Active Reading for Intercomprehension Between Sinogramic Languages
Pierric Mazodier, Yoann Goudin, Mathieu Mangeot, Valérie Bellynck

To cite this version:
Pierric Mazodier, Yoann Goudin, Mathieu Mangeot, Valérie Bellynck. Active Reading for Intercomprehension Between Sinogramic Languages. Natural Language Processing and Information Retrieval 2019, Jun 2019, Tokushima, Japan. hal-02165613v2

HAL Id: hal-02165613
https://hal.archives-ouvertes.fr/hal-02165613v2
Submitted on 4 Jul 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Active Reading for Intercomprehension Between Sinographic Languages

Pierrick Mazodier
Laboratoire LIG, équipe GETALP
Bâtiment IMAG CS 40700
38058 GRENOBLE CEDEX 9, FRANCE
pierrick@mazodier.fr

Mathieu Mangeot
Laboratoire LIG, équipe GETALP
Bâtiment IMAG CS 40700
38058 GRENOBLE CEDEX 9, FRANCE
mathieu.mangeot@imag.fr

Valérie Bellynck
Laboratoire LIG, équipe GETALP
Bâtiment IMAG CS 40700
38058 GRENOBLE CEDEX 9, FRANCE
valerie.bellynck@imag.fr

Yoann Goudin
LIDILEM – UGA / XMU
No. 422, Siming South Road, Xiamen,
Fujian, CHINA. 361005
yoann.goudin@yahoo.fr

ABSTRACT
In the domain of e-learning for a linguistic field, active reading interfaces are useful tools for learning new languages. For a learner of a given language, the goal is to assist in knowledge development in other known languages, in order to enhance their learning. An active reading program enriches the understanding of the input sentence in different ways, such as adding the pronunciation of each word or displaying a definition or a translation taken from a dictionary. By the process of analogy, the users would then be able to synthesize the information and produce by themselves a better interpretation or translation that would be, at their convenience, both syntactically and semantically appropriate.

We coin sinograms as an alternative proposal to ‘Chinese characters’ in order to distinguish the contemporary political entity from the past cultural dominant ideology over Eastern Asia and its most visible representative: the graphic system. Consequently, Sinographic languages (SL) [7] refer to the languages sharing a history including cultural loans through literary corpora and standards, the graphic system conveying them and so the lexicon still in use in these languages nowadays. Defined as such, SL include all the Sinitic languages among which standard Mandarin as the legitimate variant, but also typologically distinct languages such as Japanese, Korean and Vietnamese. In this paper, we will just deal with Japanese and Mandarin.

We propose two new features for an active reading interface that aims to help in the understanding of Sinographic languages. We focus on Japanese and Mandarin languages. The first feature identifies the reading (on’yomi, kun’yomi…) of the sinograms; the second feature displays, for certain words of the input text, their equivalent in another language that is etymologically close to the source one. In our case, the words with on’ yomi reading will be associated with their Mandarin equivalents, in their Pinyin form. This tool can be useful for learners of Japanese with some knowledge of Mandarin or the contrary.

CCS Concepts
• Computing methodologies →Artificial intelligence →Natural language processing →Language resources

Keywords
NLP; active reading; sinographic languages; intercomprehension; analogy; Japanese; Mandarin; Jibiki platform; Mecab; HanLP; onyomi; rōmaji; furigana; Pinyin

1. INTRODUCTION
In the domain of e-learning for a linguistic field, active reading interfaces are useful tools for learning new languages. For a learner of a given language, the goal is to assist in knowledge development in other known languages, in order to enhance their learning. While a straightforward translation can be useful to grasp a meaning or to get a syntactically correct output sentence, it can still be semantically incorrect. Furthermore, it does not help the learner to increase her own knowledge of the language.

Instead, an active reading program enriches the understanding of the input sentence in different ways, such as adding the pronunciation of each word or displaying a definition or a translation taken from a dictionary. By the process of analogy, the users would then be able to synthesize the information and produce by themselves a better interpretation or translation that would be, at their convenience, both syntactically and semantically appropriate.

We coin sinograms as an alternative proposal to ‘Chinese characters’ in order to distinguish the contemporary political entity from the past cultural dominant ideology over Eastern Asia and its most visible representative: the graphic system. Consequently, Sinographic languages (SL) [7] refer to the languages sharing a history including cultural loans through literary corpora and standards, the graphic system conveying them and so the lexicon still in use in these languages nowadays. Defined as such, SL include all the Sinitic languages among which standard Mandarin as the legitimate variant, but also typologically distinct languages such as Japanese, Korean and Vietnamese. In this paper, we will just deal with Japanese and Mandarin.

