

Traduction du japonais : un modèle cognitif ACT-R

Francis BISSON, Francis CORRIVEAU, Kenny CÔTÉ et Pierre-Luc ROGER

Département d'informatique, Université de Sherbrooke

1. Introduction

L'apprentissage d'une nouvelle langue est une tâche très difficile pour tout individu. La mémorisation d'une grande quantité de vocabulaire et de règles de grammaire, ainsi que la capacité de reconnaître le rôle des mots dans une phrase donnée sont des étapes importantes dans un tel apprentissage. La traduction de phrases est donc une activité complexe, qui permet de mesurer le niveau de compréhension d'une langue seconde.

À cet effet, nous présentons un modèle cognitif effectuant la traduction de phrases japonaises romanisées en français. Nous commencerons par détailler la problématique, après quoi nous exposerons le modèle implémenté dans l'architecture cognitive ACT-R. Nous donnerons ensuite les résultats obtenus avec le modèle et nous terminerons avec une analyse de ces résultats et une discussion générale du problème. Un compendium des particularités de la langue nippone sera aussi fourni en annexe.

L'objectif principal de ce papier est d'étudier la cognition humaine lors de la traduction de phrases japonaises simples.

Ces phrases dites simples contiennent toutes les parties nécessaires à la compréhension de l'énoncé, soit un sujet, un verbe et un complément. Nous faisons l'hypothèse que le sujet modélisé connaît d'avance la signification des mots qu'il lit, mais peut commettre des erreurs lors de la traduction. Cette étude permettra donc de représenter le comportement du cerveau humain lors d'un exercice de traduction.

2. Le modèle ACT-R

La traduction s'effectue en plusieurs étapes successives se représentant bien avec ACT-R, qui est dirigé par les buts. Nous avons donc un type de but pour chacune de ces étapes. Comme nous l'avons mentionné précédemment, nous supposons que le sujet modélisé connaît *a priori* la signification des mots japonais. La mémoire déclarative est donc initialisée avec le vocabulaire connu.

La traduction débute par l'identification du rôle de chaque mot dans la phrase en trois catégories : le sujet, le verbe et le complément. Étant donné la structure des phrases japonaises, le verbe est le premier

groupe de mots identifié. La lecture commence au dernier mot de la phrase et s'arrête dès qu'une particule est rencontrée. Le modèle trouve ensuite le sujet, qui se situe au début de la phrase et se termine avec une particule de sujet. Le complément est composé de tous les mots entre ces deux catégories. Un appreni pratiquant la traduction du japonais procédera à l'identification des groupes de mots dans cet ordre; le verbe est la partie la plus facile, puisqu'elle est toujours située à la fin. Le sujet, s'il est présent, sera ensuite recherché au début de la phrase.

Théoriquement, le modèle est capable d'identifier correctement le complément du verbe si le sujet de la phrase est implicite, mais nous avons préféré ne pas traiter cette partie pour des raisons que nous énoncerons plus loin. Des listes en langage LISP ont été utilisées pour permettre l'accumulation des mots dans les groupes sémantiques.

Une traduction littérale en français est ensuite effectuée pour chacun des mots. Certaines particules ont une traduction spéciale, comme la particule servant à identifier le sujet, puisqu'elles n'ont pas toujours de signification autre que le sens sémantique accordé aux mots précédant la particule.

La traduction spéciale est cependant utilisée pour remplacer certaines locutions particulières. Par exemple, la construction « de moi » (où « de » est une traduction de la

particule « *no* ») sera remplacée par « mon ». C'est donc lors de cette étape que la personne du sujet est identifiée et conservée pour la conjugaison du verbe. Nous avons choisi de représenter la conjugaison des verbes en français directement dans la mémoire déclarative, et non dans la mémoire procédurale, par souci de simplicité.

