Tokenization Guidelines for English, French, and German

Martin Volk (volk@cl.uzh.ch)

September 2, 2016

These guidelines are meant first and foremost for internal use at the Institute of Computational Linguistics of the University of Zurich. The guidelines may be modified over time. When quoting please refer to the date. For additions and comments please contact the author.

Tokenization Guidelines for English (following the Penn Treebank)

1. Dots

Rule	**Examples**
a sentence final dot is split and a separate token	He lives in New York. → He lives in New York .
a dot after a number is split (in EN); also for Roman numbers	in Windows 95. → in Windows 95 . XV. → XV .
a dot inside a number is NOT split	20.5
a dot inside or at the end of an abbreviation is NOT split (even if the abbreviation is at the end of a sentence; this is **different** from the Penn Treebank)	etc. p.m. m.p.h. i.e. e.g. This is PaineWebber Inc.
a dot after a single letter is NOT split (unless it is Roman number)	Peter M. Miller J.V. Miller Affoltern a. A.
a dot in a URL or email address is NOT split	www.uzh.ch volk@cl.uzh.ch
a sequence of 3 dots is NOT split and one token	apples, bananas …

2. Question Marks and Exclamation Marks

Rule	**Examples**
a question mark is split and regarded as sentence boundary	How big? → How big ?
an exclamation mark is split and regarded as sentence boundary	So big! → So big !

3. Commas

Rule	**Examples**
Commas as punctuation symbols in sentences are split	visit, → visit ,
Commas inside numbers are NOT split	25,000

4. Colons and Semicolons

Rule	**Examples**
a colon or semicolon at the end of a word is split	Peter says: Hello → Peter says : Hello She was an inspirational lady; she had … → She was an inspirational lady ; she had …
a colon in a sports result or a time expression is NOT split	3:1 22:28 at 9:30
a colon in a number ratio is NOT split	1:25,000
a colon or semicolon in an emoticon is NOT split	:-) ;-)

5. Apostrophes

Note: Consider different types of apostrophes (' ’)

Note: Consider that the text can be all UPPERCASE.

Rule	**Examples**
possessive **'s** is split and a separate token	Peter's → Peter 's [lemma="'s" PoS="POS"] my brother's → my brother 's
possessive **'** is split and a separate token	workers' lives → workers ' lives [lemma="'s" PoS="POS"]
a contracted word with leading apostrophe is split and a separate token	He's → He 's [lemma="be" PoS="VBZ"] I'm → I 'm you're → you 're they've → they 've [lemma="have" PoS="VBP"] you'd → you 'd [lemma="have" PoS="MD" or PoS="VBD"] let's → let 's [lemma="us" PoS="PRP"]
an apostrophe is also split when it is the symbol for minutes	35'46"N → 35 ' 46 " N
an apostrophe as single quote is a separate token	called 'The Snuff Box' → called ' The Snuff Box '
a word or a number that starts with an apostrophe is NOT split	the '80s in '95
the word **n't** with word-internal apostrophe is NOT split but a separate token	**didn't → did n't** [lemma="not" PoS="RB"] can't → ca [lemma="can" PoS="MD"] n't won't → wo [lemma="will" PoS="MD"] n't ain't → ai [lemma="be" PoS="VBP"] n't
single letters with plural **'s** are NOT split	A's
Irish family names with apostrophe are NOT split	O'Connor, O'Loughlin, O'Neill

6. Quotation Marks

Note: Consider different types of quotation marks (", ”, “, «, », ‹, ›)

Rule	**Examples**
a quotation mark is split and a separate token	“What I Learned From Frogs in Texas” → “ What I Learned From Frogs in Texas ” "50 Best Innovations" → " 50 Best Innovations "
a quotation mark is split when it is the symbol for seconds	35'46"N → 35 ' 46 " N

Note: Splitting undirected quotation marks (e.g. ") from words results in an information loss. Such quotation marks in front of a word (e.g. "Hello) indicate the **beginning** of a quote, whereas quotation marks at the end of a word (e.g. World") indicate the **end** of a quote. When we split an undirected quotation mark from a word, the mark becomes ambiguous. It is now unclear whether it starts or ends a quote (e.g. Hello " World). Therefore we consider adding a special symbol to keep the direction information (e.g. "Hello → "^ Hello; e.g. World" → World ^").

7. Hyphens, Dashes

Note: Consider different types of hyphens (short: -, long: –)

Rule	**Examples**
hyphens with numbers and measurement units (in, ft, liter, meter, mile) are split	8-in square → 8 –in square 12-ft boat → 12 –ft boat 90-mile stretch → 90 –mile stretch 2.4-liter units
hyphens with numbers and non-measurements are NOT split	21-year-old 30-minute talk 18-hole course 20th-century a 40-strong delegation
a hyphen that connects compounds or the like is NOT split	event-driven risk-free non-commodity ice-walls life-or-death

a hyphen that stands for decimals at the end of a number is NOT split	2,50.- 37,700.-
hyphens in number ranges are NOT split	2-3 hours 11-14-year-olds

8. Slashes

Rule	**Examples**
a slash in an alternative is split (this is **different** from the Penn Treebank)	September/October → September / October
a slash in an apposition is split (this is **different** from the Penn Treebank)	Lugano/TI → Lugano / TI
a slash in a URL is NOT split	www.cl.uzh.ch/volk
a slash in a number ratio is NOT split	3/5

9. Parentheses, Curly Braces, Square Brackets, Angle Brackets

Note: Consider different types of parentheses etc. (){}[]<>

Rule	**Examples**
parentheses, braces and brackets are split and are separate tokens	(EU) → (EU)

10. Measurement and Currency Units

Note: the suffixes for ordinal numbers are NOT considered measurement units (21st, 42nd, 3rd, 115th) and are thus NOT split.