2. THE JAPANESE WRITING SYSTEM AND THE READINGS OF SINOGRAMS
2.1 Japanese language vs ‘Chinese’
As mentioned above, Japanese language [6] is typologically distinct from the contemporary Sinitic languages among which Modern Mandarin [9], but historically under cultural influence including Classical Chinese literature written in sinograms as early as the 5th century. By the 8th century, in order to write down reading information and Japanese-specific words, sentences and texts, several initiatives became soon two sets of syllabaries that
are derived from the sinogramic script. In the same time, the Japanese crafted their own readings of the sinograms and by doing so emancipated from Classical Chinese. We insist on the fact that through this historical process, 'Chinese' never refers to the legitimate language spoken nowadays known as Mandarin.

2.2 Contemporary graphic typology for Japanese
As a result today, the Japanese writing system is composed of four different types of characters:

1. **kanji** (漢字) are sinograms borrowed through this historical process and sometimes from initial forms. They are syllabic, commonly used in nouns, verbs and adjectives (e.g. 東京 respectively read as To and kyō, meaning the city of Tokyo);

2. **hiragana** (平仮名) is one of the both sets mentioned above. Hiragana are 'syllabic' – moraic is more accurate [6]– signs used for particles, pronouns, adverbs, inflections (also referred as okurigana (送り仮名), etc. (e.g. そして soshite, “then”);

3. **katakana** (片仮名) refers to the second set and an alternate version of hiragana. Nowadays, they are used for non-Chinese foreign loan words (e.g. クロッカッサン kurokawasan, “croissant”).

4. **rōmaji** (ローマ字) are the Latin characters also used nowadays in Japanese contemporary texts (e.g. ABC Mart).

Furthermore, when small kanji (usually hiragana) are seen written above kanji, they are called furigana (振り仮名) and are used to indicate the reading of the sinogram below. They appear usually in texts for children or above sinograms for indicating rare readings. At last, for non-Japanese speakers, there are two main transliteration standards of rōmaji: the Hepburn and Kunreishiki (訓令式), which differs from each other only by some subtleties (e.g. しぶ for transliterating しゅふ, tsu/tu for つ/つ ...).

2.3 Reading sinograms in Japanese
Beyond this graphic typology of Japanese texts, we also have to introduce shortly how the sinograms can be read in Japanese. As already mentioned above, kanji have two types of reading:

1. **on'yomi (音読 み)**, referred here after as 'Chinese pronunciation' or on reading inherited from Chinese loans over the history, (e.g. 東京 Tōkyō, 'Tokyo');

2. **kun'yomi (訓読 み)**, referred here after as 'Japanese pronunciation' or kun reading, usually used for Japanese indigenous lexicon (e.g. 東 higashi, 'East').

The point is that beyond their belonging to differently originated lexical stocks, they also do not share the same phonological patterns. Indeed, while the latter can be polysyllabic – three for 東 higashi – the former will always conform the sinosyllabic structure shared by all the SL (cf. below 2.4).

Each kanji may have at least one reading and frequently have one or more of each of these two types of readings. In order to be complete in this overview, we have to distinguish polysinographic – composed of at least two sinograms – words with mixed readings. The are two types conventionally referred as jubakoym i (重箱読み) on+kun reading words – jū-bako ‘box’ and yutōyo mi (湯桶読み) kun+on reading words, yu-tō ‘hot pot’. At last, there is a last type referred as ateji (当て字) where kanji are used to phonetically represent native or borrowed words with less regard to the underlying meaning of the characters such as in 寿司 ‘sushi’.

2.4 On readings and the sinosyllable
As this paper deals with features shared by SL and so Mandarin and Japanese on readings, we have to discriminate several categories of on'yomi corresponding to the different times during which the under going standard reading of a given word was borrowed from Chinese by Japanese speakers and introduced into their lexicon. Through this process, a given sinogram could be borrowed several times with diachronic variation of its reading. Three different on reading categories are conventionally discriminated as follows:

1. **go’on (興音)** were the earliest readings introduced to Japan during the Nara period (e.g. 浪人 rōnin, ‘ronin’);

2. **kan’on (漢音)** followed (e.g. 外国人 gaikokuujin, ‘foreigner’);

3. **tōso’on (唐宋音) came later during Heian and Edo period (e.g. 藩団 futon ‘futon’).