Après ces étapes, le contenu de chaque groupe de mots a été traduit et réordonné; il ne reste donc plus qu'à réordonner le sujet, le verbe et le complément dans cet ordre, conformément à la structure normale d'une phrase française (contrairement au japonais, où le verbe est toujours situé à la fin de la phrase). Le modèle récite finalement la phrase traduite en entier.

Afin de satisfaire l'hypothèse que le sujet modélisé puisse commettre des erreurs lors de la traduction, nous avons utilisé le système sub-symbolique de l'architecture cognitive ACT-R. La possibilité d'introduction d'irrégularités se fait lors de l'étape de la traduction littérale, lorsque le modèle recherche une signification dans sa mémoire déclarative pour le mot *lu*.

Nous avons utilisé deux mesures de similarité entre les mots : une orthographe semblable et un sens proche. Par exemple les verbes conjugués « *yomimasu* » (lire) et « *nomimasu* » (boire) n'ont que la première lettre qui est différente. Nous considérons aussi les paires de mots gauche-droite et aller-venir comme ayant un sens proche.

Étant données ces ressemblances, il est fort probable qu'une personne apprenant le japonais se trompe lors de la sélection de la traduction appropriée pour un mot donné. Cependant, le niveau d'activation des *chunks* fournissant la bonne signification est supérieur à ceux des *chunks* d'erreur; le bruit dans la formule d'activation est donc le responsable de l'introduction de fautes.

3. Résultats

Afin de vérifier l'efficacité de notre modèle, nous avons traduit trois phrases simples soixante fois en notant le nombre d'occurrences de toutes les phrases résultantes. Dans l'exposition des résultats suivants, un astérisque (*) indique la phrase ayant la signification la plus ressemblante à la traduction réelle. Notez que le modèle ne met pas de caractères accentués dans ses mots pour éviter de perturber l'environnement d'ACT-R.

Tableau 1. Traductions de la phrase 1

« <i>watashi no tomodachi wa daigaku ni ikimasu</i> »	
« mon ami va a universite »	48 *
« mon ami vient a universite »	6
« mon ami est a universite »	1
Retrieval failure	5
Nombre d'exécutions	60

La phrase 1 contient le verbe « *ikimasu* » (aller) qui est parfois confondu lors de la simulation par « *kimasu* » (venir),

ce dernier ayant a une orthographe similaire et un sens proche. De façon assez rare, le verbe « *ikimasu* » est aussi identifié à « *imasu* » (être) puisque l'orthographe (mais non le sens) est analogue. Enfin, 8% des traductions du modèle se sont soldées par un échec du buffer *retrieval*, ce qui a mis fin prématurément au déroulement de la traduction.

Tableau 2. Traductions de la phrase 2

« <i>neko wa migi no tsukue no ue ni imasu</i> »	
« chat est a dessus de table de droite »	46 *
« chat est a dessus de table de gauche »	10
Retrieval failure	4
Nombre d'exécutions	60

La phrase 2 précise l'emplacement du chat avec le mot « *migi* » (droite) traduit correctement la plupart du temps, mais qui est parfois confondu durant la simulation par « *hidari* » (gauche). Encore une fois, 7% des traductions de la phrase ont été interrompues par un échec du buffer *retrieval*.

Tableau 3. Traductions de la phrase 3

« <i>Tanaka san wa toshokan de benkyou shimasu</i> »	
« monsieur Tanaka etudie a bibliotheque »	27 *
« madame Tanaka etudie a bibliotheque »	21
Retrieval failure	2
Nombre d'exécutions	60

Lors de la traduction de la phrase 3, le modèle a traduit très fréquemment le terme « *san* » (forme d'adresse polie) pour *madame* plutôt que *monsieur*. Les échecs du buffer *retrieval* sont survenus de façon plus rare que dans les phrases précédentes.

