Cahiers GUTenberg

TEX : LES LIMITES DU MULTILINGUISME

Michel FANTON

Cahiers GUTenberg, nº 10-11 (1991), p. 73-79.

<http://cahiers.gutenberg.eu.org/fitem?id=CG_1991___10-11_73_0>

TEX : les limites du multilinguisme

Michel FANTON

Centre d'Etudes et de Recherche en Traitement Automatique des Langues
équipe de recherche de l'INALCO associée au CNRS et au Collège de France
adresse : CERTAL - INALCO, 2 rue de Lille 75007 Paris
fax : [33] (1) 49 26 42 99

Abstract. *This paper describes the specific features of arabic typesetting and gives an account of the price to pay in developing an arabicized version of TEX.*

Résumé. Cet article présente les particularités de la typographie en langue arabe et fait le bilan des compromis nécessaires au développement d'une version arabisée de TEX.

Mots clef : TEX, arabe

1. Introduction

Bien que les écritures latine et arabe aient, semble-t-il, une origine commune, leur évolution a été très différente. Dès l'époque romaine l'écriture latine possédait deux styles fondamentaux : l'écriture manuscrite et les caractères utilisés par les graveurs pour les inscriptions lapidaires.

C'est ce style lapidaire qui, par la suite, a été utilisé et perfectionné pour confectionner les caractères mobiles de la typographie. A l'inverse, le style premier de l'écriture arabe est l'écriture manuscrite et les inscriptions monumentales ont été faites à l'imitation de ce style[1].

Il est donc tout à fait naturel que, lors de l'introduction d'ailleurs tardive de l'imprimerie[2], le style des caractères typographiques ait été également conçu à l'imitation de ce style manuscrit.

La richesse de l'imagination des calligraphes a conduit à rendre très coûteuse l'imprimerie des textes arabes à l'aide des techniques traditionnelles. Les casses pouvaient contenir jusqu'à 1000 caractères différents.

[1] Sur ces différents points cf. [MacKay85] et [MacKay86].

[2] En dehors de quelques exceptions localisées.

Plus récemment, ce problème est devenu un obstacle pour l'enseignement et la diffusion de la culture et des études ont été entreprises pour simplifier la forme et le nombre des caractères arabes afin de diminuer les coûts de production. De nombreux travaux ont été effectués mais les seuls à être parvenus à un stade industriel sont ceux du Pr. Lakhdar-Ghazal à l'IERA. C'est la solution que nous avons adoptée pour "arabiser" TEX.

L'évolution récente des techniques typographiques vers la typographie informatique : photocomposeuses pilotées par ordinateur, polices de caractères numérisés, justification et mise en page par programme, font douter certains de la nécessité actuelle d'une telle évolution pour l'écriture arabe. Ils s'appuient sur l'exemple des Japonais qui ont su développer les techniques nécessaires à la manipulation d'un système d'écriture très complexe.

L'objet de cet article est de montrer que l'utilisation de logiciels conçus pour le traitement des caractères latins ne permet d'aboutir qu'à une solution approximative même si elle est indispensable. Malgré l'adoption de caractères simplifiés, de nombreux problèmes restent en suspens. Nous allons en évoquer quelques uns.

2. Particularités de l'écriture arabe

2.1. L'alphabet arabe

L'alphabet arabe[3] comporte 29 lettres différentes, mais en prenant en compte les différents signes orthographiques le Pr. Lakhdar-Ghazal arrive à un total de 41 signes[4].

2.2. Différentes formes de consonnes

L'écriture arabe d'imprimerie étant d'origine manuscrite, les consonnes et voyelles longues prennent des formes différentes si elles sont isolées ou selon la place qu'elles occupent dans un mot : début, milieu ou fin. Elles se lient à la suivante sauf 6 d'entre elles.

[3] cf. [Lakhdar-Ghazal83] pour une étude très détaillée de cet alphabet.

[4] En éliminant cependant certains archaïsmes ou signes redondant comme le alif-suscrit ou le wasla.