Even if these readings evolved, they all conform to the sinosyllable structure as long as it was the basis of literary composition into Classical Chinese and the institution of riming. The structure is organized as Figure 1.

![Figure 1: The sinosyllable and its structure](image)

Any sinosyllable – whatever Japanese on reading included – will conform this pattern distributed between the onset and the rhyme subdivided itself as glide (or 0), nucleus, vocalic or consonant coda (or 0) plus tone (or 0).

Among the examples above, such as ね read as nin or jin, n- and j- are the onset, -i is the nucleus, and -n is the consonant coda. Compared to the kun reading of the sinogram ひito, we can see that this reading does not conform the sinosyllable distribution.

2.5 Sinogram and sinosyllable as potential of intercomprehension
Beyond this striking diversity of various graphic systems in Japanese texts and diverse reading types for kanji, either it is challenging especially for foreign learners, there are short cuts even for this latter population as soon as learning and teaching are emancipated from language ideologies. Thus, English loans easily identified through their notation in katakana might provide a huge stock of lexicon semantically transparent and very helpful for phonological awareness. Furthermore, for foreign learners already familiar with sinograms – such as those coming from China, Taiwan, Hongkong etc., kanji are – with some exceptions – not only visually accessible for their meaning. The point now is that as long as Mandarin is becoming the first SL to be learnt abroad beyond the few aeras previsously mentioned, katakana and semantic informations provided by kanji are not the only potential of intercomprehension: sinograms are also a resource through their on reading and so a feature on which learners can rely in order to transfer their reading knowledge in Mandarin into Japanese.

This is our didastic statement [4] and the theory for which our tool is designed. As far as we know, such a tool does not exist and so is our contribution.

---

1 also known as Japanese ruby characters.
3. **THE ACTIVE READING PROCESS**

### 3.1 Description of the generic process

The active reading tool [1] proceeds as follows:

1. The text received as input is sent to a morphological analyzer to obtain the lemma and grammatical information for each word.
2. Using the lemmas received from the analyzer, the tool then looks up a monolingual or bilingual dictionary to obtain lexical information about each lemma. It can be a definition or a translation.
3. The tool displays the input text by adding additional information either above or below the text, or by displaying words with different colors, or by adding a code that will display a pop-up window when the mouse hovers over the word.

### 3.2 Active reading for Japanese

#### 3.2.1 Description of the active reading

The active reading service offers to display the reading of a Japanese text in the form of *furigana* and the French translations coming from a Japanese-French dictionary. The list of the corresponding translations from the dictionary appears in a textbox when hovering the mouse over certain words. The service is available online from the jibiki.fr website.

This active reading instance is tailored for the following scenario: a learner of Japanese with a good knowledge of French. It could be enlarged to learners fluent in other languages by using different dictionaries. For example, the use of Jim Breen’s JMdict [2] could be useful for users fluent in English, Russian, German, Dutch, etc.

#### 3.2.2 The morphological analysis

The parsing and analysis of the input text is performed with the Conditional Random Field based MeCab morphological analyzer and tokenizer.

Figure 2 shows the result of the analysis by MeCab of the sentence entered by the user in Figure 3. Each line represents a token. The first left column is the token itself. Then follows a series of 8 items separated by commas. The first one is the part-of-speech + 名詞 is a noun, 助詞 is a verb, 助詞 is a particle and 記号 is a punctuation mark. The next 4 ones are used for subcategorization. Then, the next item is the lemma (eg: the 単位, 掛かりかり, the lemma is the infinitive form of the verb 掛かる). The next item is the writing form of the token in *katakana* and the last item is the reading form in *katakana* (e.g. the particle は has the writing form は ha and the reading form わ wo). For our purpose, we use the lemma and the reading form.

One can see in the example of Figure 2 that the numbers are actually parsed each figure separately by MeCab. In order to render correctly the *furigana* of the numbers, we encoded a specific function. That is especially important in Japanese because of the 数量词 (man) '10 000' counting unit which differs from the Western 10000 counting unit.

#### 3.2.3 The Japanese-French dictionary


- a digitized version of the *Cesselin* [3] Japanese-French dictionary edited by Gustave Cesselin and published in 1940. It contains more than 82,000 articles with detailed microstructure: French translations, examples, locutions, etc;
- completed with 47,630 Japanese-French or Japanese-English entries from the JMdict [2] for modern vocabulary and
- 23,486 Japanese-French or Japanese-English Wikipedia links for recent terminology.

