

# Tokenization Guidelines for ETTB 2.0

Justin Mott, Colin Warner, Ann Bies; Ann Taylor  
Linguistic Data Consortium; University of York  
{jmott|colinw|bies}@ldc.upenn.edu; at9a@york.ac.uk

April 6th, 2009

## 1 Overview

All strings separated by white space are treated as separate tokens. Also, no token can contain white space. The following contractions and related items are split into separate tokens. So, *women's*, *would've*, *cannot*, *he'll* are tokenized as **women 's**, **would 've**, **can not**, **he 'll**.

's  
've  
're  
'll  
'd  
n't  
can not  
gon na  
got ta  
lem me  
more 'n  
't is  
't was  
wan na  
wha dd ya  
wha t cha

## 2 Hyphenated Words

Most hyphenated words are split into multiple tokens. For example, *elementary-school-age* is now treated as five tokens, viz.- elementary(1) -(2) school(3) -(4) age(5). Hyphenated interjections and affixes in the following list are not split into multiple tokens. For example, *uh-oh* and *e-mail* are both single tokens: **uh-oh**, **e-mail**.

e-  
a-  
u-  
x-  
anti-  
agro-  
be-  
bi-  
bio-  
co-  
counter-  
cross-  
cyber-  
de-  
eco-  
-esque  
ex-  
extra-  
-fest  
-fold  
-gate  
inter-  
intra-  
-itis  
-less  
macro-  
mega-  
micro-  
mid-  
mini-  
mm-hm  
mm-mm  
-most  
multi-

neo-  
non-  
o-kay  
over-  
pan-  
para-  
peri-  
post-  
pre-  
pro-  
pseudo-  
quasi-  
-rama  
re-  
semi-  
sub-  
super-  
tri-  
uh-huh  
uh-oh  
ultra-  
un-  
uni-  
vice-  
-wise

### 3 Punctuation

All other punctuation not described above triggers a break in tokenization, with the exceptions outlined below. Note that for present purposes, all non-alphanumeric characters are considered ‘punctuation’.

1. Periods marking abbreviations.

Mr. Dr. Ste. Ave. etc. e.g. A.D.

2. Punctuation in web addresses.

<http://www.islamonline.net/Arabic/news/2004-12/05/images/pic05b.jpg>

<http://www.mofa.gov.sa/detail.asp?InNewsItemID=59090&InTemplateKey=print>  
rayhanenajib@menara.ma

3. Ellipses, when encoded as a string of periods. In addition, ellipses enclosed by round brackets are treated as single tokens in Sinorama.

...  
(...)

4. Complex numerals.

2.45  
20,000

5. Telephone numbers and postal codes.

tel : 02-2348-2192

6. Single quotation marks as part of words.

the 80's  
P'yongyang  
'Assad

### 3.1 Punctuation in Webtext

Since the use of punctuation in webtext differs markedly from usage in carefully-edited text, it is handled slightly differently. The tokenization of punctuation in webtext is determined by whitespace boundaries. That is, all strings of punctuation (with the exception of quotation marks) without intervening whitespace are treated as a single token. So, a sequence such as `!!!!!!!!!!!!!!!!` is treated as one token, rather than nineteen. Examples in include the following.

```
;)  
:-)  
..  
?!  
!!!!!!!!!!!!!!!!  
*****  
- _____ -
```