



NOOJ AND THE JAPANESE LANGUAGE

Claire OLIVIER

INALCO - 2009

M2 Japonais

M1 Linguistique et diversité des langues

PARTICULARITIES OF JAPANESE

- Agglutinative language
- No space between words
- 3 writing systems:
 - hiragana : for grammatical words, particles, inflectional suffixes :

あ, い, う, え, お

- katakana : for the transcription of foreign words or to emphasize a word in the text:

ア, イ, ウ, エ, オ

- kanji : Chinese characters : for nouns, verbs and adjectives' fixed root and adverbs:

強, 家, 空



GOAL

- Create a dictionary for the Japanese language.
- Find meta-linguistic methods in order to gather as much data as possible to study the language.



OUTLINE

- Specificities of the different categories of the words in Japanese
- Methods of recognition
- Dictionaries and results
- Development of the module



NOUNS

- Invariable in gender and number.
- A post-position occurs after a noun and defines its case.
- Nouns are usually written only in kanji, except in texts written for young readers: nouns are written in hiragana in that case.

Examples :

文学 *bun-gaku* literature

文学を *bun-gaku-o* literature-ACC



ADJECTIVES

- Fixed root (often a kanji followed or not by a hiragana) and a variable base in hiragana.
- Two groups : adjectives –i and adjectives –na, each with their own set of inflected forms
- An adjective always occurs before the noun that it determinates.

重い *omo-i* heavy

静かな *shizuka-na* calm



ADVERBS

- Difficult to recognize by a morphological method, because they can be written only with kanji, or only with hiragana or both.
- Recognition of adverbs derived from adjectives → they have a particular ending: *teki + ni*.

具体的な *gutai-teki-na* concrete

具体的に *gutai-teki-ni* concretely



VERBS

- Fixed root (often one or two kanjis followed or not by a hiragana) and a variable base in hiragana.
- The set of inflected forms is defined for each type of verbal ending: -u, -ru, -su, -tsu, -mu, -nu et -bu.

食べる *tabe-ru* to eat

望む *nozo-mu* to desire

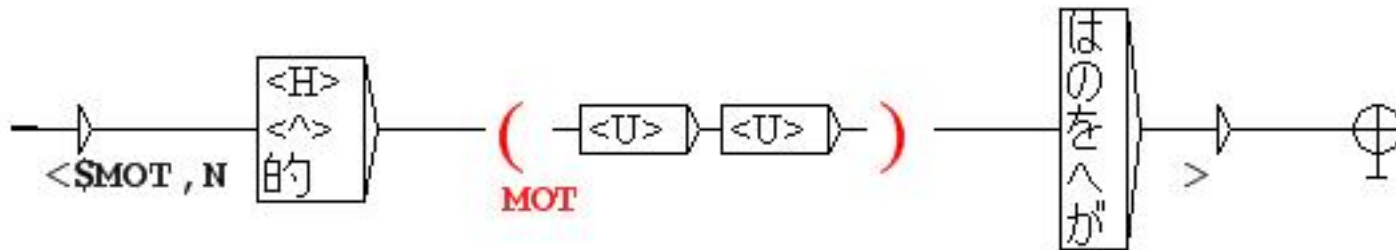
- There are 4 irregular verbs: *suru* (do), *kuru* (come), *iku* (go) et *aru* (exist).
- There are also verbs built with a *kango* (= a noun made out of 2 *kanjis*) followed by the support verb *suru*.

散歩 *sanpo* = a walk → 散歩する *sanpo suru* = to walk



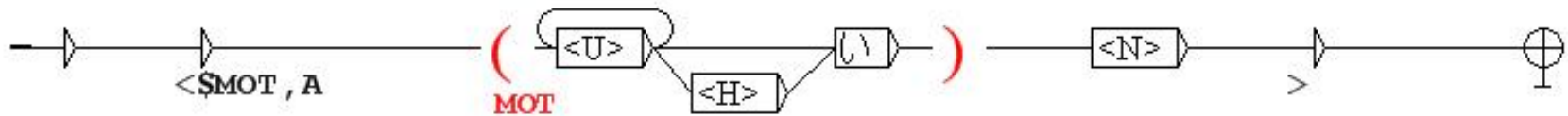
NOUNS

To find a noun built with a *hiragana*, followed by two *kanjis*, followed by a post-position:



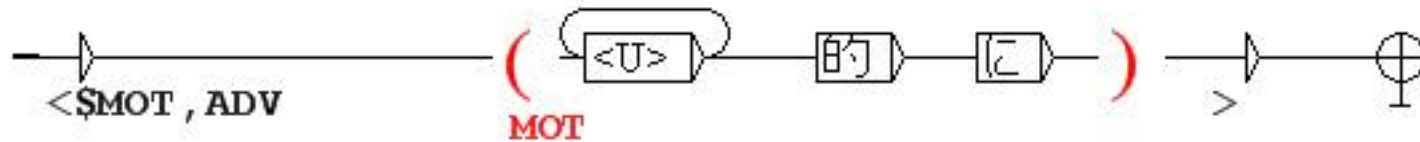
ADJECTIVES

To find the adjectives that determinate a noun already identified:



