Transcodage d'un échantillon de 1000 SMS, extraits du corpus « 88milSMS ».

Une fois l'anonymisation terminée, les SMS sont prêts à être transcodés en français standardisé afin de permettre d'éventuels traitements ultérieurs en linguistique-informatique (incluant des analyseurs morpho-syntaxiques). L'idée est de restituer l'orthographe et la grammaire afin de faciliter la compréhension, mais non pas d'« injecter » des éléments supplémentaires (*cf*. Exemples 1 à 3 ci-dessous). Tous les SMS bruts anonymisés, un échantillon de 1000 SMS transcodés et un échantillon de 100 SMS annotés sont disponibles dans le corpus diffusé. Le transcodage est utile pour le grand public, ou pour ceux qui veulent lire et comparer rapidement les SMS bruts anonymisés et transcodés. Cependant, d'un point de vue linguistique, il est extrêmement difficile de procéder à un transcodage qui convienne à tous, car les interprétations sont nombreuses et variées.

SMS brut anonymisé (n° **22446** du corpus 88milSMS) :
En fait c rien de spécial, jprends juste un peu de recul et jcomprends pas ce que jfous là, fac, psycho, montpellier, pourquoi simplement je vis, enfin bref rien de grave. Qu'est ce qui cloche chez toi?
SMS anonymisé et transcodé :
En fait c'est rien de spécial, je prends juste un peu de recul et je comprends pas ce que je fous là, fac, psychologie, Montpellier, pourquoi simplement je vis, enfin bref rien de grave. Qu'est-ce qui cloche chez toi ?

Exemple 1 : passage du SMS brut anonymisé au SMS transcodé.

Dans l'exemple 1, on n'ajoutera pas la particule de négation, ne/*n'*. On n'« injectera » pas non plus des éléments prépositionnels ou des déterminants (« *à la* fac », « *en* psychologie », « *à* Montpellier »), car le traitement automatisé demeure possible sans ces informations. En revanche, pour des formes abrégées, agglutinées, etc., on transcode en français standardisé pour qu'un analyseur morpho-syntaxique soit à même de traiter automatiquement la phrase. Dans cet exemple, la forme en apocope « fac » demeure telle quelle dans la version transcodée, car les chercheurs ont décidé de valider le transcodage en lien avec les informations apparaissant au sein du Petit Robert en ligne 2014 (PR14) : si une entrée dictionnairique existe, elle n'est pas transcodée dans sa forme entière (« fac » demeure intacte, mais « psycho » sera transcodé en « psychologie », car si l'élément « psycho- » existe effectivement dans le PR14, l'apocope qui renvoie à « psychologie » n'y figure pas). Par ailleurs, lorsque la ponctuation est présente, on rétablit les normes typographiques pour le français, ici l'espace absente avant le point d'interrogation final.

SMS brut anonymisé (n° **4789** du corpus 88milSMS) :
<PRE_4> eske t es forte en synthese ?!! Notre prof d anglais vien de nous donner une synthese de FOU a faire en FRaNcAIs !!! Allo koi
SMS anonymisé et transcodé :
<PRE_4> est-ce que tu es forte en synthèse ?!! Notre prof d'anglais vient de nous donner une synthèse de fou à faire en français !!! Allo koi

Exemple 2 : passage du SMS brut anonymisé au SMS transcodé.

Dans l'exemple 2, on ne réduit pas les marques de ponctuation (« ?!! » ou « !!! »), qui demeurent d'une forme à l'autre, et qui semblent être porteuses de sens, et on n'introduit pas non plus une ponctuation que l'on pourrait estimer absente, que ce soit à l'intérieur du SMS ou à la fin de celui-ci. En effet, la ponctuation est souvent absente dans l'écriture SMS. Pour un analyseur morpho-syntaxique, dans la mesure où la marque de changement de paragraphe (« ¶ ») permet de reconnaître le passage d'un SMS à un autre, cela ne sera pas gênant en fin de SMS. En revanche, l'absence de ponctuation interne peut s'avérer problématique pour une analyse automatique, mais les chercheurs ont renoncé à un transcodage incluant la ponctuation car l'opération aurait été très lourde à mener. Cela pourrait constituer un sujet de recherche à part entière ! Enfin, dans cet exemple, l'apocope « prof » demeure ainsi dans la version transcodée (*cf*. les explications *supra* pour l'exemple 1).

SMS brut anonymisé (n° **11326** du corpus 88milSMS) :
B, kèl intense réflexion ! Je c, t en week ! <SUR_5> a A C 2 matièr pr fèr son suG. Concer tré 5pa ièr. Bone soiré a toi é tte, bon week ? 2vé fèr gd bo ici : ke dal. Bisous.
SMS anonymisé et transcodé :
Bon/Bien/Ben, quelle intense réflexion ! Je sais, tu es en week-end ! <SUR_5> a assez de matière pour faire son sujet. Concert très sympa hier. Bonne soirée à toi et toute(s), bon week-end ? Il devait faire grand beau ici : rien. Bisous.

Exemple 3 : passage du SMS brut anonymisé au SMS transcodé.

