AUTOMATIC ADDITION OF GENRE INFORMATION IN A JAPANESE DICTIONARY

Raoul BLIN
EHESS - School for Advanced Studies in the Social Sciences
blin@ehess.fr

Abstract

This article presents the method used for the automatic addition of genre information to the Japanese entries in a Japanese-French dictionary. The dictionary is intended for a wide audience, ranging from learners of Japanese as a second language to researchers. The genre characterization is based on the statistical analysis of corpora representing different genres. We will discuss the selection of genres and corpora, the tool and method of analysis, the difficulties encountered during this analysis and their solutions.

Keywords

lexicography; corpus; genre; Japanese

Izvleček

Članek predstavlja metodo za samodejno dodajanje informacij o žanru k japonskim iztočnicam v japonsko-francoskem slovarju, ki je namenjen tako učencem japonščine kot tujega jezika kot tudi raziskovalcem. Žanrski opis je osnovan na statistični analizi korpusov različnih žanrov. Članek opisuje izbiro žanrov in korpusov, orodja in metode za analizo, težave pri analizi in rešitve zanje.

Ključne besede

slovaropisje, korpus, žanr, japonščina

1. Introduction

In the Japanese language, general lexicons and dictionaries (whether monolingual or multilingual) published to date provide very little information about the genres of the various entries. Furthermore, dictionaries that address genre generally focus on only one theme (for example the law) and cannot be used to compare different genres.

However, information regarding the genre of words is extremely important in language production since it enables the writer to choose vocabulary appropriate to the genre of the discourse or text in question. In terms of translation, such information is useful to choose the appropriate word in the target language. In addition, it will also enable the reader to define the style of a text.

Defining and manually annotating genre information for all lexical entries in a dictionary is very costly, both in terms of man-hours and financially. It is also fraught with many methodological difficulties.

The first difficulty is to provide clear criteria that can be used by lexicographers to classify the words. In a strictly scientific approach, if the criteria were qualitative, they would require evaluation and testing. A few lexicographers may use them to classify a sample of entries, after which the rate of agreement between lexicographers would be calculated. The criteria must be revised and retested until the rate of agreement is sufficiently high. This is a long and laborious process.

The second difficulty is to assign the lemma to one genre. Except for highly specialised vocabulary, words may appear in a variety of genres. Therefore, merely classifying a word as belonging to one genre to the exclusion of other genres would not be a suitable classification method. Rather, the word's classification should be graduated in terms of each genre under consideration. Once again, the evaluation procedure mentioned above becomes necessary.

The third difficulty concerns human knowledge limitations. The classification method presupposes that the lexicographer has a good knowledge of all the genres under consideration. Such lexicographers are certainly very rare. Teams of two or more researchers from different specialties are necessary. Therefore, the evaluation process mentioned above does not cover the rate of agreement between lexicographers working alone, but between teams of lexicographers. It is an additional difficulty.

A fourth challenge is to monitor variations in genre over time. This would require all the entries to be checked on a regular basis (every three or four years) and to create a team of lexicographers each time. The research and evaluation process described above would then be repeated.

As is understandable, it is not realistic to manually describe the genre of a dictionary's entries using qualitative criteria in a scientific approach. The task obviously requires an automated procedure. To this end, our idea is to base the study on a statistical analysis of corpora. By using selected corpora to represent different genres, the genre(s) of a word will be correlated to the word's frequencies in those corpora. In addition to being easy to implement, such an automated procedure based on quantitative criteria is easy to replicate and suitable for monitoring variations over time.

This process has been applied to automatically classify the genre of 16,000 entries in a Japanese-French dictionary (Blin, 2012a; abr. "DFJC"), on the basis of frequencies obtrained from a purposely-built corpus consisting of several subcorpora. In the following chapters, we will describe the process in more detail

In section 2, we will provide a brief overview of the DFJC to show what kind of data was included. In section 3, we will discuss the genres that were studied and the corresponding corpora. In section 4, we will describe the software used to count the number of occurrences, the problems encountered and their solutions (if any).

