Alexis Kauffmann

LATL (Laboratoire d'Analyse et de Technologie du Langage) – Université de Genève, 2, rue de Candolle, 1211 Genève, Suisse alexis.kauffmann@lettres.unige.ch

Résumé Nous décrivons la façon dont est formée la phrase japonaise, avec son contenu minimal, la structure des composants d'une phrase simple et l'ordre des mots dans ses composants, les différentes phrases complexes et les possibilités de changements modaux. Le but de cette description est de permettre l'analyse de la phrase japonaise selon des principes universels tout en restant fidèles aux particularités de la langue. L'analyseur syntaxique multilingue FIPS est en cours d'adaptation pour le japonais selon les règles de grammaire qui ont été définies. Bien qu'il fonctionnait alors uniquement pour des langues occidentales, les premiers résultats sont très positifs pour l'analyse des phrases simples, ce qui montre la capacité de Fips à s'adapter à des langues très différentes.

Abstract We describe here the way the Japanese sentence is structured, with its minimal content, the structure of its components, the different kinds of complex sentences and the possible modal effects. The aim of this description is to analyse the Japanese sentence in a universal way while respecting the language properties. The syntactic parser FIPS is now being developed to analyse Japanese sentences following grammar rules coming from this syntactic description. Even though the parser used to be used for Western languages only, the first results of simple Japanese sentence analysis are very positive. This shows how FIPS can adapt easily to very different languages.

Mots-clés: Analyse Syntaxique, Japonais, Grammaire

Keywords: Syntactic Parser, Japanese language, Grammar

1 Introduction

L'analyse syntaxique automatique du japonais est un sujet complexe qui a été abondamment abordé, notamment au Japon. Nous l'abordons ici dans un cadre multilingue, en utilisant l'analyseur FIPS du LATL, qui peut déjà effectuer des analyses de plusieurs langues, dont notamment l'allemand, l'anglais, le français, l'espagnol, l'italien et le grec.

L'analyseur FIPS (Wehrli, 2007) permet une analyse linguistique profonde de la phrase, en se basant sur les principes de la grammaire générative universelle (Chomsky,1995). Il permet de

représenter chaque constituant de la phrase selon une structure X-barre simplifiée de la forme [XPG X D], où XP indique le type de constituant¹, X est sa tête (qui peut être vide), et G et D sont les listes de sous-constituants de gauche et de droite, qui peuvent aussi être vides. FIPS a été écrit en Component Pascal selon une architecture à objets, avec des modules d'analyses génériques et d'autres qui implémentent chacune des règles syntaxiques spécifiques à une langue donnée. Cependant, il a jusqu'à maintenant été utilisé essentiellement pour des langues occidentales.

La problématique liée à l'ajout du japonais se situe à deux niveaux importants. Il faut d'une part, réussir à surmonter les différences essentielles entre le japonais et les langues occidentales : le japonais est écrit avec son propre système d'écriture, comprenant 2 alphabets d'une cinquantaine de caractères, environ 2000 idéogrammes usuels et d'autres plus rares; il ne comporte pas d'espaces entre les mots et peut s'écrire soit de haut en bas et de droite à gauche, soit dans le sens des langues occidentales. D'autre part, il faut pouvoir analyser les phrases japonaises selon les principes syntaxiques de la grammaire universelle tout en tenant compte des spécificités de la grammaire japonaise et des difficultés liées a son analyse: absence fréquente de formulation de certains constituants dans la phrase, scrambling², possibilité de phrases adjectivales, absence de déterminants finis ou indéfinis, rôle fort joué par les particules dans la phrase, etc.

L'intégration de la langue japonaise à l'analyseur s'est déroulée en plusieurs étapes et n'est pas encore finie. Après avoir créé un générateur morphologique pour les verbes³ et adjectifs japonais, déterminé une classification lexicale et créé un lexique japonais d'environ 15000 lexèmes, nous avons étudié plus particulièrement la structure syntaxique des phrases. Le travail a consisté Nous allons donc proposer ici une description de la façon dont sont structurées les phrases japonaises avec leurs composants et l'ordre des mots à l'intérieur de ces ceux-ci. On regardera ensuite brièvement les premiers résultats obtenus après un début d'implémentation des règles de grammaire correspondantes.