4. Analyse et discussion

En observant les résultats de la phrase 1, nous observons tout d'abord que peu importe le verbe employé, les phrases traduites proposées par le modèle ont toujours un ordre syntaxique correct (*sujet-verbe-objet*). Également, nous pouvons constater que la traduction prédite par le modèle est la majorité du temps la plus exacte. Pour expliquer l'utilisation fautive du verbe venir et plus rarement du verbe être dans la traduction, il faut se rappeler que le système sub-symbolique introduit parfois des erreurs dans la traduction de certains mots. Puisque « *ikimasu* » (aller) est similaire en orthographe à « *imasu* » (être) et similaire en orthographe et en sens à « *kimasu* » (venir), il arrive parfois que la valeur d'activation du chunk d'un de ces deux derniers verbes devienne supérieure à celle du premier à cause du bruit et soit choisie à sa place.

Pour continuer avec les prédictions du modèle au sujet de la phrase 2, il semble encore une fois que les mots de chaque traduction proposée soient toujours dans le bon ordre, et ce malgré la présence d'un objet

composé de plusieurs particules (« au-dessus | de la table | de droite »). Aussi, malgré le fait que le mot *migi* (droite) soit choisi la majorité du temps par le système sub-symbolique, il arrive parfois que le modèle le mélange avec le mot *hidari* (gauche) puisque les deux ont une sémantique similaire.

Enfin, la distribution des phrases traduites proposées par le modèle pour la phrase 3 est beaucoup plus homogène que les autres (27 « monsieur » pour 21 « madame »). Ceci survient pour deux raisons : premièrement, « *san* » est une formule d'adresse polie utilisée autant pour un homme que pour une femme (et donc ayant la même valeur d'activation de base lorsque « *san* » est recherché); deuxièmement, notre modèle ne tient pas compte du contexte et n'a donc aucun indice syntaxique lui permettant de savoir à l'aide de ses règles de productions si Tanaka est un homme ou une femme. Ceci a donc pour résultat que le modèle a donc autant de chances d'appeler Tanaka monsieur que madame.

Pour terminer, il est important de connaître la raison de la présence des *retrieval failures* survenant parfois lors de la traduction de toutes les phrases. Afin d'implémenter notre système sub-symbolique, nous avons dû introduire du bruit dans le modèle afin que la valeur d'activation des chunks soit légèrement modifiée à chaque cycle pour introduire un

facteur de hasard. Ceci amène la possibilité que le modèle choisisse la mauvaise traduction d'un mot japonais, mais rend aussi possible le cas où tous les chunks autrement valides sont maintenant sous le seuil d'activation, et donc qu'aucun ne soit sélectionné. Notre modèle ne gère pas cette occurrence et arrête donc la traduction sur le champ.

5. Conclusion

Nous avons étudié dans ce papier la cognition humaine lors de la traduction de phrases japonaises en français. Nous avons ensuite modélisé cette tâche, qui requiert un certain niveau de connaissances dans les deux langues impliquées, à l'aide de l'architecture cognitive ACT-R. La conception orientée par les buts de ACT-R nous a permis de séparer l'exercice de traduction en plusieurs sous-tâches exécutées dans un ordre séquentiel.

Des objectifs que nous nous étions fixés par rapport aux capacités du modèle, nous avons réussi à en implémenter plusieurs. En particulier, la production d'une traduction intelligible en français, la conjugaison des verbes avec la personne (déduite lors de l'identification du sujet) ainsi que le remplacement de certaines locutions particulières ont été intégrés au modèle avec succès. La simulation d'erreurs a aussi été

ajoutée au modèle, grâce au système sub-symbolique de ACT-R.

Quelques limitations du modèle cognitif peuvent cependant être soulignées. Le seul temps de verbe traité est le présent-futur et la conjugaison des verbes en français n'est pas exprimée sous forme de règles. Une meilleure façon de procéder serait de traduire le verbe en conservant son temps, puis d'utiliser des règles de productions dans la mémoire procédurale pour conjuguer correctement le verbe. Dans son état actuel, le modèle assume que les verbes doivent être conjugués au présent en français. De plus, les mots formant une proposition temporelle ne sont considérés que pour leur signification directe; leur effet sur le verbe n'est pas conservé.

