TRAVAUX DE LINGUISTIQUE JAPONAISE (1997): *Langue - Ordinateur - Mentalité* (en collaboration avec ISHIWATA Toshio - Université d'Ibaraki, Japon), Université Diderot (Paris 7), vol . X, Édition L'Asiathèque, Paris

LE JAPONAIS ET L'ORDINATEUR
- Présentation générale -

Toshio ISHIWATA* et André WLODARCZYK**

Introduction

Le japonais devient une langue internationale malgré les affirmations de certains linguistes selon qui cela ne saurait se produire pour des raisons structurelles. En effet, le japonais présente une structure très différente des langues occidentales d'autant plus que le chinois, langue appartenant à une famille linguistique encore différente, a exercé une très forte influence non seulement sur le système d'écriture du japonais, mais également sur sa grammaire et son lexique. C'est sans doute en raison de ce voisinage que le japonais est devenu une langue complexe dont l'apprentissage présente bien des difficultés pour les étrangers. D'autre part, le Japon possède sa propre et longue tradition philosophique concernant l'écriture et la grammaire, ce qui rend parfois assez difficile la compréhension des concepts que les linguistes et les informaticiens japonais manipulent aujourd'hui.

Il est donc clair que le traitement du japonais par ordinateur rencontre un bon nombre d'obstacles qui sont de nature différente de ceux rencontrés dans les recherches sur les langues occidentales. Par exemple, parmi les problèmes qui semblent passionner la plupart des spécialistes occidentaux (y compris français) en TAL intéressés par la langue japonaise, on trouve la question du traitement des caractères (kanji). Cela n'est pas seulement en raison de leur grand nombre mais aussi à cause de leur structure interne (qui est pseudo-sémantique) ainsi qu'à cause de leur articulation plutôt complexe sur les unités linguistiques du japonais. Pour ceux qui s'y intéressent tout particulièrement, nous avons prévu quelques renseignements complémentaires relatifs aussi bien à l'histoire qu'à la structure interne des caractères kanji.

La *linguistique informatique* au Japon s'est donc constituée sur la base du traitement de l'information en langue japonaise. Nous aborderons les problèmes de cette linguistique tout en nous appuyant sur ceux du traitement automatique du japonais et de la traduction automatique. Il va sans dire que les recherches menées dans ces deux grands domaines ont posé un nombre considérable de problèmes aux Japonais (le système d'écriture, le lexique et la grammaire ainsi que la reconnaissance des caractères et le traitement de la parole). Nous nous bornerons ici aux deux premières questions.

---

* Université d'Ibaraki, professeur émérite du Département de Linguistique (Japon).

** Université Stendhal - Grenoble 3, professeur au Département d'Études Orientales, 38400 St MARTIN D'HERES, e-mail: Andre.Wlodarczyk@Stendhal.grenet.fr

La Langue japonaise et son système d'écriture

Contrairement à ce qui a été dit plus haut à propos de la langue japonaise en général, sa composition phonologique est très simple. Le japonais moderne possède:

5 phonèmes vocaliques : a, e, o, i , u.
13 phonèmes consonantiques: (p, b, m), (t, d n), (s, z), (k, g, ng), r et h.
2 phonèmes-mores: N, Q.
2 semi-voyelles: Bilabiale (w) et palatale (j).

Deux phonèmes symbolisés par N et Q possèdent la qualité de more (unité phonique à structure consonantique mais au moins deux fois plus longue que l'unité phonique à structure vocalique), ce qui fait que la syllabe - dont la structure est également simple, soit ouverte (CV) soit contenant des géminées moriques entre deux syllabes (CV[N|Q]CV) - n'est pas la seule unité rythmique en japonais.

Pour noter le japonais, on utilise les kanji (caractères idéo-phonographiques d'origine chinoise et les kana (caractères japonais syllabographiques dérivés des kanji). En japonais moderne, les kanji sont très souvent utilisés pour noter les mots indépendants (concepts) et les kana - pour noter les mots adjoints (mots fonctionnels). La spécificité des kanji (y compris par rapport à la langue japonaise) réside dans les points suivants:

1. un grand nombre de caractères différents (il existe plusieurs milliers de kanji fréquents - environ 6.000)
2. une complexité de la forme graphique (un grand nombre de traits pour tracer un caractère)
3. une grande diversité de lectures (prononciations) d'un seul kanji en moyenne >2:
   (a) emprunts chinois (lectures sino-japonaises) et
   (b) les équivalents japonais (lectures japonaises)
4. un grand nombre de lectures japonaises équivalentes, les emprunts chinois présentant peu de variétés mais beaucoup d'homophonies.

En ce qui concerne le point 3, le grand nombre d'usages fait qu'il est difficile de poser des règles. Il ressort du point 4 que le lexique japonais emprunté au chinois se caractérise par un grand nombre de mots homophones. Parmi les mots notés par kanji, on observe un grand nombre de concepts importants. L'homophonie pose donc un problème fondamental à la notation de la langue japonaise. Aux 2.471 caractères retenus par la société CANON pour l'un des ses premiers dictionnaires électroniques de caractères commercialisé en 1982 (CA-2000) correspondent au total 5.342 lectures différentes, ce qui signifie qu'en moyenne un caractère peut être lu, en japonais, de plus de deux façons. Pour cette raison, pour être efficace, un système de traitement de texte doit s'appuyer, dans cette langue, sur la sélection automatique des caractères de manière non ambiguë, ce qui présuppose la connaissance de multiples règles linguistiques. Selon les sources[1] de l'Institut National de la Langue Japonaise, sur 1945 caractères d'usage courant (jôyô-kanji), 737 (38%) ont une

[1]) D'après "Zusetsu Nihongo - Gurafu de miru kotoba no sugata" (Le Japonais en schémas - la forme d'une langue vue en histogrammes), Éditions Kadogawa, 1982, Tokyo.

lecture uniquement sino-japonaise (on), 40 (2%) ont une lecture uniquement japonaise (kun), 1.168 (60%) ont les deux lectures.

