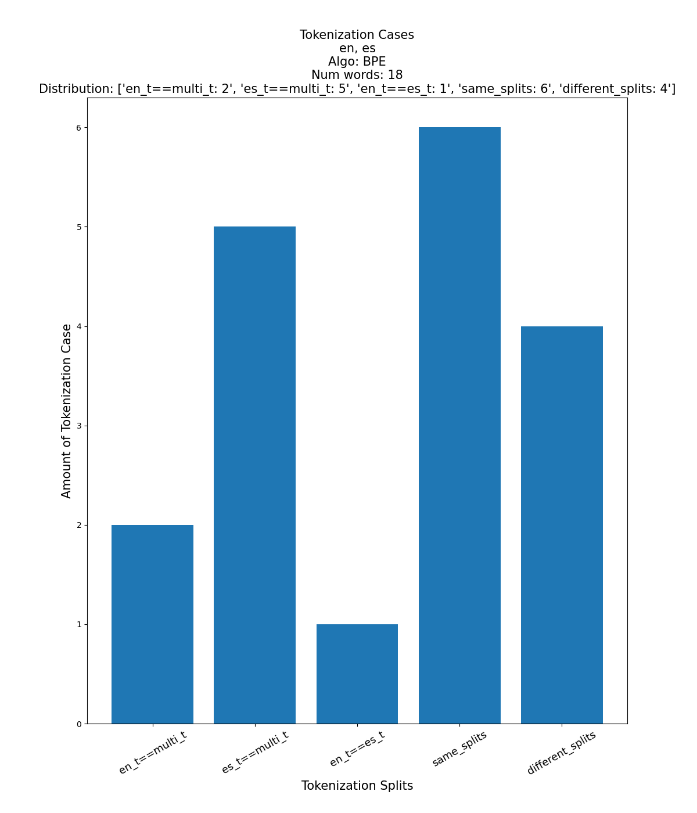
**Results Analysis – Spanish – BPE vs BPE\_SAGE**

**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 with blue bars

AI-generated content may be incorrect.

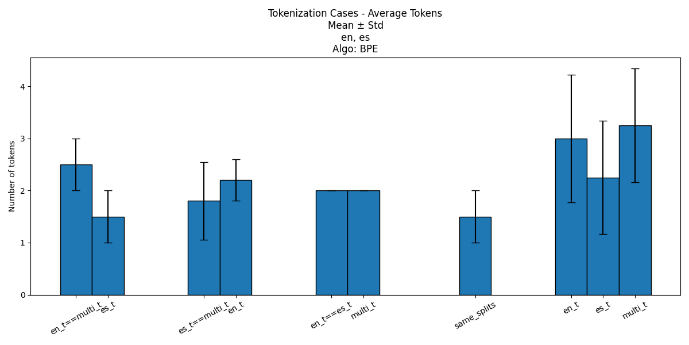
**Average Word Length:**

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 bars

AI-generated content may be incorrect.

**Frequencies:**

A graph of different colored bars

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 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:** []
* **es\_t==multi\_t:** ['papa', 'sensible', 'pan']
* **en\_t==es\_t:** ['sale']
* **same\_splits:** ['once', 'canto', 'son', 'eleven']
* **different\_splits:** ['propaganda', 'quince']

**Same Splits to Different Splits:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm** |  | **BPE\_SAGE**  **(Different Splits)** | | |
|  | **Words\Tokenizer** | **English Tokenizer** | **Spanish Tokenizer** | **Multilingual Tokenizer** |
| **BPE (Same Splits)** | once🡪  [on, ce] | [once] | [on, ce] | [o, n, c, e] |
| canto🡪 [can, to] | [can, t, o] | [canto] | [cant, o] |
| son🡪 [son] | [s, on] | [so, n] | [s, o, n] |
| eleven🡪 [ele, ven] | [el, even] | [el, eve, n] | [eleven] |

**Homographs Results:**

**Tokenization Cases:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | en\_t==multi\_t | es\_t==multi\_t | en\_t==es\_t | Same Splits | Different Splits |
| **BPE** | 2120 | 4662 | 165 | 2209 | 3898 |
| **BPE \_SAGE** | 972 | 773 | 1182 | 253 | 9874 |

**More Results:**

**CPT (Characters Per Token):**

**Note:** CPT on the whole vocabulary

|  |  |  |  |
| --- | --- | --- | --- |
|  | **English Tokenizer** | **Spanish Tokenizer** | **Multilingual Tokenizer** |
| **BPE** | 2.086 | 2.748 | 2.153 |
| **BPE\_SAGE** | 5.255 | 6.077 | 5.618 |

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

**False Friends Results:**

**Tokenization Cases:**

A graph with blue squares

AI-generated content may be incorrect. 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. 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 rectangular bars

AI-generated content may be incorrect.

**Frequencies:**

A graph of different colored bars

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 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:** ['alto', 'sale']
* **es\_t==multi\_t:** ['pie', 'quince', 'sensible', 'eleven']
* **en\_t==es\_t:** []
* **same\_splits:** ['canto']
* **different\_splits:** ['conductor', 'propaganda']

**Same Splits to Different Splits:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm** |  | **UNI\_SAGE**  **(Different Splits)** | | |
|  | **Words\Tokenizer** | **English Tokenizer** | **Spanish Tokenizer** | **Multilingual Tokenizer** |
| **UnigramLM (Same Splits)** | canto🡪 [can, to] | [can, to] | [cant, o] | [ca, n, to] |

**Homographs Results:**

**Tokenization Cases:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | en\_t==multi\_t | es\_t==multi\_t | en\_t==es\_t | Same Splits | Different Splits |
| **UnigramLM** | 2379 | 4760 | 110 | 1361 | 4444 |
| **UNI \_SAGE** | 1478 | 1788 | 1064 | 781 | 7943 |

**More Results:**

**CPT (Characters Per Token):**

**Note:** CPT on the whole vocabulary

|  |  |  |  |
| --- | --- | --- | --- |
|  | **English Tokenizer** | **Spanish Tokenizer** | **Multilingual Tokenizer** |
| **UnigramLM** | 2.666 | 3.517 | 2.776 |
| **UNI \_SAGE** | 5.199 | 5.681 | 5.52 |