Les phrases ayant un sujet implicite ne sont pas supportées par le modèle. Nous avons établi cette contrainte puisque nous considérons une seule phrase à la fois, et non un ensemble de phrases duquel nous pourrions déduire un contexte.

Nous avons aussi décidé d'ignorer l'ajout de déterminants devant les noms dans les phrases traduites en français. Cela aurait nécessité l'ajout d'une grande quantité d'information aux mots inscrits dans la mémoire déclarative, comme le genre et le nombre. Cela a pour effet de produire des traductions compréhensibles, mais pas nécessairement syntaxiquement correctes.

Pour terminer, nous souhaiterions proposer quelques améliorations possibles à apporter au modèle, en plus des lacunes identifiées précédemment. Premièrement, l'utilisation du module visuel pour émuler la lecture des mots serait un ajout intéressant, qui permettrait d'obtenir un temps de simulation plus près de la réalité après avoir réglé certains paramètres.

Ensuite, la lecture pourrait être étendue pour lire des caractères japonais plutôt que des phrases romanisées; cette partie aurait une grande incidence sur l'utilisation du système sub-symbolique dans le modèle, puisque certains caractères japonais ont plus d'une prononciation, ce qui peut évidemment engendrer plusieurs types d'erreurs.

Nous pourrions aussi ajouter le support pour des phrases plus complexes, par exemple deux phrases liées ensemble par un mot comme « mais » ou « parce que ». Cela impliquerait l'implémentation du support d'un plus grand nombre de particules, ainsi que des modifications majeures dans l'étape d'identification des rôles des mots dans la phrase pour tenir compte de la possibilité d'avoir plus d'un verbe, sujet ou complément.

Enfin, la conservation du contexte d'une phrase à l'autre serait sans doute un des ajouts les plus intéressants à apporter au modèle cognitif ACT-R que nous avons implémenté. Nous pourrions ainsi par exemple connaître le genre ou le nombre du

sujet d'une phrase suivante même s'il n'y est précisé.

En bref, malgré l'atteinte de plusieurs objectifs, nous sommes intéressés par la poursuite de cette étude en modélisation cognitive de traduction d'une langue étrangère.

Références

Bothell, D. 2005. *ACT-R 6.0 Reference Manual*. Department of Psychology, Carnegie Mellon University.

Banno, E., Ohno, Y., Sakane, Y. et Shinagawa, C. 1999. *Genki I: An Integrated Course in Elementary Japanese I*. Japan Times, Tokyo.

Banno, E., Ohno, Y., Sakane, Y., Shinagawa, C. et Tokashiki, K. 1999. *Genki II: An Integrated Course in Elementary Japanese II*. Japan Times, Tokyo.

Makino, S., Hatasa, Y.A. et Hatasa, K. 1998. *Nakama 1: Japanese Communication, Culture, Context*. Student Edition. Houghton Mifflin Company, Boston.

Breen, J. WWWJDIC. <http://www.csse.monash.edu.au/~jwb/cgi-bin/wwwjdic.cgi?1C>. Monash University, Melbourne. Site Web visité le 4 décembre 2008.

Annexe I. La langue japonaise

1. Systèmes d'écriture

Il existe 3 systèmes d'écritures en japonais (4 si on compte le romaji, l'écriture en lettres latines) : les hiragana, les katakana et les kanji. Les hiragana et les katakana sont des syllabaires composés de 46 caractères, dont le son de certains peut être modifié grâce à des diacritiques (petits symboles ajoutés aux caractères).

Les katakana sont utilisés pour écrire les mots étrangers, comme les noms de pays et de personnes, certains noms de compagnies, etc. Ces caractères peuvent aussi être utilisés pour mettre l'accent sur un mot, à la façon de l'italique en français ou en anglais. Les hiragana sont utilisés pour tout le reste.

