

尚書古活字版を対象とした訓点データベースにおける検索性の改良

The Improvements of the Searchability for Shōsho Kuntan Database

September 13th 2023

Koji Tajima, Kota Tomabechi,
Tomoaki Tsutsumi, Teiji Kosukegawa,
Tomokazu Takada

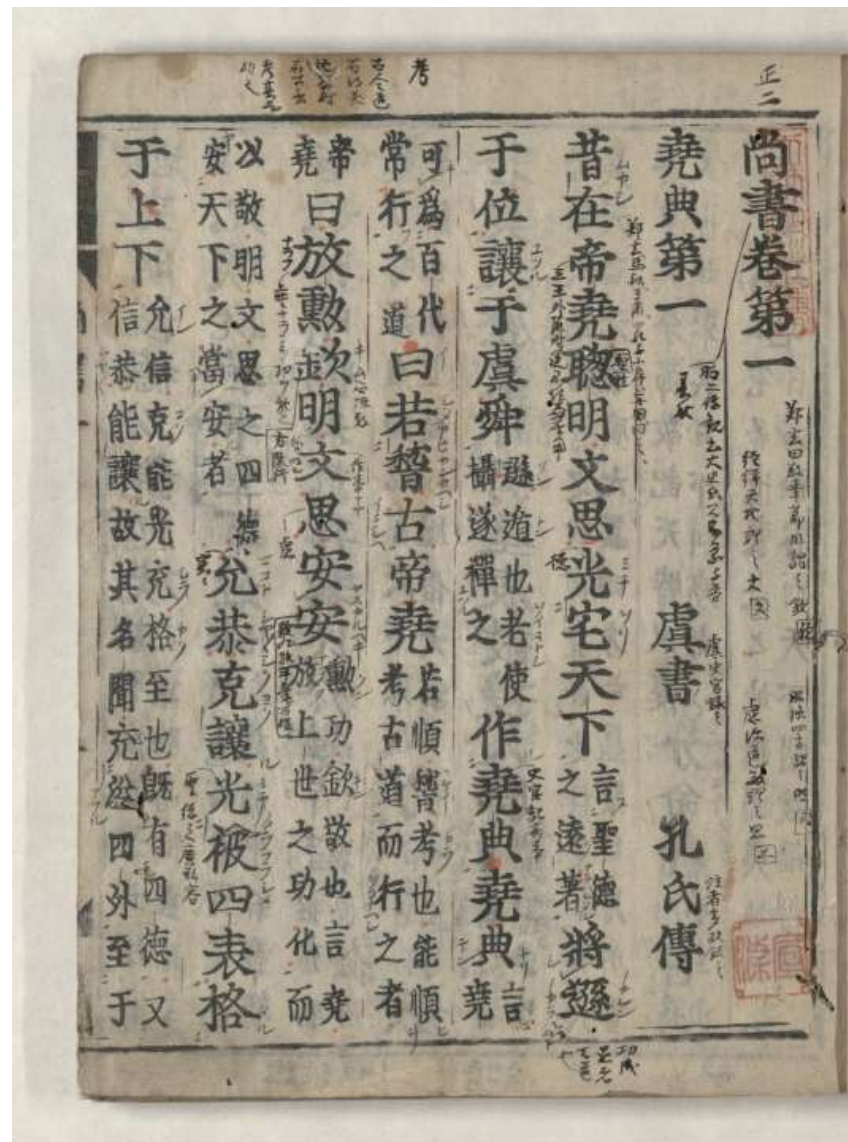


Topics

- Glossed (Kunten) Material
- Research results of the kunten material
- Glosses in the kunten material
- Challenges
- The data structure
- Data input tool
- Samples and the aggregate results

Glossed(Kunten) Material

訓点資料



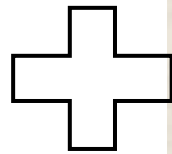
国宝 古文尚書卷第六

<https://bunka.nii.ac.jp/heritages/detail/426008>

Glossed(Kunten) Material

訓点資料

- Original text (written in classical Chinese text)
 - It is a hand written copy or printed



- Marks or memo
 - to read a original text in the local language. (Japanese, Korean)
 - Or research finding of the material



