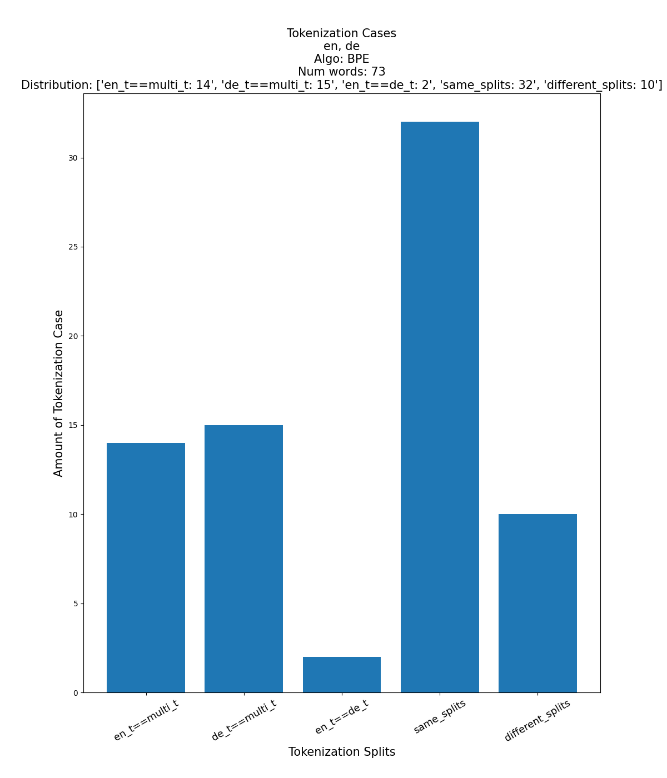
**Results Analysis – German – BPE vs BPE\_SAGE**

**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)

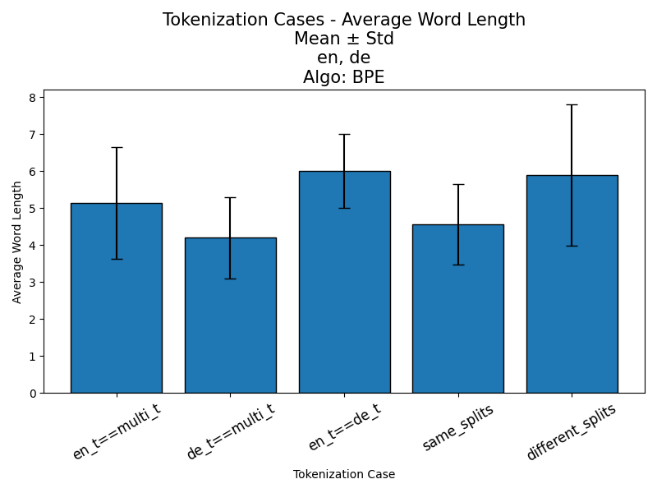
**False Friends Results:**

**Tokenization Cases:**

 A graph of a bar

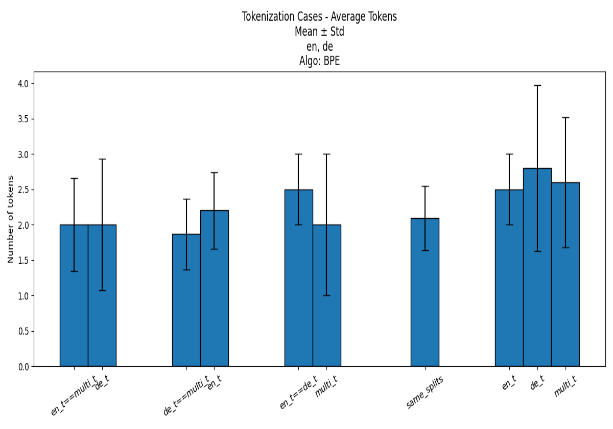
AI-generated content may be incorrect.

**Average Word Length:**

 A graph of a number of blue rectangular objects

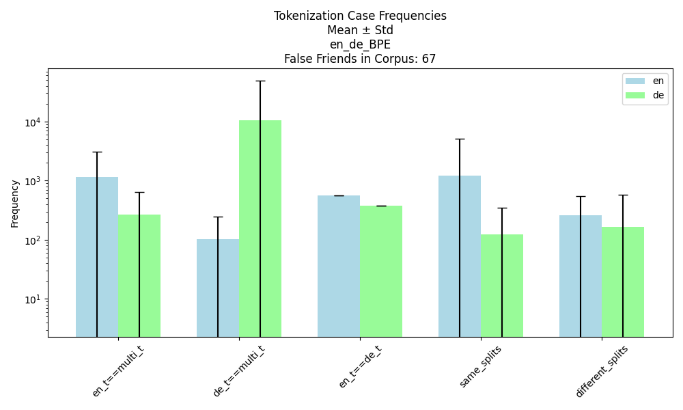
AI-generated content may be incorrect.

