Cahiers GUTenberg

TRAITEMENT AUTOMATIQUE DES LANGUES ET COMPOSITION SOUS Ω

Yannis Haralambous, John Plaice

Cahiers GUTenberg, nº 39-40 (2001), p. 139-166.

<http://cahiers.gutenberg.eu.org/fitem?id=CG_2001___39-40_139_0>

Traitement automatique des langues et composition sous Omega

Yannis HARALAMBOUS[1] et John PLAICE[2]

(1) *Atelier Fluxus Virus, 187 rue Nationale, F-59800 Lille,*
`yannis@fluxus-virus.com`
(2) *School of Computer Science and Engineering, University of New South Wales,* UNSW SYDNEY *2052, Australie,* `plaice@cse.unsw.edu.au`

Tu ne peux vivre que de
ce que tu transformes
A. DE SAINT-EXUPÉRY, *Citadelle*

Alors qu'Ω ne cesse d'évoluer et ses fonctionnalités de se diversifier et de s'élargir, on s'aperçoit que les méthodes utilisées pour permettre la composition des langues orientales peuvent aussi servir à résoudre des problèmes jusqu'alors négligés des langues occidentales. Les mêmes types d'outils servent à segmenter les propositions thaï en mots puis en syllabes et serviront aussi à déterminer si une lettre « s » d'un mot allemand écrit en gothique doit être longue ou courte... Dans ces deux cas l'outil en question est un *analyseur morphologique*, outil d'une discipline connue sous le nom de *traitement automatique des langues* (TAL en français, NLP = *Natural Language Processing* en anglais). Grâce au mécanisme d'*OTP externe* (OTP = Ω *Translation Process*) d'Ω, nous pouvons intégrer ces outils dans Ω et les utiliser en temps réel pendant la composition. Comme tout ce qui a trait aux langues dites naturelles, ces outils ne peuvent être fiables à 100 % ; ils imposent donc relecture et inspection du texte composé : Ω peut, par le biais de méthodes de colorisation ou de marquage visuel, faciliter et optimiser cette relecture, en paliant ainsi les défauts d'application des outils linguistiques.

Dans cet article nous allons étudier six cas d'utilisation d'outils ou de méthodes linguistiques, de complexité très variée et couvrant un spectre de langues assez étendu : l'anglais, l'allemand, le grec, l'arabe, le thaï et le japonais. Dans chaque cas il s'agira d'introduire certaines informations supplémentaires concernant la structure syntaxique ou morphologique du texte ou sa phonologie ; ces informations peuvent évidemment être saisies par un auteur méticuleux et patient en même temps que le texte-même : un hypothétique auteur allemand peut marquer explicitement les « s » longs et les « s »

courts, un auteur thaï peut, en théorie, indiquer les limites des mots et des syllabes. Mais cela devient vite pénible et ne devrait donc pas faire partie de la tâche de l'auteur mais plutôt de celle de l'imprimeur, voire dans certains cas de celle du correcteur ou du responsable éditorial.

0. Vérification orthographique multilingue

Cet exemple est plus général et plus basique que ceux qui vont suivre, c'est pour cela que nous l'avons appelé « exemple zéro ». Il peut être appliqué à toute langue pour laquelle il existe un vérificateur orthographique du type de `ispell` ; par exemple, à la date actuelle `ispell` [2] couvre les langues suivantes : l'afrikaans, le biélorusse, le catalan, le tchèque, le danois, le néerlandais, l'anglais, l'espéranto, le finnois, le français, l'allemand, le grec moderne monotonique, l'ivrit, l'italien, l'irlandais, le norvégien, le polonais, le portugais, le roumain, le russe, le slovène, l'espagnol, le suédois et l'ukrainien.

En utilisant `ispell` directement sur un texte écrit en TEX on s'aperçoit très rapidement que les balises TEX gênent la vérification : d'une part elles sont lues au même titre que les mots du texte et, d'autre part, même si on parvient à apprendre à `ispell` que les mots qui commencent par un backslash sont en fait des instructions et doivent être ignorées (par exemple en utilisant l'option de ligne de commande `-t` de `ispell`), il reste toujours les arguments de certaines commandes (comme `\begin{quotation}`) qui, du moins en LATEX standard, sont des mots anglais (ou latins comme *verbatim*, ou même des pseudo-mots qui n'appartiennent à aucune langue, comme *multicols*). Et puis, il y a les commentaires, les formules mathématiques ou chimiques, la musique, le code informatique (si c'est un ouvrage informatique), et ainsi de suite...

Par conséquence, quand `ispell` lit un document TEX, il lit en fait un document écrit dans deux langues : la langue du texte et... TEX. Il existe un proverbe informatique bien connu qui dit que « seul TEX peut lire du TEX », dans le sens qu'un document TEX peut être si complexe et si varié que tout parseur autre que TEX lui-même échouera. Comme tous les proverbes, ce proverbe a une part de verité : comment un vérificateur orthographique (qui est un parseur relativement naïf) pourrait-il décortiquer du code TEX et en extraire le texte, en ignorant les autres constituants du document TEX ?

C'est encore plus grave lorsque le document est multilingue : `ispell` n'est pas capable de changer de langue de vérification en cours de route. Quel cauchemar serait pire que d'être obligé de vérifier orthographiquement un dictionnaire franco-anglais en utilisant d'abord `ispell` français et ensuite `ispell` anglais...

Nous nous proposons de faciliter la vérification orthographique d'un texte en marquant les mots non trouvés dans le dictionnaire d'`ispell` (et donc suspects), tout en gardant le formattage du texte indépendant de ce marquage. Par exemple, les mots suspects peuvent être composés en rouge : cela ne change — du moins, en théorie — pas le formatage du texte : ainsi, une fois le texte vérifié, on désactive la colorisation et on imprime le texte tout à fait normalement.

Voici le code Perl du script qui a servi à cette vérification :

```
#!/usr/bin/perl

$/="";

while (<>) {
$global=$_;
open OUT, "| ispell -d francais -l >tmp.spell";
print OUT $_;
close OUT;
open IN, "tmp.spell";
while (<IN>) {
foreach $mot (split /\n/, $_) {
$MOTS{$mot} = $mot;
}
}
close IN;
foreach $mot (sort keys %MOTS) {
$resultat =~ s/$mot/\\textcolor{red}{$mot}/g;
}
print $resultat;
}
```

[À partir d'ici nous avons activé `ispell` français pour montrer le résultat d'une telle vérification : les mots « suspects » ont été soulignés.]

Supposons que ce script est appelé `verif.perl`. Pour l'utiliser sous Ω, il suffit d'écrire les lignes suivantes :

```
\externalocp\OCPverif=verif.perl {}
     \ocplist\verifier=
     \addbeforeocplist 100 \OCPverif
     \nullocplist
\pushocplist\testOCP
```

La dernière ligne est celle qui lance le processus de vérification. En l'incluant dans un groupe TEX ou un environnement LATEX on peut limiter son champ d'action. En la combinant avec un environnement LATEX linguistique on peut appliquer cette vérification uniquement au texte écrit dans cette langue.

Dans le cas d'un document multilingue on peut écrire plusieurs scripts de ce type, un pour chaque langue. Chaque environnement linguistique va alors activer et désactiver automatiquement la correction orthographique *ad hoc*. On pourrait alors choisir une couleur différente pour chaque langue, de manière à faciliter la relecture du document.

Voici donc une manière plus ergonomique de combiner un vérificateur orthographique du domaine public, comme `ispell`, avec le processus de composition d'un document sous Ω.

Malheureusement, cette méthode, ainsi que toutes celles décrites dans cet article nécessitent un système d'exploitation muni d'une ligne de commande : Unix, les divers Windows, mais pas le Macintosh (à moins d'utiliser MacOS X, ou MPW sous MacOS conventionnel). La technique Macintosh correspondante est celle des *Apple events*, qui trasmettent des messages d'une application à l'autre : on pourrait, en effet, imaginer une version Macintosh qui communique avec ses OTP externes à l'aide d'*Apple events*. Mais nous pensons que le passage à MacOS X, qui est un Unix déguisé en MacOS, serait de loin plus judicieux en termes de stabilité, possibilités et performances.

[Fin de la vérification : notez que `ispell` reconnait le nom « Unix », mais pas ceux des autres systèmes d'exploitation...]

1. L'anglais et les espaces inégales

Dans le chapitre 12 du *TEXbook* [3], Knuth décrit un aspect de la typographie anglosaxonne, à première vue tout à fait inoffensif : certains points sont suivis d'espaces plus larges que d'autres. Il s'agit des points de fin de phrase, contrairement aux points utilisés dans les abréviations. On pourrait dire que les premiers jouent un rôle syntaxique alors que les deuxièmes ne jouent qu'un rôle morphologique, donc moins important. Il est clair que

> I submitted the paper to Bull. Amer. Math. Soc. Trans. Amer. Math. Soc. refused it without any further comments.

est plus lisible que

> I submitted the paper to Bull. Amer. Math. Soc. Trans. Amer. Math. Soc. refused it without any further comments.