L'Encodage et le traitement des kanji

Il existe plusieurs méthodes de classification des kanji. Deux d'entre elles sont dignes d'intérêt: la *méthode traditionnelle* (utilisée dans les dictionnaires japonais) et la *méthode des quatre coins* (la moins ambiguë). Cependant, c'est en raison de l'impossibilité de classer les kanji de manière entièrement rationnelle que les Japonais utilisent aujourd'hui au moins 3 codages différents car malgré les nombreuses tentatives de trouver une notation symbolique des kanji pour la reconnaissance optique, le seul moyen d'éviter l'ambiguïté reste toujours l'encodage chiffré. A notre connaissance, il existe deux sources d'information très exhaustives en langues occidentales: en France, le rapport annexe à «Le Logiciel au Japon» par Michel Mariani[2] et aux États-Unis, l'ouvrage de Ken Lunde[3] intitulé «Understanding the Japanese Language Information Processing». Ces documents abordent avant tout les machines japonaises (NEC) et américaines (IBM-PC et, pour Ken Lunde, Apple Macintosh). Les stations de travail Sun[4] peuvent également être dotées d'un environnement japonais.

Étant donné leur grand nombre, les caractères kanji sont codés sur deux octets, dans le code hexadécimal Shift-JIS, le premier octet est considéré comme *octet haut* (Oh), le second comme *octet bas* (Ob). Le code d'un caractère kanji s'obtient donc en appliquant la formule suivante:

```
CodeKanji = Hex(Code(Oh) / 16) ^ Hex(Code(Oh) \ 16) ; 1er octet
          ^ Hex(Code(Ob) / 16) ^ Hex(Code(Ob) \ 16) ; 2e octet
```

Par exemple:

Le caractère 初 (*sho* ou *hajime* = début), par exemple, est représenté par les octets (èâ), c-à-d: 143 et 137.

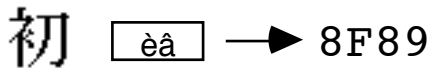
```
CodeKanji = Hex(143  / 16) ^ Hex(143  mod 16)
          ^ Hex(137  / 16) ^ Hex(137 mod 16)
```

---

2) Michel Mariani a rédigé ce rapport après une étude effectuée pendant l'été 1988 au Japon. Les informations relatives à l'encodage et au traitement des kanji sont réunies avant tout dans la partie intitulée "Introduction au traitement informatique de la langue japonaise" (46 pages) , N° 1004/NN, 11 octobre 1988, Service Scientifique de l'Ambassade de France à Tôkyô, 4-11-44, Minami-Azabu, Minato-ku, Tokyo.

3) Ken Lunde est un spécialiste des caractères de la Société Adobe. Son ouvrage de 435 pages constitue à l'heure actuelle la plus complète présentation du problème des kanji et de leur traitement (y compris les algorithmes écrits en C et très bien documentés) et a été publié par O'Reilly & Associates, Inc., 103 Morris Street, Suite A, Sebastopol, CA 95472, USA, son prix est d'environ 30 US $. Il est également possible de le commander par courrier électronique à order@ora.com. Cet ouvrage contient également un grand nombre d'adresses électroniques (en ftp) et commerciales aussi bien des sites du réseau Internet que des éditeurs de logiciels relatifs au traitement des kanji. Mentionnons enfin qu'avant d'être publié sous sa forme définitive, le livre de K. Lunde a été disponible en ftp dans plusieurs sites sous le titre "Electronic Handling of Japanese Text" (20 mars 1992) avec l'adresse de l'auteur : lunde@adobe.com.

4) Cf. (a) "Japanese Language Environment", SunTech Journal (The Independent Journal of Sun and Sparc Systems), vol. 4, N° 1, janvier 1991 ainsi que (b) SunWorld (The Independent Journal of Sun and Sparc Systems), vol. 4, N° 7, juillet 1991.



Le problème qui apparaît au premier chef concerne bien sûr la saisie. Au début, on utilisait des "imprimantes pour kanji à distance" (kanji tele-printer[5]). Elles permettaient de combiner les touches de caractères avec les touches de mode ("shifts") à 4 ou 12 degrés. Les années 1980 ont vu apparaître la méthode de saisie par conversion automatique des syllabaires kana ou des lettres de l'alphabet latin en kanji. Cette méthode s'est répandue aussi bien dans les machines dédiées que les logiciels de traitements de texte (*wapuro* - abréviation japonaise de *word processor*). De nombreux mécanismes ont été développés au cours des recherches visant à rendre la conversion en kanji de plus en plus efficace et rationnelle. C'est là l'une des causes du développement de la linguistique informatique au Japon. Par exemple, en ce qui concerne la conversion des kana en kanji, vu leur homophonie importante en japonais, on affichait une sélection à l'écran et on laissait l'utilisateur choisir mais il s'est révélé plus pratique de tenir compte de la fréquence d'emploi et de la sélectivité contextuelle plutôt que de proposer même rapidement plusieurs caractères kanji. Aujourd'hui, il est tout à fait normal de prendre en compte les syntagmes entiers lors de la conversion des kana en kanji. Les syntagmes se composent d'un ou plusieurs terme(s) indépendant(s) et d'un ou plusieurs terme(s) adjoint(s). Il va sans dire que les "syntagmes liés" (ren-bunsetsu) comportent au moins deux syntagmes simples. Grâce au traitement des syntagmes tout entiers englobant également les expressions figées, on parvient à laisser à l'utilisateur la possibilité de voir apparaître sur son écran le(s) caractères kanji qu'il recherche. Etant donné que pour distinguer entre un terme indépendant et un terme adjoint, on doit poser un ensemble de règles morphologiques, il est possible de réduire le nombre d'affichages incorrects à l'écran. Étant donné le grand nombre de caractères et la complexité de leur emploi, le mode de leur représentation et le système d'exploitation présentent des difficultés de grande échelle. Les logiciels de traitements de texte deviennent des enjeux très importants. Ainsi, les dictionnaires qui y sont destinés deviennent de plus en plus précis quant à leur composition, combinabilité et sens.