2 Travaux existants

La grammaire scolaire habituellement enseignée à l'école au Japon (Hashimoto, 1934) mélange des éléments traditionnels des grammaires occidentales et des notions plus typiquement japonaises. Des linguistes ont par la suite cherché à définir des analyses encore moins conditionnées par les principes occidentaux (St-Jacques, 1966). La plupart des grammaires japonaises modernes considèrent bien les particules comme des mots, mais pas les terminaisons verbales (Shinamori, 1997). Nous avons suivi cette optique-là. Les travaux linguistiques plus récents décrivent des analyses de phrases selon la grammaire générative (Tsujimura, 1996), en discutant des points comme la façon d'analyser les modes ou le rôle des particules.

La phrase se divise en deux parties : TP : «Time Phrase » (Syntagme temporel) et parfois CP : Complémenteur. On fait une projection fonctionnelle FP pour les adjectifs prédicatifs. Les autres constituants prennent le nom de leur tête: DP : déterminant, NP : nom, VP : verbe, AP : adjectif, AdvP : adverbe, CP : complémenteur, ConjP : conjonction de coordination, PP : pré/postposition.

² Scrambling : possibilité de réordonnancement par rapport à l'ordre de base des arguments et sujet du verbe.

³ Voir sur le site web du LATL : http://www.latl.unige.ch/french/projets/conjugueur/notes.html

Les analyseurs syntaxiques du japonais existant sont très nombreux, et nous n'en citerons ici que quelques uns. Un des plus utilisés est « KNP » (Kurohashi et Nagao, 1994). Il se base sur la grammaire de Masuoka et Takubo (Masuoka et Takubo, 1992). La classification lexicale que nous utilisons est proche de celle de cette grammaire. KNP ne crée pas une analyse syntaxique profonde mais une structure de dépendance basée sur les cas du japonais. Tsujii a créé un analyseur du japonais basé sur une grammaire HPSG (Mitsuishi J., Torisawa, K., Tsujii J., 1998), permettant une analyse de structure plus profonde. Plus récemment, de nombreux analyseurs probabilistes ont été mis au point, comme celui de Kawahara (Kawahara et Kurohashi, 2006). Les trois projets monolingues qui viennent d'être évoqués ont eu de bons résultats d'analyse, avec des résultats évalués corrects à 90% environ. Le projet MedSLT, lui, est un projet multilingue d'analyse et de traduction automatique à base de règles grammaticales qui analyse le japonais (Santaholma, 2007). Mais contrairement à FIPS, il est spécialisé uniquement dans le langage du domaine médical.

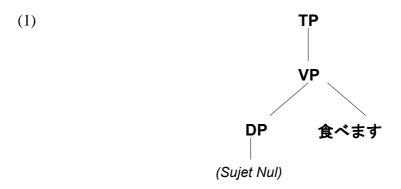
3 Structure de la phrase japonaise

3.1 La phrase minimale

On peut considérer qu'il existe en japonais trois structures typiques de phrases minimales : les phrases verbales, les phrases adjectivales et les phrases prédicatives (à copule). Le sujet n'est pas forcément formulé, le japonais étant une langue pro-drop⁴.

3.1.1 la phrase verbale minimale

Elle est constituée d'un verbe. Dans l'analyse de la grammaire, il est proposé un sujet nul qui est représenté par un syntagme nominal⁵ (Determiner Phrase: DP) vide. Il est inclus dans le groupe verbal (Verbal Phrase : VP). On métaprojette une phrase (Time Phrase: TP).



食べます。

MangePRÉSENT POLI => Je (/tu/il/elle/on/nous/vous/ils/elles) mange (/es/eons/ez/ent).

Cette phrase est grammaticale, malgré la non formulation du sujet, qui est implicité uniquement par le contexte, la terminaison du verbe ne variant ni selon la personne ni selon le

⁴ Pro-drop : se dit, en grammaire générative, d'une langue où la présence du pronom sujet devant le verbe n'est qu'optionnelle. C'est aussi le cas pour l'espagnol, l'italien et les langues slaves.

