Un analyseur morphologique pour les verbes forts en akkadien

François Barthélemy
Conservatoire des Arts et Métiers, Paris
INRIA projet ATOLL (Rocquencourt)

20 mai 2005

L'akkadien est la langue des anciens babyloniens et assyriens. Elle a été utilisée pendant 25 siècles dans tout le proche-orient ancien. Il s'agit d'une langue sémitique, ce qui signifie qu'elle a une morphologie très riche et non concaténative. Elle était notée dans l'antiquité au moyen de l'écriture cunéiforme, un système complexe et ambigu. Cette écriture note toutes les voyelles, ce qui est exceptionnel pour une langue sémitique. Transcription et transilittération dans un alphabet latin étendu sont utilisées depuis le déchiffrement de l'akkadien à la fin du XIXème siècle. Nous présentons un travail qui a pour but de réaliser l'analyse morphologique de l'akkadien depuis l'écriture cunéiforme, avec une étape intermédiaire de transcription. Nous présentons ici la première étape de ce travail qui se restreint à la morphologie verbale des verbes forts trilitères.

1 La morphologie du verbe fort en akkadien

La plupart des verbes akkadiens sont trilitères, c'est-à-dire qu'ils ont une racine composée de trois consonnes. Il existe des consonnes faibles qui ont tendance à disparaître ou à se transformer dans tout ou partie de la flexion d'un mot. Les verbes faibles sont ceux dont une, deux, voire trois des racines sont des consonnes faibles. Les verbes forts ont des racines qui ne comportent que des consonnes fortes, que l'on retrouve dans toutes les formes fléchies.

La flexion du verbe se fait par plusieurs moyens :

- la vocalisation des consonnes de la racine
- des affixes : préfixe, suffixes et infixes

Les infixes sont insérés après la première consonne racine ou après la consonne marquant une information de voie.

La morphologie verbale est très riche. Le nombre de formes d'un verbe est de l'ordre du millier, certaines étant très rares voire non attestées. Une mutiplicité d'informations interviennent :

- genre (masculin, féminin), nombre (singulier, duel, pluriel), personne
- aspect ou temps : accompli, inaccompli, parfait, permansif, impératif
- formes nominales : infinitif, participe actif, adjectif
- une voie et sous-voie, dans un système structuré à 12 paradigmes marquant de nombreuses nuances, dont certaines relèvent du lexique (à la frontière de la morphologie dérivationnelle et flexionnelle).
- un mode (de base, subjonctif, ventif)

2 Le système graphique cunéiforme

Le système cunéiforme est riche et complexe. Il comporte de l'ordre de 450 caractères qui ont chacun plusieurs interprétations possibles (parfois plusieurs dizaines), les unes syllabiques, les autres idéographiques, les dernières déterminatives. Chaque syllabe peut être notée par différents signes (jusqu'à une douzaine). Le système est donc souvent ambigu. De plus, il possède des contraintes forte, comme l'impossibilité de noter deux consonnes consécutives à l'initiale ou en finale, ce qui conduit à insérer artificiellement une voyelle brève. La notation est approximative : la longueur des voyelles n'est pas notée et des groupes de consonnes voisines phonologiquement sont souvent indistincts (par exemple {d,t,ṭ} ou {g,k,q}).

3 Structure de l'annalyseur

L'anlyseur morphologique que nous avons réalisé est expérimental. Il repose sur un système lui-même exérimental, que nous avons développé en même temps. Il s'agit d'un outil de morphologie à deux niveaux qui compile des descriptions multi-rubans en des automates et transducteurs finis. Il utilise la boîte à outil FSM de Pereira et Mohri (Bell Labs).

L'utilisation de morphologie à plusieurs rubans a été proposée notamment par Kiraz [1] pour la description de la langue syriaque (elle aussi une langue sémitique).

