«How to bring a language to the future»

Un article de Alizeh Kohari publié sur *Rest of World*, le 9 février 2021. Consulté le 19.03.2021 [en ligne].

https://restofworld.org/2021/bringing-urdu-into-the-digital-age/

En 2014, Mudassir Azeemi écrit une lettre ouverte à Tim Cook, PDG d'Apple. Il l'alerte sur les difficultés d'écrire numériquement l'ourdou. Mudassir Azeemi s'était déjà penché sur la question et avait mis au point, en 2010, un clavier ourdou téléchargeable (sous IOS). Malheureusement, bien que l'on puisse « taper » en ourdou, l'écriture qui s'affichait était retranscrite par une police entièrement différente.

L'ourdou est un dérivé de l'alphabet arabe. Il est parlé par près de 170 millions en Asie du Sud et la diaspora sud-asiatique. L'arabe s'écrit avec le script *Naskh* tandis que l'ourdou s'écrit avec le script *Nasta'līq*. Au moment où Mudassir Azeemi publie sa lettre ouverte, *Nasta'līq* était presque introuvable en ligne. Pour écrire en ourdou, il fallait soit passer par le *Naskh* ou bien l'écrire phonétiquement en caractères latins.

Mudassir Azeemi est développeur. C'est en regardant ses enfants grandir en Californie, qu'il a eu peur que ces derniers de perdre l'héritage de leur langue. Et c'est donc ainsi penché sur l'avenir de l'ourdou.

Ces premières tentatives ne sont pas rentables (-50 000 $), mais il reste déterminé. Le clavier ourdou d'Apple en 2013, écrivait, au grand dam des utilisateurs ourdou, en *Naskh*, la police arabe. La solution de Mudassir Azeemi est alors de convaincre les systèmes d'exploitation (Microsoft, Apple, Google, ...) qu'il leur faut intégrer *Nasta 'līq*.

Dans sa lettre, il écrit « The only hurdle for us is to bring the typeface that truly represent[s] the language. [...] And every language, when it is written, shines when using the typeface which it truly presented in the world. »

Dans le sous-continent, l'ourdou est le script qui diffusait le Coran, la parole de Dieu. L'ourdou serait né par Mir Ali de Tabriz, un calligraphe persan du XIVe siècle. Ali, le gendre du prophète Mahomet, serait venu à lui dans un rêve pour lui demander de dessiner des lettres ressemblant à des « ailes d'oies volantes ».

Cette insistance sur le caractère d'écriture de l'ourdou est-elle vraiment pertinente ? Le texte est un moyen d'arriver à une fin. En soit, on peut écrire l'ourdou en *Naskh*. La demande du numérisé n'est-elle pas juste un problème mineur de typographie ?

Nasta'līq est un cauchemar à coder : il se déplace de droite à gauche (comme les scripts arabes), il se déplace en pente vers le bas (plus le mot est long, plus la pente est raide), et la forme de chaque lettre dépend les lettres qui l'encadrent. Donc bien que ce soit un système alphabétique à 39 lettres, il y a des milliers de permutations. Et c'est cette difficulté-là qui pose problème.

Dès 1951, quatre ans après s'être soustrait à la domination britannique, le Pakistan (où l'ourdou est la langue officielle et a joué un rôle primordial dans le mouvement d'indépendance) crée une imprimerie nationale. Elle pourrait composer en ourdou ainsi qu'en anglais. L'ourdou joue un rôle d'édification de la nation. Bien que les principales entreprises d'impression de l'époque, telles que Monotype et Linotype, développaient des polices ourdou, les éditeurs locaux les rejetaient parce qu'elles ne correspondaient pas à l'esthétique locale *Nasta'līq*.