Les règles d'orthographe japonaises ne prévoient pas d'espaces entre les mots. Pour cette raison, fréquents sont les cas où, au moment de la saisie en syllabaire kana ou en alphabet latin, l'ambiguïté vient de la longueur des mots. Par exemple: en saisissant "teianshitaiken", il y a les trois possibilités de conversion suivantes:

1 - "teian-shitai ken" (l'affaire que l'on souhaite proposer)
proposition - *désidératif* - affaire
2 - "teian-shita iken" (l'opinion que l'on a proposée)
proposition - *passé* - opinion
3 - "teian-shi taiken" (en proposant..., l'expérience éprouvée par soi-même...)
proposition - *conjonction* - expérience

[5]) C'était leur nom commercial dans les années 1970.

Il va s'en dire que c'est l'utilisateur qui doit choisir le découpage souhaité. C'est donc à travers le lexique et la syntaxe que la spécificité du japonais se manifeste dans la saisie des caractères kanji.

TAO: grammaire et lexique

La grande différence de structure qui caractérise le japonais par rapport aux langues européennes est également la cause des nombreuses difficultés de la traduction automatique. Prenons par exemple le fait qu'en japonais il n'y a pas de distinction entre le singulier et le pluriel. Ce fait rend difficile la traduction automatique du japonais en langues européennes. Ou prenons encore comme exemple le fait que le japonais diffère considérablement des langues européennes quant à la mise en œuvre des maximes pragmatiques. Bien que ce problème soit beaucoup étudié actuellement de façon expérimentale, aucun des systèmes commercialisé n'en tient encore compte. On s'arrête au traitement de la syntaxe et de la sémantique.

L'analyse syntaxique et la génération des énoncés constituent le problème central du traitement du langage naturel au sein de la traduction automatique ou de l'intelligence artificielle. (En linguistique, on s'occupe surtout des dictionnaires, de la grammaire et plus rarement du sens).

Les recherches en traduction automatique ont commencé au Japon dans les années 60. Parmi les premiers instituts, on trouve l'Université de Kyûshû, l'Université de Kyôto, les Laboratoires d'Électricité (plus tard, ETL). Au début, on étudiait surtout les algorithmes pour la traduction automatique en même temps que la structure des langages de saisie et de sortie des textes. Le tout premier système de TAO au Japon s'appelait YAMATO. Ce système, réalisé en 1958, possédait 600 transistors et 7000 diodes ainsi qu'une mémoire de 100 Ko de DRAM, ce qui était énorme à l'époque. Plusieurs programmes de traduction de l'anglais vers le japonais ont été implémentés sur Yamato entre 1959-1961. Entre 1964-68, trois universités (Tôkyô, Kyôto et Kyûshû) se sont lancées dans la recherche des systèmes à traduire en appliquant des technologies entièrement repensées.

Ces premières expériences de laboratoire étaient de petite envergure mais bientôt les développements de grande taille ont suivi. Dans les années 80, l'Université de Kyûshû a compilé un grand dictionnaire et l'Université de Kyôto et ETL ont développé un système de traduction automatique (connu dans sa première phase sous le nom de système Mu) qui utilisait la méthode de transfert. Le système de traduction était conçu comme bilingue anglais-japonais et japonais-anglais et devait être appliqué aux résumés des articles scientifiques. Chaque direction (anglais-japonais et japonais-anglais) s'appuyait sur un ensemble de 3000 règles pour ce qui est de la grammaire et sur un lexique de 80000 entrées en ce qui concerne le dictionnaire. La vitesse de traduction atteignait 4000 mots à l'heure. La description grammaticale était réalisée en langage GRADE (GRAmmar DEscriber) spécialement conçu à cet effet. L'algorithme comportait également quelques heuristiques. Les dictionnaires prévus étaient généraux tant pour l'analyse que pour la génération.

A côté du système Mu qui a été développé dans le cadre du projet de traduction automatique national japonais dont le centre se trouve à l'Université de Kyôto, il existe à l'heure actuelle une série de grands systèmes du secteur privé dont certains ont déjà atteint la phase de commercialisation.

D'une manière générale, la plupart des laboratoires[6] japonais semblent utiliser la méthode de traduction dite "méthode par transfert" (cf. Tableau : Réalisations en TAO) car, comme l'a rappelé Makoto NAGAO, "*les langues contiennent plus d'exceptions que de cas qui pourraient être traités avec des règles clairement définies, d'une part, et présentent beaucoup de domaines qu'on ne peut pas expliciter logiquement, de l'autre*". C'est sans doute la raison pour laquelle on observe qu'au Japon, dans les recherches récentes relatives à la TAO, on ne se limite plus aux règles syntaxiques mais on introduit le mode d'exécution des transformations fondées sur le modèle dit "méthode par exemples". Il s'agit du mode de génération des énoncés les plus réussis tout en utilisant un grand nombre d'exemples de traduction et des dictionnaires de synonymes.

