BCCWJ

(Balanced Corpus of Contemporary Written Japanese)

- 100,000,000 words corpus for contemporary written Japanese
- Available from http://www.kotonoha.gr.jp/shonagon/search_form
- Produced by National Institute of Japanese Language and Linguistics (NINJAL).
- XML compliant
- POS Tagged

Creation

3 SUBCORPUS

- Publication Subcorpus (35 million words)

Books, magazines and newspapers published during 2001-2005

- Library Subcorpus (30million words)

Books catalogued at more than 13 public libraries in Tokyo after 1985.

- Special Purpose Subcorpus (35 million words)

Whitepaper text, internet text, Best selling books

Meta-data/Annotation

Meta-Data

- The source of information
  - The publication name
  - The author name
  - The publisher
  - The year of publication
  - The genre categorized by NDC (Nippon Decimal Classification)

Annotation

- Each text has the information for sentence structure
  - Article/Paragraph/Sentence/LUW/SUW/Character

    (Ex.) LUW -> 公共工事請け負い金額 (the amount billed for public construction)

    SUW -> 動き (action)、公共（public）、工事（construction）

    (Ex.) 大/雨/が/降っ/た/の/で (because of the heavy rain)

    東京/都 (Tokyo prefecture)

Usage

- Used to examine the frequency of the polite form and normal form in Japanese written text
- Normal form -> literature

  Polite form -> Philosophy, Natural Science, Industry, How-to books, juvenile literature

Report:

(https://www.ninjal.ac.jp/event/specialists/project-meeting/files/JCLWorkshop_no1_papers/JCLWorkshop2012_27.pdf)

Reference

- Takehiko, M. (2012) . The Role of Metadata in Analysis of Large-scale Corpera. Retrieved from https://www.ninjal.ac.jp/event/specialists/project-meeting/files/JCLWorkshop_no1_papers/JCLWorkshop2012_27.pdf
- National Institute for Japanese Language and Liguistics. (2012) . Balanced Corpus of Contemporary Written Japanese. Retrieved from http://www.ninjal.ac.jp/english/products/bccwj/
- Kikuo, M.(2006) . "Kotonoha, the Corpus Development Project of the National Institute for Japanese Language." Language Corpora:Their Compilation and Application pp.55-62.

  Retrieved from

  http://www2.ninjal.ac.jp/kikuo/NIJLSymp06_KM2rev.pdf

ISLRN

Title	BCCWJ
Full Title	Balanced Corpus of Contemporary Written Japanese
Resource type	Text
Source/URL	http://www.kotonoha.gr.jp/shonagon/
Format/MIME Type	Text/XML
Size/Duration	100,000,000 words
Access Medium	Online/DVD
Description	BCCWJ is a balanced corpus of one hundred million words of contemporary written Japanese. BCCWJ is one of the components of KOTONOHA. It is probably the most important of all the KOTONOHA component corpora, because it is the written register of the contemporary Japanese that is the greatest focus of interest for language researchers as well as the general public. It is also the contemporary written language that has the greatest applicability to such applications as dictionaries and teaching materials. The compilation of BCCWJ started in 2006 as a five-year project, and is supported partly by a Grant-in-Aid for Scientific Research on Priority Area from MEXT (Japanese ministry of education) : Japanese Corpus.
Version	?
Language	Japanese
Resource Creator	National Institute for Japanese Language and Linguistics
Right Holder	National Institute for Japanese Language and Linguistics
Distributor	National Institute for Japanese Language and Linguistics