Extlang ou pas extlang ?

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 26 mai 2008

https://www.bortzmeyer.org/extlang-or-not-extlang.html

Le groupe de travail LTRU <http://tools.ietf.org/wg/ltru> ("*Language Tag Registry Update*") de l'IETF est très en retard dans son projet de normalisation des futures étiquettes de langue <http://www.langtag.net/>, ces courtes chaînes de caractères qui permettent d'indiquer la langue d'un document ou de préciser la langue qu'on souhaite utiliser sur Internet. Une des raisons de ce retard est le débat sans fin sur les « "*extlangs*" », ces mécanismes permettant de représenter le concept de « macrolangue » introduit par la norme ISO 639-3 <https://www.bortzmeyer.org/iso-639-3.html>. En gros, l'égyptien doit-il se noter ar-arz ou bien arz ? Et le cantonais doit-il être zh-yue ou simplement yue ?

Dans la norme actuelle, le RFC 4646 [1], les étiquettes de langue sont formées de plusieurs sous-étiquettes, chacune identifiant la langue, l'écriture ou le pays. Les sous-étiquettes qui identifient la langue sont tirées de ISO 639-1 ou ISO 639-2, normes qui placent toutes les langues sur un pied d'égalité.

Mais le monde des langues humaines est complexe. La définition traditionnelle d'une langue est le critère de compréhension mutuelle. Si différents que puissent être les accents, le vocabulaire et l'orthographe, l'anglais d'Australie peut être compris par un irlandais (parfois avec effort). C'est donc la même langue, dont la sous-étiquette est en. De même, si proches que soient le français et le catalan, par leur origine commune, deux locuteurs de ces deux langues ne peuvent pas se comprendre. Il s'agit donc bien de deux langues distinctes, dont les sous-étiquettes sont fr et ca.

Naturellement, il existe une zone grise : le danois est-il si différent du norvégien ? L'arabe marocain du tunisien ? Le cas est d'autant plus difficile que certaines langues sont différentes à l'oral mais moins à l'écrit (c'est justement le cas de l'arabe). D'une certaine façon, il n'existe qu'une langue arabe. D'une autre, il en existe plusieurs (qui ne recoupent pas forcément exactement les frontières nationales). Parfois, l'histoire ou la politique contribuent à brouiller la perception linguistique correcte, comme dans le cas du mandarin, souvent appelé chinois par abus de langage, parce que c'est la langue de l'État chinois.

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc4646.txt

Pour tenter de modéliser ce monde très complexe et qui n'a pas été conçu rationnellement, SIL, l'organisation qui a été la principale responsable de la norme ISO 639-3 <https://www.bortzmeyer.org/iso-639-3.html> a créé un nouveau concept, celui de **macrolangue**. Une macrolangue est une langue unique, selon certains critères, et un groupe de langues selon d'autres. Les deux plus célèbres sont le chinois (avec la sous-étiquette `zh`) et l'arabe (`ar`). Le registre d'ISO 639-3 <http://www.sil.org/iso639-3/> indique donc pour chaque langue si elle est couverte par une macrolangue. Le cantonais est ainsi « couvert » par le chinois. Notons tout de suite que les langues couvertes (parfois appelées par dérision « microlangues ») ne sont **pas** des dialectes, ce sont bien des langues séparées, typiquement sans mutuelle compréhensibilité.

Comment représenter les macrolangues dans le successeur du RFC 4646 ? La première idée, dont les prémisses figurent dans le RFC 4646, était d'utiliser un concept nouveau, l'"*Extended Language Subtag*", l'**extlang**. Dans ce système, tel que prévu à l'origine, la première sous-étiquette identifiait la macrolangue (ou la langue tout court pour les cas - les plus fréquents - où il n'y avait pas de macrolangue) et un ou plusieurs "*extlangs*" la suivaient. Ainsi, l'arabe tunisien était `ar-aeb` et le zapotèque de la Sierra de Ju[Caractère Unicode non montré [2]]rez était `zap-zaa` (`zap` étant la sous-étiquette de la macrolangue zapotèque).

Pour comprendre l'intérêt de ce système, il faut voir que les logiciels qui manipulent les étiquettes de langue, lorsqu'ils ne trouvent pas une langue satisfaisante, cherchent en général en supprimant les sous-étiquettes en partant de la droite, où se trouvent les sous-étiquettes les moins significatives (ce mécanisme est décrit dans le RFC 4647). Ainsi, si on demande à un système de recherche un document en `az-Arab-IR` (l'azéri écrit dans l'écriture arabe tel qu'il est utilisé en Iran), et qu'aucun document ne correspond à cette étiquette, le logiciel peut chercher s'il a `az-Arab` (en abandonnant les spécificités iraniennes) ou même `az` (de l'azéri, quelles que soient ses autres caractéristiques). Les "*extlangs*" collaient bien à ce modèle. Une demande de `sq-als` (le tosque) serait ainsi tronquée en `sq` (la macrolangue pour l'albanais), ce qui ne serait pas une mauvaise solution de repli.