Les kanji sont des caractères d'origine chinoise utilisés en japonais. De 2000 à 3000 sont utilisés couramment dans les journaux et la littérature. Chacun de ces caractères, souvent plus complexe à dessiner qu'un katakana ou un hiragana, a une ou plusieurs significations et prononciations. Par exemple, le kanji 月 signifie « lune » et est prononcé « tsuki ». Le même symbole est utilisé dans le mot signifiant « lundi » (月曜日), mais se prononce « getsu ». Il est aussi utilisé pour les mois et se prononce « gatsu » (六月, « juin »).

Il est possible, voire très commun, d'utiliser les trois systèmes d'écriture simultanément dans une même phrase. Un même mot peut être composé de caractères hiragana et d'un ou plusieurs kanji. Les verbes, par exemple, utilisent des kanji ayant la signification désirée comme base du verbe et la terminaison (qui change selon le temps de verbe utilisé) est écrit en hiragana.

2. Particularités de la langue

2.1. Structure des phrases

La langue japonaise fait partie de la catégorie des langues dont les phrases suivent un ordre sujet-objet-verbe. Ainsi, pour dire « Je mange une pomme », il faudrait réordonner les mots dans l'ordre « Je pomme mange ».

2.2. Temps grammaticaux

Le japonais comporte plusieurs temps de verbes, mais nous nous concentrerons sur les temps présent, passé et impératif. Le présent est en fait un temps « non-passé »; le même temps

de verbe est utilisé pour le présent et le futur. Le temps réel où se produit l'action est déduit du contexte. Notons que le temps passé affecte aussi les adjectifs.

Les verbes sont divisés en deux grandes catégories, les verbes « u » et les verbes « ru », identifiés par leur dernière syllabe. La conjugaison des verbes avec les temps présent et passé est plutôt simple et il existe très peu d'exceptions. Cependant, certains verbes se terminant par la syllabe « ru » entrent dans la catégorie des verbes « u » et il n'existe pas de règle parfaite pour les classer correctement.

La conjugaison du temps impératif pour les verbes « u » présente un défi plus grand, puisque qu'il existe plusieurs règles selon la dernière syllabe des verbes. Le tableau suivant résume ces règles.

Dernière syllabe	Nouvelle terminaison	Exemple
く (<i>ku</i>) ぐ (<i>gu</i>)	いて (<i>ite</i>)	きく (<i>kiku</i> , écouter) → きいて (<i>kiite</i>)
す (<i>su</i>)	して (<i>shite</i>)	はなす (<i>hanasu</i> , parler) → はなして (<i>hanashite</i>)
む (<i>mu</i>) ぶ (<i>bu</i>) ぬ (<i>nu</i>)	んで (<i>nde</i>)	よむ (<i>yomu</i> , lire) → よんで (<i>yonde</i>)
る (<i>ru</i>) う (<i>u</i>) つ (<i>tsu</i>)	って (<i>tte</i>)	もつ (<i>motsu</i> , prendre) → もって (<i>motte</i>)

2.3. Pluriel

Le pluriel n'influence d'aucune façon les mots composant une phrase en japonais. Ainsi, il n'y a aucune différence entre les phrases « Je mange une pomme » et « Je mange des pommes ». La quantification des mots est, encore une fois, déduite du contexte.

2.4. Absence des sujets

Par souci de concision et pour éviter la répétition, les sujets dans une phrase sont souvent omis. On dira donc souvent « Mange pomme » plutôt que « Je mange une pomme » quand le sujet a déjà été identifié dans une phrase précédente. Cela ajouté au temps de verbe « non-passé » et à l'absence du pluriel peut rendre une phrase japonaise prise hors de contexte très difficile à comprendre et à analyser correctement. Nous tenterons autant que possible de minimiser ces omissions même si les phrases seraient pourtant syntaxiquement correctes.

