Cunéiforme et SMS: analyse graphémique de systèmes d'écriture hétérogènes

François Barthélemy
CNAM [1], INRIA [2]

Abstract

Dans cet article, nous établissons une analogie entre les systèmes d'écriture SMS et cunéiforme qui sont deux systèmes hétérogènes, mêlant un sous-système phonétique à d'autres sous-systèmes, notamment idéographiques. Nous partons de cette analogie pour proposer l'adaptation aux SMS de techniques de transcription développées pour l'écriture cunéiforme.

Keywords : Cunéiforme, SMS, morphologie à partition, transcription, translittération.

1. Introduction

Les courts messages de téléphonie mobile ou SMS offrent une forme de communication relativement récente qui se caractérise paradoxalement par des contraintes et une liberté remarquables. Contraintes venant de la spécification technique (taille des messages, caractères disponibles) et des interfaces de saisie (clavier petits avec peu de touche) et liberté venant d'un registre de communication familier, supportant les codes d'un groupe restreint.

Un courant d'étude s'est intéressé depuis plusieurs années aux nouvelles formes de productions linguistiques liées au téléphone mobile ou à internet (Neef & Veronis 2004), (Anis 1999). Récemment, un corpus de SMS significatif en langue française a été mis à la disposition des chercheurs (Fairon *et al.* 2007).

Une bonne partie de la spécificité des SMS provient de procédés d'écriture. Nous allons adopter une approche volontairement réductionniste en ne considérant dans cet article que cet aspect des choses. Nous allons supposer que les SMS sont une façon différente d'écrire une forme relâchée, familière et proche de l'oral de la langue française. Le but d'un système d'analyse automatique sera dès lors de transcrire les SMS en écriture habituelle et de transmettre le résultat à un système d'analyse du français.

Il s'agit donc de décrire l'écriture utilisée dans les SMS. Il est légitime de la comparer avec d'autres écritures. Parmi les différentes comparaisons possibles, celle avec l'écriture cunéiforme utilisée pour écrire l'akkadien (langue des anciens babyloniens et assyriens) nous semble particulièrement pertinente. Ce sont deux écritures hétérogènes qui comportent des sous-systèmes différents. SMS et cunéiforme mélangent une notation phonétique avec une notation conventionnelle, parfois au sein d'une même forme fléchie. Dans les deux cas, les caractères utilisés ont une pluralité d'emplois, ce qui se traduit par de nombreuses ambiguïtés et une multiplicité de graphies pour une même forme.

Nous allons utiliser cette analogie pour faire des propositions pour la transcription automatique des SMS. Elles découlent de la modeste expérience que nous avons acquise dans le traitement

[1] Conservatoire National des Arts-et-Métiers (Paris, France), barthe@cnam.fr
[2] Institut National de Recherche en Informatique, projet ATOLL (Rocquencourt, France)

des textes cunéiformes. Ces propositions ne résolvent pas, loin s'en faut, toutes les difficultés qui se présentent. Elles permettent de séparer en deux les ambiguïtés d'écriture et de reconnaître les formes utilisant plusieurs systèmes graphiques successivement.

Nous allons d'abord présenter brièvement chacune des deux écritures, puis insister sur leurs analogies pour finir avec l'esquisse d'un analyseur graphémique sous forme d'automate fini.

2. Présentation de l'écriture cunéiforme

Notre présentation de l'écriture cunéiforme concerne sa version akkadienne dans son état standard (Labat & Malbran-Labat 1948). Rappelons que l'écriture cunéiforme a été inventée vers -3300 par les sumériens pour noter leur langue agglutinante, puis adaptée pour d'autres langues dont l'akkadien et l'éblaïte (sémitiques), le hittite, le louvite (indo-européen), l'élamite et le hourrite (deux autres familles linguistiques distinctes). L'akkadien a été écrit environ de -2300 à l'an 0, avec de nombreuses différences dialectales et graphiques.

