Publié dans Revue Tranel (Travaux neuchâtelois de linguistique) 50, 115-132, 2009
qui doit être utilisée pour toute référence à ce travail

Ambiguïté des kanji et stratégie de désambiguïsation par le contexte

Nadine RAYON
LaLIC-STIH, équipe Certal (Université de Paris Sorbonne et INALCO),
Maison de la Recherche, 28 rue Serpente, F-75006 Paris
rayon.nadine@gmail.com

In this article we will examine the ambiguity of kanji, characters of Chinese origin, and one of the character sets used to write the Japanese language. We will first describe the ambiguity phenomenon by presenting the origin of this problem and its multilevel effects on the Japanese texts decoding. Then, we will expose a context based method of disambiguating kanji which draws on Japanese graphotax specificities themselves. Lastly, we will show the effectiveness of this disambiguating method when applied to computational linguistics and particularly to an automatic morphological analysis and segmentation system for Japanese texts.

0. Introduction

Avant l'importation des caractères chinois les Japonais ne possédaient pas d'écriture. L'appropriation par les Japonais de ces caractères et la constitution d'un système d'écriture propre représente un long processus dont le résultat est un système d'écriture unique, réputé pour sa difficulté de maîtrise. Si cette dernière affirmation peut être estimée relative, il reste irréfutable que le système d'écriture japonais recèle certaines spécificités qui posent notamment le problème de son décodage au niveau le plus élémentaire, l'oralisation, mais aussi aux autres niveaux de l'analyse linguistique.

Néanmoins, si ces particularités ont généré une forte ambiguïté, elles offrent également un moyen de la lever, de façon systématique, grâce au lien entre la graphotaxe[1], la morphologie et la syntaxe. Cette stratégie de désambiguïsation est une aide fondamentale pour le lecteur humain. Elle peut par ailleurs être mise en œuvre avec succès dans le cadre du traitement automatique de textes japonais.

Nous décrirons dans un premier temps le phénomène d'ambiguïté en lui-même puis les stratégies de désambiguïsation. Certains points concernant la langue japonaise seront abordés sans prétendre à l'exhaustivité et dans la mesure où cela a semblé pertinent pour la compréhension de la problématique centrale. Enfin, nous présenterons notre système d'analyse et de segmentation automatique de textes japonais basé sur ces stratégies de désambiguïsation.

[1] Règles d'organisation des caractères pour l'écriture d'une langue donnée.

1. Le système d'écriture japonais

1.1 Les types de caractères

Le système d'écriture japonais a été créé à partir des caractères chinois et est le résultat de l'adaptation de ces caractères à la langue japonaise. Il est composite et comporte deux types de caractères répartis en trois jeux de caractères.

1) Les *kanji*[2], littéralement "caractères chinois", caractères idéographiques importés de Chine.

丁 予 化 区 反 央 平 申 世 由 氷 主 仕 他 代 写 号 去 打 皮 皿 礼 両 曲 向 州 全 次 安
守 式 死 列 羊 有 血 住 助 医 君 坂 局 役 投 対 決 究 豆 身 返 表 事 育 使 命 味 幸 始
実 定 岸 所 放 昔 板 泳 注 波 油 受 物 具 委 和 者 取 服 苦 重 乗 係 品 客 県 屋 炭 度

Fig. 1: Quelques kanji

2) Les *kana*, deux séries de caractères syllabiques qui notent la même série de syllabes du japonais. Ils ont été créés par les Japonais à partir des caractères chinois. (1) Les *hiragana* résultent de la cursivisation et de la simplification des caractères chinois, (2) les *katakana* résultent de l'isolement d'une partie d'un caractère chinois.

		k	g	s	z	t	d	n	h	b	p	m	y	r	w	
a	あ ア	か カ	が ガ	さ サ	ざ ザ	た タ	だ ダ	な ナ	は ハ	ば バ	ぱ パ	ま マ	や ヤ	ら ラ	わ ワ	
i	い イ	き キ	ぎ ギ	し シ	じ ジ	ち チ	ぢ ヂ	に ニ	ひ ヒ	び ビ	ぴ ピ	み ミ		り リ		
u	う ウ	く ク	ぐ グ	す ス	ず ズ	つ ツ	づ ヅ	ぬ ヌ	ふ フ	ぶ ブ	ぷ プ	む ム	ゆ ユ	る ル		
e	え エ	け ケ	げ ゲ	せ セ	ぜ ゼ	て テ	で デ	ね ネ	へ ヘ	べ ベ	ぺ ペ	め メ		れ レ		
o	お オ	こ コ	ご ゴ	そ ソ	ぞ ゾ	と ト	ど ド	の ノ	ほ ホ	ぼ ボ	ぽ ポ	も モ	よ ヨ	ろ ロ	を ヲ	
																ん ン

Fig. 2: Les kana[3]

A ces trois jeux de caractères, on peut ajouter les chiffres arabes, qui supplantent de plus en plus les kanji numériques pour la notation des nombres, et les caractères latins, qui ne notent pas à proprement parler le

[2] On fera une distinction entre les termes "caractères chinois" et "kanji". Ce dernier sera employé pour désigner les caractères chinois en tant que composante du système d'écriture japonais, avec, comme on le verra plus tard, toutes les spécificités que cela implique.

[3] Sur la ligne supérieure, les hiragana. Sur la ligne inférieure, les katakana. La première colonne donne les kana vocaliques. Pour les autres, lire: C + V = kana (exemple: k + a = か). Les kana んet ン se lisent <n>.

japonais mais peuvent apparaître dans des textes japonais pour noter les sigles et acronymes ainsi que les mots étrangers.