The articles from JMdict and Wikipedia were imported if they appeared also in the *Super Daijirin dictionary*, in order to avoid a large number of useless entries. The resulting resource thus built contains more than 153,000 entries and 140,000 examples. It can be considered as the most reliable source for Japanese-French translations.

The result of the OCR process is never perfect. Thus, some articles coming from the *Cesselin* need corrections. The dictionary can be edited online by any user who detects an error. So far, more than 81,000 contributions were established in the 4 years of the project life.

Mecab analyzer is distributed with a generic dictionary called *ipadic* with 392,126 entries in total. Despite its consequent size, several words in our dictionary are not present in the ipadic dictionary. Thus, we decided to generate a Mecab dictionary from the content of Jibiki.fr dictionary in order to enlarge the coverage of Mecab. Our dictionary contains 22,321 entries that are not present in ipadic. The first version is still beta. On going work is on its way to produce a correct dictionary particularly concerning the subcategories of the verbs. Once this work will be completed, we will publicly release this dictionary on the Jibiki.fr project website.

![Figure 2. Parsing of a Japanese text with MeCab](image)

### 4. ENRICHING THE ACTIVE READING FOR INTERCOMPREHENSION

#### 4.1 Purpose of the new features

We will now present two new features aiming at making the active reading output display even more informative. As explained previously, Japanese language uses different writing systems and readings depending on which lexicon stock a given word belongs to: *on'yomi* for kanji whose pronunciation was borrowed from Chinese during different periods, *kun'yomi* for Japanese spellings, *katakana* for foreign loan words. The objective is to highlight the words of the text with different colors according to their stock, in order to tell them apart and get knowledge of the links between them and their various influences.

Then, the *Pinyin* is displayed below the words whose *kanji* are
on’yomi\(^2\) (see Figure 4). This information gives some clues for intercomprehension between sinogramic languages [4]. These new features are aimed to learners of Japanese with some knowledge of Mandarin or the contrary.

Figure 3 shows the new active reading interface with these new features. The top half of the window is the same as Figures 2. The result is then still displayed on the bottom half of the window with the furigana on top of the Japanese sentence and a French translation when the mouse is over a word (here the word “ōsoi”, translated in French by “tardif”) but two new features can be seen:

- The words in violet background are in kun’yomi (Yokohama, kakari, osoi);
- The words in red background are in on’yomi and below them, the Pīnyīn of the Mandarin equivalent word is displayed in red (Tōkyō, densha, fun, hou);
- The words in green background are in katakana here an English loan word (baika).

As already mentioned above, among the Chinese imperial literary standards, there was the fundamental practice of Classical riming system. If it is abolished nowadays, this institution is still the potential of intercomprehension between SL through the conformation to the sinosyllable structure. Thus, on the Figure 3, one can already see that a constant pattern emerges between Pīnyīn and rōmaji for Japanese words in on’yomi. For example, the Mandarin “dōng” is transcribed with the Japanese “tō”; the Mandarin “diàn” is transcribed with the Japanese “den”, etc.

4.3 Details of the implementation

We modified the active reading PHP program by incorporating a piece of code composed of four main parts: information extraction, decision making, Pīnyīn addition and display.

4.3.1 Program overview

The active reading's PHP code consists in the following steps, the green ones being our new features' implementation:

a. Catching user's input
b. Parsing it through MeCab morphological analyzer
c. Gathering the following elements, for each word: furigana, lemma (e.g. base form of verbs…), part-of-speech (e.g. verb, noun…), subcategory (e.g. transitive verb)
d. Converting furigana's katakana into hiragana
e. Querying the Jibiki.fr Japanese-French dictionary for retrieving translations with the REST API
f. Querying KANJIDIC with the REST API and extracting information
g. Processing it through decision function
h. Associating the result to a highlight color
i. For on’yomi word, get Pīnyīn through HanLP\(^7\) Mandarin morphological analyzer

On HTML side:

j. Displaying each word with: corresponding highlight color, furigana on top, translation in a pop-up window and, for on’yomi words, Pīnyīn below.

4.3.2 Information extraction (f)

The on/kun’yomi nature of the kanji has to be retrieved from KANJIDIC. The data is organized in XML tree structures that can be accessed with a Curl request to the Jibiki REST API. The entries to be checked are those of each kanji for each word in the text that possesses furigana. If a word also contains hiragana (as okurigana, e.g. 買い物 kaimono, ’shopping’), the XML structure of this said sinogram will be empty; we nevertheless keep the hiragana as it is. For katakana words, we will use a regular expression with a pre-established list of katakana. Instead of immediately processing the data, we store it in an array to make it available for further use and to ensure code clarity. Considering the aforementioned hypothesis of one yomi per word, the on’ and kun’yomi will be stored separately, to be later processed independently.