L'écriture cunéiforme utilise des caractères du genre de ceux de la figure 1 [1] et mélange deux systèmes graphiques distincts : un système phonétique de type syllabique et un système idéographique. Un système annexe de classification sémantique vient faciliter la lecture de certains mots.

FIG. 1: exemple d'écriture cunéiforme

Le système phonétique associe à chaque signe qui l'utilise le son d'une syllabe et peut être utilisé dans toute forme contenant cette syllabe. Les syllabes dénotées ont l'un des motifs consonnes-voyelles suivants : V, CV, VC ou CVC. Le système autorise de noter une même voyelle dans deux signes consécutif. Il autorise également de ne noter qu'une fois deux consonnes identiques successives. Par exemple **iprus** (il sépare) peut s'écrire avec les découpages suivants : **ip-ru-us** ou **ip-rus**. **Uparris** (il fait séparer) peut se scinder en : **u-pa-ar-ri-is, u-par-ris, u-par-ri-is, u-pa-ar-ris, u-pa-ri-is** ou **u-pa-ris** [2].

A cette liberté dans le découpage s'ajoutent deux autres phénomènes qui complexifient le système. Un même signe a généralement plusieurs lectures phonétiques possibles (polyphonie). Il y a souvent une forme d'approximation qui fait qu'un signe note différentes syllabes de sonorités voisines (par exemple  **gab/qab/gap/qap**). Mais souvent le signe note des syllabes qui n'ont rien de commun. C'est ainsi que le même signe  note **ud, tam, tu, par, pir**, ainsi qu'encore d'autres valeurs plus ou moins rares. Autre facteur d'ambiguïté, la même syllabe peut être notée par plusieurs signes (homophonie), jusqu'à 18 pour **tu**. Pour un état donné de l'écriture (lieu et date), le nombre est heureusement bien plus restreint mais les syllabes les plus courantes peuvent néanmoins être notées de différentes façons.

Le système logographique consiste en une association entre une séquence de caractères et un lemme. Cette séquence peut être utilisée potentiellement pour noter n'importe quelle flexion du mot. Par convention, les valeurs phonétiques sont transcrites en minuscules et les logogrammes en lettres capitales.

Le système logographique n'est pas compositionnel en ceci que le mot associé à une séquence ne peut pas se déduire des valeurs associées à chacun des caractères pris isolément. Par exemple,

[1] En fait, l'écriture cunéiforme est un système graphique en trois dimensions, l'essentiel du corpus étant formé de tablettes d'argile gravées et d'inscriptions gravées dans la pierre. La profondeur des coins, la marque de l'instrument sont des éléments essentiels de la lecture, de même que les variations d'éclairage obtenus en manipulant l'objet ou la source de lumière. Ce qui est imprimé dans le présent article est une version normalisée d'une projection en deux dimensions de l'écriture cunéiforme.

[2] En revanche, un découpage **up-ar-ris** n'est pas possible : la consonne **p** doit nécessairement vocaliser la voyelle **a** suivante.

le logogramme  transcrit **SAG-GÉME-ARAD** dénote le mot akkadien **aštapiru**, serviteur. Les sous-chaînes de ce logogramme signifient quelque chose de différent ou rien du tout (cf. figure 2). Tout au plus peut-on accepter qu'il y a une vague association sémantique entre ces logogrammes, mais qui semble difficilement utilisable pour un traitement automatique.

séquence	mot dénoté	sens
SAG-GÉME-ARAD	aštapiru	serviteur
SAG	qaqqadu	tête, capitale
SAG-GÉME	amtu	femme esclave
GÉME-ARAD	n'existe pas	