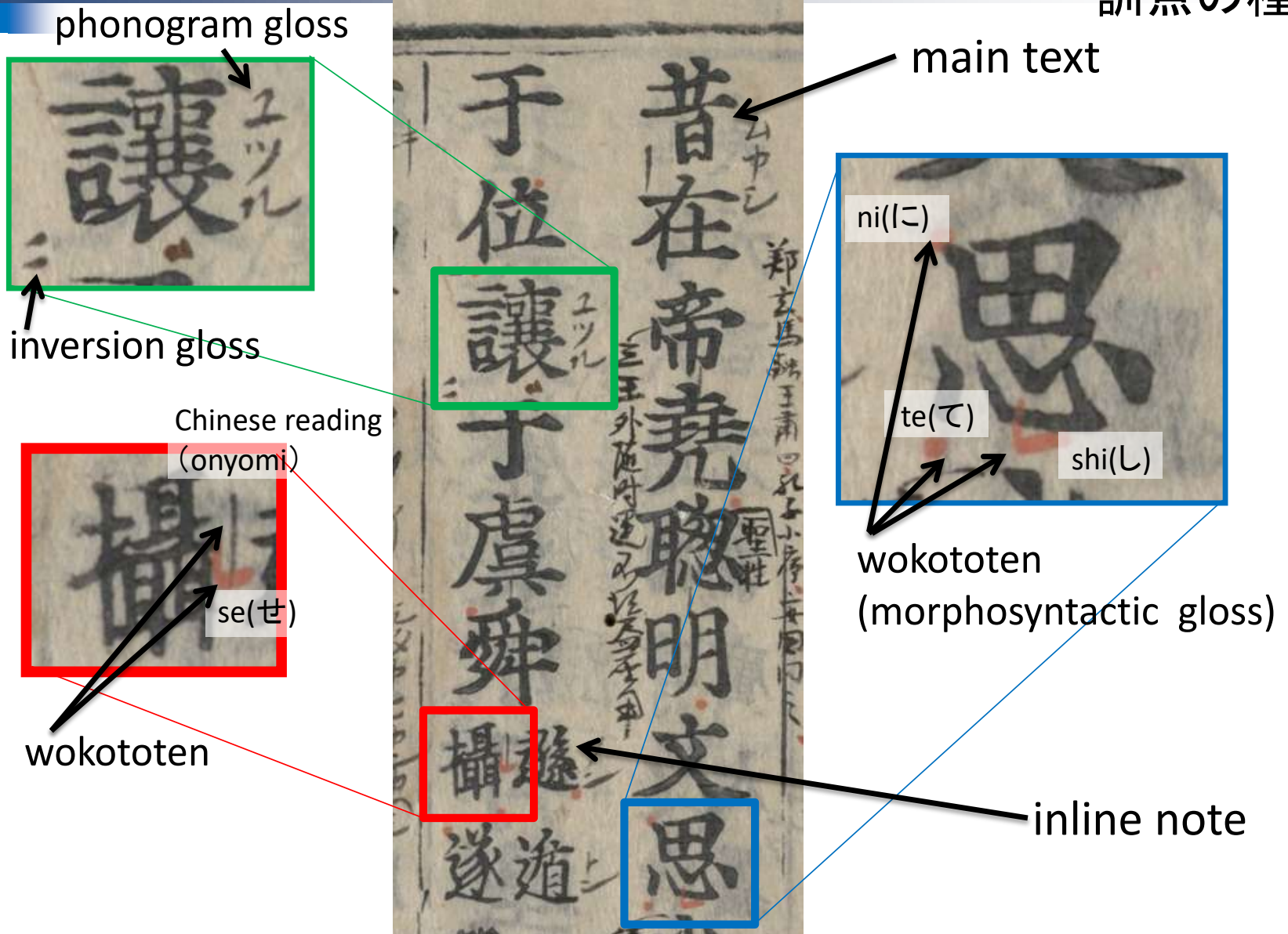


Purpose of analyzing the kunteen material

- Look up vocabulary and build an index.
(当時使用していた)語彙の分析ができる
- Understand the historical variation,
and how to use it.
訓点のつけ方の歴史的な変異がわかる
- Analyze the Japanese language of the time
when the materials were created.
加点当時の日本語を知る手掛かりになる
 - The materials contain Japanese of the Heian period (about AC800-1100).
 - The last page of the material contains information about the person who wrote it. And it also shows the writer's denomination and what they studied.

Glosses in the kunteen material

訓点の種類



Step of Understanding

- Translate from English to Japanese

I rode the airplane.

I rode the airplane.