En général, ce sont les méthodes dites "par transfert" ("transfert des arborescences" et "transfert des structures casuelles") que les centres de recherches japonais utilisent dans leurs projets visant la construction des systèmes de traduction automatique :

(1) transfert des arborescences - la représentation est l'arbre d'analyse syntaxique - ATLAS/I, KATE et ATHENE
(2) transfert des structures casuelles - la représentation est en format des structures casuelles - Science & Technology Agency System (Kagi-chô shisutemu) - ATLAS/II, VENUS, TRAP, LUTE and TAURAS.

Nous nous bornerons ici à présenter les grandes lignes d'un projet gouvernemantal (représentatif) qui utilise la méthode dite "par transfert". «Mu-Project» est l'un des projets concernant les recherches visant la conception des ordinateurs capables de comprendre le langage humain qui ont démarré en 1983, date à laquelle le gouvernement japonais avait annoncé l'intention de soutenir ce genre de recherches. Il s'agit donc du premier projet gouvernemental (Secrétariat d'État à la Science et Technologie) de cette envergure dans le domaine de l'intelligence artificielle au Japon. Il est connu sous le nom de "Projet de la cinquième génération d'ordinateurs" (Dai-go-sedai Computer Project). «Mu-Project» (Université de Kyôto et al.), tout comme «System Q» (Université de Montréal), «Ariane» (GETA - Université de Grenoble), «Eurotra» (Union Européenne), restera donc un des projets de grande envergure des années 1980. En raison de l'introduction du sens au cours de l'application des règles syntaxiques, on considérera ce genre de projets comme appartenant à la 2e génération des systèmes de TAO.

Les objectifs du projet consistaient à créer: (1) une *Base de données* (dictionnaire de terminologie scientifique, (2) un *Logiciel* : dispositif de conversion des structures syntaxiques, (3) un *Système de Traduction* combinant les deux. Seuls l'anglais et le japonais étaient concernés. En pratique, pour des raisons de commodité (familiarité des

---

6) L'un des rares examples de la stratégie de la traduction indirecte utilisant un langage intermédiaire ou interlangue est «CONTRAST» (CONtext TRAnSlaTor) - ETL à Tsukuba (Dictionnaires: (1) Unités lexicales -> concepts et (2) Concepts -> unités lexicales).

développeurs du système), le domaine d'application a été réduit aux questions relatives à la technologie électrotechnique.

Les tâches ont été réparties entre les quatre organismes participants suivants:

- JICST (Japan Information Center for Science and Technology - Nippon kagaku gijutsu jôhô sentâ): compilation des dictionnaires des mots invariables
- ETL (Electrotechnical Technology Laboratory - Denshi gijutsu sôgo-kenkyû-sho): morphologie (en analyse et en synthèse) - dictionnaire des mots variables
- RIPS (Kôgyô gijutsu-in Tsukuba jôhô keisan sentâ) - E/S des textes
- Université de Kyôto (Kyôto daigaku): système de traduction (analyse - transfert - synthèse)

Les stratégies des traitements ont été définies de la manière suivante (cité d'après le rapport[7] relatif au Projet Mu) :

(1) au lieu d'assister le traitement en cours, on doit traiter les mots et les structures inconnus du système même si les résultats de ce traitement devaient s'avérer imparfaits

(2) tous les éléments découverts au cours des recherches doivent être intégrés dans le système

(3) étant donné que la définition des règles concernant l'analyse, le transfert et la synthèse nécessite la collaboration intensive de nombreux chercheurs, on a introduit le concept de Grammaire Partielle (bubun-bunpô).

(4) afin d'aboutir, à l'avenir, à une traduction multi-lingue, on a pris soin de bien séparer les trois étapes du traitement (analyse - transfert - synthèse) les unes des autres.

(5) vu l'impossibilité d'expliquer le langage au moyen d'un simple système de règles grammaticales, le traitement des exceptions est nécessaire; ainsi, la stratégie de description des propriétés grammaticales de chaque mot a été adoptée.

(6) au cours de la saisie des informations lexicologiques, on doit tenir compte non seulement des besoins actuels mais aussi de ce qui pourra devenir nécessaire à l'avenir; d'où le compromis entre la nécessité de décrire les nombreuses propriétés grammaticales de chaque mot et celle qu'impose le traitement automatique des données.

(7) d'une manière générale, c'est la grammaire des dépendances casuelles qui est adoptée pour la phase de l'analyse et ceci n'est pas uniquement à cause du japonais; on s'accorde effectivement sur le fait que ce type de grammaire rend plus facile la détermination des sémantismes nécessaire au cours de la reconnaissance des structures casuelles.

Dans «Mu-Project», la tâche du logiciel est de convertir des structures arborescentes en d'autres structures arborescentes. Les noeuds servent de lieu de stockage des informations concernant les accords morphologiques (sur le genre, le nombre, le cas etc.), les

---

[7]) NAGAO Makoto, TSUJII Jun'ichi, NAKAMURA Jun'ichi SAKAMOTO Masayuki, TORIUMI Tsuyoshi et SATO Masayuki : "Kagaku-Gijutsu-chô Kikai-hon'yaku Purojiekuto no Gaiyô" (Esquisse du Projet de Traduction Automatique de l'Agence pour la Science et la Technologie), in «Jôhô shyori» (Traitement de l'information), Vol. 26, N° 10, Octobre 1985, Tokyo.

composantes sémantiques etc. C'est LISP qui a été choisi comme langage de programmation (cette fois, ce qui comptait plus encore que la vitesse d'exécution, c'était celle de la rédaction du logiciel).

Le langage de description grammaticale GRADE (Grammar Describer) est, lui-même, écrit en LISP (version UTILISP[8]). Le texte de ce langage contient approximativement 10.000 lignes. Il a été développé sur FACOM M382 de Fujitsu (équivalent de IBM 3081k). Mais le "Projet" utilise également l'ordinateur SYMBOLICS 3600 (machine LISP) en raison de ses capacités d'interactivité.