1.2 L'emploi des caractères

Ces différents types de caractères ont évolué au fil de l'adaptation des caractères chinois à la notation du japonais et ont maintenant des rôles distincts. Ils sont utilisés conjointement dans le style dit "kanji kana majiri bun" ("mélange de kanji et de kana").

1) Les kanji notent les éléments à caractère sémantique: principalement les substantifs et les radicaux verbaux et qualificatifs.
2) Les hiragana notent en particulier des éléments à caractère grammatical, qui existent en japonais, mais pas en chinois (particules, suffixes flexionnels).
3) Les katakana ont une sphère d'utilisation plus diversifiée. Ils notent surtout les mots d'origine étrangère (non chinoise ou coréenne), les onomatopées, mais également les termes spécialisés. Ils servent aussi à la mise en exergue (à la manière de l'italique).

Une autre particularité essentielle du japonais est qu'il s'écrit sans espace, rendant inopérante la notion de mot graphique. Par contre, depuis l'ère Meiji (1868-1912) et la réouverture à l'Occident, le système d'écriture japonais a intégré des caractères de ponctuation.

大関日馬富士が全日本力士選士権を２連覇！
2009.10.2 17:39

大相撲の第６８回全日本力士選士権は２日、東京・両国国技館で行われ、幕内トーナメントは大関日馬富士が２年連続３度目の優勝を果たし、賞金５０万円を獲得した。

日馬富士は準決勝で高見盛を突き出し、決勝では岩木山を一気に寄り切った。秋場所で４場所ぶり２４度目の優勝を果たした横綱朝青龍は２回戦で武州山に、横綱白鵬は１回戦で時天空に敗れた。十両トーナメントは白馬が制した。

http://www.sanspo.com/sports/print/091002/spf0910021740000-c.htm

Fig. 3: Exemple de texte japonais contemporain

2. L'ambiguïté des kanji

Comme on l'a déjà mentionné, avant l'introduction des kanji le japonais ne s'écrivait pas. Une fois dépassés le stade de l'écriture ornementale puis celui de l'écriture du chinois pour passer à celui de l'écriture du japonais, les caractères chinois ont donc naturellement servi à noter des unités linguistiques japonaises. Parallèlement, les Japonais ont continué à les utiliser pour noter l'important lexique d'origine chinoise importé avec les caractères[4]. Ils ont également continué à utiliser les lectures[5] d'origine chinoise de ces caractères. Ainsi, en devenant les outils de notation de la langue japonaise, les caractères chinois, que l'on peut dès lors appeler kanji, ont développé une ambiguïté qui leur est spécifique.

2.1 Pluralité des lectures sino-japonaises

Avec les caractères chinois, les Japonais ont donc importé un important lexique dit sino-japonais et aujourd'hui totalement assimilé. Moyennant quelques modifications phonologiques[6], ils ont également adopté les lectures qui avaient cours en Chine pour ces caractères. Ce sont les lectures *on* ou sino-japonaises.

L'emprunt de caractères chinois a duré une dizaine de siècles, pendant lesquels la Chine a été le théâtre de changements politiques. Les dynasties se sont succédé et les zones géographiques dominantes ont alterné, tout comme la langue des tenants du pouvoir, même si le système d'écriture restait le même. Ainsi, un caractère chinois emprunté à une époque et dans un contexte historique donnés avec une lecture pouvait être à nouveau introduit à une époque ultérieure avec une lecture, voire un sens différent. Il existe quatre systèmes de lecture sino-japonaise, qui correspondent aux principales vagues d'importation, certains étant plus vivaces que d'autres. Potentiellement un kanji peut donc avoir plusieurs lectures sino-japonaises.

2.2 Lecture sino-japonaise et lecture japonaise

Ce n'est que progressivement que les kanji en sont venus à noter la langue japonaise, amorçant la fin de l'état de diglossie[7] dans lequel se trouvait le pays jusqu'alors. Les Japonais ont d'abord lu en japonais les textes rédigés en chinois puis écrit directement en japonais.

---

4 On parlera de "lexique sino-japonais", essentiellement constitué de séquences de deux kanji, par opposition à "lexique japonais" pour le lexique autochtone.

5 La tradition fait usage du terme "lecture" pour dénommer la façon dont on oralise un caractère chinois.

6 Rappelons que le système phonologique japonais est plus réduit que le système phonologique chinois.

7 Dans un premier temps, la langue écrite était la langue chinoise.

Aux lectures sino-japonaises se sont ainsi ajoutées des lectures japonaises, les lectures *kun*. Les kanji notent donc aujourd'hui le lexique sino-japonais et le lexique autochtone.

Le choix du caractère chinois destiné à noter une unité lexicale japonaise s'est fait sur le critère sémantique.

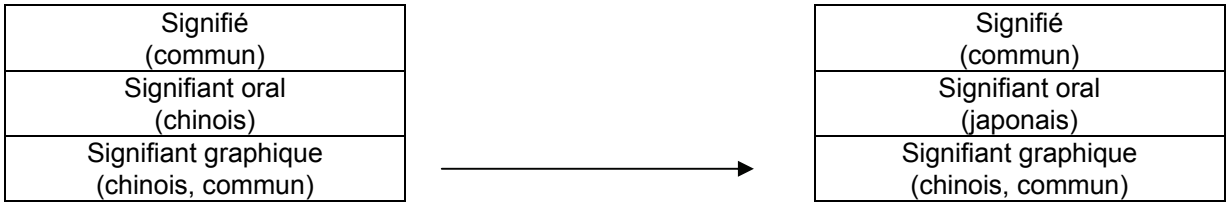

Fig. 4: Principe d'adoption d'un caractère chinois pour noter une unité lexicale japonaise

Dans l'exemple donné ci-dessous, le kanji 山 signifie /montagne/ et a la lecture japonaise *yama*. Il peut se lire également *san* (lecture sino-japonaise, dérivée de [shān]).

