە ڕەوتێکی یاساییدا.

وەرگیراو لە وتاری «مافەکانی مرۆڤ»، ویکیپیدیای کوردی

- لەتکردن یان کەرتکردنی دەق چەشنێکە لە داشکاندنی وشە و بەو لەتانە دەڵێین «**تۆکن**» (token)
- بۆ یەکەم جار لە ساڵی 1992دا ئاماژە بەو لەتانە کراوە [1]
- بنەمای زۆر ئەرکی پڕۆسەسکردنی زمان لە سەر تۆکنەکانە

تۆکن

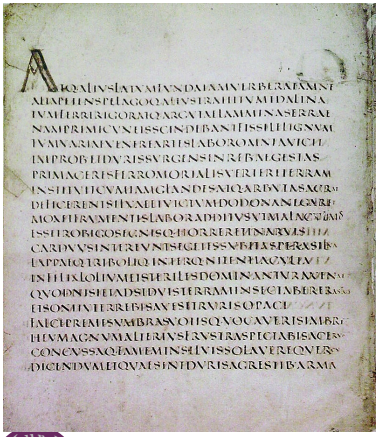
- «ڕیزبەندیی یەک لە دوای یەک و نابەتاڵی گرافیم یان فۆنێمەکان لە دەقدا» [2]
- «لەتکردنی دەق بە سەر تۆکندا بریتییە لە دۆزینەوەی سنووری وشەکان» [3]
- «تۆکن خەیاڵێکی پڕوپووچی ئەندازیارێکی کۆمپیوتەرە بۆ ئاسانکردنی کاری خۆی» [4]

لەتکردنی تۆکن: سنووری وشە

پیناسەی تۆکن بە پێی دۆزینەوەی سنووری وشە

- Dr. O'Neil presents state-of-the-ar results.
- « **دواکەوتنی شێوازەکانی بەرهەمهێنان** »
  - دواکەوتنی شێوازەکانی بەرهەمهێنان
  - دوا کەوتنی شێوازەکانی بەرهەم هێنان
  - دواکەوتن ی شێواز ەکان ی بەرهەمهێنان
  - دوا کەوتن ی شێواز ەکان ی بەرهەم هێنان
- سنووری وشە لە Scriptio continuaەکاندا
- سنووری وشە لە زمانەکانی دیکەدا
  - 我是中国人
  - つ目の入力
  - Sprachwissenschaft

پێناسە کردنی تۆکن ئەرکێکی ئاسان نییە!



ڕوانگەی ڕێنووس: پشت بە شێوازی نووسینی دەقەکە ببەستە

- داشکاندن بەپێی هێماکانی خاڵبەندی وەک «.!؟؟:؛،,;»، بۆشایی، هێڵی نوێ و تەب
  دواکەوتنی شێوازەکانی بەرهەمهێنان .
- باشی +: ئاسانە
- خراپی −
  - داهێنانی هەموو یاساکان زەحمەتە
  - پێویستی بە زانیاری لە سەر زمانەکان و ڕێنووسەکانیان هەیە
  - ناتوانێ چارەسەری سنووڕە دیارینەکراوەکانی وشە بکا
  - هەڵەی ڕێنووسی سازکردنی وشەدانەکە تێک دەدا

ڕوانگەی زمانەوانی: ڕۆڵی زمانەوانیی پێکهاتەکان وەک ڕۆڵی وشەکان لە ڕستەسازیدا

- لەتکردن بەپێی پێکهاتەکان لە وشەدا، وەک مۆرفیمەکان
- زۆر جار، تۆکنەکان بە پێی سەرچاوەیکی دیکە دەردەهێندرێن، وەک فەرهەنگ