On considère que le déterminant constitue la tête du syntagme nominal dans nos analyses. Voir : Abney, S. P. (1987). *The English Noun Phrase in its Sentential Aspect*, MIT, Cambridge MA.

nombre. Il est également possible dans la phrase japonaise que les compléments ne soient pas explicitement formulés.

3.1.2 La phrase adjectivale minimale

Elle est constituée d'un adjectif en - VV_{\circ} Cette catégorie d'adjectifs se conjugue selon le temps, un peu comme un verbe.

(2) [TP[FP]DP][AP 面白かった]]] Le sujet nul est représenté par un DP vide dans l'analyse. On fait une projection fonctionnelle (FP) pour l'adjectif prédicatif.

drôle/intéressant PASSÉ

Ici aussi le sujet est implicité. Selon le contexte, le sujet peut être par exemple une personne : « (Kaizou) était drôle » ou alors un sujet plus indéterminé: « (c') était drôle. » Les phrases minimales japonaises sont d'ailleurs souvent sujettes à des demandes de désambiguisations au cours des conversations.

3.1.3 La phrase prédicatives (à copule) minimale

Elle est constituée ainsi : soit nom (ou pronom ou syntagme nominal sans particule) + copule, soit adverbe + copule, soit quantifieur + copule, soit adjectif en - + copule. La copule peut, selon les cas, soit être équivalente au verbe être au temps correspondant avec un sujet exprimé ou implicite, soit être équivalente à « c'est », « ce sont », « c'étaient »...

Il ne peut pas y avoir de mot ou de particule de cas entre l'attribut et la copule. La copule seule ne peut pas constituer une phrase minimale :

*C'était.

En revanche, si la copule n'est pas présente, la phrase reste compréhensible. On la comprend alors comme si l'attribut était suivi d'une copule au présent.

3.2 Ordre des mots et rôle des particules dans une phrase simple

3.2.1 Ordre des mots

Le japonais est une langue à tête finale. L'ordre principal des mots dans la phrase japonaise est Sujet-Objet-Verbe.

MoiTHEMESUJET saumonOBJ mangerPASSÉ

=> Moi j'ai mangé du saumon.

L'ordre des éléments de la phrase n'est pas strict, mais en général la phrase commence par le sujet thématique, puis vient ensuite le sujet grammatical, puis le ou les objets. Les compléments circonstanciels ou adverbes phrastiques peuvent se situer n'importe où avant le verbe ou l'adjectif.

La phrase monopropositionnelle se termine par son élément minimal (verbe ou adjectif prédicatif ou attribut+copule), suivi de l'auxiliaire pour le verbe ou l'adjectif lorsqu'il s'agit d'un temps composé, puis parfois d'une ou plusieurs particules de fin de phrase. Le verbe ou l'adjectif peut être précédé par un adverbe, ou dans certains cas plus rares suivi d'une particule adverbiale.

FranceTHEME déjà snowboardOBJ fairePOTENTIEL QUESTION

=> Est-ce qu'on peut déjà faire du snowboard en France?

ou: [TP[AdvP 今] [FP [DP 東京は][AP [AdvP とても] 寒い]]です] Dans la première analyse « Tokyo » est complémenteur, dans la deuxième il est sujet.

TokyoTHEME maintenant très froid (c')est.

=> Il fait très froid à Tokyo en ce moment.

Mais l'ordre et la présence ou non des syntagmes nominaux précédant le verbe n'étant pas fixe (phénomène de scrambling), on reconnaît leur fonction essentiellement par le type de particule qui leur est attribué.

3.2.2 Rôle des particules

La particule $black{3}{\mathbb{S}}(ga)$ indique le sujet grammatical, (sauf avec un verbe au mode potentiel ou dans certaines phrases adjectivales, voir 3.5.1) tandis que $black{2}{\mathbb{E}}(wo)$ indique l'objet direct. L'objet indirect et le complément circonstanciel sont suivis par une particule postposition. Une particule de topique indique le sujet thématique. Le plus souvent, il s'agit de la particule lapha(lapha). Si le sujet thématique porte sur un groupe postpositionnel, la particule de topique s'ajoute derrière la particule postposition. Si le sujet thématique est le sujet grammatical ou l'objet direct de la phrase, alors on l'analysera comme sujet ou objet direct dans le groupe verbal (VP) et on ne le présentera pas au niveau complémenteur (CP). Dans ces cas-là, la particule de topique remplace les particules $black{3}{\mathbb{S}}$ ou $black{2}{\mathbb{C}}$.