ADVERBS

To find adverbs that are under the form –teki + ni:



DICTIONARIES

The image shows three overlapping windows of the NooJ software interface, each displaying a different dictionary file. The windows are titled 'NooJ - [mod_nom.dic]', 'NooJ - [mod_adjectif.dic]', and 'NooJ - [mod_verbe.dic]'. Each window has a menu bar with 'File', 'Edit', 'Lab', 'Project', 'Window', 'Windows', and 'Info'. The main content area of each window shows the dictionary's structure, including a header line like '#use noms.nof' and a list of entries with their grammatical classifications.

NooJ - [mod_nom.dic]
Dictionary contains 645 entries
#use noms.nof
葵, N+FLX=NO
悪用, N+FLX=NO
飴, N+FLX=NO
暗黒, N+FLX=NO
暗示, N+FLX=NO
案, N+FLX=NO
案内記, N+FLX=NO
以上, N+FLX=NO
意義, N+FLX=NO
意識, N+FLX=NO
意図, N+FLX=NO
意味, N+FLX=NO
異議, N+FLX=NO
一群, N+FLX=NO
一見, N+FLX=NO
一私人, N+FLX=NO
一糸, N+FLX=NO
一時, N+FLX=NO
一種, N+FLX=NO
一種無, N+FLX=NO
一種無上, N+FLX=NO
一女官, N+FLX=NO
一人, N+FLX=NO
一節, N+FLX=NO
一大系統, N+FLX=NO
一定, N+FLX=NO
一定普遍, N+FLX=NO
一度, N+FLX=NO
一般, N+FLX=NO
一般人間, N+FLX=NO

NooJ - [mod_adjectif.dic]
Dictionary contains 123 entries
#use adjectifs.nof
悪い, A+FLX=I
扱い, A+FLX=I
一応合理的, A+FLX=NA
一見非写実的, A+FLX=NA
一時的, A+FLX=NA
一面的, A+FLX=NA
永久的, A+FLX=NA
鋭利, A+FLX=NA
仮想的, A+FLX=NA
可能, A+FLX=NA
科学的, A+FLX=NA
確實, A+FLX=NA
学的, A+FLX=NA
肝心, A+FLX=NA
間的, A+FLX=NA
危険, A+FLX=NA
器械的, A+FLX=NA
戲曲的, A+FLX=NA
客観的, A+FLX=NA
虚無的, A+FLX=NA
共存的, A+FLX=NA
強い, A+FLX=I
狭い, A+FLX=I

NooJ - [mod_verbe.dic]
Dictionary contains 619 entries
#use verbes.nof
愛する, V+FLX=SURU
愛読する, V+FLX=SURU
扱う, V+FLX=U
暗示する, V+FLX=SURU
暗唱する, V+FLX=SURU
意識する, V+FLX=SURU
意味する, V+FLX=SURU
異なる, V+FLX=RU2
移る, V+FLX=RU2
維持する, V+FLX=SURU
違う, V+FLX=U
一致する, V+FLX=SURU
引用する, V+FLX=SURU
隠す, V+FLX=SU
運用する, V+FLX=SURU
延長する, V+FLX=SURU
応用する, V+FLX=SURU
押す, V+FLX=SU
加える, V+FLX=RU1
歌う, V+FLX=U
画策する, V+FLX=SURU
会得する, V+FLX=SURU
解釈する, V+FLX=SURU
解説する, V+FLX=SURU
回す, V+FLX=SU
概する, V+FLX=SURU
獲得する, V+FLX=SURU
学ぶ, V+FLX=BU

NooJ - [mod_...]
Dictionary contains 18 entries
科学的に, ADV
器械的に, ADV
偶然的に, ADV
芸術的に, ADV
実証的に, ADV
職業的に, ADV
人工的に, ADV
数量的に, ADV
生理的に, ADV
装飾的に, ADV
断片的に, ADV
内容的に, ADV
部分的に, ADV
文学的に, ADV
本質的に, ADV

COVERAGE



NooJ - [kagakuto_bungaku.not [Modified]]

File Edit Lab Project Windows Info TEXT

47 / 112 TUs

Characters
Tokens
Digrams
Unknowns
Ambiguities

Language is "Japanese (Japan)(ja)".
Text Delimiter is: \n (NEWLINE)
Text contains 112 Text Units (TUs).
26348 tokens including:
24752 word forms