L'œil du lecteur cerne plus rapidement les phrases lorsque celles-ci sont délimitées par des espaces plus larges que celles que l'on trouve entre les mots d'une même phrase.

Pour des raisons probablement historiques, la typographie française ne se sert pas de cette convention. Ceci est un fait bien connu de Knuth qui appelle la désactivation de cette convention `\frenchspacing` (espacement à la française) [3, p. 74]. Cette convention a également été utilisée dans d'autres languages, et en particulier en allemand — où elle est encore plus utile puisque le texte allemand fourmille de lettres majuscules et il est donc d'autant plus difficile de détecter visuellement le début et la fin de chaque phrase — malgré le fait que des typographes tels que Tschichold la dénoncent.

Sous TEX, et donc aussi sous Ω, cette convention se sert d'une propriété métrique associée à chaque caractère, le `\sffactor`. D'office dans `plain` ou LATEX, les lettres majuscules suivies d'un point produisent une espace normale, alors que les lettres minuscules suivies d'un point produisent une espace plus large. Il incombe à l'utilisateur d'intervenir lorsque ce comportement produit des résultats erronés : par exemple, lorque l'on écrit

I work at NASA. The est. arrival time is around three o'clock.

on est obligé d'intervenir deux fois : d'une part le point qui suit « NASA » est un point de fin de phrase et donc l'espace qui suit doit être plus large et, d'autre part, dans « est. » le point est utilisé pour abréger le mot « estimated » et devrait donc être suivi d'une espace-mot. La manière correcte d'écrire le code TEX correspondant serait donc :

```
I work at NASA\null. The est.\ arrival time is around three o'clock.
```

qui produira tout naturellement

I work at NASA. The est. arrival time is around three o'clock.

À notre connaissance, aucun autre système de composition ne propose cette convention[1] : on ne peut donc s'empêcher, encore une fois, de louer Knuth pour son savoir-faire et sa perspicacité. Néanmoins, avec tout le respect que

[1] Ceci dit, `emacs`, qui n'est pas un traitement de texte, mais un éditeur de texte, utilise une convention intéressante [4] : lorsqu'un point est suivi d'un espace, cet espace est automatiquement insécable ; lorsqu'un point est suivi de deux espaces au moins, il est sécable. Cela contribue à « forcer » (terme utilisé par Ken-ichi Handa, développeur de Mule (Emacs multilingue)) les utilisateurs à faire la différence entre les points d'abréviation et les points de fin de phrase. Une initiative louable et typique du créateur d'Emacs, Richard Stallman.

l'on lui doit, on est obligé d'admettre que la proposition de Knuth échoue lamentablement, non pas sur le plan technique, mais sur le plan psychologique. Quel auteur ou autrice[2] aurait la puissance mentale nécessaire pour penser à ces règles et exceptions de saisie, alors qu'il/elle est concentré(e) sur le contenu du texte qu'il/elle est en train d'écrire ? Et quel rédacteur (à part notre chère Barbara Beeton, rédactrice du *TUGboat*) a la patience et le courage nécessaires pour vérifier dans leur contexte tous les points d'un texte donné ?

On réalise, lorsque l'on a à composer un texte relativement long — voire un livre — en anglais, que cette convention, à première vue inoffensive, est en réalité un cauchemar : un livre d'environ 300 pages peut facilement contenir plus de 7 000 points, qu'il faut vérifier un par un, dans leur contexte, c'est-à-dire en relisant le texte environnant.

Comment résoudre ce problème ?

L'approche du puriste serait d'utiliser un analyseur morphologique et syntaxique, qui donnerait la nature et le rôle de chaque mot de la phrase : ainsi on saurait tout de suite quel point est une véritable fin de phrase et, par déduction, lequel ne peut pas l'être. Mais ceci pourrait être très difficile à réaliser puisque les abréviations vont probablement perturber l'analyseur ; or c'est justement dans le cas des phrases avec abréviations que ce procédé serait utile.

Notre approche est la suivante : nous partons du principe que le problème est trop difficile à résoudre (il faudrait une analyse syntaxique et sémantique poussée pour pouvoir déterminer les abréviations avec précision), et nous cherchons — du moins dans un premier temps — à trouver un compromis acceptable, tout en facilitant la correction des erreurs dues à notre système.

Le compromis est d'affirmer que *la forme abrégée d'un mot n'apparaît pas en tant que telle dans le dictionnaire*. Il est facile de trouver des contre-exemples : *term.* (abréviation de *terminal*) s'écrit comme *term* (le terme). Néanmoins, cette méthode permet de détecter correctement toutes les initiales de prénom, et grand nombre d'autres abréviations: *est.* (estimated), *St.* (Saint), *ref.* (reference), etc.

À première vue on pourrait penser qu'une recherche dans un dictionnaire (à la manière de `ispell`) suffirait à trouver les mots qui ne sont pas dans le dictionnaire. Il n'en est rien. En effet, `ispell` étant un dictionnaire qui se veut pratique et fonctionnel, il contient déjà toutes ces abréviations. Après tout, elles font partie intégrante de la langue anglaise et ne constituent en rien des erreurs.

À la place d'un dictionnaire conventionnel, nous allons utiliser un analyseur morphologique. La différence est claire : alors que `ispell` nous donne une

2 Il ne s'agit ni d'*autruche* ni de (roue) *motrice*, mais bel et bien d'*autrice*, un mot proposé par l'Académie française en l'an 2000.

information binaire (le mot est, ou n'est pas, contenu dans le dictionnaire), l'analyseur nous fournit des informations *qualitatives* : il est obligé de décrire morphologiquement chacun des mots que l'on lui fournira.

L'analyseur morphologique que nous avons utilisé est `ktext` [5] : un outil du domaine public développé par le *Summer Institute of Linguistics*. Voici un exemple d'application de cet outil : en analysant le petit texte suivant :

> The work of building Tsukuba Science City began in September 1963, when the Cabinet approved the plan as a national project aimed at strengthening Japan's position in the sciences. The city was built to relieve overcrowding in central Tokyo and also to expand educ. and res. facilities. In addition to the 44 nat. res. institutions and 3 universities, there are now nearly 140 priv. res. and educ. institutions in Tsukuba. Some are newly est., and some moved out of Tokyo and the surr. area to make a new start.

`ktext` a analysé tous les mots, et a produit un fichier du type :

```
\a %0%res%
\d %0%res%
\w res
\n .\n

\a ‘facile +NR21 +PL
\d ‘facile-+ity-+s
\cat N
\fd root_AJ
\w facilities
\n .

\a %2%in%in%
\d %2%in%in%
\cat %2%AV%PP%
\fd %2%root_AV%root_PP%
\w In
\c 1
```

Ici les signes de pourcentage séparent les résultats trouvés pour un même mot : pour le mot *In* il y a deux occurences, pour *facilities* il n'y en a qu'une seule. Dans le cas de *res.* on a zéro résultat, le mot n'est pas reconnu. De cette manière nous pouvons détecter la plupart des abréviations.

Il reste deux catégories d'erreurs potentielles :

1. les abréviations non détectées, qui nécessitent une espace-mots ;
2. les abréviations détectées, mais en fin de phrase, qui nécessitent donc une espace plus large.

Pour mieux détecter ce genre de problèmes, l'OTP qui gère `ktext`, va *marquer* les points considérés comme points d'abréviation par une barre noire bien visible, mais contenue dans une `\hbox` de largeur nulle, de manière à ne pas perturber la mise en page du texte. Voici notre texte, en utilisant l'OTP de gestion des points :

> The work of building Tsukuba Science City began in September 1963, when the Cabinet approved the plan as a national project aimed at strengthening Japan's position in the sciences. The city was built to relieve overcrowding in central Tokyo and also to expand educ and res facilities. In addition to the 44 nat res institutions and 3 universities, there are now nearly 140 priv res and educ institutions in Tsukuba Some are newly est., and some moved out of Tokyo and the surr area to make a new start.

Dans l'exemple ci-dessus, nous apercevons déjà une erreur du système : le mot *Tsukuba* étant inconnu à `ktext`, il est marqué comme abréviation. En maintenant un dictionnaire personnalisé, l'utilisateur peut éviter ce genre de bévues.

Nous espérons qu'ultérieurement, en utilisant des outils d'analyse syntaxique, nous pourrons cerner avec encore plus de précision les points de fin phrase et les abréviations : en reprenant l'exemple de l'abréviation « term. », un outil capable d'analyser syntaxiquement la phrase « he suffers from a term. disease. » nous permettra de conclure que « term. » est forcément un adjectif, et donc ne peut pas être le mot « term », et ainsi de suite.

2. L'allemand et la longueur des s

Jusqu'à il y a encore moins de 60 ans, l'allemand s'écrivait officiellement en « écriture allemande », c'est-à-dire *gothique* [6]. Pour les maisons d'édition allemandes de l'époque, le romain était une alternative, pour donner un esprit de modernité. Les grands auteurs et philosophes (Goethe et Kant, pour ne nommer que les plus importants) étaient formels sur l'utilisation du gothique pour composer leurs livres. Hitler a interdit la composition en gothique, en pleine deuxième guerre mondiale, sous prétexte que cette écriture avait... des origines juives. Il serait absurde, voire insultant vis-à-vis des victimes du nazisme, de comparer ce caprice de dictateur à ses énormes crimes contre l'hu-

manité. Néanmoins il s'agit d'une importante perte pour le patrimoine culturel allemand.

