**Balises pour l'annotation d'un échantillon de 100 SMS, extraits du corpus « 88milSMS »**


Huit balises ont été utilisées pour l'annotation de 100 SMS, extraits du corpus « 88milSMS » : TYPographie, MODification, GRAmmaire, BINettes, ABSence, LANgue, ORThographe, DIVers.

*Remarque* : dans les exemples ci-dessous, nous n'indiquons pas l'annotation complète pour la totalité du SMS, mais simplement la balise concernée, afin d'éclairer la lecture.

**1) TYPographie**
<TYP> (typographie : ponctuation, symboles mathématiques, signes diacritiques (accents, etc.) nombres, format des heures, ponctuations ou symboles inattendus, signe : &, chevrons, parenthèses, respect de la casse (majuscules/minuscules), mise en page).

Exemple : SMS n° **6885** dans 88milSMS :

Avant application de la balise <TYP> :
Zorro est arrive, sans s'presse […]

Après application de la balise <TYP> :
Zorro est <TYP_arrivé> arrive, sans s'presse […]

**2) MODification**
<MOD> modification (soit en réduction, soit en augmentation, soit en remplacement de caractères, abrègements et abréviations, acronymes, sigles et abréviations, répétition de lettres, transformations phonétiques, interjections et onomatopées…) : ht (acheter), pr (pour), c (s'est, c'est, ces…), dcd (décidé)…

Exemple : SMS n° **4360** dans 88milSMS :

Avant application de la balise <MOD> :
[…] Oui, j sui zalé ! […]

Après application de la balise <MOD> :
[…] Oui, <MOD_j'y> j <MOD_suis> sui <MOD_allé> zalé ! […]

**3) GRAmmaire**
<GRA> (accords : il viens (pour il vient), syntaxe : si j'aurais su, je serais pas venu (pour si j'avais su, je ne serais pas venu), etc.

Exemple : SMS n° **5536** dans 88milSMS :

Avant application de la balise <GRA> :
Cc tu va mieux. Mam ma dis ke tèté retmbè malade. Et bb? Bisx

Après application de la balise <GRA> :
Cc tu <GRA_vas> va mieux. Mam ma <GRA_dit> dis ke tèté retmbè malade. Et bb? Bisx

**4) BINettes**

<BIN> binettes/frimousses/émoticônes/smileys :) ^^ :p ;) :d <3 :-) xd :( :/

Exemple : SMS n° **6887** dans 88milSMS :

Avant application de la balise <BIN> :
Au dos d'son beau tornado :) elle est trop bien cette prof, chui amoureux d'elle ^^

Après application de la balise <BIN> :
Au dos d'son beau tornado <BIN> :) elle est trop bien cette prof, chui amoureux d'elle <BIN> ^^

**5) ABSence**

<ABS> (absence/ellipse : négation, pronoms, éléments manquants faciles/difficiles à identifier, etc.).

Exemple : SMS n° **19621** dans 88milSMS :

Avant application de la balise <ABS> :
[…] je met tout ça de coté et peux tout encaisser juste pour toi . […]

Après application de la balise <ABS> :
[…] je met tout ça de coté et <ABS_je> peux tout encaisser juste pour toi . […]

**6) LANgue**

<LAN> (c'est-à-dire : contact de langues, emprunts, régionalismes, néologismes, verlan, argot, etc.)