GRADE[9] permet de compiler les grammaires (règles de récriture) en un format utilisable par l'ordinateur (représentation interne). Il s'agit d'un métalangage qui peut s'appliquer à toutes les étapes du traitement : analyse, transfert et synhèse. De plus, ce métalangage permet de spécifier des règles grammaticales associées directement au dictionnaire des mots d'entrée ainsi que de contrôler le processus même de traduction automatique.

La structure de base utilisée par GRADE est l'arbre dont les noeuds sont annotés par un ensemble de paires de valeurs-propriétés. Les arbres ainsi annotés sont très riches en informations et cela à plusieurs niveaux de traitement: morphologique, syntaxique et sémantique.

Il y a bientôt dix ans (au printemps 1986), les premiers essais visant l'appréciation de la qualité de la traduction des textes japonais vers l'anglais ont été effectués sur la base des données des expériences de traduction. Au cours de ces dernières, qui se sont déroulées en trois temps, il a été pris en considération des résumés scientifiques et techniques japonais. Les résultats se sont révélés plutôt peu satisfaisants.

Tableau : Réalisations en TAO *d'après "Kikai hon'yaku Samitto" (Sommet sur la traduction automatique), pp. 200, ouvrage édité sous la direction de NAGAO Makoto, Éditions Ohm-sha, Tôkyô 1989*

[8]) UTILISP (University of Tokyo Interactive LISP). La version utilisée est capable de manipuler les caractères chinois et donne accès directe aux dictionnaires.

[9]) Il a été développé un ensemble d'utilitaires de traduction des programmes écrits en UTILISP sur M382 pour pouvoir être executés sur SYMBOLICS 3600.

Système	Société/Inst.	Lang.	Mode	Dict.	Vitesse m/H
LAMB	CANON	J->E	Transfert par arbre	2.000	1.000
ATLAS-I	FUJITSU	E->J	Grammaire(s)	53.000	60.000
ATLAS-II	FUJITSU	J->E	Grammaire(s)	50.000	60.000
HICATS/JE*	HITACHI	J->E	Gr. Dépendences	50.000	40-80.000
HICATS/EJ*	HITACHI	E->J	Gr. des Cas	50.000	40-80.000
PAROLE	Matsushita	J->E	Tranfert	5.000	?
Mu PROJECT	Kyoto Univ.	J<->E	Tranfert	80.000	4.000
MELTRAN	Mitsubishi	J<->E	Tranfert	50.000	?
PIVOT	Nihon-Denki	J<->E	Langage Pivot	50.000	60.000
MTS**	NDG***	E->J	Tranfert	?	?
PENSÉE	Oki	J<->E	Tranfert	100.000	6.000
RMT	Ricoh	E->J	Tranfert	60.000	4.500
SWP-7800	Sanyô	J->E	Tranfert	110.000	3.500
DUET-E/J	Sharp	E->J	Tranfert	60.00	5-10.000
TAURAS	Toshiba	E->J	Tranfert	50.000	7.000
SYSTRAN	Systran Co.	E<->J	Tranfert	80.000	20.000
CONTRAST	ETL	E<->J	Pivot Language	?	?

*) Système de TAO bidirectionnel japonais-anglais et anglais-japonais.
) **Machine **T**ranslation **S**ystem
***) **N**ihon **D**ata **G**eneral

Dictionnaires électroniques

La description grammaticale et sémantique du lexique doit s'appuyer sur une activité de grande envergure. Pour cette raison, au Japon, toutes les sociétés qui construisent des ordinateurs ont beaucoup investi dans ce genre de recherches et elles les poursuivent actuellement. De plus, ces centres de recherche emploient du personnel hautement qualifié, cela est surtout vrai pour l'Université Technologique de Kyushu. Bien que les dictionnaires électroniques soient utiles également en dehors de la traduction automatique, c'est surtout dans ce domaine que l'on s'efforce de saisir leur complexité.

La forme des dictionnaires électroniques varie évidemment selon le mode de travail que le système de TAO est censé utiliser : (1) par transfert ou (2) par interlangue. Il va sans dire que dans le cas de la traduction "par transfert", le dictionnaire doit être bilingue (donc un dictionnaire par paire de langues), tandis que, dans celui de la traduction "par interlangue", il suffit de construire un dictionnaire pour l'analyse et un autre pour la synthèse pour chacune des langues.

A la suite des tentatives d'extraction de l'information lexicologique à partir des dictionnaires traditionnels[10] (parfois en se servant d'une troisième langue naturelle comme

[10]) Cf. L'extraction des représentations sémantiques par graphes de dépendances des définitions du dictionnaire japonais existant à l'Université de Nagasaki.

“interlangue”), on a vu apparaître, comme il fallait s’y attendre, des tentatives de construction de dictionnaires électroniques à partir de grandes bases de textes.

Le Japan Electronic Dictionary Research Institute a été créé en 1986 aux côtés de l’ICOT (Institute for New Generation Computer Technology).

Les dictionnaires développés à l'EDR sont de 4 types:

1) *dictionnaires de mots*:
   (a) généraux: (i) japonais (200.000) et (ii) anglais (200.000)
   (b) techniques: i) japonais (100.000) et (ii) anglais (100.000)
2) *dictionnaires de concepts*:
   (a) classifications (400.000) et (b) descriptions (400.000)
3) *dictionnaires de co-occurences:*
   (a) du japonais (300.000) et (b) de l'anglais (300.000)
4) *dictionnaires bilingues*:
   (a) jap.-ang (300.000) et (b) ang.-jap. (300.000)

Tous ces dictionnaires sont reliés entre eux. Les entrées lexicales (dictionnaire de mots) sont composées de "mots de base" (représentations superficielles des mots), d'informations grammaticales (décrivant les propriétés grammaticales) et de "concepts de base" (représentés par des mots en contexte).