2.3. Ligatures entre consonnes

L'ampleur de ce phénomène dans la typographie arabe dépasse de loin ce qui se passe en typographie latine. Pour des raisons d'esthétique, des séquences de deux ou trois caractères sont *recomposées* pour n'en former qu'un. Une seule de ces *ligatures* est systématiquement faite dans toutes les circonstances. Il s'agit de ce que les arabes appellent le *lam-alif*, résultat de la combinaison de la lettre *lam* (qui correspond à un *l*) et de la lettre alif (qui est un *a* long). Les autres sont à l'initiative du typographe.

2.4. Voyelles brèves

Elles sont représentées par des signes qui se placent au-dessus et parfois au-dessous de la lettre sur laquelle elles portent. Elles ne sont pas employées dans la typographie courante : seuls les textes religieux et pédagogiques sont voyellés.

3. Codage informatique de l'alphabet arabe

3.1. Code arabe standard

Le développement de l'informatique dans les années 70 a montré l'urgence qu'il y avait à déterminer un code universel et cohérent pour l'alphabet arabe.

C'est au Pr. Lakhdar-Ghazal[5] et à ses recherches menées au sein de l'IERA que l'on doit les progrès déterminants qui ont été réalisés dans ce domaine.

La solution préconisée consiste à n'affecter qu'un code aux différentes formes de lettres. Cela facilite beaucoup la saisie des textes sur un ordinateur : l'opérateur n'a pas à déterminer la forme de lettre qu'il est en train de frapper. Cela permet également de diminuer considérablement le nombre de codes à prévoir. Cependant cela pose un certain nombre de problèmes épineux en ce qui concerne le formatage des textes à l'aide d'un ordinateur.

Les codes mis au point pour l'écriture latine ne fonctionnent pas selon ce principe : un *A* majuscule n'a pas le même code qu'un *a* minuscule. Le programme de formatage n'a pas de calcul à effectuer pour déterminer le code

[5] cf. [Lakhdar-Ghazal83].

exact du caractère à imprimer : il lui suffit de lire les données typographiques (chasse, ...) correspondant à ce code dans la table de la fonte courante.

Pour l'arabe une analyse minimale est nécessaire à laquelle on a donné le nom d'*analyse contextuelle*, car c'est par examen du contexte immédiat qu'on détermine la forme de la lettre[6].

3.1.1. Règles de base de l'analyse contextuelle

Il convient de distinguer 3 cas selon que la lettre en cours d'analyse se trouve au début, au milieu ou à la fin du mot.

Au début du mot

1. Si la lettre courante *doit être liée à la suivante*, elle prend la forme *initiale.*
2. Si la lettre courante *ne doit pas être liée à la suivante*, elle prend la forme *isolée.*

Au milieu du mot

1. La lettre courante *doit être liée à la suivante*
   (a) Si la lettre précédente *ne doit pas être liée à la suivante*, la lettre courante prend la forme *initiale.*
   (b) Si la lettre précédente *doit être liée à la suivante*, la lettre courante prend la forme *médiane.*
2. La lettre courante *ne doit pas être liée à la suivante*
   (a) Si la lettre précédente *doit être liée à la suivante*, la lettre courante prend la forme *finale.*
   (b) Si la lettre précédente *ne doit pas être liée à la suivante*, la lettre courante prend la forme *isolée.*

En fin de mot

1. Si la lettre précédente *doit être liée à la suivante*, la lettre courante prend la forme *finale.*

[6] Notons que lorsqu'on envisage le cas d'un éditeur de texte le problème est plus délicat puisqu'on prendre en compte les insertions et le suppressions de caractères.

2. Si la lettre précédente *ne doit pas être liée à la suivante*, la lettre courante prend la forme *isolée*.

4. Justification de textes en arabe

La coupure de mot est interdite en arabe. Certains attribuent cette interdiction au caractère manuscrit de l'écriture. Il nous semble plutôt que cela tienne à deux autres faits :

1. La structure sémitique de la langue sur laquelle nous ne étendrons pas ici.
2. L'absence de voyellation des textes courants.