Rule	**Examples**
a measurement unit is split and a separate token	1579m → 1579 m [lemma="meter" PoS="NN"] 500kg → 500 kg 10pm → 10 pm 10p.m. → 10 p.m. 15% → 15 % [lemma="percent" PoS="NN"]
a currency unit is split and a separate token	55CHF → 55 CHF US$15 → US$ 15 $15 → $ 15

11. Numbers

Rule	**Examples**
a number consisting of digits and fraction symbols is NOT split (this is **different** from the Penn Treebank)	1½, 5¾, 1/10th
a number consisting of digits and a non-symbol fraction is split	12 3/7
a number consisting of digits plus ordinal suffix or age suffix is NOT split	3rd 25th 1980s
a number consisting of digits and words is left as separate tokens	10.5 million
a Roman number is NOT split	XIV
a spelled out number is NOT split	sixty-four, twenty-fifth, …

12. Ligatures

Ligatures are symbols that represent letter pairs for typographical beauty. Typical ligatures are **fi** and **fl**. Ligatures are often found in text that has been extracted from a PDF document. [This is more an encoding than a tokenization issue.]

Rule	**Examples**
ligatures are converted into the corresponding letter sequences	find → find

13. Mathematical and Miscellaneous Symbols

Rule	**Examples**
the ampersand is a separate token unless it is inside an acronym	Cooper Tire & Rubber Co. American Telephone & Telegraph AT&T

Open issues:

- **Words which contain XML tags** (e.g. CO<sub>2</sub>-poor). They shall be left as one token.

Tokenization Guidelines for German

… which deviate from the English tokenization guidelines.

1. Dots

Rule	**Examples**
a dot after a number is NOT split (in DE); also for Roman numbers	in Windows 95. XV.
OLD: an acronym that is spelled with internal blanks is contracted to one token	S. A. C. → S.A.C.

2. Question Marks and Exclamation Marks

3. Commas

4. Colons and Semicolons

5. Apostrophes

Rule	**Examples**
a contracted word with a leading apostrophe is split and a separate token (like in EN)	Wie geht's → Wie geht 's
a word with an apostrophe and the suffix *sche* is NOT split	Müller'sche Meyer'schen

6. Quotation Marks

7. Hyphens, Dashes

Rule	**Examples**
a hyphen at the end of a word or a number is NOT split	Eis- und Felskletterei 12- bis 24-monatigen
a hyphen that stands for decimals at the end of a number is NOT split	21,-

8. Slashes

Rule	Examples
a slash that marks an alternative within a word is NOT split	Sportler/in Lehrer/innen

9. Parentheses, Curly Braces, Square Brackets, Angle Brackets

Rule	Examples
parentheses, braces and brackets are split and are separate tokens unless …	(EU) → (EU)
the matching symbol is inside the word	(EU)-Mitgliedschaft Berg(halb)-schuhen
a closing parenthesis after a single letter, a Roman or Arabic number is NOT split	a) ix) 12.)

10. Measurement and Currency Units

11. Numbers

Rule	Examples
a number with an internal blank is connected with an underscore	75 000 → 75_000 Tel. 076 543 271 → Tel. 076_543_271
a word that starts with digits and continues with letters is NOT split	25-jährige 43-Jährige 14tägige 8stündiger 2,5stündige ½stündige 4000er
a word that starts with letters and ends with digits is NOT split	K2, A2, CO2, Q3, B42 km2, m3, cm3

12. Ligatures
13. Mathematical and Miscellaneous Symbols

Tokenization Guidelines for French

… which deviate from the English tokenization guidelines.

1. **Dots**

Rule	**Examples**
a dot after a number is NOT split (in FR); also for Roman numbers	Windows 95. XV.

2. **Question Marks and Exclamation Marks**
3. **Commas**
4. **Colons and Semicolons**

5. **Apostrophes**

Rule	**Examples**
a contracted word with a trailing apostrophe is split and a separate token	l'eau → l' eau n'a plus → n' a plus qu'il ne s'envoie → qu' il ne s' envoi
a few words with internal apostrophes are NOT split	aujourd'hui c'est-à-dire l'on [lemma="il" PoS="CL_suj"] quelqu'un

6. **Quotation Marks**

7. **Hyphens, Dashes**

Rule	**Examples**
a hyphen between a verb and a pronoun is split; the –t marker is attached to the subsequent pronoun	regrettes-tu → regrettes –tu demandons-nous → demandons –nous affirme-t-il → affirme –t-il va-t-on → va –t-on
a hyphen between a noun and a demonstrative (ci, là, meme) is split. Lexicalized demonstratives like celui-ci, ceux-ci, celui-là are NOT split.	fois-ci → fois –ci jours-ci → jours –ci côté-la → côté –là moment-là → moment –là été-là → été –là

8. **Slashes**
9. **Parentheses, Curly Braces, Square Brackets, Angle Brackets**
10. **Measurement and Currency Units**
11. **Numbers**
12. **Ligatures**
13. **Mathematical and Miscellaneous Symbols**