Before going to the decision part, the pronunciation data has to be formatted:

- Some pronunciations are preceded or followed by a dash (\(\sim\)), indicating a restriction on their position in the word (e.g. ね – nan- (what) cannot be used at the end of a word). This information will be used to skip some cases in the decision part. We also remove the dash.

\(^2\) https://jibiki.fr/reading/pieric_f.php
\(^6\) http://www.edrdg.org/kanjidic/kanjidic.html
\(^7\) https://github.com/hankcs/HanLP
Some also include a dot (.), usually separating the radical of a verb or adjective from its inflection (e.g. 会: あう a.u, meet). We then remove the inflection and the dot.

4.3.3 Decision making (g)

Once all the data has been collected and formatted, for each word, we will compare the different possible on’yomi combinations of their kanji (and potential okurigana) with the known furigana of the word. If there is a match, we know that the word is on’yomi and we label it as such. If not, we repeat the process for the kun’yomi. If ultimately no match is found, the word will be labeled as special pronunciation.

Once the word’s reading has been figured out, the highlight color value is drawn from a correspondence array: red for on’yomi kanji, purple for kun’yomi kanji, gold for special pronunciation and green for katakana.

4.3.4 Pīnyīn addition (i)

To retrieve the Pīnyīn of the on’yomi words, we used HanLP parsing toolkit for Mandarin to retrieve the individual information for each kanji (Figure 4). HanLP is a powerful java library for processing Mandarin texts. We developed a simple java program that displays the Pīnyīn relative information for each sinogram received in input. The first line displays each sinogram separated by a comma; the next line displays the Pīnyīn with tones in numbers, the next one the Pīnyīn with tones with diacritics; the next one the Pīnyīn without tone and the other lines give more detailed information about the reading. For our purpose, we use the third line that gives us the Pīnyīn with tones represented by diacritics.

Concerning the implementation of the tool itself, as the code currently constitutes a draft of a more formalized future clean version, some improvements can be imagined for code optimization, performance improvements and trustful case handleings. This latter point can be challenging, since it requires exhaustive data sets of special cases (e.g. counters, Japanese names), most of them being virtually infinite. In the case of our feature, the extent of handled cases depends on the coverage of KANJIDIC. A more thorough (but costly) process would be to cross-check different sources. Other technical optimizations could be to display the Pīnyīn below the Japanese even on browsers that do not render the <rtc> tag correctly.

The described active reading scenario of Japanese texts could be extended at least in two different ways:

• looking up etymological information for katakana words from rich dictionaries (the Cesselin contains such information) and displaying precise information about the origin of the words. For example, the word “baiku” comes from the English word “bike”.

• Computing automatically analogies between rōmaji and Pīnyīn.

We are working also on different active reading scenarios for other languages:

• It would be interesting to include other sinogramic languages such as Korean, Vietnamese and Sinitic languages among which Taiwanese very soon.

• A study has also been driven for Berber and Arabic speaking learners of French and English [1].

• Other language families could be addressed like Romance languages, Slavic languages, Germanic languages, pidgins and creoles, etc.

6. REFERENCES


Columns on Last Page Should Be Made As Close As Possible to Equal Length

## Authors’ background

<table>
<thead>
<tr>
<th>Your Name</th>
<th>Title*</th>
<th>Research Field</th>
<th>Personal website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pierric Mazodier</td>
<td>Master Student</td>
<td>NLP</td>
<td></td>
</tr>
<tr>
<td>Mathieu Mangeot</td>
<td>Associate Professor</td>
<td>NLP</td>
<td>cv.archives-ouvertes.fr/mathieu-mangeot</td>
</tr>
<tr>
<td>Valérie Bellynck</td>
<td>Associate Professor</td>
<td>NLP</td>
<td>cv.archives-ouvertes.fr/valerie-bellynck</td>
</tr>
<tr>
<td>Yoann Goudin</td>
<td>Assistant Professor</td>
<td>Sinogramic languages</td>
<td></td>
</tr>
</tbody>
</table>

*This form helps us to understand your paper better, the form itself will not be published.

*Title can be chosen from: master student, Phd candidate, assistant professor, lecture, senior lecture, associate professor, full professor