JaLexGram v-0.14 Lexique Grammaire du Japonais

Raoul Blin

JaLexGram v-0.14

Lexique Grammaire du Japonais

2016/10/18

R.Blin
CNRS-CRLAO
blin@ehess.fr

Table des matières

Objectif....2
Formalisme....2
Les entrées du dictionnaire (395 036 entrées de base ; 25 111 dérivées)....2
Les règles (611)....3
Règle ontologiques....4
Quelques données sur JaLexGram....4
Diffusion du lexique grammaire....5
Evolutions à venir....5
Références....5

Ce texte est une description rapide du lexique grammaire JaLexGram.

Objectif

JaLexGram est un lexique grammaire en cours de construction (en 2016) destiné à l'analyse morphosyntaxique et sémantique du japonais contemporain écrit. Il contient des informations morphophonologiques, syntaxiques et sémantiques et est associé à une ontologie. Il propose en plus des traductions vers le français, exploitables pour une traduction automatique (par règle de transfert).

Formalisme

Le formalisme utilisé est une grammaire de contraintes intégrées, telle que définie dans Renaud, 2005. Pour des raisons pratiques d'implémentation, la syntaxe n'est pas exactement celle de cet ouvrage mais est complètement compatible.

L'objectif est de fournir à terme un lexique grammaire aussi compacte que possible tout en étant réaliste d'un point de vue linguistique (voir discussion dans Blin, 2009). Cet objectif sert de ligne directrice pour le choix des entrées, règles et données ontologiques.

Les entrées du dictionnaire (395 036 entrées de base ; 25 111 dérivées)

Les entrées sont des morphes. En principe, il s'agit des plus petites unités morphosyntaxiques et sémantiques possibles. Dans les faits, la décomposition n'est pas encore complètement réalisée et certaines entrées pourraient peut-être être décomposées. Le gros des entrées a été puisé dans des ressources très diverses. Un grand nombre est issu d'IPADIC (浅原 et 松本 2003) et a subi des modifications : élimination d'une partie des entrées redondantes notamment, recatégorisations.... Plusieurs listes d'entrées ont été ajoutées. Pour un nombre important de verbes et noms communs, les entrées ont été sous-divisées conformément aux entrées du Daijirin en ligne (松村 et 三省堂編修所 2006) .

Les entrées se présentent dans un format pseudo-xml. Elles sont facilement convertibles en XML ou autres. Une entrée contient au minimum les informations suivantes :

```
<entree id="..." >
<ecriture:mixte/>...   // ecriture du morphe ; constante
<trait:lemme/>...      // « identifiant unique du morphe » ; constante
<trait:sem/>...        // trait sémantique ; constante
<trait:rsem/>...       // représentation sémantique ; lambda-terme implicitement typé
<cat/>...              // catégorie syntaxique ; constante
</entree>
```

D'autres items peuvent être rajoutés mais seuls les items précédés de « trait: » ainsi que « ecriture:katakana » et « traduction:francais:val » seront pris en compte. Par exemple :

```
<entree id="193343" >
<ecriture:mixte/>貫通
```

```
<ecriture:katakana/>カンツウ
<trait:arg src="rb" commentaire="auto from glose daijirin/>(ga:(sem:humain_oi)) @ (wo:(sem:
(lemm149115xx_oi @ lemm149115_1xx_oi)))
<trait:lemme/>lemm193343xx
<trait:sem/>lemm193343xx_oe
<trait:rsem/>lbdP, P.(lbdE, (lemm193343xx_oe.E))
<trait:lecture:nbSyllabes/>2
<traduction:francais:val auteur="Claire OLIVIER ; rb" />lbdP, outrad.(P.(lbdX,X).percer).(P.
(lbdX,X).transpercer).(P.(lbdX,X).perforer)
<trait:lecture:nbmores/>4
<trait:strateLexicale source="katarigusa_1_0_1"/>chinois
<cat/>vRad_suruLex
<morphoPhono:composants:alteration/>N-N
<morph:composants:ecriture/>貫 - 通
<morphoPhono:composants:origin/>kan/go-go
<morphoPhono:synthese:origin/>kan; go
<morphoPhono:composants:nbSyllabes/>1-1
<morphoPhono:composants:nbMores/>2-2
<trait:posaccent:depuisdebut source="daijirin en ligne"/>0
<morphoPhono:composants:phoneme+accent:romaji/>ka/N-tu.R
</entree>
```

Les données de cette entrée sont les suivantes :

<ecriture:katakana/> lecture transcrite en katakana ; destinée à être utilisée pour différentes manipulations.