Un mot peut exprimer plusieurs concepts. Plus le mot est courant, plus il peut exprimer de concepts différents. Le rôle de la sémantique est de déterminer dans quel sens un mot est utilisé dans un contexte donné.

Pour décrire les concepts, les spécialistes utilisent des langages spéciaux appelés parfois "interlangues". Les expressions qui appartiennent à ces langages (voire formalismes déscriptifs) peuvent être des deux types suivants:

1) expressions *logiques* ou *formelles* (Calcul des Prédicats, Logiques modales, temporelle, intentionnelle etc)
2) expressions *descriptives* ou *souples* (réseaux sémantiques, schémas etc.)

Pour cette raison, l'EDR a opté pour un formalisme "descriptif" (plutôt que "logique") car les figures de représentation y sont graphiques et ces dernières sont censées mieux refléter le caractère "connexionniste" du sens (voire du réseau neuronique). Dans ce genre de formalisme, les mécanismes d'inférence sont du type associatif. On s'aperçoit donc que le modèle "descriptif" japonais est héritier des "réseaux sémantiques" et des "cadres" (ou "frames"). On considère que les réseaux sémantiques ne possèdent pas de sémantique formelle définie d'avance.

De plus, les réseaux sémantiques sont plus souples que les expressions logiques dans la mesure où ils admettent des raisonnements "illogiques", et telles sont souvent les expressions des langues naturelles.

Les noeuds des réseaux *a priori* peuvent contenir des informations très variées. Pour remplir ces noeuds, l'EDR a choisi d'utiliser les "frames" à cause du caractère structuré de ces derniers. En effet, les "frames" ont été inventés en tant que technique de représentation des connaissances de sens commun, non seulement des connaissances linguistiques. De

même que les réseaux sémantiques, les "frames" ont l'avantage d'être libres de toute sémantique formelle propre.

En posant ces deux principes à la base de leur dictionnaire des concepts, les Japonais ont donc préféré une approche peu contraignante par rapport au format de représentation afin de pouvoir traiter des masses considérables de données (400.000 entrées).

Ce langage conceptuel (formalisme descriptif) utilisé par l'EDR n'est cependant que partiellement universel. Son vocabulaire se compose de termes appartenant aux quatre catégories suivantes:

1. concepts de base	(BIRD, FLY, WING, AIR...)
2. relations	(agent, implement, location...)
3. attributs	(big, small...)
4. facteurs de certitude	(0, 1)

Les concepts de base sont dépendants de la langue en question mais les termes dénotant les relations, les attributs et les certitudes sont censés être universels. Les concepts sont définis comme des "classes d'images" reconnaissables hors de tout contexte. Il s'agit donc des concepts-types (par opposition aux concepts-occurences). Ces concepts sont annotés par des attributs. Les relations entre concepts sans aucune restriction "attributive" sont considérés comme des sous-classes de concepts eux-mêmes.

Par exemple: l'entrée `c#fly - agent --> c#bird` signifie qu'une sous-classe de la classe `c#bird` peut être l'agent de l'une des sous-classes de la classe `c#fly.` La classe elle-même est aussi une des sous-classes.

Voici un extrait de la classification des concepts:

1. *mono*[11]
   - 101. objets physiques
   - 102. temps et espace
   - 103. choses abstraites
   - 104. divers
2. *koto*[12]
   - 201. phénomènes/mouvement ou changement indépendants de l'intellect humain
   - 202. mouvement/changement de lieu des objets physiques ou abstraits dans un domaine physique ou abstrait impliquant la différence de temps
   - 203. action et activité/action intentionnelle
   - 204. changement (momentané ou continu), expressions comportant le progrès dans le changement et les <concepts *mono*>
   - 205. propriétés/expressions décrivant les caractéristiques (nature et condition), le naturel (nature des hommes et des animaux), les relations entre les <concepts *mono*>
   - 206. autres <concepts *koto*>.

---

[11]) lit.: "objets, choses concrètes"

[12]) lit.: "affaires, choses abstraites"

Etant donné que l'approche est conceptuelle, l'EDR utilise un langage intermédiaire appelé *interlangue*. De plus, l'EDR a choisi d'utiliser un formalisme descriptif ou souple (plutôt que logique ou formel) pour la représentation des concepts. Ce genre de formalisme est issu des recherches sur les réseaux sémantiques, cadres etc.

Les concepts de base étant définis comme des nuplets: CATEGORIE(CONCEPT, {RELATION, ACTION, OBJET,CERTITUDE}) et les concepts composés sont définis en tant qu'ensembles de nuplets. L'exemple d'un nuplet: (agent,FLY,BIRD,1).

Dans le système EDR, les concepts de base sont définis dans le dictionnaire des mots de la langue donnée. Les concepts de base sont considérés, à leur tour, comme des concepts. Par exemple, le concept composé "A big bird flies with wings in the air" est représenté par l'ensemble de n-upplets suivant:

{(big, BIRD, 1), (agent, BIRD, FLY, 1), (implement, WING, FLY, 1), (location, AIR, FLY, 1)}

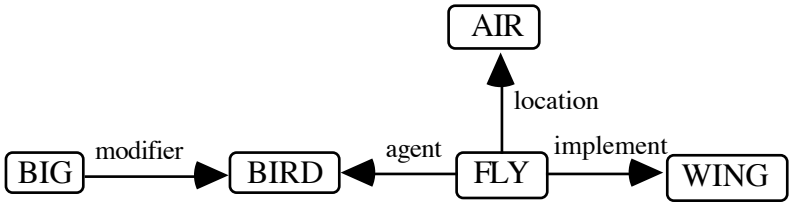


Fig. # . Représentation graphique d'un concept composé.