2.5. Niveaux de politesse

Le japonais comporte plusieurs niveaux de politesse devant être choisis selon les personnes avec qui on communique; on parlera différemment si on s'adresse à ses amis et si on s'adresse à l'empereur du Japon. Il est, par exemple, considéré impoli d'utiliser un pronom (il, tu, etc.) pour

parler d'une autre personne; il convient mieux utiliser son nom suivi d'un suffixe (san, kun, chan, etc.) Ces niveaux de politesse affectent entre autres les verbes et les suffixes attachés aux noms. Nous étudierons uniquement l'effet de deux niveaux de politesse sur les verbes, soit la forme dictionnaire (niveau familier) et la forme « masu » (niveau poli), nommée d'après la terminaison des verbes utilisant ce registre du langage.

2.6. Orientation du texte et espacements

Traditionnellement, le japonais s'écrit à la verticale de haut en bas, et de droite à gauche. De plus, il n'y a pas d'espace entre les mots et ceux-ci peuvent être coupés en plein milieu d'un changement de ligne (ou de colonne) sans avoir besoin d'ajouter quelque caractère que ce soit. Cependant, depuis l'occidentalisation du Japon, ces règles s'appliquent moins strictement; le japonais est souvent écrit à l'horizontale de gauche à droite et des espaces sont parfois insérés entre les mots.

3. Particules

Les particules sont des mots très courts (1 à 2 syllabes) qui identifient le rôle des mots précédents dans la phrase. Par exemple, la particule は (wa) indique le sujet de la phrase (qui peut, mais pas nécessairement, être le sujet du verbe). Une courte liste de particules est donnée dans le tableau ci-dessous.

Particule	Signification
は (wa)	Sujet de la phrase
が (ga)	Sujet du verbe
を (wo / o)	Objet direct du verbe
の (no)	Possession
に (ni)	Endroit dans l'espace ou dans le temps
で (de)	Endroit où se produit l'action
で (de)	Moyen de produire une action
も (mo)	Aussi
と (to)	Avec, et

Voici maintenant quelques exemples d'utilisation de ces particules (en gras et en rouge) :

Japonais : ケニ さん **は** りんご **を** 食べます。

Romaji : keni san **wa** ringo **wo** tabemasu.

Traduction littérale : Kenny (*sujet*) pomme (*objet*) mange.

Traduction : Kenny mange une pomme.

Japonais : 大学 **で** べんきょう します。

Romaji : daigaku **de** benkyou shimasu.

Traduction littérale : université (*endroit d'action*) étudier.

Traduction : [J] étudie à l'université.

Japonais : 私 **の** 友だち **と** えいが **を** 見ます。

Romaji : watashi **no** tomodachi **to** eiga **wo** mimasu.

Traduction littérale : moi (*possession*) ami(s) (*avec*) film (*objet*) regarder. (*les amis de moi*)

Traduction : Je regarde un film avec mon (mes) ami(s).

Japonais : 来年 **の** なつ **に** 日本 **に** 行きます。

Romaji : rainen **no** natsu **ni** nihon **ni** ikimasu.

Traduction littérale : année suivante (*possession*) été (*endroit dans le temps*) Japon (*endroit dans l'espace*) aller.

Traduction : [J] irai au Japon l'été prochain.

Japonais : 先生 **が** はなします。

Romaji : sensei **ga** hanashimasu.

Traduction littérale : professeur (*sujet*) parle.

Traduction : Le professeur parle.

Japonais : バス **で** 大学 **に** 行きます。

Romaji : basu **de** daigaku **ni** ikimasu.

Traduction littérale : autobus (*moyen de produire l'action*) université (*endroit dans l'espace*) aller.

Traduction : Je vais à l'université en autobus.

Japonais : すし **を** 食べました。 ごはん **も** 食べました。

Romaji : sushi **wo** tabemashita. gohan **mo** tabemashita.

Traduction littérale : sushi (*objet*) manger (*passé*). riz (*aussi*) manger (*passé*).

Traduction : [J] ai mangé du sushi. [J] ai aussi mangé du riz.