Exemple : SMS n° **43133** dans 88milSMS :

Avant application de la balise <LAN> :
if(ce_soir == film) {get_commande;} else {set_tagueule;} return "bisous"

Après application de la balise <LAN> :
<LAN> if(ce_soir == film) { <LAN> get_commande;} <LAN> else { <LAN> set_tagueule;} <LAN> return "bisous"

 

**7) ORThographe**
<ORT> (uniquement l'orthographe lexicale : erreurs de saisie, interversion de lettres, etc.)
Exemple : SMS n° **19621** dans 88milSMS :

| |
|---|
| Avant application de la balise <ORT> :<br>[…] notre couple sera tel un rosau à jamais se casser . […]<br><br>Après application de la balise <ORT> :<br>[…] notre couple sera tel un <ORT_roseau> rosau à jamais se casser […] |

**8) DIVers**
<DIV> (dans le cas où aucune autre balise ne semble convenir).

Exemple : SMS n° **4671** dans 88milSMS

| |
|---|
| Avant application de la balise <DIV> :<br>Ffghoeksjclfpzozkdkfoeeogrjzjglelsjloe<br><br>Après application de la balise <DIV> :<br><DIV> Ffghoeksjclfpzozkdkfoeeogrjzjglelsjloe |

Le SMS n° 4671 ne contient qu'un seul item lexical, et sans d'autres informations co(n)textuelles, il est très difficile de comprendre l'intention du scripteur.

**Remarques** :

*i) Doubles balises* : parfois un item lexical peut être annoté à l'aide de deux (voire plusieurs) balises.
Exemple : SMS n° **6532** dans 88milSMS :

| |
|---|
| […] <MOD_LAN> loveeeeee <MOD_LAN> uuu |

*ii) Difficulté d'attribution de balises* : parfois deux balises peuvent être difficiles à départager :
Exemple : SMS n° **4360** dans 88milSMS :

| |
|---|
| […] Bone journé. […] |

Dans cet exemple, on peut supposer que le scripteur a volontairement modifié les deux mots, mais on ne peut pas en être sûr car cela pourrait relever de l'orthographe inconnue pour ces items. Dans ces types de cas, les balises <MOD> et <ORT> peuvent être utilisées toutes les deux.

Exemple : SMS n° **11682** dans 88milSMS :

| |
|---|
| Il es rentrer  a 22h30 et jai eu ldroii au : jsui fatiguer, jai mal a la tete jvai me coucher. |

 

Dans cet exemple, on pourrait indiquer <GRA> pour « Il es rentrer » et pour « jsui fatiguer », mais on pourrait aussi supposer que le scripteur veuille économiser un appui long (sur un smartphone) afin d'obtenir l'accent aigu, et opte pour le « r » final à la place, et, dans ce cas, il s'agirait d'une modification volontaire <MOD>.

*iii) Entrées dictionnairiques* : Si un item lexical figure dans le Petit Robert en ligne 2014 (PR14), alors il ne recevra pas de balise. Par exemple, *ah*, *boum, ben, bah, bouh, ouais, frérot, lol, relou, prof, sympa, papi*, *cool*, *week-end, box-office*, etc., resteront tels quels, car ils figurent dans le PR14. (*cf*. le fichier « annotations_chercheurs » pour tous les exemples de l'extrait des 100 SMS apparaissant dans le PR14).

*iv) Ponctuation* : La ponctuation finale absente (qui est nombreuse) n'est pas indiquée dans l'annotation, ainsi que des absences de virgules, etc., au sein des SMS.

**Synthèse** :

Les chercheurs du projet ont annoté l'extrait des 100 SMS et ils ont obtenu les résultats suivants (*cf*. Figure 1). Le détail du travail se trouve dans le fichier « 100_SMS_annotations_chercheurs » :

| **Synthèse chercheurs** | | | |
|---|---|---|---|
| **TYP** | 1 | 449 | **43,59%** |
| **MOD** | 2 | 294 | **28,54%** |
| **GRA** | 3 | 84 | **8,16%** |
| **BIN** | 4 | 82 | **7,96%** |
| **ABS** | 5 | 52 | **5,05%** |
| **LAN** | 6 | 38 | **3,69%** |
| **ORT** | 7 | 30 | **2,91%** |
| **DIV** | 8 | 1 | **0,10%** |
| | Total | 1030 | **100,00%** |
| | | | |
| Mots sans modification | | 2393 | 69,91% |
| Mots étiquetés | | 1030 | 30,09% |
| | Total | 3423 | 100,00% |

Fig. 1 : Synthèse de l'utilisation des balises, pour l'extrait des 100 SMS annotés, par les chercheurs de 88milSMS.