Nasta'līq rencontrait des problèmes avec la technologie parce que cette dernière fut conçu avec l'anglais à l'esprit. Comme la machine à écrire. Comme le relève l'historien Thomas S. Mullaney dans « The Chinese Typewriter », toutes les autres langues sont considérées comme des permutations à partir de cette norme. L'hébreu est l'anglais à l'envers, l'arabe est l'anglais à l'envers et en cursive, le russe est l'anglais avec différentes lettres, siamois est l'anglais avec trop de lettres, et ainsi de suite... Peut-être que la seule langue à échapper à l'hégémonie latine fut le chinois. Étant ni alphabétique, si syllabique, le chinois a du être imaginé en dehors des limites de la technologie existante. *Nasta'līq* ne fut probablement pas suffisamment important pour se pencher avant sur la question.

Dans l'Iran moderne, *Nasta'līq* a du changé pour répondre aux limites de la technologie existante. Une identité typographique, qui fonctionne dans les limites de la presse à bloc, a lentement pris place au fil des ans. Le Farsi est la langue officielle de l'Iran. Elle fut styliquement suffisamment différente de *Naskh* arabe pour être suffisamment « persan ». Un journal de Téhéran, Iran et un de Karachi, Pakistan seront très différant l'un de l'autre.

Pourquoi l'ourdou du Pakistan n'a pas suivi le même chemin typographique que le Farsi en Iran ? Outre sa fidélité obstinée au *Nasta 'līq*, c'est peut-être la différence entre l'ourdou et les autres langues s'écrivant en *Nasta 'līq*. Une lettre est particulièrement influente sur l'apparence de l'ourdou écrit : *baṛī ye*. « and its typographic rendering is among the most challenging design questions in an Arabic font » écrit Titus Nemet dans *Arabic Type-Making in the Machine Age*.

Le *baṛī ye* ressemble à un coude plié et entraîne de nombreuses difficultés mécaniques au niveau du crénage, des placements de points et des signes diacritiques. Cette lettre, en autres, semble si insurmontable que dans les 60's, le dirigeant pakistanais Ayub Khan propose de passer à l'alphabet latin. Écrire en ourdou avec les lettres latines. C'est un refus presque immédiat. Le *Nasta'līq* est un élément prépondérant de l'identité islamique du pays. Sans progrès significatif dans l'impression de l'ourdou, les quotidiens sont écrits à la main jusqu'au 80's, calligraphes travaillant jour et nuit.

Ahmed Mirza Jamil, imprimeur à Karachi, hérite de l'amour pour le *Nasta'līq* de son père. Ce dernier s'était donné la tâche de ré-écrire lui-même à la main le Coran. Après un tiers du livre, il meurt. Ses fils reprennent son projet. Ils ont photographié les neufs chapitres manuscrits écrits par leur père, et ont coupé et collé les formations de lettres pour les reconstituer, comme un puzzle. C'est après 14 ans de travail, qu'ils ont fini les deux tiers restants du Coran de leur père. Ce travail a permis de semer des graines d'une percée dans la composition de l'ourdou.

En 1980, Mirza Jamil recense toutes les combinaisons de lettres ourdou auxquelles il a pensé, soit environ 20 000, selon les témoignages. Ces dernières vont être appelées « ligatures » et sont la base d'une nouvelle police ourdou : *Noori Nastaliq*. Ces ligatures vont résoudre la nature changeante de l'écriture. Au lieu de représenter des lettres, il représente des combinaisons entières de lettres écrites ensembles. Les 20 000 combinaisons ne sont pas exhaustives, mais la chance d'utiliser une qui ne ferait pas partie de cette liste est négligeable. Cette résolution du problème ourdou s'est inspiré de la rencontre d'un clavier chinois (engin de 48 pouces avec 500 caractères).

Mirza Jamil fut salué comme le Gutenberg du Pakistan avec sa typographie Noori Nastaliq et a même reçu les honneurs d'État.

Pendan des décennies, Noori Nastaliq fut la typographie dominante dans l'édition ourdou. Aujourd'hui, la version numérisé de celle-ci est toujours en utilisation. Mais cela reste une typographie faite dans les 80's : certaines ligatures ne s'affichent pas, les glyphes se bousculent et des amas de nuqtas mal placés gâchent le texte.