Fig. 5: Exemple pour le signifié /montagne/ et le caractère 山

Il s'agit d'un cas simple mais il est habituel de trouver des kanji qui ont plusieurs lectures japonaises. Au final, un kanji peut donc avoir plusieurs lectures sino-japonaises et plusieurs lectures japonaises. Les lectures sino-japonaises notent le lexique sino-japonais et les lectures japonaises notent le lexique japonais.

Lecture sino-japonaise/on: SHOKU (ex.: 食事 shokuji, *repas)*
Lectures japonaises/kun: **ta** (ex.: 食べる**ta**beru *manger*); **ku** (ex.: 食う**ku**u *bouffer*; **ku**rawasu 食らわす*battre*); **ha** (ex.: 食む**ha**mu *brouter)*

Fig. 6: Lectures et sens potentiels du kanji 食

En termes de nombre et type de lectures d'un kanji, une large combinatoire est possible. Certains en ont moins que d'autres et il y a bien sûr des cas, bien que rares, où un kanji n'a qu'une lecture. Ceci étant dit, il est souvent difficile de se prononcer de façon définitive sur le nombre de lectures d'un kanji: en effet, il dépend souvent de l'exhaustivité des dictionnaires et peut varier de l'un à l'autre. On reconnaît comme références communes les listes publiées par le Ministère de l'Education Nationale (§ 2.3, p. 120).

2.3 La présence des kanji

Les kanji constituent un ensemble potentiellement infini, et les recensements les plus exhaustifs en dénombrent près de 50.000. Afin de normaliser leur usage, le Ministère de l'Education Nationale publie des listes de kanji. La liste des kanji à enseigner pendant la scolarité (publiée en 1992) en compte 1006. La liste des kanji d'usage général (publiée en 1981) en compte 1945. Un adulte d'éducation moyenne est censé les maîtriser et c'est en fait un nombre minimal pour lire sans problème un texte japonais courant.

L'utilisation des kanji a eu tendance à décroître. Mais, sauf réforme radicale de la langue japonaise, les kanji sont indispensables à son écriture et à sa compréhension. On peut même ajouter que la facilitation et l'automatisation de l'accès aux kanji, dont certains furent parfois un temps abandonnés car trop complexes, que procurent les nouvelles technologies leur donnent un nouveau souffle.

2.4 Lectures, mais pas seulement

Quand on parle de la complexité et de l'ambiguïté des kanji, on aborde le plus souvent le problème par la multiplicité des lectures. C'est en effet la partie la plus visible, qui concerne l'oralisation. Mais, il convient de ne pas oublier ce qui est sous-jacent: particulièrement lorsqu'il s'agit du lexique autochtone (lectures japonaises) un kanji est aussi virtuellement la forme graphique de plusieurs unités lexicales. Sémantiquement, elles sont le plus souvent proches mais elles peuvent appartenir à des catégories lexicales différentes et donc se distinguer par leurs comportements morphologique et syntaxique.

1. Pour le kanji 煙: catégories lexicales différentes
   - 煙る <kemu(ru)> verbe
   - 煙い <kemu(i) > qualificatif
   - 煙 <kemuri> substantif
2. Pour le kanji食: catégories lexicales identiques mais traits différents
   - 食べる <ta(beru)> verbe, type shimo ichidan, colonne de flexion ba
   - 食う <ku(u)> verbe, type yodan, colonne de flexion wa
   - 食む <ha(mu)> verbe, type yodan, colonne de flexion ma

Fig. 7: Les informations qu'apporte la désambiguïsation d'un kanji[8]

Cette remarque conduit à une autre qui concerne elle aussi plutôt les lectures et le lexique japonais[9]. Dans les dictionnaires de kanji, on énumère les unités lexicales qu'un kanji peut noter. De même, les travaux qui recensent les lectures de chaque kanji prennent en compte le nombre d'unités lexicales

[8] La partie qui doit être écrite en kana est donnée entre parenthèses. Les transcriptions en caractères latins sont données entre chevrons.

[9] Les lectures sino-japonaises ne posent pas ce problème d'opposition entre lecture et unité lexicale. Il est par ailleurs impossible de dresser une liste réellement exhaustive des unités lexicales sino-japonaises dans lesquelles un kanji apparaît.

plutôt que le nombre de lectures stricto sensu. En effet, un kanji peut noter plusieurs unités lexicales pour lesquelles il va s'oraliser de la même façon, et avoir d'ailleurs un sens identique ou très proche.

Ainsi, le kanji 下 a les lectures sino-japonaises suivantes[10]: KA, GE. Il peut par ailleurs noter les dix unités lexicales japonaises suivantes: shita, shimo, moto, sa(geru), sa(garu), kuda(ru), kuda(su), kuda(saru), o(rosu), o(riru). Or, on se rend compte que pour certaines d'entre elles le kanji 下 s'oralise de la même manière:

sa (geru) (garu)	kuda (ru) (su) (saru)	o (rosu) (riru)

Fig. 8: Lectures japonaises et unités lexicales

C'est dans le sens strict d'oralisation que sera utilisé le terme lecture. En l'occurrence, le kanji 下 n'a en fait que sept lectures japonaises possibles: shita, shimo, moto, sa, kuda et o.