دواکەوتن ی شێوازەکان ی بەرهەمهێنان ۰

دوا کەوتن ی شێواز ەکان ی بەرهەم هێنان ۰

- باشی +
  - ئاسانە بەڵام نە بە قەت ڕوانگەی ڕێنووس
  - بەسوودە لەو ئەرکانەی کە پێویستیان بە بۆچوونێکی زمانەوانانە هەیە وەک لێکدانەوەی ڕستە
- خراپی –
  - پێویستی بە زانیاریی زمانەوانی لە سەر زمانەکان و ڕێنووسەکانیان هەیە
  - یاسا زمانەوانییەکان دەتوانن ناتەواو بن
  - هەڵەی ڕێنووسی سازکردنی وشەدانەکە تێک دەدا
  - ئەگەر وشەیەک لە فەرهەنگدا نەبێ و یاساکان کاریان لێ نەکا؟ ← وشەی نەناسراو

زمان پڕە لە وشە و هەموویان لە فەرهەنگدا دەرناکەون، وەک «تورتەماڵی»، «کۆڤید» یان «لاتک»

- ئەو تۆکنانەی لە وشەدان یان فەرهەنگی زمانەکەدا نەهاتوون بە ⟨unk⟩ دیاری دەکرێن
- زۆر وشەی دەگمەن یان نێوی کەس بە وشەی نەناسراو دەردەچن

دوا | کەوتن | ی | شێواز | ەکان | ی | ⟨unk⟩ | هێنان | .

- نەناسینەوەی تۆکنێک دەبێتە هۆی لە دەست چوونی زانیاری
- چۆن ئەم کێشەیە چارەسەر بکەین بەبێ ئەوەی پێویستمان بە زانیاری لە سەر هەموو زمانەکدا هەبێ؟

ڕوانگەی پێدراوە: چ پێکهاتەگەلێک باوترە، با بە تۆکنیان دانێین

- ئەم بۆچوونە چارەسەرێکە بۆ کێشەکانی بۆچوونە پێشووەکان
- پشت بە فراوانیی ئەو پیتانە دەبەستێ کە وێکڕا دەردەکەون و پێیان دەڵێن وشەڵە (subword)
- **Byte pair encoding (BPE)**: ناسراوترین و باوترین ئەلگوریتمە (algorithm) کە پشت بەم بۆچوونەوە دەبەستێ [5]
- «پاڵنەری سەرەکیی ئەم بۆچوونە ئەوەیە کە وەرگێڕانی هەندێک وشە ڕوونە، بەو مانایەی کە دەکرێت لەلایەن وەرگێڕێکی لێهاتووەوە وەربگێڕدرێن تەنانەت ئەگەر بۆ ئەو نوێش بن، بە پشتبەستن بە وەرگێڕانی یەکەکانی ژێر وشەی ناسراو وەک مۆرفیم یان فۆنێم.»

دوا | کەوتن | ی | شێواز | ەکان | ی | #بەر | هەم | هێنان | .

Neural Machine Translation of Rare Words with Subword Units

Rico Sennrich and **Barry Haddow** and **Alexandra Birch**
School of Informatics, University of Edinburgh
{rico.sennrich,a.birch}@ed.ac.uk, bhaddow@inf.ed.ac.uk

هەنگاوەکانی ڕاهێنان:

1. سازکردنی وشەدان: دەست پێدەکات بە دابەشکردنی وشەکان بۆ تاکە پیتەکان (هەر یەکێکیان هێمایەکە لە وشەدانی کۆتاییدا)
2. دۆزینەوەی جووتە هێمای فراوان لە وشەدانی ئێستادا. بەم ئەرکە دەڵێن یەکخستن (merge)
   بۆ نموونە ئەگەر پیتی «د» بە فراوانییەکی زۆرەوە لەگەڵ «ە» دەرکەوێ، «دە» وەکوو یەک هێما لە وشەدانەکە زیاد دەکا
3. زیادکردنی هێما یەکخراوەکان لە وشەدانەکە
4. دووبارەکردنەوەی هەنگاوەکانی (2) و (3) هەتا بە ژمارەی دیاریکراوی یەکخران دەگەین یان هیچ پێکهاتەیەکی نوێی هێماکان بە فراوانی پێویستەوە نەمێنێت

لەتکردنی تۆکن: ڕێکاری BPE - نموونە

```
low low low low low lowest lowest newer newer
newer newer newer newer wider wider wider new new
```