<trait:sem... > trait sémantique, à ne pas confondre avec la représentation sémantique. Ce trait sémantique est utilisé par l'ontologie.

<trait:rsem/> représentation sémantique (lambda-terme).

<trait:lemme/> identifiant, utilisé par la grammaire, notamment pour gérer les exceptions.

<trait:arg/> structure argumentale déduite des exemples fournies dans le Daijirin. La valeur du trait est un terme de traits. Les constantes ' lemm149115xx_oi' etc. sont les traits sémantiques des arguments. Les structures argumentales sont fournies pour 6 058 entrées verbales et adjectivales. C'est très inférieur au dictionnaire de Kyoudai (KAWAHARA et KUROHASHI 2006) mais l'intérêt est que les structures ont été validées manuellement[1].

<trait:strateLexicale> prend comme valeur 'japonais', 'chinois' ou 'gairaigo'. Les données sont pour l'essentiel issues du Katarigusa (茂木 et al. 2010).

<trait:lecture...> données produites automatiquement à partir de la lecture transcrite en katakana.

<morphoPhono:...> données construites automatiquement en exploitant différentes sources. Sont inexploitées.

<traduction:francais:val ...> lambda terme fournissant la traduction en français.

Des entrées peuvent être dérivées à partir d'« entrées de base » (non dérivées elles-même). Toutes les données qui ne sont pas instanciées dans l'entrée dérivée sont importées de l'entrée de référence. Ainsi l'entrée 138227_1 hérite de toutes les propriétés de l'entrée 138227 qui ne sont pas instanciées déjà dans 138227_1 :

```
<entree id="138227_1" fusionneretacquerir="138227" >
...
</entree>
```

Les règles (611)

Les règles se présentent comme suit. Seules les deux premières lignes sont obligatoires. Règle de contraintes, lecture et traduction sont des lambda-termes. U0 est le terme de traits résultant de l'opération f sur les termes de traits U1, U2, ... qui sont les termes de traits des composants c1, c2, ... du syntagme.

1 Pour une revue récente des ressources existantes, voir Marchal, 2015

```
<regle IDENTIFIANT>
c0 <-- c1 c2 ...                // regle de réécriture
U0:=f.U1.U2. ...                // contraintes, comprend la règle d'interprétation
L0:=f'.L1.L2. ...               // règle de lecture
T0:=f''.T1.T2 ...               // règle de traduction
</regle>
```

Cette représentation des règles est valable aussi bien pour la morphologie que pour la syntaxe. Par exemple, la règle d'association des préfixes du type *moto* (« ex- ») avec un nom :

```
<regle 1306111322>
nomCommun <-- classe_kyuu_moto nomCommun_morph
U0:=\siAlorsSinon
                .( (subsume.(contrainteSemSurNom:CSN1).U1) & (isInclu.SEM2.CSN1)  )
                .( (elimineSousTrait.rsem.(elimineSousTrait.contrainteSemSurNom.U1))
                        @ (siAlorsSinon
                                .(subsume.(arg:ARG2).U2)
                                .(arg:ARG2)
                                .traitNull
                                )
                        @ (arg:ARG2)
                        @ (rsem:(RSEM1.RSEM2))
                )
                .faux\
T0:=\lbdP, (T2.(lbdGN2,N2, (
                P.(lbdX, (T1.(GN2.X))).N2
                )))\
</regle>
```

Règle ontologiques

Les gloses fournies dans les dictionnaires existants et d'autres ressources encore ont permis d'élaborer en partie automatiquement un réseau de relations de synonymie, hypo-hyperonymie entre les entrées du dictionnaire. Contrairement à un dictionnaire comme le Goitaikei (Ikehara et al. 1997) qui recoure à des classements intuitifs difficiles à apprécier, les relations dans le JaLexGram sont évaluables (Blin 2012). Chaque mot ou groupe de mots synonymes constitue une classe « sémantique ». Quelques catégories sémantiques sont *ad hoc*. Les éléments de ces classes ont la caractéristique de répondre à une « intuition sémantique », mais aussi nécessairement d'avoir des comportements linguistiques spécifiques observables.

