Un analyseur morphologique multi-niveaux utilisant la jointure

François Barthélemy

CNAM, Laboratoire Cédric
INRIA, Projet Atoll
barthe@cnam.fr

Résumé

Dans cet article nous présentons un analyseur morphologique pour le verbe akkadien. Cette langue est de la famille des langues sémitiques. Les flexions du verbe font intervenir des changements internes à la racine. L'analyseur présenté ici illustre l'utilisation d'un formalisme multi-niveaux et d'opérateurs relationnels puissants, notamment la jointure. La multiplicité de niveaux intermédiaires entre les formes profondes et de surface, ainsi que les opérateurs de compositions permettent de diviser la description en contraintes relativement simples qui sont ensuite rassemblées pour s'exercer soit simultanément, soit en cascade, soit encore d'une façon mixte, c'est-à-dire simultanément pour certains des niveaux et en cascade pour d'autres. Ce mécanisme nous permet de décrire la vocalisation du radical comme un processus d'insertions successives de voyelles. Cela présente l'intérêt d'être plus simple que l'utilisation d'un schéma vocalique figé soumis à interdigitation. De plus, cela semble expliquer de façon plus économique les formes des verbes faibles.

Mots-clés : analyseur morphologique, morphologie à deux niveaux.

Abstract

In this paper, we present a morphological analyzer for the Akkadian verb. This language belongs to the semitic family. Verb inflection involves modifications within the root. The analyzer is an example of the use of a multi-level formalism having powerful relational operators, notably the join. The morphology is described using relatively simple constraints which are composed together in three possible ways : they apply simultaneously or sequentially, or even in a mixed way, where the constraints apply simultaneously on some levels and sequentially on others. This mechanism is used to describe the vocalization of verbal roots as successive vowels insertions. This process is simpler than the use of fixed vocalic patterns which implies interdigitation. Moreover, it conveniently explains the forms of weak verbs.

Keywords: morphological analyzer, two-level morphology.

1. Introduction

L'akkadien est une langue morte, utilisée dans l'antiquité en Mésopotamie et dans tout le Proche-Orient. Elle est connue par une grande quantité de documents, notamment des centaines de milliers de tablettes d'argiles. C'est une langue sémitique qui comporte une morphologie verbale riche et complexe. En plus de combinaisons d'affixes figés, la racine verbale est l'objet de transformations internes (adjonction de consonnes, changement des voyelles).

Nous utilisons la notion de marque morphologique, une unité plus petite que le morphème, pour décrire la formation du verbe de façon concaténative. Contrairement à beaucoup de descriptions de langues sémitiques, nous n'utilisons pas la notion de schéma vocalique figé, mais

un processus de vocalisation séquentiel au moyen de voyelles d'appui. Cette approche permet de simplifier la représentation des verbes faibles pour lesquels la disparition d'une consonne faible rend l'insertion de certaines voyelles d'appui inutile.

Nous utilisons un formalisme multi-niveaux relationnel partitionné. Nous poussons plus loin l'utilisation de niveaux multiples et de la notion de partition que les travaux publiés jusqu'ici. De plus, nous montrons pour la première fois une application à la morphologie de l'opération de jointure relationnelle (Kempe *et al*., 2004).

Nous allons d'abord présenter le formalisme que nous utilisons, puis nous ferons un bref survol de la morphologie verbale akkadienne. Puis nous présentons l'analyseur morphologique, en insistant sur deux points-clés : l'utilisation de la jointure et la représentation des verbes faibles qu'elle permet.

2. Morphologie à plusieurs niveaux partitionnée

Dans cette section, nous allons décrire brièvement le formalisme que nous utilisons pour décrire la morphologie du verbe akkadien. Ce formalisme a été développé spécifiquement pour cette application. Un prototype écrit en perl l'implémente en utilisant les boîtes à outils `Lextools` et `FSM` de AT&T[1], qui fournissent respectivement une implémentation des expressions régulières et des machines finies.