2.4.1 Les niveaux d'ambiguïté

Un kanji est donc potentiellement la représentation graphique de tout ou partie de plusieurs unités lexicales. Hors contexte, il est donc potentiellement ambigu à quasiment tous les niveaux: il est impossible de l'oraliser et de l'interpréter sémantiquement. Mais au-delà, il est impossible d'attribuer une catégorie lexicale à l'unité lexicale qu'il note et on ne dispose d'aucune information syntaxique ou morphologique la concernant.

En contrepartie, désambiguïser un kanji est beaucoup plus que simplement déterminer sa lecture ou même son sens. C'est identifier l'unité lexicale qu'il note et ainsi ses comportements morphologique et syntaxique (Fig. 7, p. 120).

Par ailleurs, la langue japonaise ne faisant pas usage de l'espace, la désambiguïsation, en permettant l'identification d'unités lexicales, est aussi le moyen de segmenter un énoncé japonais. On ne doit pas considérer ce dernier point comme anecdotique. En observant un texte écrit en japonais, on se convainc rapidement de son importance.

…食事を… → 食事#を
…食べた… → (食–べ)#た
…食む… → (食-む)#

Fig. 9: La segmentation par la désambiguïsation par le contexte

[10] On prend comme référence la liste des kanji usuels du Ministère de l'Education Nationale.

3. Une possibilité de désambiguïsation

On ne peut aborder la question de l'ambiguïté du système d'écriture japonais sans préciser que les langues chinoise et japonaise sont en bien des points dissemblables. Elles présentent des divergences sur les plans syntaxique (le chinois est de type SVO, le japonais de type SOV), morphologique (le chinois est à tendance isolante et monosyllabique, le japonais à tendance agglutinante et polysyllabique) et phonologique (le chinois a des tons, pas le japonais, et son système phonologique est beaucoup plus riche que celui de ses voisins insulaires).

Comme expliqué plus haut, les différents types de caractères qui constituent le système d'écriture japonais sont employés de manière spécifique. Ceci est le résultat de l'adaptation linguistiquement motivée, et rendue nécessaire par la différence entre les deux langues, des caractères chinois à la notation de la langue japonaise.

Le japonais est une langue agglutinante dont les verbes et les qualificatifs se fléchissent. Par ailleurs, elle fait usage de particules, qui par exemple sont placées après les substantifs et indiquent leur fonction dans une phrase. Ni ces flexions ni ces particules n'existent en chinois. Quand ils ont voulu écrire leur langue avec les caractères chinois, les Japonais ont donc dû trouver des solutions et ont créé les kana.

Les deux syllabaires, hiragana et katakana, sont apparus à la même époque, entre le VII^{ème} et le VIII^{ème} siècle, mais dans des contextes différents. Les katakana sont nés dans les milieux bouddhistes, de l'habitude d'annoter les textes chinois pour indiquer la lecture des caractères chinois, l'ordre syntaxique japonais ou encore ajouter les flexions et particules. Cette pratique s'est peu à peu normalisée et a abouti au système d'écriture japonais actuel (§ 1, p. 114), où ce sont les hiragana, et non plus les katakana, qui notent les flexions et autres particules.

Les raisons pour lesquelles les kana sont apparus ont ainsi généré une graphotaxe étroitement liée à la morphologie et à la syntaxe. Ce lien ouvre la voie à une stratégie de désambiguïsation des kanji par le contexte.

3.1 Les okurigana

Outre les listes de kanji, le Ministère de l'Education Nationale japonais publie des textes qui visent à normaliser l'utilisation des kana qui viennent en complément des kanji pour noter certaines unités lexicales. Il s'agit essentiellement des flexions verbales et qualificatives mais aussi de la fin d'unités lexicales telles que les adverbes. Ces textes indiquent tout simplement quelle partie de l'unité lexicale peut être écrite en kanji et quelle

autre doit être écrite en kana[11]. Dans cet emploi, ceux-ci sont appelés *okurigana* (kana accompagnants).

Les particules ne sont pas concernées par ces textes ministériels puisqu'elles sont des unités lexicales autonomes accolées à d'autres unités lexicales autonomes et non des okurigana.

3.1.1 La notion de contexte droit

Pour ne parler que d'elles, si les particules ne sont pas des okurigana, elles n'en sont pas moins indispensables à la désambiguïsation des kanji par le contexte. On élargit donc la liste des kana postposés aux kanji qui doivent être pris en compte pour obtenir la notion de contexte droit.

Un contexte droit est donc une séquence constituée de kana venant immédiatement après une séquence de kanji[12]. Il peut s'agir d'okurigana mais aussi de n'importe quelle autre séquence de kana. Dans certains cas, les contextes droits peuvent aussi contenir des caractères de ponctuation. Un contexte droit est finalement toute séquence de caractères non kanji venant immédiatement après un kanji.

3.2 La désambiguïsation par le contexte

La désambiguïsation par le contexte des kanji est en fait un procédé tout à fait courant, mis en œuvre de façon plus ou moins consciente par le lecteur humain et qui repose sur sa connaissance de la syntaxe, de la morphologie et de la graphotaxe du japonais.

La grammaire japonaise reconnaît onze parties du discours. Parmi celles-ci, ce sont les verbes et qualificatifs qui se prêtent le mieux à la désambiguïsation par contexte morphologique. Les substantifs se prêtent eux très bien à la désambiguïsation par contexte syntaxique.