Si la coupure des mots était acceptée en arabe, la lisibilité en serait considérablement affectée.

4.1. Procédés traditionnels de justification

Ne disposant pas de la coupure des mots les typographes arabes ont eu recours à d'autres techniques inspirées de la calligraphie manuelle. Nous en citerons quelques unes :

— Utilisation de caractères à forme allongée, pouvant être utilisés en fin de ligne.

— Allongement de la barre de liaison entre certaines lettres selon des critères esthétiques[7].

— Recours ou non à des ligatures qui ont, en général, pour effet de diminuer la chasse du groupe de caractères considéré.

— Acceptation d'espaces plus importants entre les mots.

4.2. Conséquences sur l'automatisation du processus de justification

Les diverses techniques qui viennent d'être évoquées ont toutes en commun, hormis la dernière qui n'est qu'un pis-aller, d'avoir une logique profondément différente de ce qui est fait pour la typographie latine.

[7] De la même façon qu'un typographe en caractères latins décide qu'une coupure est bonne ou mauvaise en fonction des traditions typographiques et de critères esthétiques.

Elles impliquent notamment que le programme de justification intervienne dans le choix des caractères à imprimer.

Une autre technique concevable serait d'accroître dans la même proportion la longueur de toutes les barres de liaison entre consonnes qui se lient lorsque la longueur du *blanc maximum* est atteinte.

Certains programmes de justification latins effectuent une certaine manipulation limitée de l'interlettrage. En arabe le problème se complique, car il faut déterminer si la séparation entre deux lettres doit s'effectuer au moyen d'une barre de liaison ou d'un espace. La barre de liaison peut être agrandie, sans trop nuire à l'esthétique, dans des proportions beaucoup plus importante que l'espace.

Les algorithmes, mis au point pour la typographie latine, sont très démunis en face d'une telle situation. TEX, en particulier, ne peut prendre en compte aucune des techniques qui viennent d'être brièvement décrites.

Il ne peut, du fait de la notion de *boite* et de *glue*, même pas manipuler l'interlettrage.

5. Conclusion

Cet article est destiné à mettre en perspective les travaux d'arabisation de TEX que nous présentons[8] dans ce même colloque et à dissiper certaines illusions sur le caractère universel de procédés qui, mis au point pour une langue donnée, s'adaptent parfois très mal à d'autres langues.

Références bibliographiques

[Fanton89] Michel FANTON, "Arabisation d'un système de photocomposition numérique", communication donnée au colloque *informatique et langue arabe* organisé par l'Institut du Monde Arabe et la Commission Economique et Sociale des Nations Unies pour l'Asie Occidentale à Paris en octobre 1989.

[Hamm75] Roberto HAMM, "Pour une typographie arabe", Sindbad, Paris, 1975.

[Knuth-Mackay87] Donald E. KNUTH and Pierre A. MACKAY, "Mixing right-to-left texts with left-to-right texts", *TUGboat*, Volume 8 (1987), no. 1, pp. 14–25.

[8] cf. O. Boughaba, S. Boutalbi & M. Fanton : "vers une version arabisée de TEX", pp. 25-44.

[Lakhdar-Ghazal83] Ahmad LAKHDAR-GHAZAL, l'alphabet arabe et les machines, in *Applied Arabic Linguistics and Signal & Information Processing, proceedings of the 1st Fall Session of Arab School on Science & Technology*, Rabat (Morocco) 1983, Volume 1 pp. 131–133, Volume 2 pp. 233–257.

[MacKay85] Pierre A. MACKAY, "Modern font design and typesetting", in *Informatics and Applied Arabic Linguistics, proceedings of the 7th Summer Session of Arab School on Science & Technology*, Zabadani Valley (Syria) 1985, pp. 2A2-1–2A2-7.

[MacKay86] Pierre A. MACKAY, "Typesetting Problem Scripts", *Byte*, February 1986, pp. 201–216.

[Microsoft88] Microsoft, "MS-DOS User's Guide. Arabic Supplement", 1988.