Avant cette interdiction, Hitler avait fortement utilisé le gothique à des fins de propagande. C'est sûrement pour cette raison que les forces d'occupation ont de nouveau interdit l'utilisation de cette écriture qui, se trouvant soudainement redoutée par tout le monde, cessa officiellement d'exister...

On ne peut qu'espérer qu'au vingt et unième siècle, libérée de ses connotations politiques, cette écriture sera de nouveau admirée pour sa grande valeur culturelle et utilisée au moins pour les textes d'époque, un peu comme le Garamond est utilisé en France ou le Caslon en Angleterre.

Quoi qu'il en soit, Ω propose d'ores et déjà un système de composition en écriture gothique.

Cette écriture a gardé la dualité de forme de la lettre « s » : « s long » ſ et « s court » ou plus précisement « s rond » ſ̆, dualité qui existait aussi dans l'écriture romaine il y a encore quelques siècles. Elle a aussi gardé un certain nombre de ligatures usuelles : ß (sz), ch (ch), ck (ck), ff (ff), fi (fi), fl (fl), ft (ft), ll (ll), ſi (si), ſſ (ss), ſt (st), tt (tt), tz (tz).

Dans les langues telles que l'anglais ou le français, les deux « s » et les ligatures ont été utilisés (et le sont en partie encore à ce jour) sans trop de problèmes : le « s » long était placé au début et à l'intérieur d'un mot, le « s » rond à la fin ; lorsqu'une ligature est disponible dans une police, on l'utilise sans modération.

L'allemand ajoute une contrainte d'ordre linguistique : les mots allemands sont très souvent des mots composés, ce qui explique leur longueur parfois excessive : un mot tel que *Forschungsinstitutsdirektor* (27 lettres) est traduit en français par trois mots et deux particules : *directeur d'institut de recherche* (30 lettres au total). Il est important que l'œil du lecteur détecte aussi rapidement que possible les limites de ces mots, de la même manière qu'il détectera les blancs qui séparent leurs homologues français.

Pour ce faire, l'allemand utilise le « s rond » quand il s'agit de la lettre finale d'une composante du mot : notre exemple s'écrit Forſchungsinſtitutsdirektor : comparez les deux « s » de Forſchungs. Ce « s rond » sépare les (sous-)mots, sans les séparer. Il en est de même des ligatures : elles ne sont pas utilisées lorsque les lettres qui auraient dû les former se trouvent à cheval entre deux composantes : comparez Brotzange qui est formé des deux mots Brot et Zange, et trotzdem où la ligature tz est entièrement contenue dans le mot trotz.

Il existe même des cas (très rares, il faut l'admettre) où le « s long » ou les ligatures servent à distinguer deux mots homographes produits par deux compositions différentes. Par exemple le mot *Wachstube* peut se prononcer « vaks-

toube » ou « vahh-chtoube » : dans le premier cas il est composé des mots *Wachs* (la cire) et *Tube* (le tube) et s'écrit Wachstube, dans le deuxième cas il est composé des mots *wach* (éveillé) et *Stube* (la chambre) et s'écrit Wachſtube. En écriture romaine, il n'y a aucun moyen de distinguer les deux mots.

Notons que certaines ligatures du gothique sont encore utilisées en allemand composé en romain : par exemple, le « ch » et le « ck » : ce dernier est célèbre puisqu'il est coupé « k-k » (ou du moins l'était avant la réforme orthographique allemande des dernières années).

Les deux « s » et le contrôle des ligatures entrent dans la catégorie des opérations typographiques essentielles pour un système d'écriture donné, mais impossibles à gérer par l'utilisateur standard d'un traitement de texte ou système de composition. Quand on est concentré sur un texte, on ne peut s'occuper de tels artifices typographiques qui demandent trop de réflexion pour devenir vraiment des automatismes. De plus, il existe des archives publiquement accessibles [9] avec des milliers de textes allemands classiques : la tentation de les composer en gothique est vraiment très forte, mais ces textes ne contiennent aucune information sur la forme du « s » ou la validité des ligatures.

Il est donc indispensable pour un système de composition de l'allemand gothique de pouvoir gérer ces particularités du système d'écriture. Mais, encore une fois, cette tâche est loin d'être triviale. Tout d'abord, l'utilisation directe d'un dictionnaire est impossible : les mots composés peuvent être arbitrairement longs et ne sont, dans la plupart des cas, pas contenus dans les dictionnaires. De plus, il n'est pas judicieux de chercher à décomposer les mots mécaniquement : souvent ils changent légèrement ou se déclinent quand ils sont composés : ainsi *Meer* + *Boden* devient *Meeresboden* (*Meer* au génitif), *Ende* + *Ziel* devient *Endziel* (perte de la voyelle finale du premier mot), etc.

Il faut donc de nouveau recourir à un analyseur morphologique. Nous avons utilisé DMM [7] (*Deutsche Malaga Morphologie*, où Malaga [8] est un langage de programmation spécialisé dans l'analyse morphologique) logiciel libre développé par Oliver Lorenz, à l'université d'Erlangen. Pour optimiser le processus, le script perl que nous utilisons comme OTP externe, n'appelle DMM que pour les mots contenant des lettres « s » ou des candidats à des ligatures, et ceci à des positions qui n'excluent pas une composition de mots. Ainsi, par exemple, inutile d'utiliser DMM pour analyser le mot Ernſt, puisque le « s » n'est qu'à une lettre de la fin du mot : il est impossible que cette lettre forme un mot en elle seule. Pour le mot Erneſtine (un prénom cher au premier auteur, puisque c'est celui de sa fille), notre système ne peut exclure d'office la possibilité d'une segmentation Erne + *Stine* et l'utilisation de DMM s'impose.

Examinons de plus près le mode de fonctionnement de DMM : pour un mot aussi compliqué que Steuereinſchätzungskommiſſion (comité de prévision des impôts), composé des quatre mots *Steuer*, *ein*, *Schätzung*, *Kommision*, nous obtenons :

```
"Steuereinschätzungskommission":
<[WordForm: "steuereinschätzungskommission",
Segmentation:
"steuer<CPD>ein<PFX>schätz<DRV>ung<FUG>s<CPD>kommission",
AltSegmentations:
<"steuer<CPD>ein<CPD>schätz<DRV>ung<FUG>s<CPD>kommission",
"steuer<CPD>ein<CPD>schätz<DRV>ung<FUG>s<CPD>kom<CPD>mission",
"steuer<CPD>ein<PFX>schätz<DRV>ung<FUG>s<CPD>kom<CPD>mission">,
POS: Substantive,
BaseForm: "steuereinschätzungskommission",
Weight: 0.64,
Gender: Feminine,
CaseNumber: Singular,
WordStructure: <[Morpheme: "steuer", Allomorph: "steuer"],
[Morpheme: "ein", Allomorph: "ein"],
[Morpheme: "schätzen", Allomorph: "schätz"],
[Morpheme: "ung", Allomorph: "ung"],
[Morpheme: "s", Allomorph: "s"],
[Morpheme: "kommission", Allomorph: "kommission"]>]>
```

Dans ces lignes, ce qui nous intéresse le plus est la segmentation

```
Segmentation:
"steuer<CPD>ein<PFX>schätz<DRV>ung<FUG>s<CPD>kommission",
```

Les codes utilisés pour ségmenter ce mot ont les significations suivantes :

- `<CPD>` : composition = racine + racine
- `<PFX>` : préfixe + racine
- `<DRV>` : racine + suffixe de dérivation
- `<FUG>` : racine + suffixe de déclinaison

Seules les segmentations de type `<CPD>` et `<PFX>` vont entraîner un « s » rond ou une absence de ligature et ce sont celles-là donc qui nous intéressent. Le mot en question ne figure pas dans le Duden, et encore moins dans les dictionnaires du domaine public. DMM nous a donc fourni une information essentielle.

En appliquant cette méthode à un véritable texte allemand classique (les *Années d'enfance* [10] de Théodore Fontane, grand admirateur par ailleurs de la France et de la culture française), DMM a fourni, sans aucune optimisation spéciale de son code, des résultats corrects à 97,2 %. Nous espérons qu'en collaboration avec Oliver Lorenz, en stockant les résultats erronés dans un dictionnaire additionel, nous pourrons améliorer largement ce résultat. Voici un extrait de ce texte, où Fontane parle d'ailleurs de ses rapports avec la France :

> Gascogne und Cevennen lagen für meine Eltern, als sie geboren wurden, schon um mehr als hundert Jahre zurück, aber die Beziehungen zu Frankreich hatten beide, wenn nicht in ihrem Herzen, so doch in ihrer Phantasie, nie ganz aufgegeben. Sie repräsentierten noch den unverfälschten Kolonistenstolz. Weil sie aber stark empfinden mochten, daß mit ihren nachweisbaren Ahnen, die bei den Fontanes als Zinngießer, potiers d'etain, bei den Labrys als Strumpfwirker, faiseurs de bas, feststanden, nicht viel Staat zu machen sei, so ließen sie die amtlich geführte kolonistische Stammtafel fallen und suchten statt dessen, auf gut Glück, nach vornehmen französischen Vetterschaften, also nach einem wirklichen oder eingebildeten Familienanhang, der, in der alten Heimat zurückgeblieben, sich mittlerweile zu Ruhm und Ansehn emporgearbeitet hatte.