2. The dictionary

2.1 The dictionary entries

The dictionary contains more than 16,000 Japanese common nouns and qualifying nouns (e.g.: *kinkyu*, 緊急, “urgent”, *ihou*, 違法, “illegal”) in Japanese, each described as the example entry in Table 1.

Table 1: Example of an entry in the Japanese-French dictionary DFJC

		White papers	**Newspapers**	**Legal texts**	**QA govn.**	**misc. QA**	**Chats**	**Total occ.**
36	*asita* 明日 (h:41); demain	11	57		5	41	215	150
<1>	<2> < 3> (<4>); <5>	<6>	<7>	<8>	<9>	<10>	<11>	<12>

Each entry includes <1> the entry number, the usual <2> reading and <3> writing of the word, <4> the list of its homographs (see section 4.2), <5> the French translation, <6-11> its frequencies in six sub-corpora/genres (see section 3.2) and <12> the total number of occurrences.

All the data used to calculate the frequencies are provided in a database distributed freely (JaLexBD[1]).

Furthermore, the dictionary provides a summary description of the sub-corpora/genres: size, list of the most frequent words, comparison of the frequencies of some morphological structures, etc.

2.2 The distributions

For each word, the dictionary provides the frequencies of the word alone and with affixes, in each of the subcorpora. Due to lack of space, the frequencies for each distribution of each word are not detailed in the paper version, but all the numbers of occurrences are provided in the JaLexBD database.

[1] http://rkappa.fr/lexic/JaLexBD/index.JaLexBD.php

We chose constructions with affixes as the most linguistically interesting ones, especially those which are less morphologically and syntactically ambiguous (see section 4.2). For example, we avoided counting strings of concatenated nouns (without particles), since this can be extremely difficult for the automatic analysis of a text. The constructions were counted as follows:

2.2.1 Word alone

The word appears without any affix.

2.2.2 Words with the adjectivising suffix *-teki* (的)

Any noun with the *-teki* (的) suffix is transformed into a *-na* adjective. N+*teki* means roughly “which has the property of N”. For example, when *-teki* (的) is attached to the noun *gainen* (概念, “concept”), it forms the word *gainenteki* (概念的), which means “conceptual”.

2.2.3 Words with the nominalising suffix *-sei* (性)

The suffix *-sei* (性) can be placed after a noun or an adjective. N/adj+*sei* is a (grammatical) noun and means roughly “the property of being N/adj”. For example, *ensyoo* (炎症) means “inflammation”, while *ensyoo* + *sei* (炎症性) means “of an inflammatory nature, type or origin”. Some of the constructions suffixed by *sei* may be lexicalized, such as 経済性 (*keizai* + *sei*, “economic efficiency, economic performance, economy”).

2.2.4 Words with the plural suffixes *-ra* (等／ら) or *-tati* (達／たち)

The suffixes *-ra* (等／ら) and *-tati* (達／たち) express the plural. For example, the common noun *gakusei* (学生 “student”) is indefinite. Depending on the context, it can be interpreted as being either singular or plural. The construction *gakusei+ra* (学生等／学生ら), however, can only be interpreted as being plural.

In contemporary Japanese, the two suffixes are both used for human nouns, like “*gakusei*” (学生). They can also be used with a humanised entity, such as *neko+tati* (猫達 “the cats”). However, this construction is generally limited to children’s language.

The difference between the two suffixes may lie in the register of the language. For example, *-ra* is known to be more formal than *-tati*. Thus, comparing the frequency of plural suffixes may provide an interesting indication of text genres.

