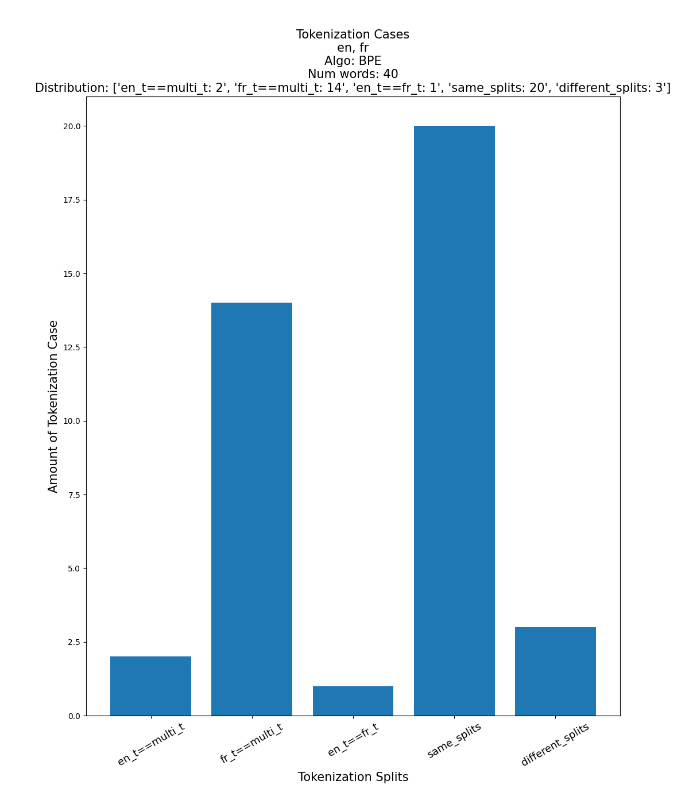
**Results Analysis – French – BPE vs BPE\_SAGE**

**Note:** The results shown in this file are for a vocabulary size of 3000. It is also important to note that the results are similar across other different vocabulary sizes (8000 and 16000)

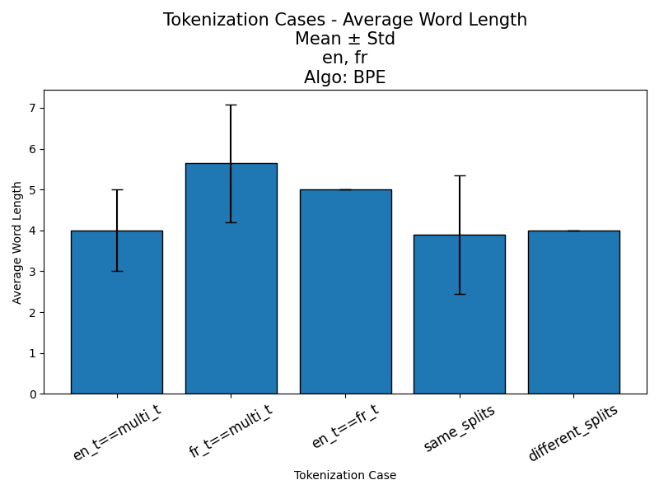
**False Friends Results:**

**Tokenization Cases:**

 A graph of a bar

AI-generated content may be incorrect.

**Average Word Length:**

 A graph of a number of blue rectangular objects

AI-generated content may be incorrect.

**Average Tokens:**

A graph of blue rectangular bars

AI-generated content may be incorrect. A graph of a number of blue rectangular objects

AI-generated content may be incorrect.

**Frequencies:**

A graph of a number of people

AI-generated content may be incorrect. A graph of different colored bars

AI-generated content may be incorrect.

**POS:**

A screenshot of a graph

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

**Word Movement 🡪 different splits:**

Note: Source distribution is BPE. The target distribution is BPE\_SAGE. We look at False Friends that moved to different splits in the target distribution

* **en\_t==multi\_t:** []
* **fr\_t==multi\_t:** ['lime', 'crayon', 'chair', 'affluent', 'caution', 'bond', 'sensible', 'douche']
* **en\_t==fr\_t:** ['groin']
* **same\_splits:** ['main', 'appoint', 'prune', 'son', 'agent', 'raisin', 'coin', 'chance']
* **different\_splits:** ['chef']

**Same Splits to Different Splits:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm** |  | **BPE\_SAGE**  **(Different Splits)** | | |
|  | **Words\Tokenizer** | **English Tokenizer** | **French Tokenizer** | **Multilingual Tokenizer** |
| **BPE (Same Splits)** | main🡪 [main] | [main] | [ma, in] | [m, a, i, n] |
| appoint🡪 [ap, point] | [appoint] | [a, p, point] | [a, p, po, i, nt] |
| prune🡪 [pr, une] | [p, run, e] | [p, r, une] | [pr, une] |
| son🡪 [son] | [s, on] | [son] | [s, o, n] |
| agent🡪 [ag, ent] | [a, ge, n, t] | [age, n, t] | [agent] |
| raisin🡪 [ra, is, in] | [rai, sin] | [r, ais, in] | [r, a, i, s, i, n] |
| coin🡪 [co, in] | [coin] | [c, o, in] | [c, o, i, n] |
| chance🡪 [ch, ance] | [chan, c, e] | [chan, ce] | [chance] |