Il est intéressant de noter, dans ce texte, l'utilisation du romain pour certains mots français (*potiers d'étain*) qui cependant ont perdu leur accent et du gothique pour d'autres mots tout aussi français (*faiseurs de bas*, ligne 8). Nous décrirons plus en détails les règles de composition du gothique dans un prochain article aux *Cahiers GUTenberg*.

Pour vérifier plus facilement la validité des choix de notre système, nous proposons un système de marquage visuel des endroits où le système a dû trancher : en activant un certain OTP, Ω va composer les « s » ou les ligatures qui nécessitent une vérification de la part de l'utilisateur dans une autre couleur ou un autre style. Voici, par exemple, le même texte où les lettres à vérifier ont été composées dans une écriture gothique plus grasse qui saute immédiatement aux yeux :

> Gaſcogne und Cevennen lagen für meine Eltern, als ſie geboren wurden, ſchon um mehr als hundert Jahre zurück, aber die Beziehungen zu Frankreich hatten beide, wenn nicht in ihrem Herzen, ſo doch in ihrer Phantaſie, nie ganz aufgegeben. Sie repräſentierten noch den unverfälſchten Koloniſtenſtolz. Weil ſie aber ſtark empfinden mochten, daß mit ihren nachweiſbaren Ahnen, die bei den Fontanes als Zinngießer, potiers d'etain, bei den Labrys als Strumpfwirker,

> faiſeurs de bas, feſtſtanden, nicht viel Staat zu machen ſei, ſo ließen ſie die amtlich geführte koloniſtiſche Stammtafel fallen und ſuchten ſtatt deſſen, auf gut Glück, nach vornehmen franzöſiſchen Vetterſchaften, alſo nach einem wirklichen oder eingebildeten Familienanhang, der, in der alten Heimat zurückgeblieben, ſich mittlerweile zu Ruhm und Anſehn emporgearbeitet hatte.

À noter que dans certains cas on peut conclure sans passer par une analyse morphologique : par exemple, dans feſtſtanden il y a nécessairement segmentation entre les deux ſt puisqu'aucun mot non-composé ne peut contenir un double ſt.

Ainsi DMM nous sert, du moins pour l'instant, comme *aide à la « gothisation » du texte* : l'auteur doit encore vérifier le résultat de cette opération à l'aide d'artifices comme celui que nous venons de décrire ; néanmoins, le temps et l'effort requis pour cette opération sont infiniment réduits. Nous allons dans l'avenir rendre DMM + Ω de plus en plus précis et essayer d'établir un moyen de mesurer le poids de chaque analyse, c'est-à-dire son coefficient de fiabilité. Dans le cas idéal, l'utilisateur pourra demander à Ω de lui soumettre uniquement les cas douteux (c'est-à-dire ceux dont le coefficient de fiabilité est au-dessous d'un certain seuil, qu'il précisera).

Une coopération internationale entre utilisateurs de ce système permettrait aussi d'établir des dictionnaires de cas particuliers dont tout le monde pourra profiter.

Nous allons également nous intéresser aux autres langues européennes qui ont utilisé l'écriture gothique et qui possédent également des règles de composition pour les deux « s » et les ligatures ; citons par exemple le danois, le néerlandais, etc.

3. Le grec et l'amour des accents

Peut-on *aimer* un signe graphique ou un phénomène grammatical ? *A priori* non, mais à entendre Elytis (prix Nobel de littérature) qui disait que « les accents sont les boucles d'oreille de cette belle femme qu'est la langue grecque », ou Pentzikis qui disait que « la Grèce du Nord s'assied sur le restant du corps grec comme un esprit doux ou un accent circonflexe »[3] on peut légitimement se poser des questions sur les rapports des Grecs et des accents. D'ailleurs,

[3] « Ἡ ἔκταση τῶν 42.614 τετραγωνικῶν χιλιομέτρων τῆς Βορείου Ἑλλάδος, μοιάζει στὸν χάρτη νὰ ἐπικάθεται σὰν πνεῦμα ἢ τόνος, ψιλὴ ἢ περισπωμένη, ἐπὶ τοῦ ὑπολοίπου ἑλλαδικοῦ σώματος », Nikos Gabriel Pentzikis, *Esprit doux ou accent circonflexe*, Éd. Agra, Athènes, 1995.

dans quel autre pays du monde, vingt ans après une réforme majeure de la langue, le livre de qualité est-il toujours fait comme si la réforme n'avait jamais eu lieu ?

Telle est la situation, du point de vue du lecteur grec ou grecophone : le livre de qualité est un livre *polytonique* (avec les trois accents, les deux esprits, le tréma et éventuellement le iota souscrit) même si dans la vie courante on est bel et bien passé au système *monotonique* (un seul accent et le tréma). Mais, quand on entre dans les coulisses du monde de l'édition, on découvre (et cela a été un véritable choc pour le premier auteur) que les accents, tant prisés et admirés par les grands de la litérature, ne sont aujourd'hui qu'un *embellissement d'imprimeur*. En effet, les auteurs, même les plus connus, écrivent en monotonique ; c'est l'imprimeur qui convertit (souvent mécaniquement, tant bien que mal), le texte monotonique en polytonique. Les erreurs d'accentuation sont corrigées, non pas par l'auteur qui, la plupart du temps, n'en connait même pas les règles, mais par un correcteur spécialisé, lors des premières épreuves. Que le lecteur français s'imagine écrire un texte français sans accents, lesquels seraient ensuite ajoutés par l'imprimeur... c'est en même temps déroutant, pénible et infiniment triste.

Quoi qu'il en soit, si les accents sont devenus une affaire d'imprimeur, et si Ω joue, du moins en partie, le rôle de l'imprimeur, alors Ω devrait naturellement être capable d'accentuer le grec. C'est une tâche non négligeable qui nécessite une analyse morphologique, mais souvent aussi syntaxique et sémantique du grec : par exemple, οἰκεῖα prend un accent circonflexe quand c'est la première personne du nominatif pluriel de l'adjectif neutre οἰκεῖον, mais prend un accent aigu lorsque c'est la première personne du nominatif singulier de l'adjectif féminin οἰκεία ; c'est une différence morphologique ; γιατί prend (suivant certaines traditions) un accent aigu lorsqu'il fait partie d'une phrase interrogative et un accent grave si la phrase est affirmative : une différence syntaxique ; ὁρῶν prend un esprit rude quand il est le génitif pluriel de ὅρος (la condition), et un esprit doux lorsqu'il est le génitif pluriel de ὀρός (le sérum) : on a identité morphologique et syntaxique mais différence sémantique.

Pour éviter le cercle vicieux « comment accentuer un texte que l'on ne peut analyser proprement que quand il est accentué ? » on peut se donner comme but de prendre le grec monotonique comme point de départ et de se limiter à la conversion du monotonique en polytonique et à la vérification de ce dernier. Il y a deux produits commerciaux [11, 12] qui font, tant bien que mal, cette conversion ; aucun projet équivalent dans le domaine public ne nous est connu, pourtant il aurait été très utile. Avant d'attaquer un projet d'une telle envergure, nous avons, en coopération avec l'équipe de Ioannis Kanellos de l'ENST de Bretagne [13], envisagé le problème de la vérification de l'accentua-

tion d'un texte et plus précisement de la *vérification des règles de l'accent grave vis-à-vis de l'accent aigu*.

En effet, l'accent grave est en quelque sorte le « mouton noir » de la famille des accents : officiellement, il a été « retiré de la circulation » déjà dans les années cinquante, lors d'une réforme orthographique moins importante que celle de 1981. Il n'est pas enseigné et, à notre connaissance, n'est pas utilisé dans l'écriture manuscrite. Il s'agit donc, depuis les années cinquante déjà, d'une « affaire d'imprimeur » par excellence.

Les règles de l'accent grave n'ont jamais été normalisées : même les grammaires les plus importantes sont divisées sur son utilisation et n'entrent pas dans le détail de toutes les situations. De par ce fait, il y a en Grèce plusieurs « écoles d'utilisation de l'accent grave » ; pratiquement, chaque maison d'édition a ses propres règles et les imprimeurs qui travaillent pour plus d'une maison doivent en tenir compte, ce qui pose aussi un problème, et non pas des moindres.