S V O → 「S」が「O」を「V」する

Change word order

Add the particle

「I」が「the airplane」を「rode」する

「私」が「飛行機」を「乗った」する

「私」が「飛行機」を「に」乗った「する」



克明俊徳以親九族

word order

克明にし俊徳を(て)
以て親す九族を

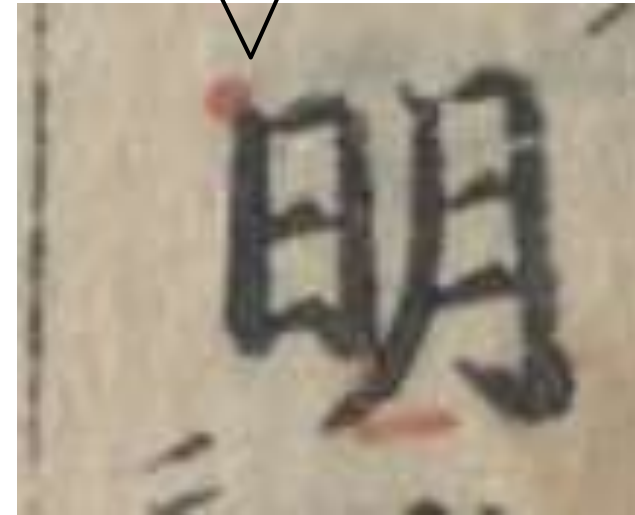
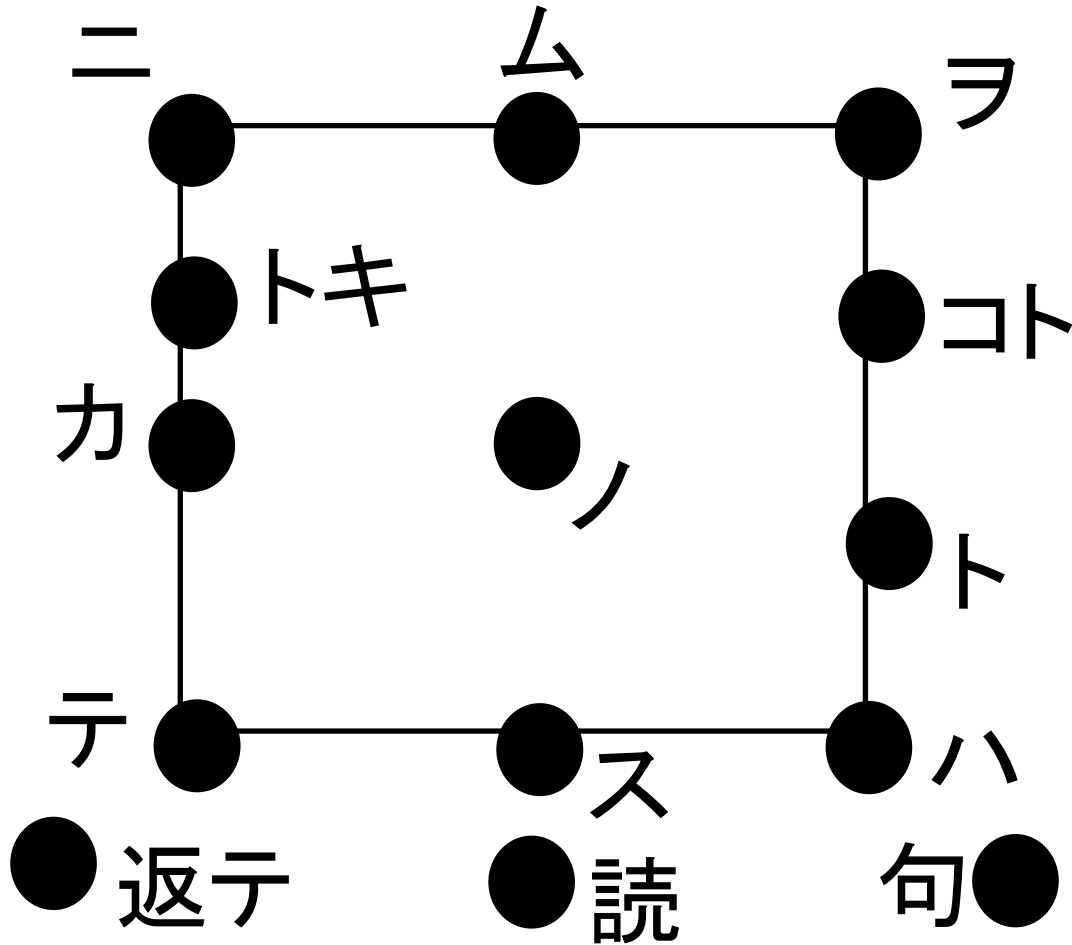
particle