**Homographs Results:**

**Tokenization Cases:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | en\_t==multi\_t | fr\_t==multi\_t | en\_t==fr\_t | Same Splits | Different Splits |
| **BPE** | 3289 | 7147 | 229 | 3820 | 4531 |
| **BPE \_SAGE** | 1181 | 1339 | 1386 | 389 | 14721 |

**More Results:**

**CPT (Characters Per Token):**

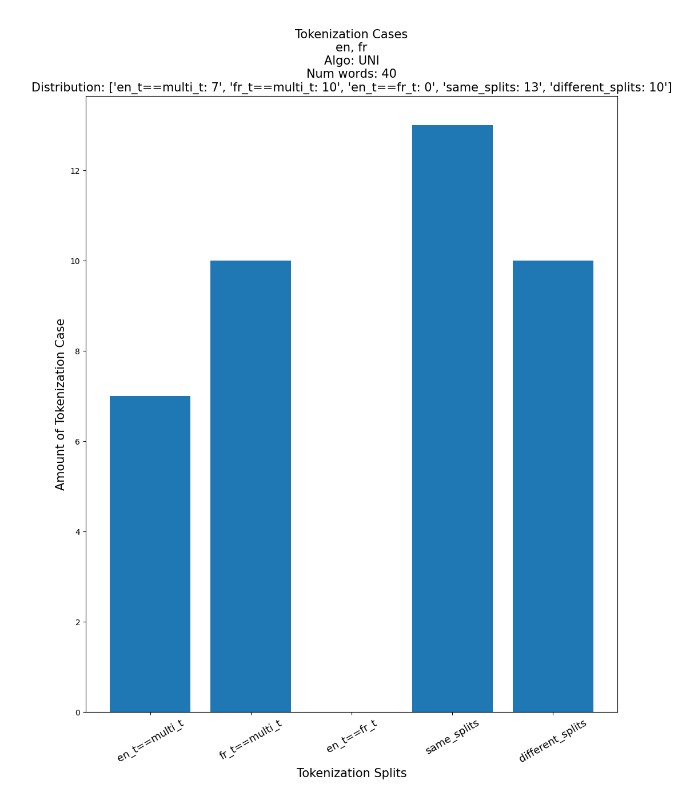
**Note:** CPT on the whole vocabulary

|  |  |  |  |
| --- | --- | --- | --- |
|  | **English Tokenizer** | **French Tokenizer** | **Multilingual Tokenizer** |
| **BPE** | 2.086 | 2.744 | 2.2 |
| **BPE\_SAGE** | 5.255 | 5.977 | 5.508 |

**Results Analysis – French – UnigramLM vs UNI\_SAGE**

**False Friends Results:**

**Tokenization Cases:**

 A graph with blue bars

AI-generated content may be incorrect.

**Average Word Length:**

A graph of a number of blue rectangular objects

AI-generated content may be incorrect. A graph of a number of blue rectangular objects

AI-generated content may be incorrect.

**Average Tokens:**

A graph of blue rectangular bars

AI-generated content may be incorrect. A graph of blue bars

AI-generated content may be incorrect.

**Frequencies:**

A graph of a number of people

AI-generated content may be incorrect. A graph of a number of people

AI-generated content may be incorrect.

**POS:**

A screenshot of a graph

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

**Word Movement 🡪 different splits:**

Note: Source distribution is UnigramLM. The target distribution is UNI\_SAGE. We look at False Friends that moved to different splits in the target distribution

* **en\_t==multi\_t:** ['appoint', 'ours']
* **fr\_t==multi\_t:** ['affluent', 'caution', 'ail']
* **en\_t==fr\_t:** []
* **same\_splits:** ['far']
* **different\_splits:** ['billion', 'prune', 'raisin', 'pain', 'groin']

**Same Splits to Different Splits:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm** |  | **UNI\_SAGE**  **(Different Splits)** | | |
|  | **Words\Tokenizer** | **English Tokenizer** | **French Tokenizer** | **Multilingual Tokenizer** |
| **UnigramLM (Same Splits)** | far🡪 [f, ar] | [f, a, r] | [f, ar] | [far] |

**Homographs Results:**

**Tokenization Cases:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | en\_t==multi\_t | fr\_t==multi\_t | en\_t==fr\_t | Same Splits | Different Splits |
| **UnigramLM** | 3463 | 5888 | 291 | 2201 | 7173 |
| **UNI \_SAGE** | 1648 | 5036 | 941 | 1650 | 9741 |

**More Results:**

**CPT (Characters Per Token):**

**Note:** CPT on the whole vocabulary

|  |  |  |  |
| --- | --- | --- | --- |
|  | **English Tokenizer** | **French Tokenizer** | **Multilingual Tokenizer** |
| **UnigramLM** | 2.666 | 3.168 | 2.833 |
| **UNI\_SAGE** | 5.199 | 6.007 | 5.611 |