3.2.1 Inventaire des contextes droits

L'algorithme de désambiguïsation repose dans un premier temps sur l'identification d'une articulation "kanji-non kanji" (c'est-à-dire kanji-contexte droit). Une fois cette articulation repérée, il faut identifier le contexte droit afin de désambiguïser le kanji. Le travail préparatoire à cette identification est l'inventaire des contextes droits pertinents. Un contexte droit est pertinent s'il peut être associé de manière univoque à un ensemble de traits morpho-syntaxiques et ainsi désambiguïser un kanji. Autrement dit, ce contexte droit ne se trouve après un kanji que dans un cas unique et identifié.

---

[11] Il convient de signaler qu'en théorie, tout ce qui peut se noter en kanji peut se noter en kana. Mais dans la pratique, l'usage exclusif des kana est quasiment inexistant.

[12] Dans le reste de l'article, on utilisera le terme kanji mais il est entendu qu'il peut s'agir de séquences de kanji.

3.2.3 Graphotaxe et morphologie: le cas des verbes

Les verbes japonais ne se fléchissent ni en genre ni en nombre. Ils sont classés en trois types morphologiques réguliers (yodan, kami ichidan, shimo ichidan) et deux types irréguliers (sahen, kahen) qui ne comptent qu'un verbe chacun. Cette classification a été établie à partir des différences de flexion.

Selon la grammaire traditionnelle japonaise, un verbe est constitué d'un radical et d'une désinence qui forment ce qu'on appelle une base. Il existe six bases verbales (mizen, ren'yô, shûshi, rentai, katei, meirei).

Certaines de ces bases sont autonomes et apparaissent telles quelles dans un énoncé. D'autres doivent, ou peuvent, être suivies de suffixes flexionnels ou de particules conjonctives. Les bases isomorphes se distinguent par leur fonction (déterminante: équivalent à une subordonnée relative, rentai; conclusive: fin de phrase, shûshi) ou les suffixes et particules qui les suivent.

Mizen (négative)	Non autonome	延び <no(bi)>
Ren'yô (connective)	Non autonome	延び <no(bi)>
Shûshi (conclusive)	Autonome	延びる <no(biru)>
Rentai (déterminante)	Autonome	延びる <no(biru)>
Katei (conditionnelle)	Non autonome	延びれ <no(bire)>
Meirei (impérative)	Autonome	延びよ/ろ <no(biyo/ro)>

Fig. 10: Paradigme pour le verbe kami ichidan 延びる<no(biru)>

Sur le plan graphotaxique, le radical se note en kanji et les désinences, suffixes flexionnels et particules en kana[13]. Dans le tableau qui suit (Fig. 11, p. 125), on voit que la forme verbale 延びる <no(biru)> est constituée d'un radical et d'une désinence. Le radical 延 <no> s'écrit en kanji et la désinence びる <biru> s'écrit en kana. On voit aussi que la forme verbale 延びた <no(bita)> est constituée d'un radical, d'une désinence et d'un suffixe flexionnel. Le radical 延 <no> s'écrit en kanji, la désinence び <bi> s'écrit en kana et le suffixe flexionnel た <ta> (suffixe du passé) s'écrit lui aussi en kana.

[13] Il est bien entendu que dans l'écrasante majorité des cas on utilise des hiragana. Néanmoins, quand la précision n'est pas absolument nécessaire, on parlera de kana.

Forme verbale		
Verbe (base verbale)		
Radical	Désinence	Suffixe ou particule
Kanji	Kana [kana]	Kana*
延<no>	びる <biru>	
延<no>	び <bi>	た <ta>

Fig. 11: Règles graphotaxiques pour les verbes

Une des désinences possibles pour un verbe kami ichidan est la séquence de kana *Ciru*[14]. Pour le verbe 延びる <no(biru)>, ce sera びる <biru>. En outre, cette séquence de kana n'apparaît que dans ce cas de figure. On peut donc affirmer qu'un kanji suivi de la séquence de kana びる <biru> note le radical d'un verbe kami ichidan, à colonne de flexion ba[15], à la base shûshi ou rentai. En ce qui concerne la segmentation qui en découle, on identifie une articulation entre le radical et la désinence ainsi qu'à la fin du verbe, juste après la désinence.

3.2.3 Graphotaxe et syntaxe: le cas des substantifs

Ce n'est pas le comportement morphologique des substantifs qui est intéressant pour la désambiguïsation par le contexte, mais leur comportement syntaxique. En effet, dans un énoncé, ils sont dans la très grande majorité des cas suivis d'au moins une particule. Ils s'écrivent essentiellement en kanji alors que les particules s'écrivent toujours en kana. Tout comme pour les verbes, on est donc en présence d'une configuration "kanji-non kanji" où les non kanji sont des contextes droits inventoriables et associables à des traits morpho-syntaxiques.

Pour prendre un exemple des plus simples, un kanji suivi du hiragana を <o> note un substantif complément d'objet direct et le hiragana lui-même note la particule de l'accusatif (Fig. 9, p. 121).

4. Une application au traitement automatique du japonais

Dans la mesure où la désambiguïsation des kanji concerne non seulement leurs lectures mais aussi un ensemble de traits morpho-syntaxiques, elle est exploitable de façon tout à fait efficace dans le cadre du traitement automatique du japonais et notamment de l'analyse morphologique automatique, comme le démontre le système d'analyse automatique des séquences de kanji auquel aboutissent nos travaux.

---

[14] *C* pour consonne.

[15] La colonne de flexion est déterminée par la consonne initiale de la désinence.

4.1 Un système d'analyse sans dictionnaire

A l'image de l'humain qui fait appel à sa connaissance de la langue pour désambiguïser un kanji, ce système fonctionne sans aucun dictionnaire ou base de données et n'a recours qu'à une liste de règles.