FIG. 2: logogrammes composés de sous-séquences de SAG-GÉME-ARAD

Un logogramme peut représenter potentiellement toute flexion du lemme auquel il est associé, mais utilisé seul, il dénote une forme typique. Il est possible de préciser la forme réelle au moyen de compléments. Par exemple, le nombre typique est le singulier. Le logogramme seul note généralement le singulier. Pour noter un autre nombre (duel ou pluriel), il existe des logogrammes que l'on peut ajouter en fin de forme. Pour d'autres types de flexion (déclinaison, conjugaison), il est possible de compléter le logogramme avec un *complément phonétique*, c'est à dire une notation d'une partie caractéristique de la forme au moyen de signes phonétiques. Par exemple, le cas est noté par une voyelle en fin de forme. Le logogramme **mātu** (le pays) suivi du signe phonétique **ti** doit être lu **māti** (génitif). De même, le complément phonétique peut être utilisé pour désambiguïser parmi différents lemmes associés à un même logogramme. Par exemple le signe  signifie à la fois le dieu soleil (**UTU, šamšu**), le jour (**U_4, ūmu**), la couleur blanche (**BABBAR, pūṣu**). Un complément phonétique peut préciser quel sens choisir.

L'écriture comporte des caractères qui ont pour fonction d'indiquer le champs sémantique de la forme qui le suit ou qui le précède. On appelle cela un *déterminatif*. Par exemple  précède les noms divins et  les noms propres d'hommes. Dans la transcription, les déterminatifs sont notés surélevés par rapport à la ligne d'écriture, comme des exposants.

Il y a en tout trois fonctions possibles pour un caractère : phonétique, idéographique ou déterminative. Or les caractères ne sont pas spécialisés pour une fonction : ce sont les mêmes caractères qui tiennent les différents rôles. Par exemple le caractère  sert de déterminatif des noms divins, il possède les valeurs phonétiques **an** et **il** et les valeurs idéographiques **DINGIR**, (**ilu**, dieu) et **AN**, (**šamû**, ciel). Les multiples ambiguïtés sont généralement résolues par le contexte et cela fonctionne relativement bien.

Les différents systèmes peuvent cohabiter au sein d'une même forme, avec redondance, dans le cas des compléments phonétiques, ou sans redondance dans le cas de noms propres qui mêlent souvent idéogrammes et notations phonétiques, et parfois même déterminatif imbriqué. Un exemple d'écriture de nom propre est détaillé dans la figure 3.

3. Présentation de l'écriture SMS

Dans cette section nous allons faire un bref tour d'horizon de procédés d'écritures utilisés dans les SMS en nous appuyant essentiellement sur l'étude de Fairon, Klein et Paumier (Fairon *et al.*

caractère					
type	déterminatif	phonétique	phonétique	déterminatif	logogramme
valeur	*P*	niq_2	pa	*DINGIR*	IM
transcription		niq	pa		dû

FIG. 3: Graphie attestée du nom propre **niqpadû**

2007) dont nous tirons la plupart de nos exemples.

Le procédé d'écriture le plus fréquent consiste à utiliser l'écriture normale du français qui est un système conventionnel doté d'une orthographe permettant de distinguer les homophones. Ce système est partiellement phonétique, un son pouvant être associé à une séquence de caractères, avec de notables ambiguïtés dans les deux sens, un son pouvant être noté par différentes séquences (par exemple o, au, eau) et une même séquence pouvant avoir plusieurs prononciations (par exemple e dans femme, lemme, geler). Une autre caractéristique du système d'écriture est la présence importante de signes muets, non prononcés, notamment en fin de forme, marquant sa flexion. Nous appellerons ce système *l'écriture orthographique*.

Un deuxième procédé consiste à utiliser une écriture phonétique où chaque caractère ou séquence note un phonème. Par exemple **kel** pour **quelle**. Cette écriture peut parfois noter des liaisons.