1. سازکردنی وشەدان: دابەشکردنی وشەکان بۆ تاکە پیتەکان. دەقەکە بە پێی بۆشایی داشکێنە و کۆتاییی وشەکان بە هێمایەکی تایبەت دیاری کە وەک «</w>» یان «_»

کۆرپس

```
'l o w </w>': 5
'l o w e s t </w>': 2
'n e w e r </w>': 6
'w i d e r </w>': 3
'n e w </w>': 2
```

وشەدان

```
d, e, i, l, n, o, r, s, t, w, </w>
```

لەتکردنی تۆکن: ڕێکاری BPE – نموونە

1. سازکردنی وشەدان

کۆرپس

```
'l o w </w>': 5
'l o w e s t </w>': 2
'n e w e r </w>': 6
'w i d e r </w>': 3
'n e w </w>': 2
```

وشەدان

```
d, e, i, l, n, o, r, s, t, w, </w>
```

2. یەکخستنی جووتە فراوانەکان

'e', 'r' دەبن بە یەک وشەڵە ⇐ er

فراوانیی جووتەکان

```
('e', 'r'): 9
('r', '</w>'): 9
('w', 'e'): 8
('n', 'e'): 8
('e', 'w'): 8
('l', 'o'): 7
('o', 'w'): 7
('w', '</w>'): 7
('w', 'i'): 3
('i', 'd'): 3
('d', 'e'): 3
('e', 's'): 2
('s', 't'): 2
('t', '</w>'): 2
```

لەتکردنی تۆکن: ڕێکاری BPE - نموونە

1. سازکردنی وشەدان
2. یەکخستنی جووتە فراوانەکان
3. زیادکردنی هێما یەکخراوەکان لە وشەدانەکە

کۆرپس

```
'l o w </w>': 5
'l o w e s t </w>': 2
'n e w er </w>': 6
'w i d er </w>': 3
'n e w </w>': 2
```

وشەدان

```
d, e, i, l, n, o, r, s, t, w,</w>,er
```

4. هەنگاوی 2 و 3 دووپات بکەوە

فراوانیی جووتەکان

```
('er', '</w>'): 9
('n', 'e'): 8
('e', 'w'): 8
('l', 'o'): 7
('o', 'w'): 7
('w', '</w>'): 7
('w', 'er'): 6
('w', 'i'): 3
('i', 'd'): 3
('d', 'er'): 3
('w', 'e'): 2
('e', 's'): 2
('s', 't'): 2
('t', '</w>'): 2
```

4. هەنگاوی 2 و 3 دووپات بکەوە

```
er</w> ⇐ 'er', '</w>'
```

کۆرپس

```
'l o w </w>': 5
'l o w e s t </w>': 2
'n e w er</w>': 6
'w i d er</w>': 3
'n e w </w>': 2
```

وشەدان

```
d, e, i, l, n, o, r, s, t,
w, </w>, er, er</w>
```

فراوانیی جووتەکان

```
('n', 'e'): 8
('e', 'w'): 8
('l', 'o'): 7
('o', 'w'): 7
('w', '</w>'): 7
('w', 'er</w>'): 6
('w', 'i'): 3
('i', 'd'): 3
('d', 'er</w>'): 3
('w', 'e'): 2
('e', 's'): 2
('s', 't'): 2
('t', '</w>'): 2
```

4. هەنگاوی 2 و 3 دووپات بکەوە

ne ⇐ 'n', 'e'

کۆرپس

```
'l o w </w>': 5
'l o w e s t </w>': 2
'ne w er</w>': 6
'w i d er</w>': 3
'ne w </w>': 2
```

وشەدان

```
d, e, i, l, n, o, r, s, t,
w,</w>, er</w>, er, ne
```

فراوانیی جووتەکان

```
('ne', 'w'): 8
('l', 'o': 7
('o', 'w': 7
('w', '</w>': 7
('w', 'er</w>': 6
('w', 'i': 3
('i', 'd': 3
('d', 'er</w>': 3
('w', 'e': 2
('e', 's': 2
('s', 't': 2
('t', '</w>': 2
```

ڕێکاری BPE لە پاش 10 یەکخستن دەگا بەمانە:

کۆرپس

```
'low</w>': 5
'low e s t </w>': 2
'newer</w>': 6
'wider</w>': 3
'new </w>': 2
```

وشەدان

```
d, e, i, l, n, o, r, s, t, w, </w>
er, er</w>, ne, new, lo,
low, newer</w>, low</w>, wi, wid
```

فراوانیی جووتەکان

```
('wid', 'er</w>'): 3
('low', 'e'): 2
('e', 's'): 2
('s', 't'): 2
('t', '</w>'): 2
('new', '</w>'): 2
```

لەتکردنی تۆکن: ڕێکاری BPE

- لە کاردا، زۆربەی مۆدێلەکان و ئامێرەکان 32 بۆ 64 هەزار وشەڵەیان لە وشەداندا هەیە بۆ نموونە، GPT2 50257 وشەڵەی هەیە
- وشەڵەکان دەتوانن مۆرفیمەکان بن وەک «یەک» یان وشە باوەکان وەک «هەیە» یان هەر پێکهاتەیەکی باو
- لە پاش ئەژمارێکی دیاریکراوی یەکخستن لە سەر کۆرپسێکی گەورە، وشەدانەکە بۆ لەتکردنی تۆکنەکان بە کار دێ
- لەتکردنی تۆکنەکان بۆ دەقێکی نوێ بە پێی ئەو وشەڵانە دەکرێ کە فێریان بووین کەوا بێ، گرینگ نییە دەقە نوێیەکە چی تێدایە. وشەدانەکە تەنیا پشت بە کۆرپسەکە دەبەستێ.
- ئەم ڕێکارە لە زۆر مۆدێل و ئامێردا بە کار دێ، بە تایبەت لە مۆدیلەکانی زمان (language models)
- BPE تەنیا ڕێکار لە سەر وشەڵەکان نییە Unigram [6] و WordPiece [7]

- دەربڕینی فرەوشە (multiword expression) پێکهاتووە لە چەند بەش کە وێکڕا واتایەکی جودایان هەیە
  - بڕیاریان دا ← بڕیار یان دا
  - خشکە و پشکە ← خشکە و پشکە
  - گوێ لە مست ← گوێ لە مست
  - وەرە وەر ← وەرە وەر
  - bi-can-û-bên ← bi can û bên
- چۆن ئەم دەربڕینانە لەت بکرێن؟
  - بەکارهێنانی فەرهەنگی تایبەت بە دەربڕینە فرەوشەییەکان
  - بەکارهێنانی ڕوانگەی پێدراوە بە ڕەچاوکردنی بۆشایی؟

- مۆدێلە نۆرالْەکان (neural models) ی زمان سیستەمگەلی تەنیشت بە تەنیشت (end-to-end) ن لە پڕۆسەس کردنی زماندا و جێگەی ڕێڕەوە نەریتییەکانیان گرتووەتەوە.
- چەمکی ئێستای لەتکردنی تۆکن: پشتبەستن بە پێدراوە لە باتی یاسای زمانەوانی
  - دابەشکردن بۆ یەکەی بچووکتر و ناتایپۆگرافی وەک وشەلْەکان
  - تۆکنەکان لەگەلْ ڕێنووس یان یەکەی زمانەوانیدا یەک ناگرنەوە ⇐ تۆکن ≠ وشە
  - لەتکردنی تۆکنەکان بە پێی ڕێنووس یان زمانەوانی ← «لەتکردنی پێشەکیی» (pre-tokenization)
  - تۆکنە نەریتییەکان ← «پێش‌ـتۆکن» (pre-token)

- «یەک بەش کە بە کەللەڕەقی لە کۆنەوە ماوەتەوە لەتکردنی تۆکنە» [8]
- «شکۆی هەتاهەتایی بۆ هەر کەسێک کە بتوانێ لەتکردن وەک هەنگاوێکی پێویست لە مۆدێلە زمانییە گەورەکاندا (large language models, LLMs) بسڕێتەوە.» (Andrej Karpathy)
- بێ ئینسافی لە حاندی فرەزمانی [9]

		computational linguistics		**زمانەوانیی کۆمپیوتەری**
GPT-3	5	computational linguistics	32	زمانەوانیی کۆمپیوتەری
GPT-4	4	computational linguistics	22	زمانەوانیی کۆمپیوتەری
GPT-4o	4	computational linguistics	12	زمانەوانیی کۆمپیوتەری
		$ 🙂		$$$$ 😕