Cette approche sans dictionnaire permet de répondre au problème des mots inconnus (mots qui ne figurent pas dans un dictionnaire). Ce problème est de plus en plus important compte tenu de l'accroissement de la masse textuelle et de la diversité de ses origines. Il l'est encore davantage quand il s'agit des kanji, ensemble potentiellement infini. En effet, une règle efficace permet de traiter non pas une, mais un ensemble d'unités lexicales. En l'occurrence cela pourra être un ensemble de verbes (les verbes en びる <biru> ou même en *ℭiru*) ou encore l'ensemble des substantifs dans telle ou telle fonction (les substantifs COD, suivis de la particule を <o>).

La plupart des analyseurs automatiques du japonais ont recours à des méthodes statistiques d'apprentissage sur corpus annotés manuellement et à des dictionnaires. C'est le cas de Chasen, l'analyseur morphologique le plus largement utilisé, et de son prédécesseur Juman. Il n'est pas question de remettre en cause l'efficacité de ces analyseurs, ni même de la comparer à notre système. Néanmoins, dans la mesure où ils utilisent des dictionnaires, ils n'apportent pas de réponses au problème des mots inconnus. Notre approche, qui tente d'y répondre en s'appuyant sur une description intentionnelle et sur les informations extractibles de l'énoncé lui-même, est complémentaire à la leur.

4.1.1 Au-delà des types de caractères et de la segmentation

La segmentation par type de caractères est une des méthodes utilisée pour la segmentation d'énoncés japonais et qui pourrait pallier le fait que le japonais s'écrit sans espace et rend inopérante la notion de mot graphique. Mais, en se basant uniquement sur ce critère, on s'expose à de nombreuses erreurs. Pour reprendre le cas des verbes, la segmentation se fera à tort entre le radical, noté en kanji, et la désinence, notée en kana. Il est nécessaire de prendre en compte le contexte droit du kanji pour identifier un verbe et déterminer la segmentation correcte.

Par ailleurs, notre système va au-delà de la segmentation et opère un étiquetage morpho-syntaxique des unités identifiées. Cet étiquetage est impossible si on prend comme unique critère le type des caractères et que l'on ignore les contextes morphologiques et syntaxiques.

4.2 La formalisation: des contextes droits aux règles

Les règles sont en fait une formalisation des contextes droits et de leurs corrélats morpho-syntaxiques (Fig. 12, p. 127). La prémisse de ces règles correspond à la chaîne de caractères à analyser: le kanji et le contexte droit. Elle est constituée de symboles génériques ou terminaux. Les symboles

génériques représentent un ensemble de caractères (p.ex. le K représente tous les kanji) et les symboles terminaux sont des caractères. La première position de la prémisse représente le kanji qui précède immédiatement les kana. C'est en effet le seul pertinent pour la présente analyse.

La conclusion de ces règles rassemble les traits morpho-syntaxiques afférents à la prémisse.

4.2.1 Les traits morpho-syntaxiques

Pour les catégories lexicales variables, à savoir les verbes et les qualificatifs, les traits sont la catégorie lexicale (cl), le type (tv) et la base (bv), auxquels s'ajoute la position de la fin du verbe ou du qualificatif (f).

Pour les substantifs, il n'y a qu'un trait, la catégorie lexicale (cl) à laquelle s'ajoute la position de la fin du substantif (f).

1. Corrélation contexte droit/traits morpho-syntaxiques: verbe
   Un kanji suivi du contexte "びる" <biru> note toujours un radical verbal kami ichidan, colonne de flexion ba, base shûshi ou rentai. Le verbe se termine après la désinence (deux caractères après le radical).
   - Règle: K びる:{cl=vrb tv=kami bv=shuren f=2};
2. Corrélation contexte droit/traits morpho-syntaxiques: substantif
   Un kanji suivi du hiragana を <o> note toujours un substantif qui se termine juste avant la particule.
   - Règle: K を:{cl=subst f=0};

Fig. 12: Formalisation des contextes droits sous forme de règles

4.3 L'analyse morphologique

L'algorithme du système d'analyse morphologique reprend celui de la désambiguïsation par le contexte décrit précédemment. L'essentiel du système ne réside d'ailleurs pas dans cet algorithme mais dans les connaissances linguistiques qu'il met en œuvre et la façon dont elles sont formalisées et organisées.

L'algorithme d'analyse comporte quatre étapes principales: (1) le repérage des successions kanji-non kanji, (2) la comparaison du contexte droit et des prémisses des règles, (3) l'étiquetage de l'unité lexicale que note le kanji et (4) la segmentation.

4.3.1 L'organisation des données

Les règles sont collectées dans un fichier où elles n'ont pas à être ordonnées. Ceci en permet une maintenance facile par n'importe quel utilisateur. Néanmoins, quelques contraintes sont posées afin que deux règles ne soient jamais applicables au même cas: les règles doivent être soit disjointes soit en relation de cas particulier.

- Si on considère que les prémisses dénotent chacune un ensemble de chaînes de caractères telles qu'on peut les rencontrer dans un énoncé japonais, une règle R1 est disjointe d'une règle R2 si la prémisse P1 de la règle R1 est disjointe de la prémisse P2 de la règle R2. Une prémisse P1 et une prémisse P2 sont disjointes si les ensembles de chaînes que dénotent les prémisses P1 et P2 n'ont aucun élément en commun.
- Une règle R1 est le cas particulier d'une règle R2 si la prémisse P1 de la règle R1 est un cas particulier de la prémisse P2 de la règle R2. Une prémisse P1 est un cas particulier d'une prémisse P2 si l'ensemble de chaînes que dénote la prémisse P1 est inclus dans l'ensemble de chaînes que dénote la prémisse P2.