Notre outil se situe dans la filiation de la morphologie à deux niveaux fondée par (Koskenniemi, 1983). Cette approche est basée sur une description mettant en relation des formes abstraites (dites lexicales) avec des réalisations de surface, au moyen de règles contextuelles. Ces règles qui décrivent notamment des phénomènes phonologiques ou graphémiques, décrivent une correspondance terme à terme des composants des deux niveaux. La description est de nature relationnelle : elle peut se lire de l'abstrait vers le concret (génération) ou du concret vers l'abstrait (analyse). Elle est non-déterministe, si bien qu'une même réalisation peut correspondre à plusieurs formes lexicales et réciproquement. Les différentes règles décrivent des contraintes qui s'appliquent simultanément. Les systèmes de règles sont compilés en un transducteur fini à états.

Des travaux successifs dont le plus achevé est (Kiraz, 2001) ont étendu le modèle de la morphologie à deux niveaux de deux façons : d'abord, en permettant d'utiliser plus de deux niveaux. Au lieu de décrire des relations binaires, on décrit des relations n-aires avec n formes différentes. Ces formes peuvent être reliées par un modèle linéaire où des formes intermédiaires sont utilisées entre le concret et l'abstrait, permettant de diviser la description en étapes successives. Elles peuvent être reliée par un modèle en étoile où plusieurs dimensions abstraites orthogonales sont croisées pour obtenir une réalisation. Kiraz utilise notamment ce modèle en étoile pour décrire la morphologie du syriaque, avec un niveau abstrait qui se décline en trois dimensions : la racine consonantique, les affixes figés, le modèle vocalique.

La seconde extension de la morphologie à deux niveaux consiste à utiliser la notion de partition pour définir les correspondances termes à termes entre niveaux. Les règles de Koskenniemi définissent une correspondance caractère par caractère, avec éventuelle insertion d'un pseudo-caractère 0 lorsqu'il n'y a pas de correspondance effective avec un caractère de l'autre niveau. La morphologie partitionnée permet de mettre en correspondance non pas des caractères, mais

[1] Ces boîtes à outils sont diffusées gratuitement pour un usage académique à l'URL http ://www.research.att.com/sw/tools/lextools/.

des chaînes (éventuellement vides).

Par exemple, la correspondance entre la forme abstraite `in+possible+té` et la forme de surface `impossibilité` sera donnée comme suit :

règles à la Koskenniemi

i	n	+	p	o	s	s	i	b	0	l	e	+	t	é
i	m	0	p	o	s	s	i	b	i	l	i	0	t	é

partitions

in	posssible	té
im	possibili	té

On appelle partition, chacune des correspondances partielles `(in,im)`, `(possible,possibil)` et `(té,té)`.

Les règles de notre langage sont une extension des expressions régulières dont les composants de base sont des partitions. Voyons un exemple :

```
...<[tparf],t,ta>(<[infixe],_,_>|<[red],_,_>)...
```

Chaque partition est notée entre piquants < et >. Les points de suspension dénotent n'importe quelle séquence de partitions, éventuellement vide. `[tparf]` est un symbole composé de plusieurs caractères. `[infixe]` et `[red]` sont deux classes de symboles. Leur utilisation dans une expression régulière est équivalente à la disjonction des membres de la classe. Dans d'autres règles, nous utilisons des classes telles que `[voyelle]`, `[consonne]`, etc. Le caractère `_` (soulignement) est utilisé pour représenter n'importe quelle chaîne à l'intérieur d'une partition. L'exemple que nous donnons est un système à trois niveaux, avec un niveau abstrait, un niveau intermédiaire et un niveau concret. Il peut se lire : *à l'intérieur d'une forme donnée, un infixe de parfait (*`[tparf]`*) réalisé au niveau intermédiaire par un* `t` *se réalise concrètement en* `ta` *s'il est suivi d'un infixe ou d'une gémination (*`[red]`*).*

Les expressions régulières utilisent des partitions qui sont implicitement choisies dans un ensemble fini préalablement déclaré. Par exemple, si nous écrivons `<_,[lettre]*a,_>`, cela ne dénote pas l'ensemble infini des partitions dont le second membre est une chaîne de lettres terminée par un `a`, mais le sous-ensemble fini des partitions appartenant à l'ensemble des partitions possibles et qui ont un second composant se terminant par un `a`.