Nasrullah Mehr, calligraphe, explique les inconvénients d'une approche de la composition basée sur les ligatures « you can do 42 thousand ligatures, you can do a hundred thousand, you can do one million — it won't end. [...]Words from other languages crop up, new words come up ; you cannot remain true to a language this way ».

À contrario de son père, Nasrullah Mehr, Zeeshan est plus passionné par les mathématique que par la calligraphie. Père et fils se sont mis à développer une police numérique : *Mehr Nastaliq*. C'est la première typographie ourdou à être entièrement développée localement.

Ils sont des pionniers dans les polices numérisées ourdou. Le *Nasta'līq* est à la traîne dans l'espace numérique et c'est difficile de la faire exister dans cet espace déjà existant et défini. Là où il existe des milliers de polices numériques latines, les polices numériques ourdou sont trop nombreuses. Les typographies *Nasta'līq* dans les rues du Pakistan sont très probablement peinte à la main.

Ils sont partis de zéro et on commencé leur travail par une recherche Google « what are fonts ? how do we develop them ? ». Zeeshan Mehr explique qu'ils ont acheté un logiciel, mais qu'ils ont mis du temps à comprendre comment il fonctionnait et donc à démarrer. La typographie fut construite en collaboration avec un conseil technologique gouvernemental pendant 10 ans, Mehr Nastaliq utilise 500 caractères (écrits à la main par Nasrullah, et une infime fraction des 20 000 glyphes répertoriés par Jamil). La police pèse 60 Ko. Elle ne ralentit pas les sites web et se charge rapidement. On peut même allonger les lettres et ajouter des signes diacritiques.

Les Mehr ont maintenant une entreprise, MehrType, et souhaitent exploiter leur expertise. Ils ne veulent pas seulement faire une bonne police *Nasta'līq*, mais une florilège de bonnes polices.

Nasrullah travaille sur la transformation de l'écriture d'Abdul Majeed Parveen Raqam, fondateur de l'école Lahori du XXe siècle de *Nasta'līq*. Cette calligraphie est surnommée l'écriture nationale du Pakistan. 75 ans après sa mort, son écriture est convertie en police numérique et ainsi prolonge sa durée de vie.

Au moment où Mudassir Azeemi a contacté Tim Cook en 2014, les technologies de composition avaient plus ou moins rattrapé les subtilités du *Nasta'līq*. Mehr Nastaliq était en cours de développement et le script était difficilement trouvable en ligne.

Ce manque de police facile d'accès pousse les poètes ourdou à écrire à la main ou avec des logiciels spécifiques et spécialisés. Les poèmes sont alors importés en fichier image puis partagés sur les réseaux. Le même procédé est utilisé pour les sites Internet. La translittération latine de l'ourdou fut grandement utilisée également, que ce soit dans les livrets de poésie par SMS devenus populaires, ou bien même dans l'enseignement scolaire. Certains élèves faisaient leur devoir en translittération latine. « Writing a language in another script is like trying to drop off your skin and trying to have a new one » déplore un enseignant.

Que faudrait-il faire pour que *Nasta'līq* soit facile d'accès en ligne ? On pensait que la solution était la prise en charge au niveau du système. Ce consensus croissant à déclencher une vague d'activisme de la part des consommateurs et un appel à un « humanisme linguistique » a été fait.

Une semaine après sa lettre ouverte à Tom Cook, Mudassir Azeemi a reçu un appel d'un représentant d'Apple. Reconnaissant sa plainte, Apple promet de se pencher sur le problème. Trois ans après, la police *Noto Nastaliq* devient la police ourdou par défaut sur les produits Apple.