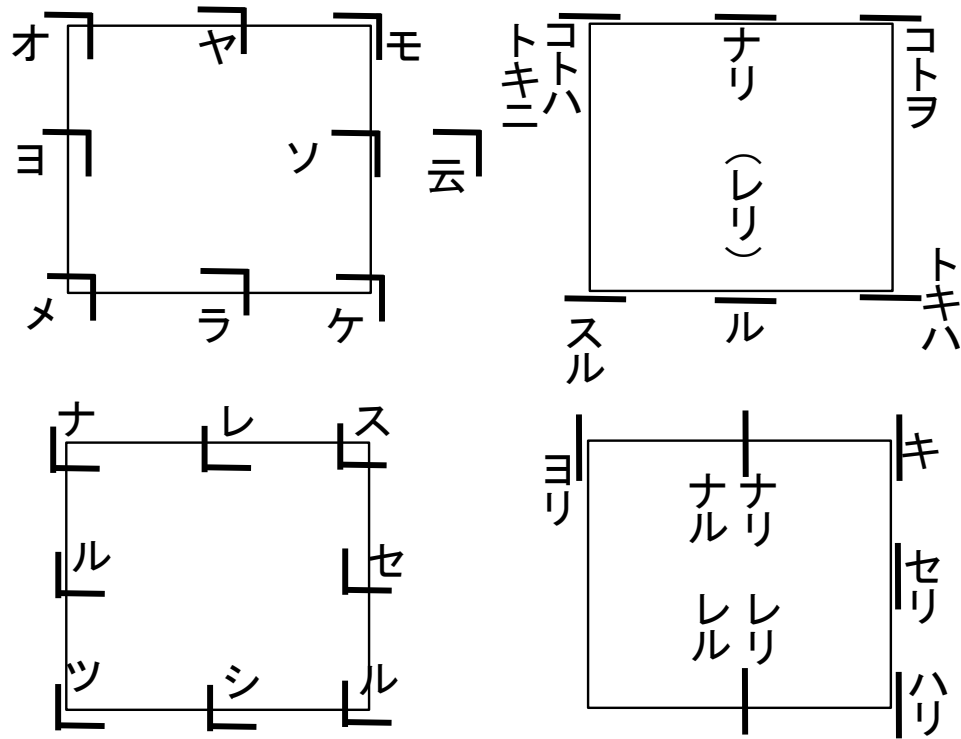
克俊徳を明にして
以て九族を親す

書き下し文

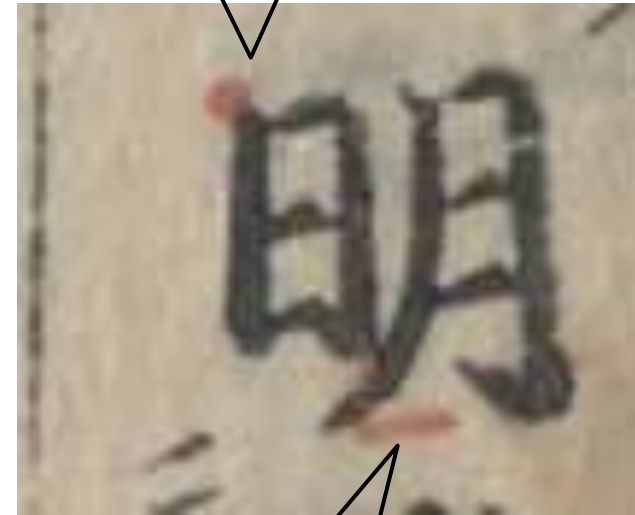
Wokototen mapping



Wokototen mapping



=



シ



克明俊徳以親九族

word order

克明にし俊徳を(て)
以て親す九族を

particle



克俊徳を明にして
以て九族を親す

書き下し文



Features of the Database

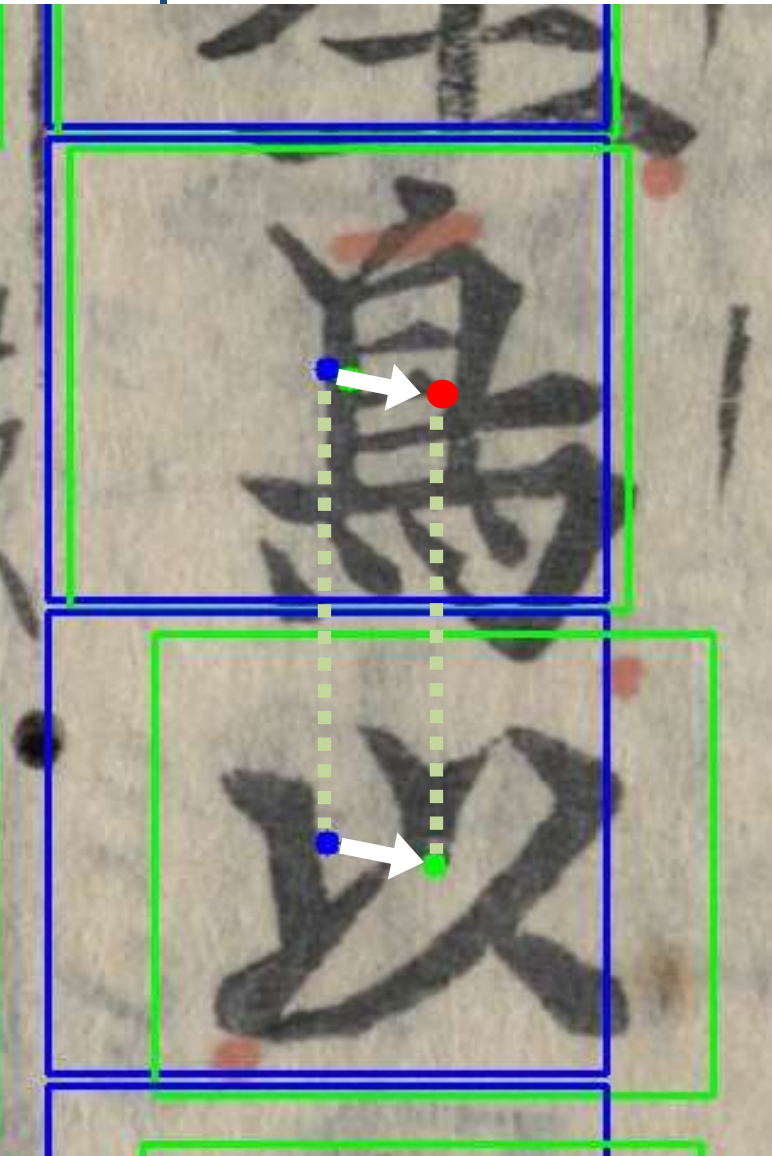
- Search the main text in
 - plain text (plain Japanese) or
 - detail kunten location and design.
- Search results shows images of the materials.
 - Created an algorithm to automatically cut out characters.
 - Use IIIF Curation Viewer to obtain images.



Character cutout

- Tried to detect character position by OCR.
 - Only 65.2% of character could be detected.
- We considered a different method.
- Our character positions detection method is the following process.
 1. Limit the detection range
 2. Combining character elements