Un autre système phonétique consiste à utiliser comme valeur phonétique le nom du caractère tel qu'on le prononce : **ka** pour **K**, **té** pour **T**. On peut par exemple écrire **GT**, **j'étais**. Ce système admet certaines approximations comme dans cet exemple les nuances entre différentes voyelles. Le signe **1** peut être utilisé aussi bien pour **un** que pour **in**. Nous appellerons ce système phonétique syllabique alors que le précédent sera dénommé phonétique alphabétique. Notons que ce système est très insuffisant pour noter l'intégralité de la langue française. Même si au prix d'approximation, l'on ne compte que 10 voyelles et 20 consonnes, il faudrait de l'ordre de 200 signes pour noter toutes les combinaisons consonne+voyelle, autant pour les combinaisons voyelle+consonne. On est loin du compte, d'autant que de nombreux caractères sont pour ainsi dire inutiles dans cet emploi (la ponctuation, y, w, etc).

Une valeur phonétique syllabique associé à un nombre peut concerner une séquence de plusieurs chiffres comme notamment 10 pour **dis**. On peut imaginer des emplois analogues de 20 ou 100. Dans ce dernier cas, il n'y aurait pas de gain de longueur par rapport à une notation orthographique. Mais on sait que le principe d'économie n'est pas le seul guide de l'auteur de SMS. Quoi qu'il en soit, nous classerons ces séquences parmi les procédés de l'écriture phonétique syllabique.

L'écriture syllabique tend à intégrer les pronoms personnels **j'** et **t'** dans le verbe, comme par exemple dans **GT** pour **j'étais**.

Autre procédé, une écriture de type idéographique où un signe ou une séquence de signes note un mot, comme @ pour oreille.

Les smileys notent une information qui est de nature para-verbale, qui vient commenter l'information textuelle qui l'accompagne. La question se pose de la transcription qui doit en être faite. D'un point de vue graphique, on peut les considérer comme des séquences figées de caractères.

Un autre système consiste à ne noter que les consonnes d'une forme. Par exemple **bcp** pour **beaucoup**. Ce système d'écriture a été utilisé pour certaines langues, notamment sémitiques. Nous appellerons ce système *écriture consonantique*.

L'attribution de certaines formes à l'un ou l'autre des systèmes n'est pas toujours évident. Par exemple le signe **+** est-il un idéogramme désignant le mot **plus** ou un signe de l'écriture syllabique associé à la syllabe **plus** ? Si l'on trouve une forme comme **sur+** pour **surplus** la deuxième hypothèse se verra confirmer. On ne peut pas exclure non plus que le signe ait les deux valeurs, donnant dans les cas où ce signe est isolé la même transcription avec deux interprétations graphiques différentes.

LOL et MDR sont-ils de la même nature non verbale que les smileys ou sont-ils plutôt des idéogrammes ? Un locuteur doit rarement articuler **Laughing Out Loud**, en revanche **Mort de Rire** semble plausible.

4. Analogies entre les deux écritures

Cunéiforme et SMS sont deux systèmes d'écriture très différents mais qui présentent certaines analogies pas si fréquentes, qui peuvent nous donner des pistes pour l'analyse graphémique.

Commençons par dresser un tableau de ces analogies.

– ce sont des systèmes graphiques hétérogènes regroupant plusieurs sous-systèmes différents.
– différents systèmes d'écriture peuvent être utilisés pour une même forme. C'est le cas pour les noms propres et les compléments phonétiques en cunéiforme. En SMS, une forme comme **pac** (pour **passé**) conjugue un début en écriture orthographique avec une dernière lettre en phonétique syllabique. De même les simplifications de consonnes doubles peuvent s'interpréter comme un passage local en phonétique alphabétique.
– les graphèmes ne sont pas, en général, dédiés exclusivement à un des systèmes. Il y a là une source d'ambiguïté.
– les sous-systèmes phonétiques syllabiques ont un caractère approximatif qui fait qu'un caractère couvre un champ de valeurs proches plus qu'une valeur.
– au sein d'un même système, certains caractères peuvent avoir plusieurs interprétations. Dans le cas du SMS phonétique, certaines valeurs d'origine anglo-saxonne sont utilisées, comme par exemple **2** pour **tou** ou **oo** pour **ou**.
– certains éléments de l'écriture ne s'insèrent pas dans la structure linéaire de la phrase. C'est le cas des déterminatifs en Akkadien et des smileys en SMS. Bien que de nature différente, ils apportent une information qui ne se traduit généralement pas par des mots.
– Il n'y a pas de façon unique d'écrire un mot, mais une pluralité de choix à la disposition de celui qui écrit.