Voici une version de ces règles pour le grec moderne (d'autres règles ou exceptions mineures peuvent venir s'ajouter suivant le cas) :

1. RÈGLE : l'accent grave se place sur la dernière syllabe du mot, l'accent aigu se place sur la pénultième ou l'antépénultième.
2. EXCEPTION À 1 : l'accent grave devient aigu lorsque le mot est suivi d'un des signes de ponctuation suivants : point, point supérieur, virgule, point d'exclamation, point d'interrogation, deux-points. Les guillemets, parenthèses, crochets, tirets longs et points de suspension sont ignorés lorsque l'on applique les règles d'accentuation.
3. EXCEPTION À 1 : lorsqu'un mot accentué sur la finale ou l'antépenultième est suivi d'une particule enclitique il récupère l'accent de l'enclitique sur sa dernière syllabe, sous forme d'un accent aigu : τὸ δικό μου, ἡ μέθοδός μου (à noter que dans le deuxième exemple, le mot a effectivement deux accents).
4. RÈGLE : le mot τί prend toujours un accent aigu.
5. RÈGLE : les mots ποιός/ά/ό, γιατί prennent un accent aigu lorsqu'ils sont interrogatifs et un accent grave sinon.
6. EXCEPTION À 2 ET RENFORCEMENT DE 5 : pour indiquer sans ambiguité le cas où γιατί suivi d'une virgule se trouve être affirmatif et non pas intérrogatif, il prendra un accent grave : αὐτὸ ἔγινε γιατὶ, ἐνῶ τὸ ἤξερε, δὲν ἤθελε νὰ τὸ πιστέψει.

L'outil développé en collaboration avec l'ENSTB vérifie ces règles (ainsi que certaines règles plus basiques destinées à détecter les erreurs de frappe), de manière interactive. Nous travaillons sur une version intégrée à Ω sous

forme d'OTP externe, qui pourra corriger automatiquement certains cas qui ne posent aucun doute, et composer de manière spéciale (par exemple, en rouge) les cas ambigus où il incombera à l'utilisateur d'Ω de trancher. Encore une fois, un vérificateur/correcteur interactif du monde WYSIWYG devient sous Ω un vérificateur/correcteur semi-interactif : les corrections ou les cas suspects sont *soumis* à l'auteur de manière discrète pour ne pas perturber la mise en page, mais suffisamment claire pour attirer son attention.

Ce vérificateur a été appliqué avec succès aussi bien sur des textes saisis originalement en grec polytonique, ou convertis du grec monotonique par un des deux logiciels du marché, ou reconnus optiquement.

L'Institut de Traitement de la Langue et de la Parole [14], à Athènes, a développé un analyseur morphologique du grec (monotonique). L'utilisation de cet outil sous Ω nous permettra une vérification/correction plus approfondie des accents du grec, ainsi que peut-être une amorce du projet d'accentuation automatique. *À suivre...*

4. Concilier la grammaire arabe et l'ordinateur

Apprendre l'arabe classique est un véritable plaisir pour un mathématicien : la grammaire de cette langue est une formidable structure mathématique avec objets et règles. On a constamment l'impression que les règles ne sont pas des règles empiriques *a posteriori*, mais des décisions conscientes et hautement rationelles de quelque mathématicien de l'illustre école mathématique arabe — ce qui n'est pas le cas de la plupart des autres langues, notamment de certains langages de programmation...

Malheureusement ce superbe édifice qu'est la grammaire arabe est terni par l'ordinateur. En effet, l'arabe a été codé dans Unicode [15, 20] (et dans les autres codages qui l'ont précédé) de la même manière qu'il est utilisé sur machine à écrire, c'est-à-dire en donnant plus d'importance à la graphie qu'aux objets grammaticaux.

Voici quelques exemples et les solutions que nous avons adoptées pour rétablir la situation.

4.1. Le hamza

On trouve dans le codage Unicode cinq caractères appelés : « lettre hamza » (ء), « lettre alif avec hamza supérieur » (أ), « lettre waw avec hamza » (ؤ), « lettre alif avec hamza inférieur » (إ), « lettre yeh avec hamza supérieur » (ئ). Les

noms de ces caractères prêtent à confusion : il s'agit en fait de *cinq représentations différentes d'une seule et unique lettre arabe*, le *hamza*. Le choix de la représentation se fait suivant des règles très strictes, en fonction des voyelles courtes du hamza lui-même et des lettres environnantes. L'arabe possède trois voyelles courtes : le *fatḥa* (que nous allons noter a), le *kasra* (noté i) et le *damma* (noté u). Il a également un signe appelé *sukun* pour noter l'absence de voyelle courte (nous allons le noter $\emptyset$).

Les règles se divisent en trois ensembles, selon que le hamza est au début, au milieu ou à la fin du mot. Voici un tableau résumant les règles de représentation du hamza (à l'horizontale, la voyelle courte précédant le hamza, à la verticale, la voyelle courte du hamza) :

	début de mot	milieu de mot				fin de mot			
		a	i	u	$\emptyset$	a	i	u	$\emptyset$
a	1	1	4	2	1	1	4	2	0
i	3	4	4	4	4				
u	1	2	4	2	2				
$\emptyset$	1	1	4	2	0				

où 0, 1, 2, 3, 4 dénotent les cinq caractères Unicode représentant le hamza : « lettre hamza », « lettre alif avec hamza supérieur », « lettre waw avec hamza », « lettre alif avec hamza inférieur », « lettre yeh avec hamza supérieur », resp.

Il serait donc plus naturel, vis-à-vis de la logique de la langue arabe, d'utiliser un code générique pour « hamza », et de laisser au logiciel le calcul de la représentation graphique appropriée. Quand les voyelles sont explicitement indiquées par l'utilisateur (c'est le cas des textes classiques comme le Coran, et c'est ce que préconise Ahmed Lakhdar-Ghazal [16], directeur de l'Institut d'Arabisation du Maroc et créateur du système d'écriture simplifiée qui porte son nom et qui est amplement utilisé dans le maghreb [17]), ce calcul est immédiat, il suffit en effet de consulter la table ci-dessus. Ω intègre déjà cette fonctionnalité (introduite pour la première fois par Klaus Lagally dans ArabTEX [18]).

La situation se complique quand la voyellisation est partielle, c'est-à-dire lorsque l'utilisateur n'indique pas explicitement les voyelles. De plus, la propriété de « début de mot » n'est pas toujours triviale à vérifier pour l'ordinateur puisqu'il y a souvent des particules qui, tout en étant des mots à part, sont jointes au mot : par exemple dans يانّس وجون il y a en fait trois mots : يانّس, و et جون, le deuxième étant la particule *et*, collée au mot qui la suit. Les règles de représentation du hamza s'appliquent au mot en soi, et non pas à l'agglo-

mérat « particule(s) + mot ». Il faut donc que l'algorithme sache distinguer les mots des particules.

De nouveau, on a besoin d'un analyseur morphologique. Il existe depuis peu un tel outil commercial [19] mais qui, malheureusement, ne fonctionne pas en mode *batch* et donc ne peut s'intégrer à Ω. Nous sommes en contact avec des équipes développant de tels outils et projettons de les tester et intégrer dans Ω aussitôt que possible. Ces outils permettront aussi une post-voyellisation du texte, ce qui peut être très utile à un lectorat ne connaissant pas encore suffisamment l'arabe pour la faire mentalement : ce cas est à rapprocher de la méthode d'introduction de *furiganas* dans le texte japonais que nous allons décrire à la section 6.

4.2. L'article défini et le wasla

L'article défini alif-lam de l'arabe a une propriété spéciale : au début d'une phrase la lettre alif est prononcée alors qu'à l'intérieur de la phrase elle est phonétiquement assimilée avec la voyelle courte de la dernière lettre du mot précédent : ainsi الكتابُ الكبير sera prononcé *alkitaabulkabir* et non pas *alkitaabu alkabir*. Pour marquer graphiquement l'assimilation de l'alif on utilise un signe diacritique spécial, le *wasla* : الكتابُ ٱلكبير. Inversement, l'absence de wasla signifie que le alif est effectivement prononcé ; certains grammariens préconisent l'utilisation d'une voyelle fatḥa dans ce cas : الكتابُ ٱلكبير, utiliser les codes « lettre alif » et « voyelle courte fatḥa » pour représenter cela est de nouveau douteux : après tout, le alif est une voyelle longue et n'a donc nul besoin d'être suivie d'une voyelle courte supplémentaire !

Ω pourrait introduire ou vérifier l'existence et la validité du wasla. Notons qu'il y a d'autres mots qui prennent un wasla : ce phénomène est appelé « hamza instable ».

4.3. La nounation

Le terme *nounation* provient de la lettre *noun* : quand on veut indiquer l'indétermination, on ajoute à la voyelle courte finale d'un mot la valeur phonétique d'un noun, representée par un dédoublement de la graphie de la voyelle courte : كتابُ devient كتابٌ, etc. Unicode et les codages qui l'ont précédé définissent des versions avec nounation des trois voyelles courtes arabes ; du point de vue grammaire il est plutôt naturel d'utiliser un code spécial unique pour indiquer la nounation. Cette différence serait sans importance s'il n'y avait pas quelques phénomènes grammaticaux qui interviennent en rendant l'approche d'Unicode linguistiquement inappropriée.