**Average Tokens:**

 A graph of blue bars

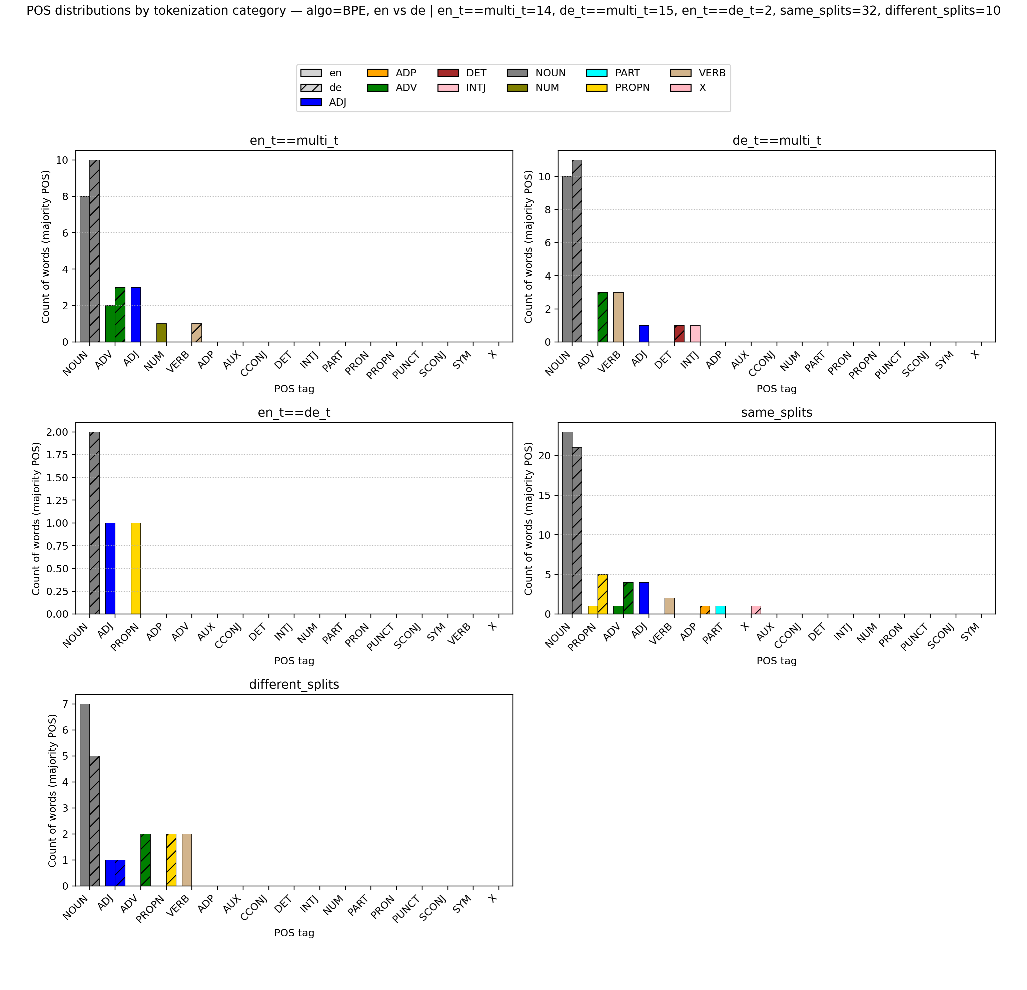
AI-generated content may be incorrect.

**Frequencies:**

 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.

**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:** ['bald', 'most', 'lunge', 'provision', 'stark']
* **de\_t==multi\_t:** ['lied', 'fund', 'mitten', 'pest', 'gut', 'spinner', 'hell', 'sage']
* **en\_t==de\_t:** ['schmuck', 'roman']
* **same\_splits:** ['heft', 'dank', 'also', 'beamer', 'hose', 'spanner', 'gaffer', 'probe', 'mode', 'sense', 'slip', 'jammer', 'tier', 'brand', 'pole', 'rind']
* **different\_splits:** ['happen', 'promotion', 'stadium']

**Same Splits to Different Splits:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm** |  | **BPE\_SAGE**  **(Different Splits)** | | |
|  | **Words\Tokenizer** | **English Tokenizer** | **German Tokenizer** | **Multilingual Tokenizer** |
| **BPE (Same Splits)** | heft🡪 [he, ft] | [h, ef, t] | [he, ft] | [h, e, f, t] |
| dank🡪 [d, ank] | [dan, k] | [dank] | [d, ank] |
| also🡪 [also] | [a, ls, o] | [als, o] | [also] |
| beamer🡪 [be, amer] | [beam, er] | [be, a, mer] | [b, e, ame, r] |
| hose🡪 [ho, se] | [hose] | [h, os, e] | [h, o, se] |
| spanner🡪 [sp, ann, er] | [spa, n, n, er] | [span, n, er] | [spann, er] |
| gaffer🡪 [g, af, fer] | [g, aff, er] | [ga, f, fer] | [g, a, ffer] |
| probe🡪 [pro, be] | [p, r, ob, e] | [pro, be] | [prob, e] |
| mode🡪 [mo, de] | [m, o, de] | [mo, de] | [mode] |
| sense🡪 [sen, se] | [sense] | [sen, se] | [se, ns, e] |
| slip🡪 [s, li, p] | [s, l, i, p] | [s, li, p] | [s, lip] |
| jammer🡪 [j, am, mer] | [jam, mer] | [j, a, m, mer] | [j, a, mm, er] |
| tier🡪 [t, ier] | [tie, r] | [tier] | [t, ier] |
| brand🡪 [br, and] | [bra, nd] | [brand] | [b, rand] |
| pole🡪 [po, le] | [pole] | [p, o, le] | [po, l, e] |
| rind🡪 [r, ind] | [r, in, d] | [rin, d] | [r, i, n, d] |