Il en ressort que les phénomènes de *typographie* sont les plus saillants, suivis par les *modifications* (substitutions, réductions, ajouts, etc.). La balise qui concerne la *grammaire* arrive en troisième position, suivie, dans l'ordre, par les *binettes*, l'*absence*, la *langue*, l'*orthographe* et la balise *divers*. Il est également intéressant de constater que 70 % de l'extrait des 100 SMS ne subit aucune modification. 113 commentaires figurent dans le fichier annoté, en partie pour décrire les entrées dictionnairiques apparaissant dans le PR14.

 

Quatre groupes étudiants ont également effectué le même travail (cf. le fichier « 100_SMS_annotations_etudiants »). Ensuite, une étudiante a effectué une harmonisation des résultats des quatre groupes. La synthèse de ces résultats est présentée dans la Figure 2 :

| Synthèse étudiants | | | |
|---|---|---|---|
| MOD | 1 | 271 | 30,83% |
| TYP | 2 | 263 | 29,92% |
| ORT | 3 | 113 | 12,86% |
| BIN | 4 | 83 | 9,44% |
| GRA | 5 | 72 | 8,19% |
| ABS | 6 | 39 | 4,44% |
| LAN | 7 | 30 | 3,41% |
| DIV | 8 | 8 | 0,91% |
| | Total | 879 | 100,00% |

Fig. 2 : Synthèse de l'utilisation des balises, pour l'extrait des 100 SMS annotés, par des étudiants de Master en Sciences du Langage.

En comparant les deux figures contenant les résultats chiffrés, on constate des différences importantes d'attribution de balises. De même, des attributions divergentes existent entre groupes d'étudiants (*cf.* les tableaux récapitulatifs en fin de fichier « 100_SMS_annotations_etudiants »). Cela peut être dû aux raisons suivantes, mais pas seulement : certains encodeurs étudiants ont utilisé la balise <ABS> pour une absence de ponctuation, alors que cela doit être codé avec la balise <TYP>, <ABS> étant réservé pour les ellipses de pronoms, de négation, etc. De même pour la balise <ORT> qui doit être uniquement réservée pour l'orthographe (lexicale) et non pas grammaticale. Dans ce dernier cas, on utilisera <GRA>. Un exemple comme « quil » doit être codé en <TYP> et non en <MOD>. Ces « erreurs » d'étiquetage posent tout de même la question de différenciations importantes au niveau des choix des annotateurs et peuvent poser un réel problème dans l'étiquetage d'un grand corpus.

Les écarts entre les différents traitements montrent à quel point il est extrêmement difficile de proposer une annotation standardisée. Lors du projet *sud4science LR*, les chercheurs ont invité les acteurs des collectes précédentes, dans le cadre de *SMS4science*, lors de deux journées d'étude à Montpellier (« Harmonisation/standardisation des méthodes de traitement de corpus écrits de type SMS. Anonymisation, transcodage, annotation. », 14-15 novembre 2011, MSH-M), à présenter leurs balises pour l'annotation de leurs corpus de SMS. Une harmonisation générale a ensuite permis aux chercheurs *sud4science* de réduire le nombre de balises précédemment utilisées, afin d'envisager le balisage éventuel du corpus 88milSMS. Par la suite, ils ont décidé de fournir ici un échantillon d'annotation de 100 SMS, mais de renoncer à l'annotation de l'ensemble du corpus 88milSMS, précisément car, d'une part, la tâche serait gigantesque, et, d'autre part, les chercheurs ne seraient pas nécessairement en accord avec le choix des balises. Cet échantillon permet de fournir des pistes de recherche, mais il est important que chacun ait accès au corpus anonymisé, sans que des initiatives de balisage supplémentaire leur soient imposées.

 