— Le premier phénomène est purement grammatical : pour représenter la nounation de la voyelle courte fatḥa, la grammaire fait plus que simplement dédoubler la voyelle : il y a une lettre alif qui est ajoutée à la fin du mot. Ainsi, la nounation de وَلَد est وَلَدًا. Pour coder ce mot sous Unicode, on est obligé d'utiliser le code de la lettre « alif » ordinaire, ce qui est théoriquement une erreur puisqu'ici il s'agit d'une lettre muette d'origine purement grammaticale. L'histoire ne s'arrête pas là : la règle que nous venons d'énoncer a trois exceptions : la lettre alif muette n'est pas utilisée lorsque le mot se termine par un ta marbuta, un alif suivi d'un hamza, ou un alif surmonté d'un hamza. Dans ces cas, on se contente de la voyelle fatḥa dédoublée.
— Le deuxième phénomène est d'origine typographique : la tradition typographique est de placer le fatḥa dédoublé *avant* (c'est-à-dire à droite de) l'alif muet, sur la lettre qui le précède. Mais certains imprimeurs la placent *sur* l'alif muet. Coder ces deux représentations du même phénomène grammatical de manière différente, tout simplement parce que le codage est basé sur la graphie plus que sur le fond grammatical, est une approche plutôt malheureuse.

Puisqu'Ω est désormais capable d'appliquer toutes ces règles, l'arabophone conscient de la grammaire de sa langue peut dorénavant la saisir de manière plus logique et plus cohérente que celles des systèmes commerciaux, encore basés sur les contraintes de la machine à écrire. Le fait que les utilisateurs d'ArabTEX utilisent une approche similaire depuis déjà presque une décennie rend notre proposition encore plus réaliste. Bien sûr, celle-ci n'est qu'une des manières possibles de composer du texte arabe sous Ω : les utilisateurs qui préfèrent le mode « orienté graphie » de la saisie WYSIWYG pourront toujours le faire avec la même facilité.

5. Le thaï et la non-existence du mot

Un grand savant français a déclaré que « le mot n'existe pas, seule existe la phrase ». Pour le typographe cela n'est heureusement pas vrai (unephrasequineséparepaslesmotsseraitunvéritablecauchemar...) sauf dans le cas de certaines langues comme le thaï. Une phrase thaï consiste en une suite de lettres concaténées et absolument rien ne permet au non-thaïphone de distinguer les mots. Car les mots existent bel et bien, mais les Thaïlandais ne se sentent en rien obligés de les séparer.

Si le lecteur a trouvé amusant le dessin du chapitre H (H comme *hyphenation* = césure) du *TEXbook* [3], ce dessin qui montre le lion de TEX en bourreau, prêt

à décapiter le mot supercalifragilisticexpialidocious, c'était sans doute parce que ce mot n'est qu'une rarissime exception, une singularité de l'espace des mots...

L'exception devient la règle lorsqu'on passe au thaï : en effet les phrases thaï peuvent être très longues (imaginez du Proust traduit en thaï...) et on est bien obligé de les couper. Sans parler du problème des blancs entre les phrases qui peuvent, du fait de leur rareté, subir des étirements énormes : n'oublions pas que tout l'art de TEX est dans l'équilibrage des blancs entre les mots, mais quand ces blancs n'existent pas, même TEX reste les bras liés.

À cela s'ajoute une autre donnée : exceptionnellement on a le droit de couper aussi entre les syllabes. Techniquement parlant on a donc deux ensembles de césures possibles, de priorités sensiblement différentes. Ci-dessous, un texte thaï, composé dans la police Ω Serif Thai (corps 10/16), dessinée par Tereza Tranaka en 1998, lors de son séjour au Laboratoire Électrotechnique (ETL) de Tsukuba, Japon. Cette police a aussi été utilisée dans le livre d'Unicode version 3, pour représenter la table de l'écriture thaï [20, p. 430]. Le texte est composé trois fois : sans information paratextuelle,

> งานที่กำลังดำเนินอยู่ในปัจจุบันได้แก่งานตกแต่งประดับประดารอบๆ บริเวณงาน ใช้งบประมาณ 9 ล้านเหรียญ ซึ่งกล่าวได้ว่าเป็นงานตกแต่งที่ยิ่งใหญ่ชิ้นหนึ่งตั้งแต่ประเทศออสเตรเลียเคยจัดสวนบันเทิงประกอบไปด้วยเครื่องเล่นประเภทเทคโนโลยีชั้นสูงตั้งอยู่บนเนื้อที่ 5 เฮกเตร์และใช้งบสร้างถึง 50 ล้านเหรียญ ด้านการติดต่อกับนานาชาติและบริษัทต่างๆ จนถึงขณะนี้มีประเทศที่สนใจร่วมงานถึง 31 ชาติในจำนวนนี้ มีประเทศสหราชอาณาจักร สหรัฐอเมริกา โซเวียตรัสเซีย สาธารณรัฐประชาชนจีน แคนนาดา ญี่ปุน และเยอรมันตะวันตกรวมอยู่ด้วย.....ยอดจำนวนของบรรดาผู้มาร่วมงานจากนานาชาติมี 40 แห่ง

avec marquage explicite des mots :

> งาน|ที่|กำลัง|ดำเนิน|อยู่|ใน|ปัจจุบัน|ได้|แก่|งาน|ตก|แต่ง|ประดับ|ประดา|รอบๆ บริเวณ|งาน ใช้|งบ|ประมาณ 9 ล้าน|เหรียญ ซึ่ง|กล่าว|ได้|ว่า|เป็น|งาน|ตก|แต่ง|ที่|ยิ่ง|ใหญ่|ชิ้น|หนึ่ง ตั้ง|แต่|ประเทศ|ออสเตรเลีย|เคย|จัด|สวน|บันเทิง|ประกอบ|ไป|ด้วย|เครื่อง|เล่น|ประเภท|เทคโนโลยี|ชั้น|สูง|ตั้ง|อยู่|บน|เนื้อ|ที่ 5 เฮ|กเตร์|และ|ใช้|งบ|สร้าง|ถึง 50 ล้าน|เหรียญ ด้าน|การ|ติด|ต่อ|กับ|นา|นา|ชาติ|และ|บริ|ษัท|ต่างๆ จน|ถึง|ขณะ|นี้|มี|ประเทศ|ที่|สน|ใจ|ร่วม|งาน|ถึง 31 ชาติ|ใน|จำนวน|นี้ มี|ประเทศ|สหราช|อาณาจักร สหรัฐ|อเมริกา โซเวียต|รัสเซีย สาธารณรัฐ|ประชาชน|จีน แคน|

นาดา ญี่ปุ่น และเยอรมันตะวันตกรวมอยู่ด้วย.....ยอดจำนวนของบรรดาผู้มาร่วมงานจากนานาชาติมี 40 แห่ง

et avec marquage explicite des mots et des syllabes :

งาน|ที่|กำ|ลัง|ดำ|เนิน|อยู่|ใน|ปัจ|จุ|บัน|ได้|แก่|งาน|ตก|แต่ง|ประ|ดับ|ประ|ดา|รอบๆ บริ|เวณ|งาน ใช้|งบ|ประ|มาณ 9 ล้าน|เหรียญ ซึ่ง|กล่าว|ได้|ว่า|เป็น|งาน|ตก|แต่ง|ที่|ยิ่ง|ใหญ่|ชิ้น|หนึ่ง ตั้ง|แต่|ประ|เทศ|ออส|เตร|เลีย|เคย|จัด|สวน|บัน|เทิง|ประ|กอบ|ไป|ด้วย|เครื่อง|เล่น|ประ|เภท|เทค|โน|โล|ยี|ชั้น|สูง|ตั้ง|อยู่|บน|เนื้อ|ที่ 5 เฮ|กเตร์|และ|ใช้|งบ|สร้าง|ถึง 50 ล้าน|เหรียญ ด้าน|การ|ติด|ต่อ|กับ|นา|นา|ชาติ|และ|บริ|ษัท|ต่างๆ จน|ถึง|ขณะ|นี้|มี|ประ|เทศ|ที่|สน|ใจ|ร่วม|งาน|ถึง 31 ชาติ|ใน|จำ|นวน|นี้ มี|ประ|เทศ|สหรา|ชอา|ณา|จักร สหรัฐ|อเมริ|กา โซ|เวียต|รัส|เซีย สา|ธา|รณ|รัฐ|ประ|ชา|ชน|จีน แคน|นา|ดา ญี่|ปุ่น และ|เยอ|รมัน|ตะ|วัน|ตก|รวม|อยู่|ด้วย.....|..|ยอด|จำ|นวน|ของ|บรร|ดา|ผู้|มา|ร่วม|งาน|จาก|นา|นา|ชาติ|มี 40 แห่ง

Pour segmenter une phrase en mots et un mot en syllabes, nous utilisons des outils linguistiques développés par l'équipe de Virach Sornlertlamvanich, au Centre National d'Électronique et de Technologie des Ordinateurs (NECTEC) à Bangok [21]. Pour l'instant, ces outils insèrent encore simplement des `\discretionary` dans le texte : nous prévoyons une version d'Ω avec *plusieurs niveaux de priorité de césure* pour la fin de cette année 2001.