Pour que ces contraintes ne pèsent pas sur l'utilisateur amené à modifier ou ajouter des règles, une procédure de vérification se charge du contrôle. De même, une autre procédure s'assure que les règles sont syntaxiquement correctes.

4.3.2 Extraction et calcul des traits et informations complémentaires

En cas d'identité entre un contexte droit et la prémisse d'une règle, le kanji en cours d'analyse est étiqueté avec les traits morpho-syntaxiques de la conclusion de cette règle. Certains traits ne figurent pas dans la conclusion parce qu'ils sont calculables à partir d'autres traits, comme l'indique le tableau ci-dessous.

Extraits de la conclusion		*Calculés à partir de la conclusion*
Catégorie lexicale	→	Frontière radical-désinence de la forme de surface
Type de flexion	→	Colonne de flexion
Base	→	Forme lemmatisée
Segmentation à droite	→	Frontière radical-désinence de la forme lemmatisée

Fig. 13: Calcul des traits

4.3.3 Le traitement des lectures

On a pu constater que, ni parmi les traits fournis par la conclusion ni parmi les traits calculés, on ne trouve de lecture de kanji alors que c'est souvent le prisme à travers lequel on considère le problème de l'ambiguïté des kanji. La raison en est simplement qu'il n'y a aucun moyen de systématiser et de formaliser la relation entre kanji et lecture. Autrement dit, s'essayer à déduire la lecture d'un kanji de sa forme est au mieux hasardeux lorsqu'il s'agit des lectures sino-japonaises et inutile lorsqu'il s'agit des lectures japonaises. Par ailleurs, on ne peut attribuer une lecture à un ensemble de kanji à partir de traits morpho-syntaxiques communs. Par exemple, pour reprendre le cas de 延びる <no(biru)>, on ne peut pas affirmer que tous les kanji qui notent un radical verbal kami ichidan à colonne de flexion ba se lisent *no*. De même qu'il

faut apprendre par cœur les lectures de chaque kanji, un traitement automatique incluant ces lectures ne peut se faire qu'au cas par cas.

Ceci étant dit, si on connaît les lectures d'un kanji, définir les traits morpho-syntaxiques d'une instance de ce kanji permet presque toujours de choisir celle qui est correcte. Ainsi, si on a déterminé que le kanji 延 note un radical verbal kami ichidan à colonne de flexion ba, alors on sait qu'il se lit et ne peut se lire que *no*.

Malgré ce dernier argument, on pourrait objecter de la validité d'une désambiguïsation des kanji qui ne traite pas les lectures. Mais tout apprenant du japonais, même locuteur natif, sait qu'il n'est pas nécessaire de savoir en oraliser les kanji pour analyser et comprendre un énoncé japonais. On peut même rédiger un texte sans connaître l'oralisation des kanji employés. Les niveaux de connaissance des kanji (passifs: lecture et compréhension; actifs: écriture et oralisation) ne sont pas interdépendants.

En termes de traitement automatique, notre système a démontré que même s'il ne traite pas les lectures, il obtient des résultats largement satisfaisants et utilisables en l'état dans des cadres plus larges que celui de l'analyse morphologique. La segmentation d'un énoncé japonais est totalement indépendante de son oralisation. Tant qu'elle ne concerne ni la phonétique ni la phonologie, l'analyse linguistique l'est aussi.

4.4 Résultat de l'analyse automatique

Notre système d'analyse automatique prend en entrée un texte écrit en japonais[16]. Il produit un étiquetage des séquences de kanji et une segmentation partielle de ce texte.

- Entrée
  江戸時代から現在まで多くの学者たちが考えた。
- Sortie

1-4	江戸時代	subst				
7-8	現在	subst				
12-13	多く	adverbe				
15-18	学者たち	subst				
20-21	考えた	verbe	shimo	agyo	renyo	考-える

Fig. 14: Exemple d'analyse automatique par le contexte

[16] Le corpus de base utilisé pour les tests dont les résultats sont donnés est constitué de textes employés pour l'enseignement du japonais langue étrangère à l'université et qui présentent un large éventail de phénomènes linguistiques. D'autres tests ont été effectués sur des textes de presse.

4.4.1 Les ambiguïtés non levées

Notre système d'analyse présente un taux d'erreur de 1% et un taux de silence de 5%. Le taux de séquences de kanji qui restent ambiguës est de 18%. Il est dû au fait que ce système est non déterministe et préfère un étiquetage multiple à un étiquetage unique présentant un risque d'erreur.

Par ailleurs, quand l'ambiguïté n'est pas levée, ce n'est que partiellement. L'étiquetage, bien que multiple, fournit un certain nombre d'informations exactes.

71-73	意味−し	verbe	kami	sagyo	renyo	意味−しる
" "	"	"	yodan	"	"	意味−す
" "	"	"	sahen	"	"	意味−する

Fig. 15: Un cas d'ambiguïté partielle

Dans l'exemple présenté ci-dessus, le résultat de l'analyse d'une forme verbale, seuls le type du verbe et la lemmatisation restent ambiguës. La segmentation, la catégorie lexicale, la colonne de flexion et la base sont univoques. D'autre part, une grande proportion des cas d'ambiguïté non levée concerne l'opposition substantif-qualificatif nominal, deux catégories lexicales indiscernables sur le plan morphologique. Or, il s'agit là d'un débat qui a divisé les linguistes les plus éminents.

En tout état de cause, la stratégie de désambiguïsation reste efficace et les résultats obtenus peuvent être exploités en l'état pour des tâches ultérieures.