The process of counting plural suffixes is based on the hiragana transcription, i.e. ら (*ra*) and たち (*tati*). This restriction is justified by the fact that the Chinese character (等), which can be used to write the suffix *-ra*, is ambiguous, since 等 can also be used to write *nado* (“etc.”). In order to prevent possible errors, we did not count the occurrences of 等, but only the transcriptions in hiragana. In order to unify the counting procedure, we also restricted the counting of *-tati* to the hiragana transcription. This restriction certainly comes at the expense of *-tati,* since *-tati* is written more frequently in Chinese characters. Thus, the results in the DFJC certainly minimise the number of occurrences of *-tati*, although we cannot say to what extent it is minimised.

3. Corpora and genres

The corpus is divided into seven sub-corpora. Each sub-corpus has specific characteristic(s) that we will refer to as “genre”. These characteristics mainly stem from their source. For example, “journalistic genre” (i.e. “journalistic corpus”) will refer to the collection of texts retrieved from newspaper websites. The details of the other sub-corpora are outlined below.

We consider that such a “genre” definition is explicit enough and does not require the laborious evaluation process described in the introduction.

The description may be supplemented with other characteristics, but they will not have been used to build the corpora and define the genres.

We can distinguish between two types of text: monologues and dialogues. A “dialogue” refers to a text which provides an answer to a question, or which is constructed to be answered by someone other than the author. A “monologue” refers to a text which is not constructed to be followed by an answer, and which does not provide an answer itself.

We also distinguish between reviewed and non-reviewed texts. The assumption is that the variety of morpho-syntactical structures and vocabulary is wider in non-reviewed texts. When reviewing is part of the production process, it is expected that the author and reviewer will agree to some (at least implicit) conventions about acceptable (or authorized) language. The author must produce a text corresponding to this agreement. If not, the reviewer will correct the text according to these constraints. In principle, there are no such limitations in non-reviewed texts.

If the set of authors is limited, the variety of structures and vocabulary is limited to the skills of those authors. Indeed, the variety is expected to be poorer in comparison with corpora produced by an unlimited set of authors. This is why we distinguished between textual corpora produced by a limited and an unlimited set of authors.

We should also take the writing time into account. We assume that the variety of vocabulary and morpho-syntactic structure is greater when writers have unlimited time to write.

Table 2 below presents a summary of those characteristics for all the corpora. In addition, it shows the size of the sub-corpora and their frequencies.

Table 2: Characteristics of the corpora used

(Conventions: “+” yes for all the texts of the corpus; “–” no for all the texts of the corpus; “+/–” depending on the text of the corpus)

		No. of sentences	**frequency of updates**	**reviewing**	**only one theme for the corpus**	**limited set of authors**	**limited set of readers**
monologue	White papers	105 520	one year	+	–	+	+
	Daijirin Dictionary	176 809	partial, long-term	+	–	+	–
	Newspapers	167 819	1 day	+	–	+	–
	Legal texts	34 688	1 partial year	+	+	+	+
dialogue	Q&A govn.	54 901	1 full year	+	–	+	+
	Q&A misc.	136 946	1 partial day	+/–	–	–	–
	Chats	23 547	1 full day	+/–	–	–	–

3.1 Selection criteria for the corpora

We applied the criteria below to select the sub-corpora/genres.

3.1.1 Representativeness of a subcorpus with respect to its genre

In order to obtain a better representativeness, we built the sub-corpora as follows. Our strategy differs from the well-known corpus BCCWJ (Maruyama, 2009) in many respects.

Whenever possible, we used the complete collection of texts from a source rather than using a sample. For example, we collected all the White Papers of 2009, 2010 and 2011, whereas the BCCWJ contains only samples of collections.

For the same reason and also unlike the BCCWJ, all the texts of the sub-corpora are complete. We did not use any samples.

The corpus is strictly limited to written language. Transcriptions of spoken language, such as the Minutes of the Diet (included in the BCCWJ) are excluded.

3.1.2 Development of the corpus

The DFJC, published in 2012, is the first step of the project to observe the development of genres over time. As such, it was necessary to choose genres/corpora that would change over time.