Face aux multiples ambiguïtés relevées dans le système graphique, les assyriologues ont des habitudes de travail consistant à utiliser deux niveaux différents dans l'analyse des textes : celui de la translittération et celui de la transcription. La translittération consiste à décrire une lecture de chaque caractère parmi les différentes lectures possibles, celle qui est choisie dans l'interprétation proposée pour le texte. Cette lecture consiste en un choix du sous-système graphique (idéogramme, syllabique ou déterminatif) et un choix parmi les valeurs dans le sous-système.

De cette translittération, on peut déduire sans ambiguïté les caractères qui la composent, alors que dans l'autre sens, depuis les caractères, un choix est opéré.

Les valeurs syllabiques sont écrites en minuscule, les idéogrammes en majuscule, les déterminatifs en exposants. S'il y a plusieurs signes pour écrire une syllabe ou un idéogramme, ces différentes façons sont ordonnées par fréquence décroissante et un indice numérique précise le rang du caractère rencontré, avec par défaut la valeur 1. La figure 4 donne un exemple de translittération.

FIG. 4: exemple de translittération de caractères

La transcription propose une interprétation du texte où les éléments manquants (par exemple les consonnes redoublées ou la longueur des voyelles) sont restitués, les logogrammes sont remplacés par le mot akkadien qu'ils dénotent, les déterminatifs n'apparaissent pas, les voyelles artificiellement redoublées dans la graphie syllabique sont transcrites par une seule voyelle. La transcription est donc une texte akkadien écrit en alphabet latin étendu, selon les principes de grammaire généralement reconnus.

5. Propositions pour l'analyse graphémique des SMS

Dans cette section, nous allons esquisser un outil visant à transcrire du texte SMS en orthographe usuelle, et plus précisément l'échelon morpho-lexical de cet outil. L'hypothèse simpliste que nous faisons est que la langue notée dans les SMS est du français oral et donc, une fois le décryptage de l'écriture réalisé, des outils standards permettent d'analyser cette langue. Evidemment, il ne s'agit que d'une hypothèse de travail temporaire, permettant de progresser.

La base d'une transcription est nécessairement un lexique de formes fléchies parmi lesquelles il s'agit de *reconnaître* les formes écrites en langage SMS. Pour l'écriture orthographique, la re-

graphèmes	e	x	em	p	l	es
phonèmes	ɛ	gz	ã	p	l	

FIG. 5: exemple de mise en relation des graphèmes et des phonèmes

connaissance est directe. Pour l'écriture consonantique, un transducteur fini très simple permet au choix de rajouter optionnellement des voyelles partout dans la forme ou de supprimer toutes les voyelles du lexique pour permettre un appariement. Les choses sont un peu plus délicates si toutes les consonnes ne sont pas notées. Par exemple **qq** pour **quelque**.

Pour les écritures phonétiques, il faut disposer d'une description phonétique de chaque forme fléchie comme par exemple dans le DELAP (Laporte 1990) ou bientôt dans Morphalou (Romary *et al.* 2004) [3].

Il ne suffit pas d'avoir pour chaque forme fléchie la ou les formes phonétiques correspondantes, il faut que ces deux formes soient alignée pour que l'on puisse passer d'un système graphique à l'autre à l'intérieur même d'une forme. Par exemple, pour reconnaître la forme **hazard**, il faut nécessairement interpréter le **h** et le **d** comme relevant du système orthographique, le **z** du système phonétique, les **a** et le **r** indifféremment de l'un ou l'autre système. Il faut donc que l'on puisse passer d'un système à l'autre tout en conservant la linéarité, c'est à dire qu'on puisse vérifier que l'ensemble de la forme est couverte et que seulement elle est notée. Dans cet exemple, il faut donc non seulement un lexique qui associe **hasard** et **azaʁ** au lemme **hasard**, mais également un appariement de sous-formes : **(h,)(a,a)(z,z)(a,a)(r,ʁ)(d,)**.