Il est possible (mais peut être pas très probable) qu'une approche plus classique par des motifs de césure, génerés par `patgen`, etc. aurait donné des résultats similaires. En choisissant les outils de notre ami Virach, nous avons opté (a) pour la précision des méthodes linguistiques vis-à-vis du manque d'intelligence inhérente de `patgen`, (b) pour l'existant : pas besoin de ré-inventer la roue puisque ces outils existent, fonctionnent parfaitement bien et s'intègrent sans la moindre difficulté dans Ω (l'équipe de Virach les a même adaptés à Unicode UTF-8 spécialement en vue de leur utilisation sous Ω). Mais ce choix présente aussi des inconvénients : lorsque l'on insère des commandes telles que `\discretionary` au plein milieu des mots, on casse le système de crénage entre les lettres.

Le système de crénage de TEX est d'office inadapté pour être utilisé pour le thaï. En effet, cette écriture, comme la plupart des écritures asiatiques du sud-est ou indiennes, possède un certain nombre de voyelles et autres signes qui se placent sur ou sous les lettres, à la manière des accents français ou des voyelles hébraïques. Informatiquement parlant, ces signes sont des « caractères » (au sens Unicode) qui se placent *entre* les caractères qui représentent les lettres.

Compte tenu de leur position graphique, ces signes ne devraient, la plupart du temps, affecter en rien le crénage entre les lettres. Mais comment créner sous TEX deux caractères quand un autre caractère se trouve intercalé ?

Ce problème nous pousse à redéfinir de fond en comble le système de césure et de crénage d'Ω, afin de le rendre capable de faire face aux problèmes d'une langue comme le thaï (qui n'est pas un cas isolé, loin de là). Mais avant de le faire, nous avons introduit deux primitives qui nous permettent de donner une solution provisoire, pas aussi élégante qu'un nouvel algorithme de césure et de crénage, mais drôlement efficace : il s'agit des primitives `\leftghost` et `\rightghost`. *Ghost* en anglais signifie *fantôme* ; en écrivant `\leftghost␣A` on obtient le « fantôme de la lettre A », c'est-à-dire un caractère invisible et de chasse nulle qui ne se manifeste que par son comportement : en effet, le fantôme de A réagit, en termes de crénage, comme la véritable lettre A lorsqu'il est précédé (ou suivi) d'une autre lettre (réelle ou fantomatique). Plus précisement, le *fantôme gauche* de A se comporte comme le côté gauche de A, et le *fantôme droit* de A se comporte comme le côté droit de A, toujours en termes de crénage, et uniquement en termes de crénage.

Pour créner donc deux lettres A et B, informatiquement séparées par une commande (comme `\discretionary`) ou par d'autres caractères, mais graphiquement concaténées, il suffit de faire suivre la première du fantôme gauche de la seconde.

Pour le thaï cela est fait par les OTP *ad hoc*. Cela produit une quantité de code énorme, mais ce code n'est jamais visionné par aucun humain, puisqu'il est généré par un ou plusieurs OTP et tout de suite « digéré » par Ω.

En développant ce système de composition du thaï on s'est aperçu qu'il manquait toujours de flexibilité et il en manquait cruellement, au point où certaines lignes débordaient très clairement, malgré toutes les césures proposées au système. Mais qui connait TEX sait que souvent il suffit de très peu pour obtenir des résultats surprenants : cela a été le cas ici, quand on a décidé de transgresser les règles de composition thaï et d'autoriser un infime interlettrage. Que le lecteur soit rassuré : l'interlettrage en question était de l'ordre de `0pt plus 1pt` pour un corps de 10 points (US). Comme d'un coup de baguette magique tout est rentré dans l'ordre : pour preuve, le texte thaï de notre exemple utilise cet interlettrage, heureusement invisible à l'œil nu, mais dont les effets sont eux bien visibles et agréables.

Cet interlettrage, combiné avec les lettres fantômes, est une première dans le monde de TEX : pour la première fois on a un blanc variable (une *glue* ou un « ressort ») et en même temps le crénage entre les lettres !

Nous n'allons pas présenter ici les fonctionnalités ni l'utilisation du système thaï d'Ω. Nous nous contenterons simplement d'affirmer que ce système vise à

résoudre les problèmes de composition du thaï de la manière la plus naturelle possible (« naturelle » pour un thaïphone, éventuellement technophobe ou TEXnophobe), tout en respectant au maximum les traditions typographiques thaï.

La réécriture de certaines parties d'Ω relatives à la césure et au crénage, contribuera davantage à l'accomplissement de ce projet.

6. Le japonais et les annotations supralinéaires

La langue japonaise utilise conjointement quatre systèmes d'écriture :

1. les idéogrammes d'origine chinoise (les *kanji*),
2. un syllabaire pour les mots d'origine japonaise (les *hiragana*),
3. un syllabaire pour les mots d'origine étrangère (les *katakana*) et
4. l'écriture latine (les *romāji*) avec en outre un accent barre pour les voyelles longues.

Les *kanji*, utilisés en Chine depuis le XX^{e} siècle av. J.-C., ont été introduits au Japon seulement au IV^{e} siècle av. J.-C. Ce qui a été une transformation radicale est le fait que, lors de l'introduction des *kanji*, les Japonais ont également emprunté aux Chinois le vocabulaire chinois correspondant. Ainsi, par exemple, les Japonais n'avaient qu'un seul mot, avant l'introduction de l'écriture, pour le concept de « milieu » : le mot *naka* ; quand le caractère 中 (qui signifie entre autres « milieu ») est arrivé au Japon, il est venu accompagné du mot chinois pour le même concept, qui est *chū*. Par conséquence, aujourd'hui on utilise aussi bien *naka* que *chū*, selon le contexte et selon un certain nombre de conventions.

La phonologie japonaise étant radicalement différente de la phonologie chinoise, les tons et toutes les autres subtilités phonétiques chinoises ont tout simplement disparu et on trouve aujourd'hui en japonais une foule d'homophones. On est ainsi arrivé à la situation actuelle du japonais, où (a) un caractère peut être prononcé, suivant le contexte, de plusieurs manières souvent tout à fait différentes, (b) des dizaines de caractères peuvent avoir exactement la même prononciation. À cela s'ajoute le fait que le nombre des *kanji* n'est pas limité : tout Japonais doit connaître les 1945 caractères de base, mais la lecture d'un livre peut nécessiter plusieurs milliers de caractères, une police standard en contient 7 000, les deux codages JIS en contiennent plus de 12 000 [22] et le projet informatique *GT-Mincho* de l'Université de Tokyo [24] a déjà répertorié et vectorisé quelques 68 000 caractères.

On voit donc la nécessité d'avoir plusieurs systèmes d'écriture : ils servent à désambiguïser le message parlé ou écrit. Un message oral peut être écrit de manière unique en utilisant des *hiragana* ou *katakana*, mais pour déterminer le sens de chaque mot il faut avoir recours aux *kanji*. Inversement, le message écrit est composé de concepts mais ne contient souvent pas d'information précise sur sa représentation orale : un verbe, un adjectif ou un nom utiliseront le même *kanji* tant qu'ils transportent le même concept : il faut recourir à l'oral pour les déterminer avec précision.

Il ne s'agit donc pas seulement d'utilisation de plusieurs systèmes d'écriture, mais d'utilisation *simultanée* de ces systèmes : la manière la plus rigoureuse d'écrire du japonais est d'utiliser des *kanji* combinés avec des *hiragana* ou *katakana* (ces deux écritures syllabiques sont appelées des *kana*). Typographiquement cela se fait en écrivant des *kanji* sur lesquels on ajoute des *kana* de petite taille, à la manière des signes diacritiques occidentaux ou des voyelles hébraïques. Ces *kana* de petite taille sont appelés des *furigana* ou *rubi*. Les *furigana* ont un grand nombre d'applications différentes, souvent en apportant de l'information non triviale.

Mais comme il arrive souvent dans les langues vivantes (comme par exemple dans le cas de l'accent tonique du russe ou des voyelles arabes), le message écrit se passe des systèmes de désambiguïsation, puisque le lectorat est capable d'en capter le sens sans besoin d'aide externe : les journaux japonais ne contiennent que rarement des *furigana*, pour des noms propres ou des noms de lieux.

Cette situation change, quand on change de lectorat, que ce soit au Japon (les enfants et adolescents ne connaissent qu'un nombre limité de *kanji*) ou à l'étranger (les étrangers apprenant le japonais). Pour ces catégories de lecteurs les *furigana* sont extrêmement utiles.

Nous proposons un système d'introduction automatique des *furigana*, basé sur un analyseur morphologique du japonais, du nom de `chasen` (形態(けいたい)) [25] et un dictionnaire des *kanji* établi par Jack Halpern [26].