Pour chaque langue, il y a deux dictionnaires de concepts:

(1) dictionnaire de classification des concepts
(2) dictionnaire de description des concepts

Le dictionnaire de classification des concepts utilise l'idée d'héritage, ce qui permet de réduire le volume des dictionnaires.

Les dictionnaires bilingues mettent en rapport les concepts des deux langues. Les termes descriptifs utilisés sont sélectionnés parmi ceux qui permettent de décrire les relations universelles.

Les informations qui permettent de mettre en rapport les concepts des deux langues sont appelées “étiquettes de correspondances inter-langues”. Elles sont des 4 types suivants:

- équivalence
- synonymie
- hyperonymie
- hyponymie

En réalité, il y très peu d’équivalences entre les langues (kita = north, sentô = public bath, furan = incubation). Pour cette raison, toute ressemblance (définie comme proximité) est considérée comme équivalence.

La relation de synonymie indique que les entrées mises en rapport diffèrent au point de ne pas pouvoir être considérées comme équivalentes. Par exemple: *abbey - daikyôkai*, *abbey - daiteitaku.*

Les super-ensembles sont indiqués lorsque la même entrée-source renvoie à plusieurs entrées-objets.:

bôshi - casquette
bôshi - chapeau

Les sous-ensembles sont indiqués lorsque plusieurs entrées-sources renvoient à la même entrée-objet.:

chiheisen - *horizon*
suiheisen - *horizon*

Des outils de consultation sont prévus pour les utilisateurs finals.

Quelques autres problèmes

Parmi les problèmes traités dans les domaines de la reconnaissance des caractères et du traitement de la parole, il y en a beaucoup qui se réfèrent au matériel mais les nombreuses questions méthodologiques qui subsistent laissent également beaucoup de place aux logiciels. Au Japon, le domaine de la reconnaissance des caractères abonde en problèmes liés au traitement des kanji. Ce qui rend problématique la reconnaissance automatique des textes japonais, c'est surtout le fait que les kanji se composent de nombreux traits et possèdent des formes très variées et complexes, néanmoins ces derniers temps les recherches dans ce domaine ont beaucoup progressé. Malgré l'abondance d'approches, il n'y a pas encore de système qui fonctionnerait de manière parfaite. En ce qui concerne la synthèse et la reconnaisssance de la parole, les recherches sont assez avancées. Les résultats les plus spectaculaires ont été obtenus dans les instituts de recherche en TAO pour les besoins de la téléphonie en réalisant des démonstrations de la traduction des conversations.

Mentionnons enfin que de nouvelles théories linguistiques du japonais sont apparues dans le sillage et en profitant des expériences du traitement automatique de cette langue. Les deux modèles les plus remarquables sont les grammaires de GUNJI Takao et MIZUTANI Shizuo.

GUNJI Takao a profité des résultats des recherches consacrées à l'anglais utilisant les grammaires d'unification (notamment HPSG) pour décrire un ensemble de faits de la langue japonaise de façon tout à fait originale. Sa grammaire, dont l'original a été publié en anglais, est connue par son acronyme JPSG[13]. Comme sa grande soeur, elle permet de traiter les langues dont l'ordre linéaire est relativement libre. Il est intéressant de noter que l'ICOT, dans sa dernière phase d'existence, a utilisé JPSG au cours du développement de sa version de Prolog (cu-Prolog) - le premier langage pour clauses de Horn à contraintes symboliques.

D'un autre côté, MIZUTANI Shizuo a construit son modèle formel de la grammaire japonaise en s'appuyant sur les concepts élaborés par la tradition linguistique indigène. Il s'agit bien entendu et avant tout des phénomènes syntaxiques, mais la particularité de

[13]) Japanese Phrase Structure Grammar - A Unification-based Approach, by GUNJI Takao, D. Reidel Publishing Company, Dordrecht 1987.

cette grammaire consiste en ce que les constituants sont très originaux et que les structures ne sont pas seulement arborescentes mais parfois aussi des treillis. La fait d'avoir développé un appareil conceptuel considérable a permis de poser plusieurs noeuds uniques (un par niveau de l'arbre) sur la même branche; de sorte que l'arbre devient un chemin unique parsemé d'étiquettes. Cela s'est révélé très intéressant notamment parce que beaucoup de ce que nous sommes obligés de reconnaître comme ambiguités syntaxiques se laissent éliminer sans quitter le domaine de la syntaxe. Nous avons cru bon de rendre cette grammaire accessible aux linguistes occidentaux en la traduisant en français[14], mais déjà une nouvelle présentation de cette grammaire, fondée sur la théorie des ensembles, a vu le jour au Japon.

Avant de clore ce survol des problèmes que la langue japonaise a posés à l'informatique nous aimerions signaler le fait suivant. En Europe, on traduit - cela ne surprendra personne - le plus souvent de l'anglais mais le japonais est - cela est peu connu - la deuxième langue dont les Européens traduisent le plus souvent à l'heure actuelle. Or, la TAO n'a toujours pas de réponse précisement parce que la traduction automatique (ou même seulement "assistée par ordinateur") du japonais en langues européennes exige que l'ordinateur mémorise une quantité considérable de connaissances (en plus des règles grammaticales).