To date, statistical studies in Japan about the Japanese language have been designed for static corpora. Those corpora were created once and for all and no updates were planned. Ever since the first statistical study on a large corpus of written Japanese was conducted by the National Institute for Japanese Language and Literature in the 50’s (Yamazaki, 2006), no corpus has been updated or compared to previous versions to observe its development over time.

To break away from this method, all the corpora used for the DFJC are textual collections that will be updated within a few years (or within a few months in some cases). The corpora of newspapers, chats, questions to the government, miscellaneous Q&A and White Papers will be entirely renewed in one year. Some of the sub-corpora, such as commercial dictionaries and fundamental legal texts (Constitution, etc.) should not change for a long time, but as long as they are distributed without changing (and excepting cases where they are distributed explicitly as historical texts), they should be understandable. Thus, even if such texts have not been renewed for a while, the language used represents the current language during the period of their distribution.

The sub-corpora for the DFJC consist of texts produced mostly between 2008 and 2011, which represents a span of 4 years. We had first planned to re-compile the texts annually, but the size of the corpora was not sufficient to obtain a good representativeness. Consequently, we estimate that a span of 3 to 5 years would be a good compromise.

3.1.3 Accessibility

In order to build the corpora, we also had to make do with limited financial and human resources. Scanning or manually retyping texts, as has been done for the BCCWJ, was out of the question. The solution was therefore to use the Internet. To this end, the corpora/genres selected were taken from collections of texts accessible on the Internet. However, we did not have enough technical (and financial) resources to build a corpus as large as the one described by Kawahara and Kurohashi (2006) which contains 470 million sentences. Even if the texts are easy to access on the Web, we had to limit the selection to a relatively small set of genres/corpora (710,000 sentences)

3.2 Detailed presentation of the sub-corpora

3.2.1 White papers

This is a collection of White Papers published in 2009, 2010 and 2011. Due to their thematic variety, this corpus cannot be considered as representative of one discipline. However, we assume that the production conditions are homogenous: the texts are written by a restricted set of authors (specialists). We also assume that these texts are reviewed.

3.2.2 *Daijirin* dictionary

To our knowledge, the language genre of dictionary texts has not yet been studied, despite the fact that it is certainly original and subject to many editorial conventions.

This corpus has not yet been completed. In the online *Daijirin* dictionary (Matsumura, 2006; http://dic.yahoo.co.jp), we only used the pages corresponding to the lemmas of the DFJC. As such, this corpus only contains 16,000 entries even though *Daijirin* contains 230,000 entries in total. Furthermore, *Daijirin* cannot be considered as being representative of all Japanese dictionaries. Therefore, this corpus is a poor representation and has been used as a test corpus only. In the future, the totality of the *Daijirin* and other online dictionaries will be used.

3.2.3 Newspapers

This corpus contains the online versions of three newspapers, covering the editions from April through December 2011. The newspapers are Asahi (www.asahi.com), Nippon Keizai (www.nikkei.com) and Nikkan Kougyou (www.nikkan.co.jp). The first two newspapers are widely distributed and have a significant place among newspapers. Furthermore, newspapers are very important in the everyday lives of Japanese people (almost all Japanese households are subscribed to a newspaper). We plan to incorporate more newspapers in the future.

3.2.4 Legal texts

The legal corpus is divided into two parts. The first part is the compilation of all official legal texts produced in 2008, 2009 and 2010 (law 法律, *hooritsu*; Cabinet Office Ordinance, 内閣府令, *naikakuhurei*; decree 命令, *meirei*). The second part is the compilation of six legal codes (Constitution, 憲法, *kenpoo*; civil law, 民法, *minpoo*; commercial law, 商法, *shoohoo*; criminal law 刑法, *keihoo*; civil procedure, 民事訴訟法, *minzi soshoohoo* and criminal procedure 刑事訴訟法, *keizi soshoohoo*).The second part will not be updated, whereas the first one is renewed every year.