Des règles contextuelles peuvent être utilisées, qui ne sont que des raccourcis syntaxiques pour écrire des expressions régulières. Un contexte est une description de ce qui peut entourer une partition. Il se divise en contexte gauche et contexte droit, chacun d'eux étant donné sous forme d'une expression régulière de partitions. On distingue classiquement trois types de règles :
– règle de restriction de contexte : une règle de ce type permet d'exprimer qu'une certaine partition n'apparaît que dans certains contextes ;
– règle de coercion : elle permet d'exclure certaines réalisations pour un certain schéma de partition, dans certains contextes ;
– une règle composite est une règle qui cumule les deux types de contraintes (restriction de contexte et coercion).

Les expressions régulières sont compilées indépendamment en transducteurs multi-bandes partitionnés. (Barthélemy, 2005b) et (Barthélemy, 2005a) décrivent les techniques de compilations employées, qui sont dérivées des techniques classiques de (Kaplan et Kay, 1994), ainsi que des opérations permettant de composer ces transducteurs. Chaque transducteur décrit une relation n-aire régulière partitionnée. Les opérations ensemblistes (union, intersection, différence) ainsi que les opérations de chaîne (concaténation, clôture sous concaténation) sont définies pour les transducteurs, de façon cohérente par rapport aux notions correspondantes sur les relations.

Deux opérations plus inhabituelles viennent compléter les calculs sur les transducteurs. Il s'agit

de la projection et de la jointure. Ces deux opérations sont les opérations de l'algèbre relationnelle utilisée classiquement dans le domaine des bases de données. La projection consiste à *oublier* certains composants d'une relation n-aire. Par exemple, si l'on projette la relation ternaire définie par l'exemple donné ci-dessus sur les première et dernière composantes (oubli de la seconde composante), on obtient une relation binaire reliant directement le niveau abstrait et le niveau de surface.

L'opération de jointure est plus complexe. Elle consiste à identifier certaines des composantes de deux relations différentes, pouvant être d'arités différentes. Le résultat a généralement une arité plus grande que ses opérandes. Deux cas particuliers de jointure correspondent aux opérations d'intersection et de composition de transducteurs. L'intersection est la jointure qui identifie toutes les composantes de même rang de deux machines ayant la même arité. La composition est une jointure dans laquelle la dernière composante du premier opérande est identifiée avec la première de la seconde composante. Mais contrairement à la composition de transducteurs, les composantes identifiées ne disparaissent pas du résultat.

La jointure est l'opération centrale qui permet de composer deux descriptions partielles en une seule machine finie. Par exemple, en faisant la jointure sur une composante de deux descriptions à deux niveaux, on obtient une machine unique à trois niveaux.

Notre langage permet d'utiliser les opérations ensemblistes, de chaîne, la projection et la jointure pour créer de nouvelles machines à partir de celles que nous décrivons par des expressions régulières. Un fichier source comprend plusieurs parties : des déclarations, notamment l'alphabet et les structures de traits utilisées. Puis des ensembles d'expressions régulières, chaque ensemble étant structuré autour d'un ensemble fini de partitions possibles et d'une déclaration des ordres permis pour les partitions. Enfin vient la description des machines finies calculées au moyen des expressions régulières et des opérateurs fournis.

3. Morphologie du verbe akkadien

La morphologie verbale en akkadien est complexe. Comme pour les autres langues sémitiques, une petite partie de la flexion suit un schéma concaténatif classique, avec des préfixes et suffixes figés, marquant notamment le nombre, le genre, le cas et le mode. Mais la plus grande partie du système consiste en des transformations internes de la racine. La racine est composée fondamentalement d'un jeu de consonnes, généralement au nombre de trois, appelées *consonnes radicales*. Des modifications interviennent pour composer sur cette base un élément morphologique que nous appellerons *base verbale*. Ces modifications sont les suivantes : ajout d'un affixe avant ou après la première consonne, gémination d'une consonne radicale, insertion d'une voyelle pour vocaliser la première ou la seconde radicale.

