TOKENIZATION RULES

Text elements annotated for MUC-7 Named Entity and Coreference tasks must consist of one or more complete tokens. This document explains where boundaries of tagged strings are meant to be located when there is NO explicit whitespace between alphanumeric characters and a punctuation mark or other special character.

1 Punctuation marks and special characters are normally considered separate tokens.

1.1 Examples with period (ellipsis, sentence-end punctuation)

"...Jaguar company in Britain."

```
...<ENAMEX TYPE="ORGANIZATION">Jaguar</ENAMEX> company in <ENAMEX
   TYPE="LOCATION">Britain</ENAMEX>.
```

1.2 Examples with hyphen or dash

"Chicago-based"

```
<ENAMEX TYPE="LOCATION">Chicago</ENAMEX>-based
```

"U.S.-based"

```
<ENAMEX TYPE="LOCATION">U.S.</ENAMEX>-based
```

"U.S.-Japan trade negotiations"

```
<ENAMEX TYPE="LOCATION">U.S.</ENAMEX>-<ENAMEX TYPE="LOCATION">Japan</
   ENAMEX> trade negotiations
```

"an Eaton-Sumitomo joint venture"

```
an <ENAMEX TYPE="ORGANIZATION">Eaton</ENAMEX>-<ENAMEX
   TYPE="ORGANIZATION">Sumitomo</ENAMEX> joint venture
```

"PHILADELPHIA--A new recycling center has been built."

```
<ENAMEX TYPE="LOCATION">PHILADELPHIA</ENAMEX>--A new recycling center has
   been built.
```

1.3 Examples with apostrophe (possessives)

"California's"

```
<ENAMEX TYPE="LOCATION">California</ENAMEX>'s
```

"Guiness' Schenley Industries"

```
<ENAMEX TYPE="ORGANIZATION">Guiness</ENAMEX>' <ENAMEX
   TYPE="ORGANIZATION">Schenley Industries</ENAMEX>
```

1.4 Examples with other punctuation

"(IBM)"

(<ENAMEX TYPE="ORGANIZATION">IBM</ENAMEX>)

""IBM stock fell today," he said" [note the double quote preceding IBM]

```
"<ENAMEX TYPE="ORGANIZATION">IBM</ENAMEX> stock fell today," he said
```

1.5 Examples with special characters

[this example no longer valid, but still illustrative until a substitute can be found]

"US$10"

```
<ENAMEX TYPE="LOCATION">US</ENAMEX><NUMEX TYPE="MONEY">$10</NUMEX>
```

2 When a proper name or number contains an internal punctuation mark or other special character, the word containing that character is treated as just one token.

2.1 Examples with period

2.1.1 A period that marks an abbreviation is considered part of the abbreviation token, even when the abbreviation appears at the end of a sentence.

"U.K. industry"
```
<ENAMEX TYPE="LOCATION">U.K.</ENAMEX> industry
```

"Microtest Inc."
```
<ENAMEX TYPE="ORGANIZATION">Microtest Inc.</ENAMEX>
```

"Spokane, Wash."
```
<ENAMEX TYPE="LOCATION">Spokane</ENAMEX>, <ENAMEX TYPE="LOCATION">Wash.</
   ENAMEX>
```

"Limousines are manufactured in the U.K."
```
 Limousines are manufactured in the <ENAMEX TYPE="LOCATION">U.K.</ENAMEX>
```

2.1.2 A period used as a decimal marker is considered integral to the number token.

"$5.10"
<NUMEX TYPE="MONEY">$5.10</NUMEX>

2.2 Examples with hyphen or dash (see also section 3, below)

"F. Gregory Fitz-Gerald"
```
<ENAMEX TYPE="PERSON">F. Gregory Fitz-Gerald</ENAMEX>
```

"Prudential-Bache Securities"
```
<ENAMEX TYPE="ORGANIZATION">Prudential-Bache Securities</ENAMEX>
```

"Allen-Bradley Co. and Hewlett-Packard Co. have undertaken a joint marketing and development program linking A-B manufacturing automation equipment with HP Unix-based computers."
```
<ENAMEX TYPE="ORGANIZATION">Allen-Bradley Co.</ENAMEX> and <ENAMEX
   TYPE="ORGANIZATION">Hewlett-Packard Co.</ENAMEX> have undertaken a joint
   marketing and development program linking <ENAMEX TYPE="ORGANIZATION">A-
   B</ENAMEX> manufacturing automation equipment with <ENAMEX
   TYPE="ORGANIZATION">HP</ENAMEX>...
```

"one-hundred percent"
```
<NUMEX TYPE="PERCENT">one-hundred percent</NUMEX>
```

2.3 Examples with slash

"The venture will be called Quality Spring/Togo Inc."
```
The venture will be called <ENAMEX TYPE="ORGANIZATION">Quality Spring/
   Togo Inc.</ENAMEX>
```

"10/13/89"
```
<NUMEX TYPE="DATE">10/13/89</NUMEX>
```

2.4 Examples with other punctuation (special uses of apostrophes)

"McDonald's burger company"

```
<ENAMEX TYPE="ORGANIZATION">McDonald's</ENAMEX> burger company
```

"'87"

```
<TIMEX TYPE="DATE">'87</TIMEX>
```

2.5 Examples with special characters

"S&P 500 Index"

```
<ENAMEX TYPE="ORGANIZATION">S&P</ENAMEX> 500 Index
```

3 Hyphen at end of line

3.1 When a hyphen is used at the end of a line to separate a single word into two parts, the word is treated as a single token.

"Phila-
delphia"

```
<ENAMEX TYPE="LOCATION">Phila-
delphia</ENAMEX>
```

3.2 If, however, the word is naturally hyphenated and the hyphenated word just happens to be broken at the hyphen at the end of a line, the parts of the word are treated as separate tokens.

"Chicago-
 based"

```
<ENAMEX TYPE="LOCATION">Chicago</ENAMEX>-
based
```