3.2.5 Written questions submitted to the government

A compilation of all the written questions (from the Diet) submitted to the government between 2008 and 2010.

3.2.6 Miscellaneous questions and answers

A compilation of websites taken from oshiete.goo.ne.jp. This site is equivalent to the website Chiebukuro used for the BCCWJ. Each page contains an open question and possibly one or more answers. In some cases, there is no answer. Questions can address any subject matter.

3.2.7 Chats

This corpus is made up of pages from different chat websites. Due to financial concerns, this corpus is small. The procedure for collecting the pages included in this corpus will be changed in order to obtain more information.

It must be pointed out that the fundamental difference between this corpus and the corpus of miscellaneous questions and answers is the temporal constraint in terms of production. In miscellaneous question-answer dialogues, there is no (implicit) constraint on the time interval between the question and the answer. On the other hand, the response is usually immediate where chats are concerned.

4. Tools and analytical method

This section presents the software used, the problems encountered, their solutions and their impact on results.

4.1 Software

The number of occurrences of the words in the corpora was counted using the free software SAGACE version 4.2.0 [2] (Blin, 2012b). This tool is designed for searching patterns, but it does not analyse entire sentences. Patterns are defined as strings of words (or characters). A word can be a single word or any word of a pre-defined category. In the latter case, the category must be listed in the lexicon which is associated with SAGACE. SAGACE is only executed from the command line. To launch the query, the required pattern and the search parameters are described in a request form (a simple text file) interpreted by SAGACE.

We have chosen this software mainly on account of its ease of use. Unlike symbolical parsers, it is not necessary to develop a grammar , which requires time to carry out maintenance operations. It is sufficient to create a lexicon listing the words of the various categories. This is important since the maintenance and modification of a rule-based grammar can be a complex operation. Furthermore, there is currently no free and open grammar for Japanese. SAGACE also differs from statistical parsers (such as Mecab 3) on account of the fact that it does not require training and manual evaluation. In order to obtain good results with such tools, it is necessary to perform training and manual evaluation for each genre of text. Such a procedure is very costly.

[2] http://crlao.ehess.fr/japonais-coreen/corpus/sagace/sagace.html

Furthermore, SAGACE is autonomous and does not require any other software. It performs all the necessary functions for the analysis: a request interface, search engine and results interface. In addition, untagged plain text is sufficient. Thus, no pre-analysis is required.

As mentioned above, the advantage of SAGACE is that it is easy to use. The drawback, however, is that errors of analysis are (perhaps) more frequent than with other parsers. In order to limit the risk of errors, we selected less ambiguous patterns to be searched. As a result, not all the occurrences were counted. The frequencies indicated in the dictionary are thus slightly lower than the real frequencies. We are unable to assess the difference.

4.2 Difficulties of analysis and solutions

The automatic analysis of the Japanese language is subject to a few difficulties, which are well known in the field of Natural Language Processing. In this section, we will discuss how they have been solved (or not) using SAGACE, and what impact this solution had on the results.

4.2.1 Lemmatisation

Words are not separated graphically in written Japanese. Even if the parser analyses the entire sentence, morpho-syntactic errors may occur.

Only a semantic and pragmatic parser can prevent errors, but no such tools currently exist.

To limit the risk of errors with SAGACE, we first applied the traditional “longest match method”. Secondly, we restricted the number of searched patterns to the ones with a low risk of ambiguity (even if it is not zero). Overall, the searched pattern includes the contiguous words before and after the target structure. For example, when searching occurrences of a noun suffixed by 性 (*sei*), we used a pattern including a particle or punctuation mark on the left, and a particle, punctuation mark or copula on the right:

$$\left\{\begin{matrix}\text{particle}\\ \text{punctuation}\end{matrix}\right\}\ \text{NOUN 性}\ \left\{\begin{matrix}\text{particle}\\ \text{punctuation}\\ \text{copula}\end{matrix}\right\}$$

As an example, this is the description of the pattern in the request form:

>0 cat:particle | punctuation | XX	// 1
=0 cat:LEXEME /-affich:trait:lemme /-count	// 2
=0 性	// 3
=0 cat:particle | punctuation | copula	// 4

These lines are interpreted as follows:

(1) the first element of the pattern is anywhere (“>0”) in the sentence. It is either a particle, a punctuation mark, or a mark to indicate the beginning of sentence.