Dans cet exemple, les choix de transcodage s'avèrent complexes à plusieurs endroits : 1) le « B » initial, que les chercheurs qualifient d'« abréviation sémantisée » peut renvoyer à plusieurs possibilités, non exhaustives, comme : « Bon », « Bien », « Ben », ou encore l'initiale d'un prénom, etc. ; pour cette raison, le SMS brut anonymisé est également maintenu dans la diffusion des SMS transcodés, car, selon le c(o)ntexte, la lettre renvoie à différentes possibilités ; ici, on ne peut trancher ; 2) « t » transcodé en « tu es » : pour des sociolinguistes et/ou des spécialistes de l'oral, il peut paraître non souhaitable d'effectuer le choix : « tu es » ; cependant, les chercheurs rappellent ici leur souhait de traitement automatique et un analyseur morpho-syntaxique ne pourra reconnaître « t » ou si l'on remplace cela par « t 'es », une confusion s'introduira avec d'autres structures comme, par exemple, dans « tu t'es trompé » ; 3) d'autres éléments elliptiques (« **Le** concert **était** très sympa… ») ne seront pas introduits, car ceux-ci apportent une dimension interprétative supplémentaire ; 4) le segment « Bone soiré a toi é tte, bon week ? » peut renvoyer à au moins deux interprétations : « Bonne soirée à toi et à toutes » ou bien « Bonne soirée à toi et à toute » (sous entendu « à tout à l'heure ») ; pour la partie « bon week ? », cela correspond à « as-tu passé un bon week-end ? » mais les chercheurs ne souhaitent pas introduire un segment important de phrase, car la forme interrogative choisie par le scripteur aurait pu être autre qu'une inversion sujet-verbe ; 5) le pronom « il » doit être inséré afin que l'analyseur morpho-syntaxique puisse opérer ; 6) « que dalle » n'est pas maintenu, car cela ne figure pas dans le PR14 ; cela est remplacé par « rien ». Ce choix peut ne pas convenir à tous, d'où l'importance de maintenir le lien visible entre la consultation du SMS brut anonymisé et le SMS anonymisé.

Six étudiants ont travaillé sur le transcodage dans le cadre de leur Master 1 de mars à juin 2013[1]. Ils ont étudié la faisabilité d'une méthode d'alignement des SMS pour faciliter le passage du SMS brut anonymisé au SMS transcodé en français standardisé, et ils ont proposé un modèle pour une interface en ligne afin de faciliter le travail de l'annotateur humain. Le modèle d'alignement incluant une interface s'intitule AlignSMS (*cf.* Lopez *et al*. 2014).

Dans le fichier téléchargeable des 1000 SMS, deux transcodages ont été effectués : l'un par les chercheurs, et l'autre par un groupe d'étudiants de Master en Sciences du Langage. Les étudiants ont appliqué les consignes des chercheurs concernant le transcodage (ne pas transcoder si l'entrée existe dans le PR14 ; ne pas injecter d'éléments interprétatifs, par exemple). En revanche, dans certains cas, ils ont pris l'initiative intéressante de maintenir des néologismes :

> « Nous avons repéré de nombreux néologismes dans les messages étudiés, par exemple « textote-moi » ou « bipe-moi ». Dans ces cas particuliers, nous avons choisi de n'opérer aucune modification car nous avons estimé que remplacer « textote-moi » par « envoie-moi un texto » relèverait assurément d'une interprétation plus que d'un transcodage. »[2]

Pour les mots en langue étrangère, les étudiants n'ont pas nécessairement traduit en français, estimant que le transcodage se distingue de la traduction :

> « Nous pouvons citer également les nombreux anglicismes, par exemple « okay », « check moi quand tu sors » et autres mots issus de langues étrangères (« besos », « chica »). S'agissant d'un travail de transcodage et non de traduction, dans ces cas aussi, nous avons choisi de ne pas intervenir afin de ne pas dénaturer les messages. De la même manière, les SMS écrits dans une autre langue n'ont pas été traduits en français. » (idem, Dossier étudiant).

Les étudiants concluent que l'écriture SMS étudiée à travers le travail de transcodage « contient une réelle forme de créativité […] riche en images et vivacité expressive [et qui] est imprégnée des personnalités des émetteurs et récepteurs des SMS (de leur passé, de leur humour, de leur quotidien), des effets de modes, des contextes actuels. »

Pour le traitement du corpus 88milSMS, les chercheurs ont revu leurs prévisions à la baisse, à cause du temps très important nécessaire pour l'étape d'anonymisation (21 mois de stages étudiants). Lors de la diffusion du corpus anonymisé, le téléchargement prévoit un échantillon transcodé de 1 000 SMS. Une fois la méthode de transcodage choisie, des transcodages supplémentaires pourront être effectués, notamment par d'autres étudiants en Sciences du Langage, en Informatique ou dans d'autres disciplines, dans le cadre de travaux de validation de cours et de recherche.

[1] Aghiles Lounes, Tarik Zaknoun, Zakaria Mokrani, Reda Bestandji, Takfarinas Sider, Ahmed Loudah, Master I Informatique, Spécialité : « Informatique pour les sciences » sous la direction de Mathieu Roche, à l'université de Montpellier 2. Reda Bestandji a poursuivi le travail initial en faisant un stage d'un mois, en juin 2013.

[2] Dossier étudiant, L. Dalle, J. Faisant, M. Jaffal, V. de Martino, Master 1, Sciences du Langage parcours DiMIP (EAD), Université Paul-Valéry Montpellier 3, décembre 2013.