Le système verbal se décompose en sous-systèmes appelés *voies* (ou *schèmes*) selon deux axes indépendants, caractérisés par des éléments de nature morphologique. Le premier axe distingue quatre cas, identifiés par des chiffres romains :

– voie I (ou *G*) : il s'agit de la forme simple où la base verbale n'a pas de consonne ajoutée en initiale et où il n'y a pas de gémination systématique ;
– voie II (ou *D*) : la seconde radicale est géminée et la première voyelle de la forme verbale (que ce soit dans un préfixe ou dans la base verbale) prend la couleur `u` ;
– voie III (ou *Š*) : la consonne `š` est ajoutée à l'initiale de la base verbale et la première voyelle de la forme prend la couleur `u` ;
– voie IV (ou *N*) : la consonne `N` est ajoutée à l'initiale de la base verbale.

Le second axe distingue trois cas, selon la présence ou l'absence d'un infixe dans la base verbale. Cet infixe est inséré après la première consonne de la base verbale, qui est selon le cas une consonne ajoutée (voies III et IV), ou la première radicale (voies I et II).
– voie 1 : pas d'infixe ;
– voie 2 : infixe `t` ;
– voie 3 : infixe `tan`.
N'importe quelle voie du premier axe peut se croiser avec n'importe quelle voie du second axe, sauf la combinaison des cas IV et 2 qui n'existe pas.

Si la définition du système de voies est facile à décrire en termes de structure morphologique, il est plus difficile d'en décrire le sémantisme. La voie IV sert à exprimer le passif de la voie I pour les verbes transitifs. Les voies II et III fonctionnent en parallèle. Pour une racine donnée, souvent une seule des deux est utilisée, et donne un sens particulier à cette racine. Ces deux voies ont souvent une connotation factitive. La voie II peut aussi exprimer la multiplicité des actions ou des objets. Souvent, le verbe en voie II ou III a simplement un autre sens qu'en voie I. La voie 1 est le système de base. La voie 3 exprime les fonctions itérative et habituelle. La voie 2 exprime le passif pour les voies II et III. Pour la voie I, dont le passif est exprimé par la voie IV, la voie 2 exprime la réciprocité, la réflexivité.

Au sein de chaque voie, il existe trois formes nominales (participe actif, adjectif et infinitif), une forme pronominale (permansif) et quatre formes verbales (accompli, inaccompli, parfait, impératif). Nous utiliserons le terme de *temps* bien qu'il soit un peu impropre puisque l'opposition accompli/inaccompli est plus aspectuelle que temporelle. Le parfait est utilisé dans certaines propositions subordonnées. Les formes nominales ont une morphologie nominale, avec un suffixe marquant le genre, le nombre et le cas.

Les différentes oppositions du système des temps sont marquées par des moyens différents selon les voies. Par exemple, l'opposition accompli/inaccompli est marquée par une gémination de la seconde radicale aux voies I et IV et par une vocalisation (`a` ou `i`) aux voies II et III. Un autre exemple : le parfait est toujours caractérisé par l'ajout d'un infixe `t`, mais cet ajout se fait à une base verbale bâtie sur la forme de l'inaccompli aux voies I et IV et sur la forme de l'accompli aux voies II et III. La vocalisation de l'accompli et de l'impératif suit trois schémas distincts selon le découpage suivant : voie I.1 ; voies I.2, I.3 et IV.3 ; voies II, III et IV.1. Pour le participe actif, il y a deux types de marques : une voyelle longue `ā` à la voie I.1 ; un préfixe figé `mu` aux autres voies. Ces quelques exemples illustrent la variabilité des marques et le fait que les voies se découpent en différents sous-ensembles selon les cas.

Une même marque morphologique peut être utilisée pour différents usages. Par exemple, l'infixe `t` peut marquer une voie 2 ou le temps du parfait. Il peut être redoublé dans le cas d'un parfait en voie 2. La gémination de la seconde radicale peut noter le temps inaccompli ou la voie II. Une voyelle `i` après la deuxième radicale peut être significative (par exemple, elle caractérise l'accompli aux voies II et III) ou dépourvue de signification, comme par exemple pour le participe actif où elle est une simple voyelle support insérée quand une voyelle est nécessaire pour des raisons graphiques.