Show Text Annotation Structure

史実というものは文学を離れては存在することが困難なように思われる。単なる年代表のようなものとはかく、いわゆる史実が歴史家の手によって一応合理的な連鎖として記録される場合は結局その歴史家の「創作」と見るほかはない。「日本歴史」というものはどこにも存在しなくて、何某の「日本歴史」というものだけが存在するのである。ところが必要な鎖の輪が欠けているために実際は関係のよくわからぬ事件が、史家の推定や臆測《おくそく》で結びつけられる場合が多いであろう。それでいわゆる歴史と称するものは、ほんとうの意味での記録としてずいぶんたより少ないものと考えられるのである。事がらごく最近に起こった場合でもその事からの真相が伝えられることは存外むづかしいものである。たとえばマッキンレーが初めて大統領に選ばれたときに馬鈴薯《ばれいしょ》の値段が暴騰したので、ウィスコンシンの農夫らはそれをこの選挙の結果に帰した。しかし実は産地の早魃《かんぱつ》のためであった。近ごろの新聞には、亭主《ていしゅ》が豆腐を一人で食ってしまって自分に食わせないという理由で自殺した女房のことが伝えられた。まさかこれほどではないまでも歴史の中にはこれに類するものが存外にたくさんあるであろうと想像される。

また一方で歴史と称するものは通例王者、勝利者、支配者の歴史であって、人間の歴史としてははなはだ物足りないものである。少なくとも人間の歴史はただその中に偶然的に織り込まれているに過ぎない。歴史を読むのみではわれわれは祖先民族の生活も心理もきわめておぼろげにしかうかがうことができない。

この欠陥を補うものはまず第一に個人の日記、随感録のごときものである。そういうものが後代に愛読され尊重されるのは、必ずしもそれが「文章」であるためではなくて、それが「記録」であるためであろう。殿上の名もない女官がおぼつかない筆で書いた日記体のもので、それが忠実な記録であるために実証的の価値があり同時にそこに文学としての価値を生じるものと思われる。

第二にはいろいろの物語小説の類である。その中に現われる人物が実際にあったとか、なかったとかいう事はほとんど問題にも何もならないことであって、それらの仮想人物によって代表された人間の定型と、叙述された事件の定型はたしかに存在したのである。これはあらゆる「史実」よりもはるかに確実であって疑う余地を存しない。これはその書が後代の偽作でない限り言われることである。作者がいかにも豊富なる想像力の所有者であってもその時代を偽り描くということは到底不可能な仕事だからである。それで、ちょうど、ある弾丸の描く弾道はまた同時に他のすべての可能な弾道を代表するように、一遊星の軌道はまさしく天体引力の方則を代表するように、光源氏《ひかるげんじ》や葵《あおい》の上《うえ》の行動はまさしくその時代の男女の生活と心理の方則を代表するものとも考えられる。こういう意味において、源氏物語や落窪物語《おちくまものがたり》のようなものは、中等学校の歴史教科書よりも、文化国の大新聞の記事よりも、はるかに忠実な記録であり実証的な資料として役立つものである。おもしろいことには、そういう価値の多少がまたほとんど直ちに普通にいわゆる文学的芸術的価値の多少と一致するように思われるのである。

歴史は繰り返す。方則は不変である。それゆえに過去の記録はまた将来の予言となる。科学の価値と同じく文学の価値もまたこの記録の再現性にかかっていることはいうまでもない。

それのみではない。科学が未知の事象を予報すると同様に、文学は未来の新しい人間現象を予想することも可能である。

想像力の強い昔の作者の予想した物質文明機関で現代にすでに実現されているものがはなはだ多い。電燈でも、飛行機でも、潜水艇でもまたタンク戦車のごときものすら欧州大戦よりずっと以前に小説家によって予想されている。市井の流行風俗、生活状態のようなものはもちろん、いろいろな時代

Cancel

COVERAGE

- Nouns = 723 occ
 - Verbs = 619
 - Adjectives = 123
 - Adverbs = 19
 - Copula = 6
 - Demonstrative words = 9
 - Grammatical words = 12
 - Dates = 5
- Dictionaries totalize 1,438 words.
- 13,317 characters are recognized (out of 26,311), ie 50,6% of the corpus.



DEVELOPMENT OF THE MODULE

- Identify nouns written only with *hiragana* : difficult → how to distinguish the inflectional suffix in *hiragana* of an adjective from a word written also in *hiragana* that follows directly this adjective?
- Identify nouns derived from verbs or adjectives : these nouns have a particular derivational suffix : -*sa*, -*ke* for nouns derived from adjectives and the basis -*i* of the verb for nouns derived from verbs.

Ex : *omo-i* (heavy) → *omo-sa* (heaviness)

kae-ru (to come back) → *kae-ri* (a come back)



DEVELOPMENT OF THE MODULE

- Recognize nouns made of only one kanji.
- Recognize other types of adverbs.
- Develop further the dictionary of grammatical words.
- Complete the dictionary of copula.
- Develop the inflectional grammar of the verbs.



DEVELOPMENT OF THE MODULE

- Recognize nouns made of only one kanji.
- Recognize other types of adverbs.
- Develop further the dictionary of grammatical words.
- Complete the dictionary of copula.
- Develop the inflectional grammar of the verbs.

THANK YOU FOR YOUR ATTENTION