3.3 Structure des groupes nominaux/postpositionnels

Les syntagmes nominaux(DP) peuvent commencer par un déterminant possessif, démonstratif ou interrogatif, ou, plus rarement, un pronom démonstratif ou interrogatif. Il n'existe pas de

déterminant fini ou indéfini, les noms n'ont pas de genre et ne s'accordent pas selon le nombre. On peut ensuite avoir un adjectif, une subordonnée relative ou un groupe postpositionnel au cas possessif précédant le nom.

Il existe trois classes morphologiques d'adjectifs : en - VVV, en - 7 et en - O. Les adjectifs en - VVV ont leur terminaison qui varie selon le temps. Les adjectifs en - 7 reçoivent le suffixe - 7 lorsqu'ils se trouvent devant un nom. Ils sont invariables. Les adjectifs en - O reçoivent la particule - O (no) lorsqu'ils se trouvent devant un nom. Ce sont en fait des noms qui prennent une valeur d'adjectif lorsqu'on leur donne le cas possessif.

On trouve ensuite le nom. Il est souvent suivi d'une particule. Mais, avant la particule, il peut être suivi d'un deuxième nom pour former un mot composé, ou d'un titre honorifique ou professionnel pour un nom propre.

Tanaka Monsieur

Un syntagme nominal peut être également formé de plusieurs syntagmes nominaux coordonnés par des particules conjonctions de coordination. On peut aussi lui associer un quantifieur qui peut se placer à plusieurs endroits possibles dans la phrase.

3.4 Phrase complexe

On ne peut en japonais utiliser plusieurs verbes dans une même phrase qu'en utilisant certaines structures spécifiques de subordination, coordination ou énumération.

3.4.1 coordination

Il faut tout d'abord remarquer que les conjonctions de coordination utilisées entre des syntagmes nominaux et des propositions ne sont général pas les mêmes. Par exemple, la particule conjonction de coordination \succeq (to) qui signifie « et » ne peut s'employer que pour relier des syntagmes nominaux, et non des phrases. Par contre, on peut coordonner deux propositions avec des particules équivalentes à « mais », comme la particule de coordination \mathcal{D}^{ς} (ga). La première proposition a alors la structure d'une phrase monopropositionnelle sans particules de fin de phrase, suivie de la particule de coordination puis souvent d'une virgule japonaise « , », puis de la deuxième proposition.

3.4.2 Subordination

On peut faire une phrase avec une particule conjonction de subordination entre deux propositions de la même manière qu'avec une particule conjonction de coordination.

3.4.3 Subordonnée relative ou complétive

Il en existe trois types:

• subordonnée complétive précédant un verbe.

Celle-ci se construit avec la particule de subordination complétive \succeq (to), qui est équivalente à « que » en français. On écrit la proposition subordonnée sans particule de fin de phrase, suivie de \succeq puis du verbe auquel la subordonnée est rattachée.

(10)[TP [VP [DP]][CP [TP [VP [DP]] 行きたい]]と]言った]] Les têtes des constituants sont finales. 行きたい と 言った。

vouloir aller que avoir dit => (II) a dit qu'(il) voulait y aller.

subordonnée relative

La construction de la subordonnée relative ne nécessite pas de particule : après la proposition subordonnée sans particule de fin, on ajoute le nom. On ne sait pas à priori quel pronom relatif il faudrait imaginer entre la subordonnée et le nom : « qui », « que », « dont »... Mais en général on le devine aisément grâce au contexte.

(11) [DP [NP [CP [TP [VP [DP]行きたい]]] 人]] 行きたい 人

vouloir aller personne => les gens qui veulent y aller

• subordonnée complétive exprimant la nominalisation d'un verbe

Cette subordonnée se construit exactement comme la précédente. Elle ne peut être employée qu'avec quelques mots comme \P (la chose, l'action) ou $\mathcal{E} \subset \mathcal{E}$ (le lieu), ou alors avec la particule de nominalisation \mathcal{O} à la place du nom. Elle transforme le verbe en un nom, qui doit se comprendre un peu comme un infinitif français.