C'est le cas des noms communs suffixables de 元 (*moto* , « ex, ancien ») (Blin 2016). L'exemple suivant signifie que les *yakuza* (« yakuza, mafieux ») sont classés comme entités susceptibles d'être « datées » par le préfixe *moto* (« ex-yakuza »); que les entités « datables » avec *moto* sont toujours des 'humains' (hyponyme de *hito* ou *ningen*), et que les humains sont des 'personnes morales'. La catégorie de 'personne morale' est elle aussi une catégorie *ad hoc* qui s'est imposée pour caractériser un grand nombre de noms ayant en commun d'être arguments des mêmes verbes. Il ne s'agit pas de 'personnes morales' au sens juridique usuel.

```
<onto auteur="rb"  commentaire="Yakuza"/>lemm156060xx_oi =< prefixableDeMOTO_oi
<onto auteur="rb"/>prefixableDeMOTO_oi =< humain_oi
<onto auteur="rb"/>humain_oi =< personneMorale_oi
```

Quelques données sur JaLexGram

Nombre d'entrées de base : 395 036

Nombre d'entrées dérivées : 25 111

Nombre de règles : 611

Nombre de règles d'ontologie : 172 236

Nombre d'entrées verbales ou adjectivales comportant une information sur la structure argumentale : 6 058

Nombre de traits : 41

Nombre de catégories : 269

Nombre d'entrées traduites : 239 609 . La traduction d'une majorité de noms propres vaut simplement la lecture transcrite en katakana.

Répartition des strates lexicales (pour les seules entrées comportant l'information sur la strate lexicale) :

strate lexicale:chinois:30505 (42%)

strate lexicale:japonais:26392 (36%)

strate lexicale:gairaigo:16324 (22%)

Nombre d'entrées comportant une indication de la position de l'accent : 72554

Principales accentuation (<position accent> nombre) : <1> 15595

<3> 8032

<0> 7179

<2> 7025

<4> 3000

<5> 1300

...

Diffusion du lexique grammaire

Le JaLexGram est destiné à terme à être diffusé librement. Dans l'immédiat, il peut être partagé dans le cadre de collaborations. Pour toute information, contacter blin@ehess.fr .

Evolutions à venir

...Innombrables. Une évolution majeure serait l'introduction d'une base de connaissances.

Références

Blin, Raoul. 2009. *Introduction à la linguistique formelle*. TIC et sciences cognitives. Paris: Hermès science publications : Lavoisier.

———. 2012. « sokuryou meisi no imi - tougo teki tokutyou ni kan suru kousatu [syntaxe et sémantique des noms de grandeurs (en japonais)] ». In *Lexicon Forum no.6*, Kageyama Taro, 173-202. Hituji Shobo.

———. 2016. « The nominal prefix in Japanese 元 moto- (”ex-”) ». https://hal.archives-ouvertes.fr/hal-01279394.

Ikehara, Satoru, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, et Yoshihiko Hayashi. 1997. *Goi-Taikei - A Japanese Lexicon*. Iwanami Shoten. Tokyo.

KAWAHARA, Daisuke, et Sadao KUROHASHI. 2006. « Case Frame Compilation from the Web using High-Performance Computing ». *IPSJ SIG Notes* 2006 (1): 67-73.

Marchal, Pierre. 2015. « Acquisition de schémas prédicatifs verbaux en japonais ». Paris: INALCO.

Renaud, Francis. 2005. *Temps, durativité, télicité*. Peeters. Bibliothèque de l’Information Grammaticale 60. Louvain, Paris.

松村明, et 三省堂編修所. 2006. 大辞林. 第 3 版 éd. 三省堂. http://ci.nii.ac.jp/ncid/BA78867520.

浅原正幸, et 松本裕治. 2003. « IPADIC ユーザーズマ ニュアル （version 2.7.0) ». In . 奈良先端科学技術大 学院大学 情報科学研究科 松本研究室.

茂木俊伸, 山口昌也, 桐生りか, et 田中牧郎. 2010. « 語種辞書『かたりぐさ』　利用マニュアル ». http://www2.ninjal.ac.jp/lrc/index.php?%B8%EC%BC%EF%BC%AD%BD%F1%A1%D8%A4%AB%A4%BF%A4%EA%A4%B0%A4%B5%A1%D9%2F%CD%F8%CD%D1%A5%DE%A5%CB%A5%E5%A5%A2%A5%EB.