

# Package: pipian (via r-universe)

July 16, 2024

**Type** Package

**Title** Tiny Interface to CaboCha for R

**Version** 0.3.9

**Maintainer** Akiru Kato <paithiov909@gmail.com>

**Description** A tiny interface to 'CaboCha'; a Japanese dependency structure parser. The main goal of this package is to implement a parser for that XML output.

**License** MIT + file LICENSE

**URL** <https://github.com/paithiov909/pipian>,  
<https://paithiov909.github.io/pipian/>

**BugReports** <https://github.com/paithiov909/pipian/issues>

**Depends** R (>= 4.0.0)

**Imports** dplyr, igraph, purrr, Rcpp, readr, rlang (>= 0.1.2), stringi, utils

**Suggests** knitr, rmarkdown, roxygen2, spelling, testthat (>= 3.0.0)

**LinkingTo** Rcpp

**Config/testthat/edition** 3

**Encoding** UTF-8

**Language** en-US

**RoxygenNote** 7.3.1

**SystemRequirements** CaboCha

**Repository** <https://paithiov909.r-universe.dev>

**RemoteUrl** <https://github.com/paithiov909/pipian>

**RemoteRef** HEAD

**RemoteSha** 58e5c40819848cd0766b94a682df11b212e49aad

## Contents

ngram_tokenizer	2
pack	2
ppn_cabocho	3
ppn_make_graph	4
ppn_parse_xml	4

<b>Index</b>	<b>6</b>
--------------	----------

---

ngram_tokenizer	<i>Ngrams tokenizer</i>
-----------------	-------------------------

---

### Description

Make an ngram tokenizer function.

### Usage

```
ngram_tokenizer(n = 1L)
```

### Arguments

n                   Integer.

### Value

ngram tokenizer function

---

pack	<i>Pack prettified data.frame of tokens</i>
------	---

---

### Description

Packs a prettified data.frame of tokens into a new data.frame of corpus, which is compatible with the Text Interchange Formats.

### Usage

```
pack(tbl, pull = "token", n = 1L, sep = "-", .collapse = " ")
```

### Arguments

tbl                   A prettified data.frame of tokens.  
 pull                  Column to be packed into text or ngrams body. Default value is 'token'.  
 n                     Integer internally passed to ngrams tokenizer function created of audubon::ngram\_tokenizer()  
 sep                   Character scalar internally used as the concatenator of ngrams.  
 .collapse            This argument is passed to stringi::stri\_join().

**Value**

A data.frame.

**Text Interchange Formats (TIF)**

The Text Interchange Formats (TIF) is a set of standards that allows R text analysis packages to target defined inputs and outputs for corpora, tokens, and document-term matrices.

**Valid data.frame of tokens**

The prettified data.frame of tokens here is a data.frame object compatible with the TIF.

A TIF valid data.frame of tokens are expected to have one unique key column (named 'doc\_id') of each text and several feature columns of each tokens. The feature columns must contain at least 'token' itself.

**See Also**

<https://github.com/ropensci/tif>

---

ppn\_cabocha

*Execute cabocha command*

---

**Description**

Execute 'cabocha -f3 -n1' command using system2, then return the paths to the temporary XML files.

**Usage**

```
ppn_cabocha(text, rcpath = NULL)
```

**Arguments**

text	A character vector to be parsed with CaboCha.
rcpath	String; path to the 'mecabrc' file if any.

**Value**

Paths to the CaboCha XML output are returned.

**Examples**

```
## Not run:  
ppn_cabocha(enc2utf8("\u96e8\u306b\u3082\u8ca0\u3051\u305a"))  
  
## End(Not run)
```

---

ppn\_make\_graph            *Cast dependency structure as an igraph*

---

### Description

Cast dependency structure as an igraph

### Usage

```
ppn_make_graph(df)
```

### Arguments

df                        Output of pipian::ppn\_parse\_xml.

### Value

An 'igraph' object is returned.

### Examples

```
xml <- ppn_parse_xml(system.file("sample.xml", package = "pipian"))
ppn_make_graph(xml)
```

---

ppn\_parse\_xml            *Parse XML output of CaboCha*

---

### Description

Parse XML output of CaboCha

### Usage

```
ppn_parse_xml(
  path,
  into = c("POS1", "POS2", "POS3", "POS4", "X5StageUse1", "X5StageUse2", "Original",
           "Yomi1", "Yomi2"),
  col_select = seq_along(into)
)
```

### Arguments

path                      String; output from pipian::ppn\_cabochoa.

into                      Character vector; feature names of output.

col\_select                Character or integer vector; features that will be kept in the result.

*ppn\_parse\_xml*

5

**Value**

A data.frame.

**Examples**

```
head(ppn_parse_xml(system.file("sample.xml", package = "pipian")))
```

# Index

`ngram_tokenizer`, 2

`pack`, 2

`ppn_cabocha`, 3

`ppn_make_graph`, 4

`ppn_parse_xml`, 4