Pendant un moment, la crise existentielle en ligne de l'ourdou a eu l'impression d'être évitée. Zeerak Ahmed (concepteur de software pakistanais) est plus septique. Dans un article de blog de l'époque, il écrivait que l'intégration de la police *Nasta'līq* avait été soit timide, soit négligente. En effet, les mots étaient si petits qu'ils étaient presque illisibles sur écran. Il soulève ironiquement le problème du mot ourdou « complet » : il ne rentre pas complètement dans l'espace texte des messages Apple.

Selon lui, le problème principal n'est pas la non-disponibilité des polices *Nasta'līq* mais les interfaces systèmes qui sont construits principalement pour le latin. « We're just at the point where we're beginning to make sure our interfaces work throughout the world — that the moment you switch to a right-to-left language, things don't break» dit Zeerak Ahmed.

Si le problème de *Nasta'līq* est théoriquement résolu, ils s'en posent de nouveaux avec les nouvelles avancées : « We don't have the data needed to build artificial intelligence on top of existing technology: voice, handwriting recognition, dictionaries, autocorrect» dit Zeerak Ahmed.

Selon lui, les données déjà existantes en ourdou sont pleines d'erreurs. L'absence de logiciel de correction orthographique entraîne une prolifération de fautes de frappe. « All Urdu software is broken because the underlying data is broken » dit-il. Le machine-Learning est un désastre : « We simply took a bunch of Urdu digests and digitized them. It's just blobs of text. You don't know when it's from, where it was published, who wrote it ».

Zeerak Ahmed travaille à créer une dataset en ourdou qui pourra supporter le machine-learning.

L'ourdou écrit n'est pas seulement des caractères arabes, il intègre de la ponctuation latine, des mots anglais écrit en latin, mais aussi des mots écrit en *Naskh*. Donc la banque de 20 000 combinaisons de Mirza Jamil est de loin d'être complète.

Les problèmes et le retard de l'ourdou à s'intégrer dans les logiques inhérentes aux polices de caractères, ont retardé certains progrès dans la langue ourdou selon Zeerak Ahmed. « The one major discipline that has shown up in the world since the eighties is information technology, and that seems to be the one discipline that, coincidentally, doesn't seem to have any technical vocabulary in Urdu. There's no way to prove it at a causation level, but I think it says a lot about what incomplete digitization does to the development of a language ».

La dataset de Zeerak Ahmed a atteint suffisamment de données pour être le point de départ du machine-learning en ourdou. Et il a bien conscience de la responsabilité qui découle d'une telle technologie. En effet, la technologie modifie le langage des gens par ses contraintes et limites.

« Because it means, if I feed my system bad grammar, it's going to suggest bad grammar, and people are going to write in bad grammar and that bad grammar is slowly going to be ingrained. If I take terminology and slowly ingrain it, it will become part of our vernacular » dit-il. Sa dataset est open-source.

L'ourdou n'est pas l'unique langue à être confrontée à une crise existentielle numérique. Le linguiste András Korna a mené une étude en 2013, qui démontre que sur 7 000 langues utilisées dans le monde (dont 2 500 considérées en disparition), 5 % d'entre elles devenaient des langues pleinement fonctionnelles en ligne. L'expansion d'Internet précipite une extinction linguistique massive.

Pour que l'ourdou soit le plus possible intégré au web, il faut s'attaquer à plusieurs problématiques : que les calligraphes traditionnelles fassent parti de la conversation, se dégager de l'ombre projetée par la technologie latino-centrée et sur la pression populaire sur les principales plateformes technologiques. La plupart de locuteurs ourdou ne sont pas conscients de la manière dont la technologie façonne et a façonné la langue.

Zeerak Ahmed conclus : « I don't think it's an eventuality that all languages will one day be represented fully digitally. It's totally plausible we get to a point where it makes more sense to shift the culture than to shift the underlying tech ». Par exemple : en 2020, la communauté scientifique a modifié les directives officielles pour nommer les gènes. Puisque Microsoft Excel les intégrait comme étant des dates. Pour résumer, la tradition scientifique a fléchi devant le formatage informatique. La même chose pourrait arriver sur le plan culturel.