Formally, the description of the elements of the patterns are formulas written in a language close to propositional language. The interpretation is very similar to the interpretation of propositional logic. For example, the description of the first element can be formally interpreted as follows: the element is any word belonging to the category (“cat:”) defined as the union (“|”; disjunction) of three basic categories (propositional constant): category of particles (“particle”), category of punctuation (“punctuation”) and category “XX” (which is a singleton containing the mark indicating the beginning of a sentence). All the basic categories are listed in the lexicon associated with SAGACE. Using this description language, it is possible to “create” new categories by merely combining basic categories listed in the lexicon associated with SAGACE and without modifying this lexicon.

(2) the second element is contiguous (“=0”) with the precedent one. It is the word to be counted (“/-count”); it is any word of the category named LEXEME in the lexicon used by SAGACE.

(3) the third element is contiguous (“=0”) with the previous one. It is “ 性 ”.

(4) the third element is contiguous (“=0”) with the previous one. It is either a particle, a punctuation mark or the copula.

A more detailed description of the pattern syntax is available in the manual and online tutorials of SAGACE. All requests are provided in the DFJC.

4.2.2 Homography

Some words have the same graphic form but a different reading, and perhaps a slightly different meaning. For example, two homographic words transcribed as 魚 are almost synonymous, but have a different reading, *uo* and *sakana*. In the corpus, in order to know which reading is being referred to, a semantic (including pragmatics) analysis must be performed, but such an analysis tool is not currently available. A statistical analyser may solve the problem, but the results are not absolutely certain and the tool requires training.

For a great number of regular common nouns, there is a homographic proper noun. For example, *mori* (森, “forest”) and *hayasi* (林, “wood”) are also used as a last name. These last names are very common. Fortunately, in the corpora that have been used, they frequently appear with specific affixes, such as honorific suffixes (san, “miss, mister”) for human proper nouns. As such constructions don’t agree with the pattern we use, most of them have been excluded from the counting operation. Despite these precautions, it is possible that some occurrences may have been counted as proper nouns. Thus, the frequencies of the entries which can also be used as common nouns may be slightly higher than the real frequencies. We assume that this is a minor problem with no significant impact on the frequencies.

In the DFJC, we tagged all the entries which can also occur as a proper noun. To this end, we used a list of 320,000 proper nouns, extracted from the mecab-naist-dic and some other resources. The list contains most Japanese personal names, place names and company names. It does not include Chinese proper nouns. This list is not very long, but it should suffice to fulfil the purpose.

4.2.3 Multiple transcriptions

All Japanese words can be transcribed in various ways, by combining three character sets: hiragana, katakana and Chinese characters. Standard dictionaries provide the standard transcription. In fact, there is no “official” or “academic” standard. The so-called standard transcription could be defined as “the one which most closely resembles government prescriptions, among the most used transcriptions”. For example, the word *môsikomi* (“application”) is lexicalized as 申し込み or 申(し)込み, depending on the dictionary. The parentheses indicate that the characters can be omitted (but are still pronounced). Some dictionaries used for Natural Language Processing provide all the most common transcriptions and consider them as entries. For example, mecab-naist-jdic 4 provides five transcriptions/entries for the word *môsikomi* (“application”): もうしこみ, モウシコミ, 申しこみ, 申込み and 申し込み. Such lexicalisation has many flaws: it is very redundant and not exhaustive.