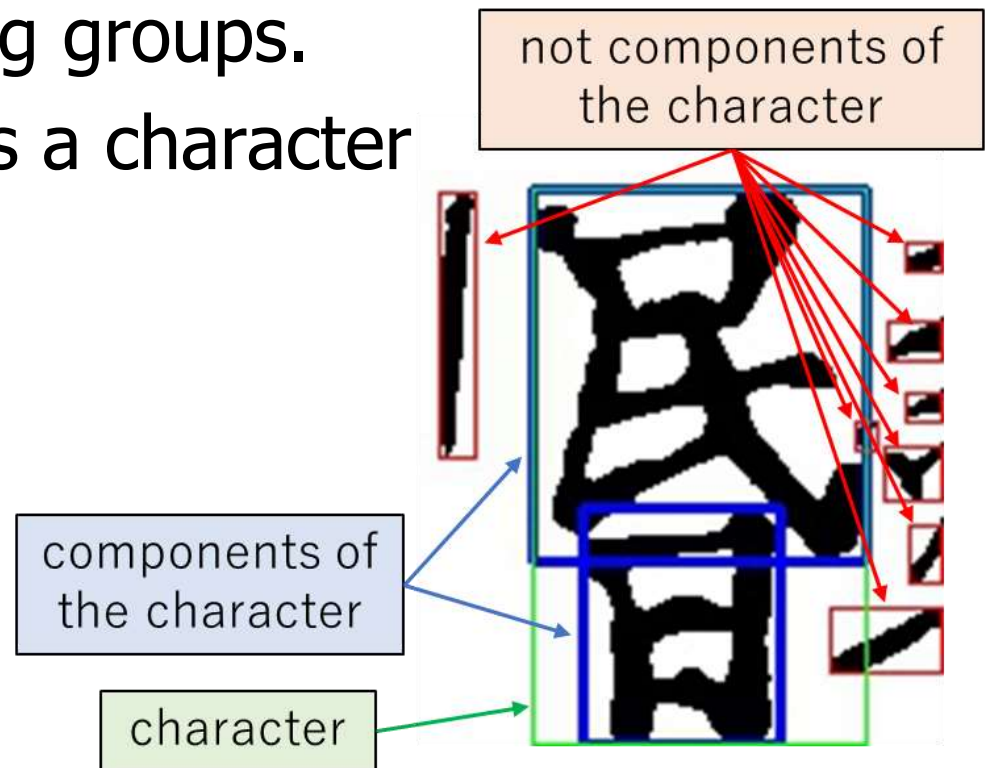
(1) Limit the detection range



- Set base character position and size using the printed data.
 - **Blue: Detection range and its center**
- Correct the character position. It lines up with the previous character position.
 - Red : Detection result and its center.**
 - Green : Detection range and its center after correction.**

(2) Combining character elements

- Extract black pixels in the area defined in (1)
- Group adjacent pixels
- Cut away groups of less than a certain size.
- Merge remaining groups.
- The results is as a character





Demonstrations

尚書（古活字版 第三種本）

訓点情報データベース

Kunten Database

このデータベースについて
About this database

使い方
How to use this database

ポリシー
Policy

お問い合わせ
Contact Us

国立国語研究所
NINJAL

Simple search

: This page is used to Search the database by search words (vocabulary or Kanji + Sentakana).

search words:

Detailed search page

: This page is used to search for elements that meet more than one row of search requirements by setting the requirements on the point's position, appearance (color) and shape, etc.

Frequency search page

Choose Coordinates :

All

(-3,-3)	(-2,-3)	(-1,-3)	(0,-3)	(1,-3)	(2,-3)	(3,-3)
(-3,-2)	(-2,-2)	(-1,-2)	(0,-2)	(1,-2)	(2,-2)	(3,-2)
(-3,-1)	(-2,-1)	(-1,-1)	(0,-1)	(1,-1)	(2,-1)	(3,-1)
(-3,0)	(-2,0)	(-1,0)	(0,0)	(1,0)	(2,0)	(3,0)
(-3,1)	(-2,1)	(-1,1)	(0,1)	(1,1)	(2,1)	(3,1)
(-3,2)	(-2,2)	(-1,2)	(0,2)	(1,2)	(2,2)	(3,2)
(-3,3)	(-2,3)	(-1,3)	(0,3)	(1,3)	(2,3)	(3,3)

Choose the style of the element : 朱 墨 All

Choose the mark of the element : · | L - o ∞ • 7 \ / Ω ☆ All

Position	Style	Mark

Add new row

Choose the style of the gojun element : 朱 墨 All

○レ ○ー ○二 ○三 ○上 ○下 ○中 ○四 ○五 ○(-レ) ○乙 ○甲 ○All

Gojun element style	Gojun element mark

Add new row

Input Kanji to search :

明

Choose the start book :

1

Choose the end book :

9

Input Kanji to search :

明

Gojun element style

Gojun element mark

Delete

Search

[BACK](#)

Statement

Book Start	Book End	Kanji
1	9	明





No	Coordinate	Style	Mark
----	------------	-------	------

No	Gojun Style	Gojun Mark
----	-------------	------------

257 results

No	Page	Place	Character	Link	Image
卷1	1才	03行06文字	明	Link	
卷1	1才	06行07文字	明	Link	
卷1	1才	07行03文字	明	Link	

Demonstrations

卷1	1ウ	01行04文字	明	Link	
卷1	1ウ	01行12文字	明	Link	
卷1	1ウ	02行25文字	明	Link	
卷1	1ウ	03行04文字	明	Link	



Conclusion

- This study purposes to create the database of the “gloss” on the classical Chinese texts (kuntan materials).
- We digitize "Shangshu" (old type print version) for an example.
 - And shows the result of quantitative analysis of wokototen marks.
- A system was completed to automatically cut out and show the image of the materials.