Nous proposons d'utiliser les techniques de morphologie à partitions pour modéliser un tel système. Il s'agit d'une variante de la morphologie à deux niveaux (Koskenniemi 1983) proposée par Black (Black *et al.* 1987) puis développée par Kiraz (Kiraz 2001) et l'auteur de cet article (Barthélemy 2007). Nous avons notamment utilisé ce modèle pour réaliser un analyseur morphologique de l'akkadien (Barthélemy 2006) qui comprend un module expérimental d'analyse du cunéiforme (non publié).

Les principales caractéristiques de ce modèle sont qu'il est multi-niveaux (pas nécessairement deux), qu'il est possible de définir plusieurs degrés de granularité dans la description, qu'il se compile en automates finis clos sous intersection et différence ensembliste.

Le caractère multi-niveaux nous intéresse pour associer des graphies utilisant le système SMS à un lexique apparié de formes fléchies avec leur phonétique. De plus, conformément à notre analyse, il nous semble judicieux de distinguer le niveau de la translittération et celui de la transcription, ce qui nous amène finalement un système à quatre niveaux. La multiplicité des degrés de granularité permet de distinguer entre le symbole d'une part et l'association phonème-graphème qui apparie des chaînes de longueurs variables : graphème, digramme, trigramme, etc, mais aussi parfois deux phonèmes pour un graphème (voir un exemple figure 5) et des graphèmes ou séquences muets.

L'utilisation des opérations algébriques permet d'ordonner les procédés graphiques. Par exemple, il est possible de calculer un ensemble de formes expliquées par le système orthographique, puis un autre ensemble de formes expliquées par un système mixte et de retirer ensuite, par une combinaison de projections, compositions et différences ensemblistes, les explications phonétiques de formes déjà reconnues par le système orthographique. Cela permet de limiter l'ambiguïté sans aucune perte en précision.

Prenons un exemple qui explicite les quatre niveaux et cette possibilité de définition différentielle. Dans la translittération, nous allons noter le caractère, le sous-système graphique utilisé (**orth** pour orthographique, **alph** pour phonétique alphabétique, **syll** pour phonétique syllabique, **cons** pour consonantique, **logo** pour logographique), et la valeur pour les cas où il y a ambiguïté (par exemple **1** en phonétique syllabique pouvant avoir les valeurs **ɛ̃** ou **œ**).

[3] La version phonétique de ce lexique de forme fléchies est en cours de développement au moment où nous écrivons ces lignes.

graphie sms	a	l	l	é
translittération	orth_a	orth_l	orth_l	orth_é
forme fléchie	a	l	l	é
phonétique	a	l		e

graphie sms	a	l	l	é	
translittération	orth_a	orth_l	orth_l	alph_é_e	
forme fléchie	a	l	l	e	r
phonétique	a	l		e	

FIG. 6: exemples d'analyses multi-niveaux de la forme **allé**

La figure 6 donne deux analyses de la forme **allé**. On est intéressé par les deux interprétations décrites de même que celles correspondant aux formes fléchies **allez, allais, allait, allaient**. En revanche, avoir une seconde interprétation de **allé** comme graphie de la forme **allé** n'apporte rien. C'est ce genre de double interprétation que l'on peut éliminer (ou non) en utilisant des opérations algébriques. Par ailleurs, l'union permet de représenter l'ambiguïté de façon satisfaisante.

6. Conclusion

En résumé, l'étude de l'analogie entre les écritures cunéiformes et SMS nous amène à faire deux propositions. La première consiste à réaliser une translittération comme étape intermédiaire menant à la transcription en écriture normale. Cela permet de diviser les ambiguïtés en deux ensembles : celles qui relèvent de la multiplicité des usages des caractères, qui sont traitées par la translittération, et celles qui relèvent de l'homophonie approximative des formes, qui est traité dans une seconde étape. Un troisième type d'ambiguïté vient de l'homographie de formes différentes. Ce niveau-là n'est pas pris en compte dans notre système puisqu'il n'est pas spécifique à l'écriture SMS.