Mais, à l'examen plus attentif des "*extlangs*", plusieurs problèmes sont survenus. D'abord, la constatation que le repli en tronquant par la droite ne donnait pas toujours des résultats corrects. Rien ne garantit que les langues couvertes par une même macrolangue sont mutuellement compréhensibles, bien au contraire. Donc, le repli « aveugle » obtenu en supprimant les sous-étiquettes en partant de la droite ne va pas être satisfaisant. Ensuite, les "*extlangs*" compliquent le modèle puisqu'ils représentent un nouveau cas à spécifier, à implémenter et à expliquer. Enfin, les "*extlangs*" peuvent porter un message erroné, celui comme quoi une langue particulière du groupe serait la langue principale. Ce message peut être souhaité (la plupart des arabophones sont très attachés à la référence à une langue arabe unique), ou pas mais il est toujours délicat, puisqu'il place les langues dans une hiérarchie. De nombreuses discussions, parfois houleuses, avaient animé le groupe de travail autour de ces problèmes.

C'est ainsi que LTRU <http://tools.ietf.org/wg/ltru> avait, en décembre 2007, renoncé à utiliser <http://www.ietf.org/mail-archive/web/ltru/current/msg08935.html> les "*extlangs*" et changé <http://www.ietf.org/mail-archive/web/ltru/current/msg08954.html> les "*Internet-Drafts*" pour mettre en œuvre cette décision. Dans cette nouvelle version, le registre des langues gardait l'information comme quoi une langue était couverte par une macrolangue mais les étiquettes de langue étaient uniquement formées avec la langue elle-même. Le tosque était donc noté `als`, le zapotèque de la Sierra de Ju[Caractère Unicode non montré]rez `zaa` et le mandarin `cmn`.

Mais, à l'IETF, les choses ne se passent pas toujours aussi simplement. Cinq mois après ce changement, les problèmes de la nouvelle version apparaissent, souvent soulevés par des gens qui n'avaient

2. Car trop difficile à faire afficher par LaTeX

guère contribué au travail concret, et rien dit à l'époque de l'abandon des "*extlangs*". Dans les cas où le groupe autour d'une macrolangue comporte une langue dominante (ce qui est le cas du chinois avec le poids du mandarin, ou celui de l'arabe avec la référence qu'est l'arabe standard, mais pas du zapotèque), de nombreux documents ont déjà été étiquetés avec l'étiquette qui est devenue celle de la macrolangue. Par exemple, les ressources en mandarin ne devraient plus être notées `zh` comme avant mais `cmn`. Que faire avec les innombrables documents en mandarin qui avaient été étiquetés `zh` en suivant le RFC 4646 ?

Idéalement, le seul registre des langues suffirait à trouver des bonnes solutions. Mais, en fait, les décisions à prendre par le logiciel dépendent souvent de critères non linguistiques. Par exemple, si un utilisateur a configuré comme langue préférée le breton, il n'est pas déraisonnable de lui servir du texte en français. Non pas que les deux langues soient proches (la première est une langue celte et le second une langue romane) mais parce que, en pratique, presque tous les gens qui comprennent le breton comprennent également le français. Mais on ne peut pas mettre cette information (de taille importante et qui change souvent) dans le registre ! Il faudrait donc se résigner, ce qui ne semble pas facile, à ne pas avoir dans le registre d'information permettant un repli « intelligent ».

Après, là encore, un débat très chaud, le groupe LTRU a finalement fait marche arrière `<http://www.ietf.org/mail-archive/web/ltru/current/msg10469.html>` le 26 mai 2008, revenant aux "*extlangs*". Cela nécessite de revenir à la précédente version des "*Internet-Drafts*", de revoir documents et implémentations. Et sans garantie que cela ne recommence pas dans quelques mois...

Ma vue personnelle est que le système est instable : on peut bâtir un bon argumentaire pour les deux solutions (contre les "*extlangs*", voir le texte de Mark Davis `<http://www.ietf.org/mail-archive/web/ltru/current/msg10349.html>`), aucune n'est parfaite, car les langues humaines n'ont pas été conçues pour faciliter la tâche de l'IETF. Il aurait fallu adopter une solution et s'y tenir mais les mécanismes de décision à l'IETF ne rendent pas cela facile.

`https://www.bortzmeyer.org/extlang-or-not-extlang.html`