3.4.4 Enumération verbale

On peut en japonais faire une phrase avec plusieurs verbes juxtaposés. Dans une telle phrase, seul le dernier verbe est réellement un verbe conjugué, le ou les autres étant obligatoirement à une forme participiale. Si la forme participiale est le gérondif en - (c'est comme si on avait «et » ou «et puis » entre les groupes verbaux.

(13)[TP [ConjP [TP[VP [DP 私は] 起きて]] [TP[VP[DP シャワーを] 浴びて]] [TP[VP [DP 朝ご飯を] 食べました]]]] On groupe les propositions comme si elles étaient coordonnées avec un ConjP.

私は 起きて、シャワーを浴びて、 朝ご飯を 食べました。

JeTHEMESUJET se leverGEROND, doucheOBJ prendreGEROND, petit-déjeunerOBJ mangerPASSÉ POLI =>Je me suis levé, j'ai pris ma douche et puis j'ai mangé mon petit-déjeuner.

3.5 Changements modaux : Potentiel, Passif

En japonais le verbe peut être conjugué à différents modes. La structure de la phrase se trouve changée, par rapport à la structure d'une phrase dont le verbe est à l'indicatif, si le verbe est au mode factitif, au factitif passif, ou notamment au potentiel ou au passif.

3.5.1 Mode potentiel

Le mode potentiel exprime la capacité à accomplir une action, un peu comme le modal « can » en anglais. Il existe deux manières différentes d'analyser le mode potentiel.

Dans la première analyse on considère que le mode potentiel a une structure passive, traduisible littéralement par « pouvoir être+participe passé », et que donc la particule $\mathring{\mathcal{D}}^{\varsigma}(ga)$ indique bien le sujet grammatical.

AnthonyTHEME tennisSUJET fairePOTENTIEL

=> Pour Anthony, le tennis peut être fait.=> Anthony sait jouer au tennis.

Cette analyse s'adapte bien au fait que la forme traditionnelle potentielle des verbes du 2ème groupe est la même que leur forme passive. De plus, elle ressemble à l'analyse des phrases adjectivales, qui utilisent également les particules $l\sharp(ha)$ et $\hbar^{\sharp}(ga)$.

AnthonyTHEME jambeSUJET douloureuse

=>Anthony, sa jambe est douloureuse.=>Anthony a mal à la jambe.

Cependant, cette méthode ne marche que pour les verbes transitifs, pas pour les intransitifs. Dans la deuxième méthode d'analyse, qui fonctionne pour les verbes transitifs ou intransitifs, on considère que le mode potentiel à une structure active, traduisible littéralement par « pouvoir +verbe », et que $\mathring{\mathbb{Z}}$, qui remplace $\mathring{\mathbb{Z}}$ (wo), indique l'objet direct et non le sujet grammatical dans les phrases au mode potentiel.

AnthonyTHEMESUJET tennisOBJET fairePOTENTIEL

=> Anthony peut faire du tennis. => Anthony sait jouer au tennis.

Cette analyse permet aussi de comprendre l'usage de \mathcal{E} à la place de \mathcal{D}^{S} dans les phrases au mode potentiel, qui est de plus en plus accepté à l'oral. Elle peut également être utilisée pour certaines phrases adjectivales exprimant la capacité où \mathcal{D}^{S} n'indique vraisemblablement pas le sujet grammatical.

Il faut par ailleurs ajouter qu'il existe des verbes qui sont par définition au mode potentiel (ex : 50), ou factitif, alors que la plupart des verbes ont leur infinitif défini à l'indicatif.

3.5.2 Mode passif

passif direct

Au passif direct, pour un verbe transitif, l'objet direct de la phrase active correspondante devient le sujet. Il est suivi de \mathfrak{B} , qui peut être remplacé par la particule de topique \mathfrak{L} . Le complément d'agent, s'il est exprimé, est suivi de \mathfrak{L} (ni), ou $\mathfrak{L} \mathfrak{L} \mathfrak{I} \mathfrak{I}$ parfois.

(le)livreSUJET Takahashi par lirePASSÉ PASSIF

passif indirect

Au passif indirect, c'est l'objet indirect de la phrase active qui devient sujet, comme c'est parfois le cas en anglais.