**Homographs Results:**

**Tokenization Cases:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | en\_t==multi\_t | de\_t==multi\_t | en\_t==de\_t | Same Splits | Different Splits |
| **BPE** | 3929 | 4377 | 681 | 5573 | 3334 |
| **BPE \_SAGE** | 1442 | 1588 | 1223 | 490 | 13151 |

**More Results:**

**CPT (Characters Per Token):**

**Note:** CPT on the whole vocabulary

|  |  |  |  |
| --- | --- | --- | --- |
|  | **English Tokenizer** | **German Tokenizer** | **Multilingual Tokenizer** |
| **BPE** | 2.086 | 3.411 | 2.415 |
| **BPE\_SAGE** | 5.255 | 5.932` | 5.574 |

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

**False Friends Results:**

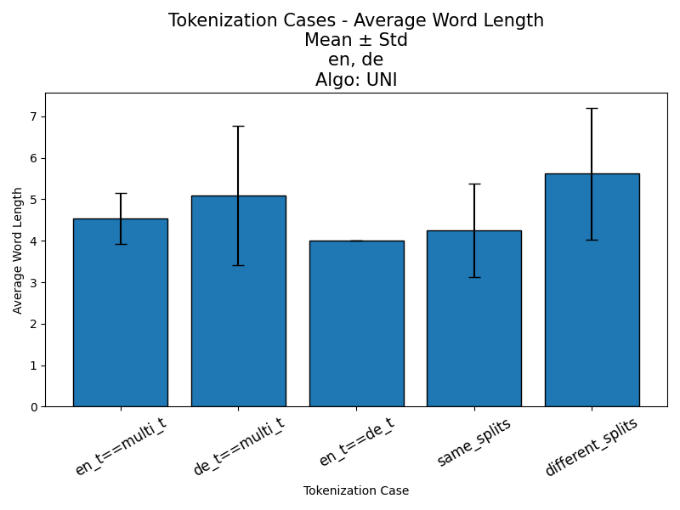
**Tokenization Cases:**

A graph of blue rectangular bars

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.

**Average Tokens:**

A graph of a graph of a number 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 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:** ['trunk', 'tier', 'still', 'happen', 'heft']
* **de\_t==multi\_t:** ['mitten', 'provision', 'spot', 'billion', 'fund', 'stark', 'spinner', 'promotion', 'hut', 'beamer', 'mist', 'sense', 'rate', 'mode', 'die', 'schmuck']
* **en\_t==de\_t:** []
* **same\_splits:** ['rat', 'fast', 'roman', 'pole', 'probe']
* **different\_splits:** ['spanner', 'stadium', 'links', 'evergreen', 'gaffer']

**Same Splits to Different Splits:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm** |  | **UNI\_SAGE**  **(Different Splits)** | | |
|  | **Words\Tokenizer** | **English Tokenizer** | **German Tokenizer** | **Multilingual Tokenizer** |
| **UnigramLM (Same Splits)** | rat🡪 [rat] | [r, a, t] | [rat] | [r, at] |
| fast🡪 [fast] | [fast] | [fa, s, t] | [f, a, s, t] |
| roman🡪 [roman] | [roman] | [rom, an] | [rom, a, n] |
| pole🡪 [pol, e] | [pole] | [p, ol, e] | [pol, e] |
| probe🡪 [pro, be] | [p, rob, e] | [probe] | [pr, ob, e] |

**Homographs Results:**

**Tokenization Cases:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | en\_t==multi\_t | de\_t==multi\_t | en\_t==de\_t | Same Splits | Different Splits |
| **UnigramLM** | 2833 | 6857 | 291 | 2640 | 5273 |
| **UNI \_SAGE** | 2270 | 1959 | 1121 | 900 | 11644 |

**More Results:**

**CPT (Characters Per Token):**

**Note:** CPT on the whole vocabulary

|  |  |  |  |
| --- | --- | --- | --- |
|  | **English Tokenizer** | **German Tokenizer** | **Multilingual Tokenizer** |
| **UnigramLM** | 2.666 | 4.132 | 3.036 |
| **UNI \_SAGE** | 5.199 | 5.034 | 5.536 |