Un autre intérêt de l'étape de translittération est de faciliter la classification et le recensement des usages d'un caractère. Il permet également d'ordonner les alternatives en considérant comme plus ou moins bonnes les différentes interprétations alternatives d'une graphie.

Après la translittération, notre seconde proposition est d'utiliser pour les SMS les techniques que nous avons employées pour le cunéiforme, à savoir la morphologie multi-niveaux à partitions. Cette technologie permet une analyse pertinente des formes écrites au moyen de plusieurs systèmes d'écriture radicalement différents. Les opérations algébriques fournies par ce modèle permettent la modularité et la hiérarchisation des descriptions. L'implémentation sous forme d'automates finis est efficace.

Nous ne savons pas s'il est possible avec la technologie actuelle de représenter toutes les formes graphiques pouvant être mise en correspondance avec un lexique de forme fléchies en un seul automate fini. Il se peut que la taille dépasse les possibilités des systèmes existants. Seule une expérimentation permettra de le dire. Notons que même si ce n'est pas le cas, un système scindé en plusieurs automates peut demeurer parfaitement utilisable.

Une expérimentation suppose d'abord la construction d'un lexique où les formes graphiques et phonétiques sont alignées. Puis un catalogue des différents emplois des différents caractères utilisés dans les SMS.

Nos propositions répondent à une partie des questions que pose la transcription des textes SMS, mais bien évidemment pas à toutes. Parmi les questions importantes que nous n'avons pas traitées, il y a celle de la segmentation des formes. C'est un problème délicat, qui d'ailleurs prolonge l'analogie que nous faisons, puisqu'il n'y a pas de séparation entre les mots en écriture cunéiforme et pas toujours en écriture SMS.

Par ailleurs, la liberté de ce système d'expression et l'apport de la culture de chaque auteur à sa

façon de s'exprimer défient les possibilités d'inventaire.

References

ANIS J. (1999), *Internet communication et langue française*, Hermès (Paris).

BARTHÉLEMY F. (2006), "Un analyseur morphologique utilisant la jointure", in *In Traitement Automatique de la Langue Naturelle (TALN)*, Louvain, Belgique.

BARTHÉLEMY F. (2007), "Using Mazurkiewicz Trace Languages for Partition-Based Morphology", in *ACL*, Prague (Republique Tchèque).

BLACK A., RITCHIE G., PULMAN S. et RUSSELL G. (1987), "Formalisms for morphographemic description", in *Proceedings of the third conference on European chapter of the Association for Computational Linguistics (EACL)* : 11–18.

FAIRON C., KLEIN J. R. et PAUMIER S. (2007), *Le langage sms, étude d'un corpus informatisé*, Presses universitaires de Louvain, Louvain-la-Neuve (Belgique).

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

KOSKENNIEMI K. (1983), "Two-Level Model for Morphological Analysis", in *IJCAI-83*, Karlsruhe (Allemagne) : 683–685.

LABAT R. et MALBRAN-LABAT F. (1948), *Manuel d'Epigraphie Akkadienne*, Paul Geuthner (Paris), 6eme edition, 1995.

LAPORTE E. (1990), "Le dictionnaire phonémique DELAP", in *Langue Française*, vol. 87.

NEEF E. G. D. et VERONIS J. (2004), *Journée d'Etude de l'ATALA : Le traitement automatique des nouvelles formes de communication écrite*, `http ://www.up.univ-mrs.fr/veronis/je-nfce/index.html`.

ROMARY L., SALMON-ALT S. et FRANCOPOULO G. (2004), "Standards going concrete : from LMF to Morphalou", in *Workshop on Electronic Dictionaries, Coling*, Genève, Suisse.