L'analyse d'une forme nécessite souvent la prise en compte de plusieurs marques différentes, à la fois positives et négatives, c'est-à-dire permettant de réfuter une ou plusieurs alternatives. Prenons un exemple. La forme `uparris` est un accompli à la voie II.1. La forme est marquée par le `u` initial (un préfixe marquant la personne, originellement `i` ou `a` ayant changé de couleur pour marquer la forme) et par la gémination du `r`. Le `u` seul pourrait également marquer une voie III. La gémination pourrait marquer un inaccompli. Une fois la voie établie, la voyelle finale

`i` marque un accompli. Cette voyelle pourrait parfaitement se retrouver à cette même position pour un inaccompli à la voie I (la forme serait `iparris`[2]). De même, le `i` pourrait apparaître dans un participe actif, mais alors le préfixe `u` ne pourrait pas exister.

En résumé, la morphologie utilise un certain nombre de marques : préfixes, suffixes, augment de la base verbale préfixe ou infixe, vocalisation de la base verbale, mais une analyse nécessite la prise en compte de plusieurs marques pour déterminer un même trait morphologique. Notons qu'il existe une ambiguïté résiduelle, certaines formes étant identiques et ne pouvant se distinguer que par le contexte (par exemple le parfait en voie II.1, l'accompli en voie II.2 et l'accompli en voie II.3, `uptarris`).

4. Structure de l'analyseur morphologique

La description abstraite d'une forme verbale se fait au moyen de traits morphologiques prenant leur valeur dans des domaines finis. Ces traits sont les suivants : la personne, le genre, le nombre, la voie, le temps (ou la forme nominale), le cas (ou l'état construit), le mode, la racine représentée par un trait pour chaque radicale, la classe vocalique. Les deux derniers éléments relèvent du lexique, les autres relevant de la flexion. Comme nous l'avons dit, il serait vain de tenter d'établir une relation univoque entre ces traits et des marques morphologiques et donc d'ordonner ces traits en fonction de l'ordre des marques dans la forme. Nous avons donc fixé un ordre conventionnel arbitraire entre ces traits, ce qui définit le niveau lexical de notre analyseur.

Entre le niveau lexical et le niveau de surface, nous utilisons un niveau intermédiaire fondé sur la notion de marque morphologique. Une marque morphologique est une séquence de caractères contigus dans tous les contextes où elle apparaît. Une même marque peut connaître des variations de surface. Par exemple, l'infixe `tan` peut se réaliser en `ta, tan, tana, tar` et d'autres séquences où le `n` s'assimile à la consonne qui le suit. Nous utilisons dans l'analyseur une notion de *type de marque* qui consiste en une distinction entre marques selon leurs relations avec les traits morphologiques. Par exemple, l'infixe `t` correspondra à deux types de marques selon que cet infixe marque la voie ou le parfait.

Les types de marques utilisés sont les suivants :
– des affixes figés : préfixe personnel, préfixe de participe actif, suffixe personnel, suffixe pronominal, suffixe de mode ;
– des composants de base verbale : première, deuxième et troisième radicale, infixe de parfait, infixe `t` de voie, infixe `tan`, préformante `š`, préformante `n`, gémination de voie, gémination de temps, voyelle signifiante 1 et 2.

Ce sont ces types de marque, dans l'ordre d'apparition dans la forme verbale, qui constituent le niveau intermédiaire de l'analyseur. Voici un exemple qui montre les trois niveaux d'analyse d'une forme verbale.

lexical	masculin singulier 3 voie III.3 inaccompli indicatif racine prs						
intermédiaire	préfixe personnel	préformante š	infixe tan	première radicale	seconde radicale	voyelle 2	troisième radicale
surface	u	š	tana	p	r	a	s

Figure 1. Un exemple d'analyse

[2] En fait cette vocalisation ne serait pas possible avec la racine `prs`, mais elle le serait avec d'autres racines comme par exemple `pqd`.

Notons qu'il y a une correspondance terme à terme entre les deux niveaux du bas, entre chaque type de marque et sa réalisation, alors que le niveau supérieur concerne l'ensemble de la forme.

Le passage du niveau intermédiaire au niveau de surface se fait en plusieurs étapes successives, en utilisant d'autres niveaux intercalés. D'abord, chaque type de marque est transcrit dans une forme canonique qui est l'affixe figé dans son ensemble ou une forme minimale pour les éléments de la base verbale. Ensuite, des voyelles d'appui sont insérées dans la base verbale suivant des règles de nature composite, morphologique, phonologique et graphique. Enfin, des transformations de surface telles que l'assimilation, la dissimilation, la contraction, sont appliquées. Chacune des étapes peut être conditionnée par les traits morphologiques du niveau abstrait.