Au niveau le plus abstrait (niveau lexical), notre description comporte 4 rubans :

- un ruban de service pour des informations d'affichage (nous ne l'écrivons pas dans l'exemple qui suit)
- un ruban pour la racine
- un ruban pour la vocalisation aspectuelle, qui comporte zéro, une ou deux voyelles utilisées pour vocaliser la racine
- un ruban pour les éléments infixes introduits après la première consonne racine ou intercalés entre le préfixe personnel et la première racine.

Par ailleurs, chaque forme est décomposée en sept segments, certains pouvant être vides : 1, le préfixe personnel ; 2, le préfixe de voie (ou paradigme), 3,4,5 : les trois racines avec leurs vocalisations et infixes éventuels ; 6 : le suffixe personnel ; 7 : le suffixe de mode. L'utilisation de segments a également été proposée par Kiraz pour la morphologie du syriaque.

Une première description associe les quatre rubans lexicaux à une forme intermédiaire comportant toutes les consonnes plus les voyelles des affixes invariables (par exemple, les suffixes et préfixes personnels) et les voyelles aspectuelles. Cette description est composée d'une multiplicité (30) d'expressions régulières locales, composées par des opérations d'intersection et d'union.

Un système de règles à deux niveaux relie ensuite cette description intermédiaire à une forme entièrement vocalisée. Un autre système de règles à deux niveaux, appliqué en cascade, réalise des transformations phonologiques de surface (par exemple, des assimilations). Cela donne une forme de surface, en transcription.

Voyons un exemple : le verbe de racine PRS à la voie IV-3, inaccompli, troisième personne du pluriel, masculin.

segment	1	2	3	4	5	6	7
racine voyelles aspectuelles infixes		 ntn	p	r a	s		
forme intermédiaire 1	i	ntn	p	ra	s	ū	
forme intermédiaire 2	i	ntana	p	ra	s	ū	
forme transcrite	i	ttana	p	ra	s	ū	

Un autre système relie la transcription à la forme de surface en cunéiforme. Ce nouveau système segmente différemment les formes, selon un critère non plus de structure morphologique, mais de structure graphique : chaque caractère cunéiforme définit un segment. Cette segmentation est incomptible avec la précédente. Par exemple, le redoublement de la seconde racine, caractéristique de la voie II, quand elle est notée dans l'écriture, l'est toujours dans deux segments différents alors qu'il s'agit du même élément morphologique. Ce système comporte trois rubans : un pour le caractère cunéiforme, un pour la transcription, un ruban intermédiaire contient une interprétation des caractères parmi toutes leurs valeurs possibles.

4 Etat actuel de l'analyseur

La première partie de l'analyseur qui relie via deux étapes intermédiaire des informations lexicales à une forme transcrite est opérationnelle sous forme d'un transducteur à 7 rubans compilé en un automate fini. Nous n'avons pas de lexique de racine complet. La description n'est pas parfaite, mais le travail qui reste à faire relève de la finition et de la mise au point.

La seconde partie qui va de la transcription au cunéiforme est à l'état d'ébauche. Nous n'avons pas encore de répertoire exhaustif des caractères et de leurs valeurs. Nous ne traitons pour le moment que des valeurs syllabiques.

Le plus gros problème qui demeure est de nature technique : nous ne savons pas relier les deux parties du système à cause de la différence entre leurs façons de segmenter les formes. Nous cherchons à définir une opération de composition entre transducteurs à partitions ayant des partitionnements différents.

De façon générale, le travail sur la morphologie de l'akkadien nous a conduit à travailler sur les formalismes réguliers de description morphologique et les méthodes de compilation de ces formalismes en machines finies. C'est un exemple intéressant qui pose des questions quand à l'utilisation des rubans et des partitionnements.

Les prochaines étapes concernant le traitement de l'akkadien sont la complétion de la description du cunéiforme, puis le traitement des verbes faibles et enfin, des noms et adjectifs.

Références

[1] George Anton Kiraz. 2001. *Computational Nonlinear Morphology.* Cambridge University Press.