لەتکردنی تۆکن بۆ کوردی

له نووسینی کوردیدا، بۆشایی و نیشانەی نیوبۆشایی (ZWNJ) بۆ جوداکردنەوەی وشەکان بەکار دێن، بەڵام کێشەی زۆریان هەیە:

- یەکلانەبوونەوەی ڕێنووس:
  - 18ê ,18-ê ,18'ê
  - بەکار دێ، بەکاردێ، بە کار دێ، بە کاردێ
- نووساندنی زۆری وشەکان:
  - لەوێشدایە ← له وێ ش دا یه
- وشە لێکدراوەکان:
  مردوومراو
  خواخراوبۆکردگ
  «له کوردستانەوە کۆتایی بە پێوەندییەکە هێنا.» ← له ... ەوە؟ کۆتایی هێنان؟

لەتکردنی تۆکن بۆ کوردی لە KLPTدا

- لەتکردنی تۆکن لە KLPTدا پشت بە فەرهەنگێکی وشە و شیکاریی مۆرفۆلۆژی دەبەستێ
  - گونجاوە بۆ لەتکردنی زمانەوانی
  - یارمەتی دۆزینەوەی سنووری وشە دەدا
  - ڕێخۆشکەرە بۆ ئەرکی دیکە لە زمانەوانیی کۆمپیوتەریدا
- نموونەی لەتکردن:

دەق: بەرهەمهێنانی شێوازەکانی دواکەوتنی
دوای لەتکردن: دوا ـ کەوتن ـ ی شێواز ـ ەکان ـ ی بەرهەم ـ هێنان ـ ی

- لەم وتارەدا باسی کراوە: [10]
- KLPT: https://github.com/sinaahmadi/klpt

ڕێکاری KLPT: نموونە

1. کۆکردنەوەی وشە
2. وشە لێکدراوەکان بدۆزەوە
   کوردیی ناوەندی: 8180 وشە (1513 لێکدراو)
   کوردیی سەروو: 9970 وشە (1507 لێکدراو)
3. هەموو فۆرمە ئاڵۆز و نائاڵۆزەکانیان ساز بکە

نموونەی کوردیی سەروو

```
"bi-can-û-bên":
        "bicanûbên"
        "bi canûbên"
        "bican ûbên"
        "bi can ûbên"
        "bicanû bên"
        "bi canû bên"
```

نموونەی کوردیی ناوەندی

```
"ئاخر-و-ئۆخر":
    "ئاخرو ئۆخر"
    "ئاخروو‌ئۆخر"
    "ئاخر و ئۆخر"
    "ئاخروئۆخر"
    "ئاخر وئۆخر"
```

- لیستی پێشگر، پاشگر و کلیتیکەکان:
  پێشگر: لە، وە، دە، ڕە، بە
  پاشگر: ەکە، ەکان، ان، گەل
  جێناوە لکاوەکان: م، ت، ی، مان، تان، یان
- نموونەی شیکاریی مۆرفۆلۆژی:

بەرزترەکە ← بەرز + تر + ەکە

- ژمارەی مۆرفیمەکان:
  سۆرانی: 161 پاشگر و 11 پێشگر
  کورمانجی: 46 پاشگر و 17 پێشگر

ڕێکاری KLPT: نموونە

1. **پێش-پڕۆسەس کردن:**
   یەکخستنی نووسینی پیتەکان
   زیادکردنی بۆشایی لە دەوری خاڵبەندی و ژمارەکان
2. **لەتکردنی وشە لێکدراوەکان:**
   دۆزینەوەی وشە لێکدراوەکان لە فەرهەنگدا
   جوداکردنەوەیان بە نیشانەی ژێرهێڵ _
3. **لەتکردنی وشە:**
   جیاکردنەوە بە بۆشایی
   دۆزینەوەی وشەکان لە فەرهەنگدا
4. **شیکاریی مۆرفۆلۆژی:**
   دۆزینەوەی پێشگر و پاشگرەکان
   جیاکردنەوەی مۆرفیمە لێکدراوەکان