Voici un petit texte japonais (tiré d'un article des auteurs sur Ω, publié dans le dernier numéro du magazine d'informatique japonais **bit** [27]) :

> Ωは，組版する文書にフィルタを選択的に適用することができる（選択的とは，どんなフィルタをいつでも使ったり止めたりできるということである．たとえば言語ごと，構成要素ごとに違うフィルタを適用することができる）．この能力は， NLP （自然言語処理）の手法を用いて，組版の新たな地平を拓くこととなった．実際，文書の加工に特別な自然言語向けソフトウェアを必要とす

る場合があるだろう．単にその言語の単語のコーパスで済む場合も，本当の形態素解析を要する場合もあるだろうけれど．

et voici le même texte muni de *furigana* :

Ωは，組版(くみはん)する文書(ぶんしょ)にフィルタを選択的(せんたくてき)に適用(てきよう)することができる（選択的(せんたくてき)とは，どんなフィルタをいつでも使(つか)ったり止(と)めたりできるということである．たとえば言語(げんご)ごと，構成要素(こうせいようそ)ごとに違(ちが)うフィルタを適用(てきよう)することができる）．この能力(のうりょく)は，NLP（自然言語処理(しぜんげんごしょり)）の手法(しゅほう)を用(もち)いて，組版(くみはん)の新(あら)たな地平(ちへい)を拓(ひら)くこととなった．実際(じっさい)，文書(ぶんしょ)の加工(かこう)に特別(とくべつ)な自然言語向(しぜんげんごむ)けソフトウェアを必要(ひつよう)とする場合(ばあい)があるだろう．単(たん)にその言語(げんご)の単語(たんご)のコーパスで済(す)む場合(ばあい)も，本当(ほんとう)の形態素解析(けいたいそかいせき)を要(よう)する場合(ばあい)もあるだろうけれど．

Nous voyons que, dans la plupart des cas, un *kanji* est surmonté d'un, deux ou trois *kana*. Mais dans certains cas (par exemple 場合(ばあい)) on voit un groupe de *kanji* surmonté d'un groupe de *kana*. Explication : les *kanji* sont souvent combinés avec d'autres *kanji* pour former des mots. Dans ce cas on parle de « complexes de *kanji* ». Pour des raisons pédagogiques, nous avons gardé dans le cas des complexes de *kanji* la correspondance visuelle entre un *kanji* et les *kana* qui lui correspondent en centrant les derniers sur le premier. Mais dans certains cas un complexe de *kanji* se dote d'une *prononciation entièrement nouvelle* qui n'a plus aucun rapport avec les prononciations de chaque *kanji* le constituant. On parle alors de complexe de *kanji irrégulier* ; dans ce cas, on centre l'ensemble des *kana* sur l'ensemble des *kanji*. La césure de ces constructions est un des problèmes de la composition du japonais.

Mais revenons à notre analyseur morphologique : en analysant ce texte il nous fournit des informations du type de celles de la figure 1. Nous y voyons que pour chaque complexe de *kanji* `chasen` fournit une « lecture », c'est-à-dire une représentation orale qui lui semble plus judicieuse que les autres, compte tenu du contexte. Mais cette représentation orale est globale pour le complexe : *a priori* il n'est pas du tout clair quel *kana* correspond à quel *kanji*, ni même s'il agit d'un complexe irrégulier, auquel cas cette correspondance est impossible par définition.

Notre outil consulte pour chaque *kanji* un dictionnaire [26] pour obtenir la liste de toutes les lectures de chaque caractère : il forme toutes les combinaisons possibles de lectures et les compare à la lecture donnée par l'analyseur

kana et *kanji*	*kana* (écriture)	*kana* (phonétique)	*romāji*
\section	\section	\section	\section
{\OMEGA	{\OMEGA	{\OMEGA	{\OMEGA
と	ト	ト	*to*
自然	シゼン	シゼン	*shizen*
言語	ゲンゴ	ゲンゴ	*gengo*
処理	ショリ	ショリ	*shori*
}	}	}	}
EOS			
\OMEGA	\OMEGA	\OMEGA	\OMEGA
は	ハ	ワ	*wa*
,	,	,	,
組版	クミハン	クミハン	*kumihan*
する	スル	スル	*suru*

FIGURE 1: Extrait de la sortie de `chasen`.

morphologique. Si une de ces combinaisons est identique à la lecture de l'analyseur, alors nous savons exactement quel sous-ensemble de *kana* correspond à chaque *kanji*. Si par contre aucune de ces combinaisons ne donne la lecture de l'analyseur, nous avons à faire à un complexe irrégulier. Nous composons donc les *kana* en fonction de ces résultats.

La composition des *furigana* est en partie encore un problème ouvert : ceux-ci entrent dans le contexte général des annotations supralinéaires. On a deux lignes de texte qui interagissent : lorsque les *kana* associés à un caractère le dépassent en largeur, faut-il éloigner ce caractère de ses co-caractères environnant ? Cela dépend bien sûr du dépassement, mais aussi des éventuels *kana* qui se trouvent sur les autres caractères. Que faire en fin ou en début de ligne ? Que faire lorsque le caractère est suivi de texte dans un autre système d'écriture ? Une multitude de problèmes non résolus : en matière de *furigana* et plus généralement d'annotations supralinéaires, Ω a encore du « pain sur la planche »...

Mais le jeu en vaut la chandelle : grâce à notre système d'« auto-furiganisation », tout texte japonais disponible sur support informatique devient si non lisible du moins plus accessible aux catégories des lecteurs mentionnées plus haut (enfants, adolescents, étrangers), dont nous-même faisons partie.

Conclusion

Dans cet article nous avons développé brièvement certaines applications de méthodes linguistiques à la composition sous Ω. Nous en avons implementé et testé certaines, mais le fait est que toutes ces applications constituent autant de directions de recherches futures. Dans l'avenir nous allons encore intensifier l'intégration de méthodes de traitement automatique des langues dans Ω dans le but d'en faire un outil de plus en plus intelligent et diversifié.

Références bibliographiques

[1] A. DE SAINT-EXUPÉRY, *Citadelle*, Gallimard, 1948.

[2] `http://fmg-www.cs.ucla.edu/geoff/ispell.html`

[3] D. E. KNUTH, *The TEXbook*, Addison Wesley, 1986.

[4] `http://www.gnu.org/manual/emacs/html_chapter/emacs_25.html#SEC214`

[5] `http://www.sil.org/computing/catalog/ktext.html`

[6] Y. Haralambous, Druckſatz in gebrochenen Schriften, Comptes-rendus de la réunion DANTE'00, Clausthal, 2000.

[7] `http://www.linguistik.uni-erlangen.de/~orlorenz/DMM/DMM.html`

[8] `http://www.linguistik.uni-erlangen.de/~bjoern/Malaga.de.html`

[9] `http://www.gutenberg.aol.de/`

[10] `http://www.gutenberg.aol.de/fontane/kinderjr/kinderjr.htm`

[11] `http://www.tonismos.gr/`

[12] `http://www.magenta.gr/gr/mon2pol/gr_mon2pol.htm`

[13] `mailto:ioannis.kanellos@enst-bretagne.fr`

[14] `http://www.ilsp.gr/`

[15] Y. HARALAMBOUS, *Unicode, XML, TEI, Ω and Scholarly Documents*, Comptes-rendus de la conférence Unicode 2000, Amsterdam, 2000.

[16] A. LAKHDAR-GHAZAL, *Arabe Standard Voyellé — Code Arabe*, Institut d'Études et de Recherches pour l'Arabisation, Rabat, 1988.

[17] Y. HARALAMBOUS, *Simplification of the Arabic Script: Three Different Approaches and their Implementations*, Springer Lecture Notes on Computer Science 1375, Comptes-Rendus de *Electronic Publishing, Artistic Imaging, and Digital Typography 7th International Conference on Electronic Publishing, EP'98. Held Jointly with the 4th International Conference on Raster Imaging and Digital Typography, RIDT'98*, St. Malo, 1998.

[18] `http://www.informatik.uni-stuttgart.de/ifi/bs/research/arab_e.html`

[19] `http://www.sakhr.com/Technologies/MMMP.htm`

[20] *The Unicode Standard Version 3.0*, The Unicode Consortium, Addison-Wesley, Reading, 2000

[21] `http://www.links.nectec.or.th`

[22] K. LUNDE, *CJKV Information Processing, Chinese, Japanese, Korean & Vietnamese Computing*, O'Reilly, Cambridge, 1998.

[23] `http://www2.gol.com/users/jpc/Japan/Kanji/history.htm`

[24] `http://www.l.u-tokyo.ac.jp/KanjiWEB/00_cover.html`

[25] `http://chasen.aist-nara.ac.jp/`

[26] `http://www.csse.monash.edu.au/~jwb/kanjidic.html`

[27] Y. HARALAMBOUS et J. PLAICE, 組版・文書処理システム Ω, **bit** 4, 2001, Kyoritsu Shuppan, Tokyo.

Les articles [6], [15], [17], [27], peuvent être consultés au format PDF sur `http://www.fluxus-virus.com/fr/research.html`