5. Vocalisation de la base verbale

La principale difficulté de la morphologie du verbe akkadien vient des voyelles de la base verbale. Suivant la proposition de (Malbran-Labat, 2001), nous distinguons trois types de voyelles :
– les voyelles significatives qui dépendent de la racine verbale : nous les appellerons *voyelles catégorielles* ;
– les voyelles significatives qui ne dépendent pas de la racine verbale. Ce seront les *voyelles temporelles* ;
– les voyelles non significatives, que nous dénommerons *voyelles d'appui*.

Une voyelle est considérée comme significative si elle est utilisée pour déterminer un ou plusieurs traits morphologiques. Autrement dit, si elle intervient dans l'analyse. Une voyelle d'appui n'a pas d'autre rôle que de permettre d'énoncer ou d'écrire une forme qui sans cela contiendrait trop de consonnes. Une voyelle significative constitue à elle seule une marque morphologique. Il existe deux types de marques, le premier pour les voyelles vocalisant la première radicale, le second pour les voyelles vocalisant la seconde radicale. Les voyelles d'appui quant à elles apparaissent dans notre analyse comme faisant partie de la réalisation de leur consonne support.

Dans l'analyseur, les voyelles significatives apparaissent d'entrée alors que les voyelles d'appui sont ajoutées dans un second temps. Les règles d'insertion des voyelles d'appui sont difficiles à exprimer au moyen de contraintes simultanées alors qu'elles s'expriment naturellement au moyen de contraintes s'exerçant en séquence.

Le verbe akkadien ne peut pas comprendre une séquence de plus de deux consonnes consécutives. Toute séquence de trois consonnes ou plus doit être scindée par insertion d'une ou plusieurs voyelles d'appui. Le choix de la consonne vocalisée (le première ou la seconde de la séquence de trois) dépend du type de marque de la consonne. Un infixe `t` doit être vocalisé en priorité, ou en l'absence d'un tel infixe, la préformante š. Ensuite viennent la première radicale, la préformante N et la seconde radicale. Cet ordre de priorité est implémenté par une séquence de règles contextuelles appliquées en séquence. La première règle insère un `a` après un infixe `t` si celui-ci apparaît dans une séquence de 3 consonnes. La deuxième règle réalise l'insertion d'un `a` après la préformante š si celle-ci est dans une séquence de trois consonnes ou plus. Cette deuxième règle étant appliquée après la première, elle n'a pas à se préoccuper de la présence ou non d'un infixe `t` dans la séquence.

Les règles que nous évoquons font référence à deux notions : d'une part le type de marque ; d'autre part la réalisation des marques. Par exemple, la vocalisation d'un š ne sera pas la même si ce š est une préformante ou une radicale. C'est le type de marque qui détermine le choix de la consonne vocalisée. D'un autre côté, l'application de la règle est déterminée par l'existence d'une séquence de trois consonnes, qui se détecte dans une réalisation de la séquence de mar-

ques. Les règles sont appliquées en cascade pour ce qui est des réalisations et en parallèle pour ce qui est des séquences de marques.

Ce sont des règles à trois niveaux utilisant une opération de jointure. Voyons concrètement le cas des deux premières règles.

```
<[infixe],t,_> <=> <[infixe],t,ta> /
    <_,([lettre]*)[cons],_> __ <_,[cons]([lettre]*),_>
    <_,([lettre]*)[voy],_> __ <_,[cons],_><_,[cons]([lettre]*),_>
<[preff],[sh],_> <=> <[preff],[sh],[sh]a> /
     <_,([lettre]*)[voy],_> __ <_,[cons],_><_,[cons]([lettre]*),_>
```

Ces règles sont de type composite (à la fois restriction de contexte et coercion). La première ligne de la règle donne un motif et sa réalisation nécessaire et exclusive. Ensuite viennent une ou plusieurs lignes décalées qui spécifient chacune un contexte dans lequel la règle s'applique. Chaque règle décrit une relation ternaire, avec en première composante le type de marque, en seconde composante, la forme avant insertion, et en troisième composante, la forme après éventuelle insertion (si le contexte l'exige). L'application en cascade consiste à identifier la première composante de chaque règle et la troisième de la première règle avec la seconde de la deuxième règle (cf. figure 2). Il s'agit d'une opération de jointure $\bowtie_{1=1,3=2}$. Le résultat de cette opération est une relation d'arité 4.