Toshio ISHIWATA et André WLODARCZYK

ANNEXE : Structure et évolution des kanji

Pour des raisons aussi bien structurelles qu'historiques (étymologiques), les rapports entre une langue donnée et son "système d'écriture" sont souvent complexes surtout lorsque les utilisateurs de ces deux systèmes ont choisi de noter principalement les unités lexicales et non phoniques de leur langue. Il est intéressant d'observer que les écritures à base "idéographique" sont - surtout à leur apparition - une sorte de spéculation vis à vis du sens. Elle sont représentation graphique (voire "symbolique") du monde. En témoignent la distinction entre pictogrammes (wen, "dessins") et leurs dérivés (zi, "enfants des dessins") et les débats de l'Antiquité chinoise[15] à propos de l'écriture chinoise que désigne le mot composé *wenzi* (écriture, littéralement: "dessins et leurs dérivés"). Mentionnons également l'oeuvre de Xu Shen (30-124), lexicographe et réformateur chinois pour sa théorie des *six espèces* (liu shu) qui a permis de dégager les six

---

[14]) «Description systématique de la grammaire japonaise» par MIZUTANI Shizuo, ouvrage traduit du japonais par Reiko SHIMAMORI et André WLODARCZYK, suivi de "Application de la grammaire de mizutani au traitement informatique" par André WLODARCZYK dans Travaux de Linguistique Japonaise, vol. IX, Université de Paris 7, Paris 1991 (Adresse de la rédaction: UFR d'Asie Orientale, 2, Place Jussieu, 75251 PARIS CEDEX 05)

[15]) Par exemple: le problème de la *rectification des noms* (zhengming) chez Confucius (551-479 av. J. Chr.) ou celui de leurs *dérivés* (zi) chez Zheng Xuan (127-200 av. J. Chr.), cf. Léon VANDERMEERSCH, "Écriture et langue graphique en Chine", DÉBAT, N° 62, 1990, Galimard, Paris.

classes[16] de caractères suivantes: *pictogrammes* (xiangxing, lit. figuration de la forme), *déictogrammes* (zhishi, lit.: indication de la chose), *syllogigrammes* (huiyi, lit.: combinaison des sens), *morphophonogrammes* (xingsheng, lit. forme clé et phonétique), *emprunts* (jiajie, lit.: empruntés) et *doublets* (zhuanzhu). Mais même dans une écriture dite idéographique, les caractères qui se réfèrent directement aux concepts (signifiés) sont très rares: sur 9.353 caractères récensés par Xu Shen, 500 seulement sont de purs pictogrammes. Probablement, ce fut l'invention de la notion de *clé* (élément constitutif servant aussi de critère de classement) qui a fait perdre la nature "idéographique" à l'écriture chinoise. Au départ, Xu Shen a posé 540 clés. Ce nombre a varié plusieurs fois depuis ce temps au cours des réformes[17] successives. La dernière a fixé le nombre de clés à 214. Ces clés se combinant le plus souvent par deux, trois ou quatre ont permis de constituer un inventaire considérable des caractères dont le nombre varie bien entendu selon le volume du dictionnaire, de 6.000 à 45.000.

La date exacte de l'introduction du système chinois d'écriture au Japon reste inconnue mais on peut dire que depuis le 6e siècle au moins l'écriture de l'Empire des Han[18] (ch.: hanzi, j.: kanji) n'était pas inconnue au Japon. Les premiers lettrés japonais étaient confrontés aux deux problèmes linguistiques suivants:

(1) comment utiliser les caractères chinois pour noter la langue japonaise ? [problème de la **notation**]

Avec l'écriture chinoise, il était possible de noter le japonais de deux façons: (a) - soit en exploitant le rapport de *synonymie* entre les deux langues chinoise et japonaise; c'est-à-dire: en établissant des correspondances entre les caractères chinois et les mots japonais (par exemle: yama "montagne", nagaru "couler") soit (b) en utilisant les rapports d'homophonie des deux langues à travers les ideogrammes, c'est-à-dire: en établissant des correspondances entre les caractères chinois et les syllabes japonaises (par exemle: ya-ma "montagne", na-ga-ru "couler"). La conscience de ce qu'un caractère chinois pouvait être utilisé alternativement soit avec sa charge sémantique unitaire soit en tant que marque d'une valeur phonétique[19] monosyllabique a conduit les Japonais à élaborer plusieurs syllabaires dérivés de l'écriture chinoise. Deux d'entre ces syllabaires sont toujours en usage au Japon, car ils accompagnent les caractères jusqu'à aujourd'hui. Il s'agit notamment du *katakana* (lit.: noms apparents partiels des lettres) et du *hiragana* (lit.: noms apparents complets (des lettres)).

(2) comment lire les textes chinois à la japonaise ? [problème de la **traduction**]

Pour déchiffrer le sens des textes originaux chinois directement en leur langue (sans les traduire), les Japonais ont inventé une technique permettant d'annoter les caractères chinois au moyen d'ajouts graphiques servant de points de repère («kunten»). Le but de ces annotations conventionnelles consistait à donner la lecture japonaise directe d'un texte original chinois en indiquant l'ordre des mots ainsi que certains morphèmes grammaticaux japonais correspondants.

C'est notamment l'utilisation des caractères chinois dans le but de noter phonétiquement les mots japonais qui a conduit à l'élaboration des syllabaires japonais *kana* (noms apparents) qui, dans l'esprit des lettrés japonais de l'époque Heian (794-1192), s'opposaient aux *mana* (noms réels); ces derniers n'étant rien d'autre que les caractères chinois qui conservaient leur sens d'origine.

---

16) Les traductions françaises des classes de caractères de Xu Shen sont de Léon VANDERMEERSCH, op. cit. ci-dessus.

17) Réformateurs et nombre de clés : Xu Shen (30-124) - 540, Zhang Can (VIIIe s.) - 160, Li Congzhou (entre X -XIIIe s. ?) - 89, Zao Huiqian (1351 -1395) - 360, Mei Yingzuo (1570-1615) - 214.

18) La dynastie des Han régna en Chine entre 206 av. et 220 apr. J.C.

19) Les caractères chinois étaient employés d'abord arbitrairement puis conventionnellement pour noter les valeurs phonétiques des morphèmes japonais.