نموونە

«٢٥ کەسیان ڕێ کەوتن»

1. 25 کەسیان ڕێ کەوتن
2. 25 | کەسیان | __ڕێ-کەوتن__
3. 25 | _کەسیان_ | __ڕێ-کەوتن__
4. 25 | _کەس_ | _یان_ | __ڕێ-کەوتن__

- **قەبارەی بچووکی فەرهەنگ:**
  کاریگەری لە سەر کوالێتیی لەتکردن
- **ئاڵۆز بوونی واتای پێکهاتەکان:**
  بۆ نموونە وشەی «لاوین» کە دەکرێ «لاو» + «ین» یان وەک یەک وشە بێ، بەڵام سیستەمەکە بە '_لاو_ین' لەتی دەکا
- **پێکهاتە پەرژ و بڵاوەکان:**
  وەک لە «بڕیارێکی گەورەیان دا»دا، هەرچەند «بڕیار دان» کردارێکی لێکدراوە بەڵام وشەی دیکەیان کەوتووەتە نێوان
- **کێشەی کارایی:**
  - گەران لە فەرهەنگ کاتگیرە
  - ڕوانگەی پێدراوە؟

«دواکەوتنی شێوازەکانی بەرهەمهێنان لەم ئابووریانەدا دەگەڕێتەوە بۆ: نەبوونی هۆیەکانی تەکنیکیی تازەی هاورده تا بەرهەمهێنەکان بەکاری بێنن.»

KLPT

دواکەوتن ی شێواز ەکان ی بەرهەمهێنان
لەم ئابووریانەدا دەگەڕێتەوە بۆ : نەبوونی
هۆ یەکان ی تەکنیکیی تازە ی هاوردە
تا بەرهەمهێنەکان بەکاری بێن ن .

(4000) BPE

دو اکەوت نی شێواز ەکانی بەرهەم هێنان
لەم ئابووری انەدا دەگەڕ ێتەوە بۆ : نەبوونی
هۆی ەکانی تەک نی ک یی تاز ەی هاو رد
ە تا بەرهەم هێن ەکان بەک اری ب ێنن .

هەڵسەنگەندی لەتکردنی تۆکن پێوەندی بە ئەرکەوە هەیە.

سەرچاوەی هەموو وێنەکان ویکیمیدیایە: https://commons.wikimedia.org/wiki/Main_Page

[1] Jonathan J. Webster and Chunyu Kit.
Tokenization as the initial phase in NLP.
In *COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics*, 1992.

[2] Eric de la Clergerie and Lionel Clément.
MAF: a morphosyntactic annotation framework.
Actes de LTC, pages 90–94, 2005.

[3] David D Palmer.
Tokenisation and sentence segmentation.
Handbook of natural language processing, pages 11–35, 2000.

[4] Sabrina J Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y Lee, Benoît Sagot, et al.
Between words and characters: a brief history of open-vocabulary modeling and tokenization in NLP.
arXiv preprint arXiv:2112.10508, 2021.

[5] Rico Sennrich, Barry Haddow, and Alexandra Birch.
Neural machine translation of rare words with subword units.
In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.

[6] Taku Kudo.
Subword regularization: Improving neural network translation models with multiple subword candidates.
arXiv preprint arXiv:1804.10959, 2018.

[7] Mike Schuster and Kaisuke Nakajima.
Japanese and Korean voice search.
In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5149–5152. IEEE, 2012.

[8] Kris Cao and Laura Rimell.
You should evaluate your language model on marginal likelihood over tokenisations.
arXiv preprint arXiv:2109.02550, 2021.

[9] Aleksandar Petrov, Emanuele La Malfa, Philip HS Torr, and Adel Bibi.
Language model tokenizers introduce unfairness between languages.
arXiv preprint arXiv:2305.15425, 2023.

[10] Sina Ahmadi.
A tokenization system for the Kurdish language.
In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 114–127, 2020.