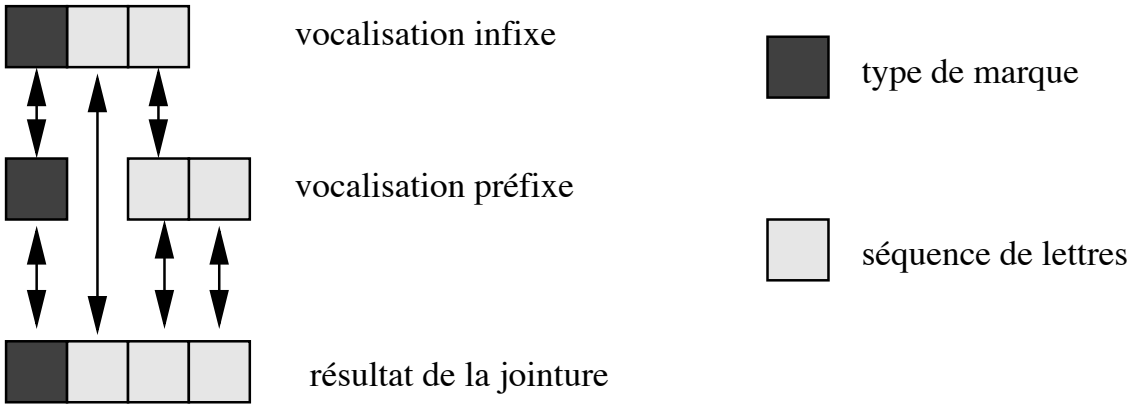

Figure 2. Opération de jointure

6. Traitement des verbes faibles

Les verbes faibles sont des verbes dont une (ou plusieurs) des consonnes radicales sont faibles, c'est-à-dire qu'elles ont tendance à disparaître des formes de surface. Les consonnes faibles en akkadien sont 'alif[3](noté '), y et w. Selon le contexte, une telle consonne faible apparaîtra en tant que consonne (ex : `uša"il`, racine š'l), disparaîtra complètement (`*'epiš>epiš`, racine 'pš), sera assimilée à une consonne adjacente (ex : `uše'peš>ušeppeš`, racine 'pš) ou laissera comme trace de sa présence un allongement d'une voyelle adjacente (ex. : `*i'puš>īpuš`, racine 'pš). Les verbes faibles sont très fréquents en akkadien : de l'ordre de la moitié des formes verbales d'un texte sont faibles.

[3] Le 'alif akkadien correspond à 5 consonnes distinctes du proto-sémitique : 'alif, 'ayn, h, ḫ et ǵ.

Le traitement des verbes faibles qui est souvent proposé consiste à dériver la forme faible de la forme de surface correspondante des verbes forts. Nous proposons un autre traitement, qui consiste à considérer que la disparition des consonnes faibles rend inutile l'insertion de certaines voyelles d'appui. En revanche, les voyelles significatives sont systématiquement préservées par les verbes faibles. Dans la plupart des cas, la consonne faible disparaît sans laisser de trace ou en allongeant une voyelle existant indépendamment. Dans quelques cas, la consonne faible se réalise en une voyelle dont elle dicte la couleur.

Notre traitement des verbes faibles fonctionne relativement bien : la plupart des cas sont traités de façon satisfaisante par la règle simple d'allongement de la voyelle adjacente. Une autre règle permet une disparition sans trace de la faible dans certains contextes inter-consonantiques. Il reste enfin quelques exceptions traitées par des règles spécifiques.

Dans notre modèle, la disparition d'une consonne faible peut s'effectuer à chaque insertion de voyelle d'appui. Cela signifie que pour chaque étape d'insertion, il faut un traitement des verbes faibles. Ce traitement est à peu près le même à toutes les étapes. Prenons quelques exemples.