MitsukoSUJET mauvais sushiOBJ vendre PASSÉ PASSIF

=> Mitsuko, on lui a vendu des mauvais sushis.

Au passif indirect, on peut aussi former un passif avec un verbe intransitif.

Mitsukothemesujet pluie par tomber PASSÉ PASSIF

=> Mitsuko a été mouillée par la pluie./ Mitsuko a pris la pluie.

4 Implémentation et premiers résultats

Après avoir fini la description des phrases japonaises, nous en avons déduit une quarantaine de règles d'analyse syntaxique qui sont en cours d'implémentation.

L'absence d'espaces entre les mots japonais était une première difficulté à surmonter. Ce problème a pu être réglé en effectuant une analyse morphologique simultanément à l'analyse syntaxique. Une autre difficulté réside dans le fait que le japonais est une langue à tête finale avec un ordre des constituants très peu stable, contrairement aux langue occidentales. On a donc adapté l'algorithme d'analyse pour générer un syntagme temporel avant de rencontrer le verbe, qui se trouve toujours en fin proposition, en tenant bien compte des particules des groupes nominaux situés avant le verbe.

Nous avons pour l'instant réussi à analyser efficacement les phrases minimales (voir exemples 1 à 3), les groupes nominaux sans subordonnée relative (ex. 8 et 9) et la plupart des cas de phrases simples (voir exemples 5 à 7). Les premiers résultats sont donc très satisfaisants. Il subsiste cependant des problèmes concernant l'analyse des phrases complexes et des subordonnées relatives. Il faudra par ailleurs améliorer la prise en charge des noms propres en insérant dans le lexique des listes importantes de noms propres japonais, ou de noms propres étrangers retranscrits en japonais.

5 Conclusion

La description proposée des propriétés grammaticales de la langue japonaise et l'implémentation des règles correspondant à ces propriétés ont permis de rendre possible une analyse syntaxique des phrases japonaises par l'analyseur multilingue FIPS. Avec une analyse profonde et un lexique riche, celui-ci peut retrouver des relations de dépendances peu configurationnelles. Il peut gérer le phénomène du scrambling grâce au traitement des particules, et il permet de gérer les têtes finales grâce aux paramètres de son schéma x-barre, en fixant le spécifieur et les compléments à gauche et la tête à droite. De plus, une analyse selon la grammaire universelle, applicable à toutes les langues, est plus exploitable pour la traduction automatique du japonais avec une autre langue qu'une analyse dans une grammaire trop spécifique au japonais. On espère donc par la suite utiliser les analyses du japonais pour de la traduction automatique.

Remerciements

Je voudrais remercier Izumi Tahara et Christopher Laenzlinger pour l'aide qu'ils m'ont apportée.

Références

Наянмото S.(1934) 国語法要説 Tokyo, Meiji Shoin.

SAINT-JACQUES B.(1966) Analyse structurale de la syntaxe du japonais moderne. Librairie C.Klincksieck.

Kuwae R. (1980) Manuel de Japonais, L'Asiathèque.

Masuoka T et Takubo Y.(1992) 基礎日本語文法, Kurishio Shuppan, Tokyo.

Kurohashi S. et Nagao M. (1994) A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. *Computational Linguistics*, 20(4):507–534.

CHOMSKY N. (1995) The Minimalist Program, Cambridge, Mass., MIT Press.

TSUJIMURA N. (1996) An introduction to Japanese Linguistics, Blackwell Publishers.

Shinamori R. (1997) Grammaire Japonaise Systématique, Jean Maisonneuve.

MITSUISHI J., TORISAWA, K., TSUJII J.(1998) HPSG-Style Underspecified Japanese Grammar with Wide Coverage. Actes de *COLING—ACL 98*, pp. 876--880 Montreal, Canada.

KAWAHARA D. et KUROHASHI S. (2006): A Fully-Lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis, Actes de *HLT2006*, pp.176-183, New York. USA.

Wehrli E.(2007): Fips, a "Deep" Linguistic Multilingual Parser, Actes de ACL2007, Prague.

Santaholma M.(2007). Grammar sharing techniques for rule-based multilingual NLP systems. Actes de *NODALIDA* 2007, Tartu, Estonia.