5. Conclusion

Les conditions dans lesquelles est né le système d'écriture japonais en ont fait un des plus complexes qui soient. Les kanji recèlent une large part de cette complexité et ils sont un exemple particulièrement intéressant d'ambiguïté. D'une part, on retrouve cette ambiguïté à tous les niveaux de l'analyse linguistique. Elle rend difficile jusqu'à l'oralisation d'un texte. D'autre part, ce qui a créé cette ambiguïté (l'adoption pour noter une langue d'un système d'écriture inadéquat) a aussi créé les conditions permettant de la lever. En effet, confrontés à l'obligation d'adapter le système d'écriture chinois à leur langue, les Japonais ont créé un système d'écriture propre dont la graphotaxe est étroitement liée à la morphologie et à la syntaxe.

Exploitée dans le cadre du traitement automatique du japonais, cette stratégie de "désambiguïsation par le contexte", basée sur des phénomènes de surface récurrents, s'avère particulièrement intéressante, notamment pour une analyse automatique sans dictionnaire.

Bibliographie

Abe, S. (1989): 常用漢字の送り仮名 *Jôyôkanjihyô no okurigana* (Les okurigana des kanji d'usage général), Tôkyô (Meiji shoin).

Alleton, V. (1967): L'écriture chinoise. Paris (PUF).

Coulmas, F. (1988): Overcoming diglossia: the rapprochement of written and spoken Japanese in the 19th century. In: Pour une théorie de la langue écrite, 191-201. Paris (CNRS).

Coulmas, F. (2003): Writting systems: an introduction to their linguistic analysis. Cambridge (Cambridge University Press).

Coyaud, M. (1985): L'ambiguïté en japonais écrit. Paris (PAF).

Coyaud, M. (1988): La pertinence en graphémique. In: Pour une théorie de la langue écrite, 157-163. Paris (CNRS).

Coyaud, M. (1989): Grammaire du japonais standard. Paris (PAF).

Griollet, P. (1985): La modernisation du Japon et la réforme de son écriture. Paris (POF).

Griollet, P. (1994): L'orthographe du japonais et les "études nationales". In: Cipango, 3, 7-36. Paris.

Hadamitzky, W. & Durmous, P. (1984): Kanji to kana, manuel et dictionnaire de l'écriture japonaise. Berlin (Ostasien-Verlag).

Haguenauer, C. (1951): Morphologie du japonais moderne. Paris (Klincksieck).

Hashimoto, S. (1969): 助詞· 助動詞の研究 *Joshi-jodoshi no kenkyû* (Etude sur les particules et les auxiliaires). Tôkyô (Iwanami).

Kabashima, T. (1979): 日本の文字 *Nihon no moji* (L'écriture japonaise). Tôkyô (Iwanami shinsho).

Kindaichi, H. (1957): 日本語 *Nihongo* (La langue japonaise). Tôkyô (Iwanami shinsho).

Kitahara, Y. (1981): 日本語助動詞の研究 *Nihongo jodoshi no kenkyû* (Etude sur les auxiliaires japonais). Tôkyô (Taishukan).

Kurohashi, S. & Nagao, M. (1998): Japanese morphological analysis system JUMAN, Department of Informatics (Kyôto University, en japonais).

Matsumoto Y. *et al.* (2002): Morphological analysis system Chasen version 2.2.9 manual, Technical report (NAIST).

Miller, R. A. (1967): The Japanese language. Chicago (University Press of Chicago).

Ministère de l'Education Nationale (1981): 常用漢字表 *Jôyôkanjihyô* (Liste des kanji d'usage général), Tôkyô (Ôkurashô Insatsukyoku).

Ministère de l'Education Nationale (1981): 送り仮名のつけ方 *Okurigana no tsukekata* (Guide d'utilisation des okurigana). Tôkyô (Ôkurashô Insatsukyoku).

Nelson, A.N. (1974): The modern reader's Japanese-English dictionary. Tôkyô (Tuttle).

Ôno, S. & Shida, T. (1977): 文字 *Moji* (Ecriture). Tôkyô (Iwanami shoten).

Ozaki, Y. *et al.* (1991): 大字源 *Daijigen* (Dictionnaire de kanji). Tôkyô (Kadokawa shoten).

Rayon, N. (2005): Analyse morpho-graphémique pour la catégorisation automatique des séquences de kanji dans des textes japonais. In: The Prague Bulletin of Mathematical Linguistics, 83, 47-58. Prague.

Rose-Innes, A. (1943): Beginner's dictionary of Chinese-Japanese characters. Tôkyô (Maisonneuve).

Seeley, C. (1991): A history of writing in Japan. Leiden (Brill).

Shimamori, R. (1991): Des particules japonaises. Tôkyô (Taishukan).

Shimamori, R. (1994): Grammaire japonaise systématique. Paris (Maisonneuve).

Shirota, S. (1998): 日本語形態論 *Nihongo keitairon* (Morphologie du japonais). Tôkyô (Hitsuji Shobô).

Tamaoka, K., Kirsner K., Yanase, Y., Miyaoka, Y. & Kawakami, M. (2002): A Web-accessible database of characteristics of the 1.945 basic Japanese kanji. In: Behavior Research Methods, Instruments, & Computers, 34 (2), 260-275.

Tamaoka, K., Kirsner, K., Yanase, Y., Miyaoka, Y. & Kawakami, M. (2000): Database for the 1.945 basic Japanese kanji.

Tokieda, M. (1952): 日本文法 *Nihon bunpô* (Grammaire japonaise). Tôkyô (Iwanami zensho).