Le parfait voie III.1 du verbe `epēšu` se construit par étapes :
– marques morphologiques : préfixe personnel + préfixe š+ infixe t + radicale 1 + radicale 2 + voyelle temporelle + radicale 3 ;
– forme sans voyelle d'appui : `ušt'piš` ;
– il y a une séquence de plus de 3 consonnes `št'p` comportant un infixe : la première étape de vocalisation insère un `a` d'appui après le `t`. Résultat : `ušta'piš` ;
– il y a maintenant une voyelle adjacente à la faible qui peut disparaître et allonger cette voyelle. Résultat : `uštāpiš` ;
– les quatre autres étapes de vocalisation n'ajoutent pas de voyelles parce qu'il n'y a plus de séquence de 3 consonnes ;
– une règle phonétique impose le changement de couleur du `a` en `e`. Résultat : `uštēpiš`. C'est la forme de surface ;
Pour un accompli voie III.1 du même verbe, le traitement est comparable, mais la première étape de vocalisation n'ajoute pas de voyelle parce qu'il n'y a pas d'infixe dans la séquence de trois consonnes qui persiste (le forme est `uš'piš`). C'est la deuxième étape qui vocalise le préfixe š, ce qui introduit la voyelle à allonger (dérivation `uša'piš>ušāpiš>ušēpiš`).

7. Conclusion

Le système actuel regroupe 52 expressions régulières dont 16 règles contextuelles. Il décrit de façon adéquate les verbes forts trilitères. La description des verbes faibles est moins achevée. Il est d'ailleurs très difficile d'obtenir un résultat parfait compte tenu de la variabilité de certaines formes faibles. Plus qu'un système d'analyse exhaustive des formes apparaissant dans les textes, nous souhaitons avoir une description normative analogue à celle des grammaires courantes de la langue.

Le système dans sa forme la plus développée comporte 17 niveaux, dont un utilisé pour représenter des structures de traits. Il est obtenu par 14 jointures successives de descriptions à deux ou trois niveaux. Le pivot du système est le niveau qui décrit une forme comme une séquence de types de marques. Ce niveau intervient dans 12 des 14 jointures. Le système est compilé en un transducteur fini à 740 000 états. Il est non-déterministe, ce qui lui permet de modéliser l'ambiguïté de certaines formes verbales.

L'exemple que nous avons décrit dans cet article montre la puissance de la morphologie multi-

niveaux partitionnée. Celle-ci peut apporter un plus grand confort d'utilisation que les formes traditionnelles de morphologie à états finis (morphologie à deux niveaux et systèmes de réécriture). L'emploi d'un système à partition se justife particulièrement dans des cas de mise en correspondance de représentations très différentes, pour lesquels une correspondance terme à terme est arbitraire. Par exemple, lors de la mise en correspondance d'un morphème et d'une structure de trait, il n'y a pas lieu d'établir une correspondance entre la première lettre et le premier trait. Un autre exemple : le passage d'une écriture conventionnelle à une écriture phonétique. Plus l'écriture est éloignée de la phonétique, plus la notion de partition est intéressante. C'est le cas notamment des écritures idéographiques et syllabiques.

Références

BARTHÉLEMY F. (2005a). « Multi-Partition Finite-State Morphology ». In *soumission en cours*.

BARTHÉLEMY F. (2005b). « Partitioning Multitape Transducers ». In *International Workshop on Finite State Methods in Natural Language Processing (FSMNLP)*. Helsinki, Finland.

KAPLAN R. M. et KAY M. (1994). « Regular Models of Phonological Rule Systems ». In *Computational Linguistics*, 20 :3, 331–378.

KEMPE A., CHAMPARNAUD J.-M. et EISNER J. (2004). « A note on join and auto-intersection of n-ary rational relations ». In *B. Watson and L. Cleophas, editors, Proc. Eindhoven FASTAR Days*. Eindhoven (Netherlands).

KIRAZ G. A. (2001). *Computational Nonlinear Morphology*. Cambridge University Press.

KOSKENNIEMI K. (1983). « Two-Level Model for Morphological Analysis ». In *IJCAI-83*. Karlsruhe, Germany, p. 683–685.

MALBRAN-LABAT F. (2001). *Manuel de langue akkadienne*. Publications de l'institut Orientaliste de Louvain (50), Peeters.