The multiplicity of transcriptions is not a problem per se, since the author’s choice of one transcription among many can help to characterise a written style. It rather represents an editorial problem when publishing a paper dictionary: listing all the transcriptions takes up a lot of space, even though many of them have such a low frequency that they are insignificant. The DFJC provides only the most common transcriptions. For some words, the frequency is the sum of the frequencies of two transcriptions. For example, when a noun contains the so-called honorific prefix *o*, we do not separate the transcription in kana from the transcription in Chinese characters. For example, the number of occurrences of the entry *otearai* (お手洗, “toilet”) is the addition of the number of occurrences of the two transcriptions お手洗 and 御手洗.

Variations of transcriptions are not only obtained by combining different systems. Some words have two or more transcriptions in Chinese characters. In most of these cases, two transcriptions exist: an “academic” transcription and a “popular” transcription. For example, the “academic” transcription of *tamago* (“egg”) is 卵. The popular transcription is 玉子. In the DFJC, the two (or more) transcriptions are clearly separated and constitute independent entries.

4.2.4 Homography *and* homophony

For some words, there are other words that are both homophonic and homographic. This is more common with monosyllabic (one kana) words. It can also occur with the kana transcription of a word. For example, “tooth” and “blade” are homophonic: *ha*. They are usually written in Chinese characters (resp. 歯 and 刃). However, when they are written in kana in a corpus, it is necessary to perform a semantic analysis to determine what word is being referred to. For the DJFC, we chose to count only dictionary lemmas, which are mainly in Chinese characters. We assume that existing entries in hiragana do not have such homophonic and homographic equivalents.

5. Conclusion and outlook

In this paper, we explained the process we have implemented to automatically characterise the genre(s) of 16,000 words in a Japanese-French dictionary. We plan to repeat this work regularly, about every three or four years, using the same process. There is room for improvement, and we wish to improve at least two points. Firstly, the number of entries will be increased. In particular, we will add verbal nouns. Secondly, as explained above, some corpora must be changed: the dictionary will be supplemented and we need a more reliable source for chats.

We also plan to conduct the same study on inflected words, such as verbs and adjectives. Despite the fact that SAGACE is not well designed for manipulating inflected words, a large-scale test (Blin, 2012c) showed, however, that the same method can be applied with the same tool. A more detailed (and manual) assessment of the results is required.

If the results are good enough, we will apply the process to locutions, including locutions with inflected words.

References

Blin, R. (2012a). Dictionnaire de fréquence du japonais contemporain - 16 000 noms (Youfeng.). Paris.

Blin, R. (2012b). *SAGACE v4.2.0*. CNRS. Retrieved from http://crlao.ehess.fr/japonais-coreen/corpus/sagace/manuel/Manuel.pdf

Blin, R. (2012c). Fréquences des verbes japonais dans un corpus de grande taille. Blin. Retrieved from http://rkappa.fr/sagace/tutoriel/sagace4-2/data/ListeDesFrequencesDesVerbesJaponais.pdf

Kawahara, D., & Kurohashi, S. (2006). *Case frame compilation from the web using high-performance computing*. Presented at the 5th International Conference on Language Resources and Evaluation.

Maruyama T. (2009). “Gendai nihongo kakikotoba keikin koopasu” monitaa kaihatu deeta (2009nendoban) sanpuringu houhou ni tuite [About the method of sampling in the “Balanced Corpus of Contemporary Written Japanese” (v.2009)]]. National Institute for Japanese Language and Linguistics

Matsumura, A. (2006). *Daijirin Second Edition*. Tokyo: Sanseido.

Yamazaki, M. (2006). Kokuritu kokugo kenkyuuzyo no goi tyousa no rekisi to kadai [Thematics and history of the lexical surveys of the National Institute of Japanese Language]. *12th Workshop “Thematics and history of the lexical surveys of the National Institute of Japanese Language”* (pp. 168–186). Tokyo University.