Normalisation du Corpus

Français

SNF-sms4science

01.03.2013

Table des matières

Règles de normalisation du corpus

Français

1. Tokens

Dans les SMS nous ne distinguons pas à priori des mots, mais des *tokens*. Ceux-ci sont générés de manière automatique par le logiciel. Le logiciel reconnaît des éléments séparés par des espaces, des apostrophes ou des traits d'union, ainsi que la majorité des émoticons et les signes de ponctuation.

Nous ne normalisons donc pas des mots, mais des tokens. Il est possible que le nombre de tokens dans les SMS ne corresponde pas au nombre de tokens au niveau normalisé, ce qui ne doit pas nous préoccuper.

Cependant il est aussi possible que l'identification automatique des tokens ne soit pas correcte. Ces « erreurs » doivent être corrigées à la main.

1.1. Un token dans le SMS → deux tokens normalisés

Dans les SMS, les pronoms sont souvent agglutinés proclitiquement au verbe (*jsuis*). A un token dans le SMS correspondent alors deux tokens normalisés (*je* | *suis*). Ce réajustement peut être effectué facilement lors du travail de normalisation : dans le casier qui se trouve en dessous du token du SMS, on introduit les deux mots correspondants à la norme, séparés par un espace.

1.2. Plusieurs tokens dans le SMS → un token normalisé

Ce cas-ci est également facile à traiter. Quand au niveau SMS il y a plusieurs tokens ([*pomme*] [*de*] [*terre*]) qui correspondent à un seul token normalisé, ce dernier est introduit dans le **premier** casier en dessous des tokens du SMS ([*pomme de terre*] [ ] [ ]) et **les deux autres casiers restent vides**.

1.3. Fausse tokenisation au niveau SMS

Quand au niveau SMS il y a une faute de tokenisation (p.ex. [*ajourd'*] [*hui*] en deux tokens ; [*n*] [*8*] pour *nuit* en deux tokens), il faut le noter. Pour ce faire, on choisit dans la partie *SMS Details* le Statut *needs retokenizing*. Cette faute sera corrigée à la main plus tard.

Quand on indique *needs retokenizing*, on laissera les casiers sous les éléments correspondants vides (pour des raisons pratiques).

1.4. Récapitulation

Nous distinguons les cas suivants :

1. 1 token dans l'original, 2 ou plus tokens au niveau normalisé
   (1) [*jsuis*] → [*je suis*]
   pas de retokenization
2. a) Plusieurs tokens dans l'original, 1 token au niveau normalisé
   (2) [:] [-] [>] → [:->]
   (3) [*s*] [*8*] → [*s8*]
   marquer comme *needs retokenizing* et indiquer où se situe le problème (ces tokens ont été séparés par le logiciel à cause de l'occurrence de signes de ponctuation et/ou de chiffres et non pas par l'auteur du message).
   b) Plusieurs tokens dans l'original (séparés par des espaces dans le message d'origine), 1 token au niveau normalisé :
   (4) [*pomme*] [*de*] [*terre*] → [*pomme de terre*] [] []
   inscrire le mot entier dans le premier casier et laisser vides les autres. Pas de commentaire.
3. 1 token dans l'original (erroné), 2 ou plusieurs tokens au niveau normalisé
   (5) *Dè kon* [*rentre-ver*] *6 h* → [*rentre*] [-] [*ver*]
   marquer comme *needs retokenizing* et indiquer où se situe le problème
4. Cas spéciaux
   (6) [*egypteB-)*] → [*egypte*] [*B-)*]
   marquer comme *needs retokenizing* et indiquer où se situe le problème

Informations supplémentaires

Dans la partie *Glossing*, sous les deux lignes à éditer (*Message* et *Gloss*), il y a des informations supplémentaires qui ont l'étiquette *PoS* (*Part of Speech*) et *Lang* (*Language*). Ce sont des informations qui sont partiellement générées automatiquement. Y appartiennent l'indication si un token est un signe de ponctuation (*PUN*) ou un émoticon (*EMO*). Si dans ces indications il y a des fautes, on introduira une note dans le champ *Note* dans la partie *SMS Details*.

Cas particuliers :

Les **pronoms enclitiques** qui sont séparés du verbe avec un trait d'union constituent un token à part. La séquence [*peux*] [-] [*tu*] n'a pas besoin de réajustement (le trait d'union faisant partie des signes de ponctuation reconnus par le logiciel). Le ***-t-* de liaison** dans la séquence [*a*] [*-t-*] [*il*] est également repris tel quel.

Les trois tokens **[*est*] [-] [*ce*]** restent inaltérés, car la séquence est analysable en tant que verbe plus pronom. Par contre, nous considérerons comme un seul token les éléments suivants :

(7) [*est*] [-] [*ce*] [*que*] → [*est-ce que*]

Ces quatre tokens seront regroupés dans un, car il s'agit d'une particule interrogative lexicalisée. On marquera alors *needs retokenizing*.

2. Majuscules/minuscules

Tous les mots sont écrits comme ils se trouvent dans un dictionnaire français (Le Petit Robert (PR) : http://pr.bvdep.com/ nous sert de référence).

Concrètement, cela veut dire que les noms propres (incl. acronymes), ainsi que l'item *Dieu* s'écrivent avec majuscules. Des entités politiques sont traitées comme des noms propres en français : *la République française, le Conseil d'État* et aussi *l'État* (en cas de doutes, on consultera le Petit Robert). Les titres comme *le Président Hollande, le Premier ministre* etc. prennent des majuscules.

Conformément au Petit Robert, les adjectifs de pays (*français*) s'écrivent avec minuscules. Par contre, quand il s'agit d'une personne (*le Français*, *la Française*) ou du pays (*la France*), on emploiera des majuscules (ce qui ne vaut pas pour la langue : *le français*). Tout le reste prend des minuscules. Cela vaut aussi pour les mots en **début de phrase**. Si la phrase commence p.ex. par un pronom, celui-ci prendra une minuscule.

Exception : les substantifs empruntés à l'allemand prennent des majuscules.

3. Apostrophe

Certains auteurs de SMS utilisent des apostrophes là où des sons sont omis dans le langage familier.

(8) a. [*t'es*] [*où*] [*?*]

Ces apostrophes qui ne correspondent pas à la norme française sont remplacées par la lettre (les lettres) élidée(s).

Cet exemple doit alors être normalisé par

b. [*tu es*] [*où*] [*?*]

4. Ellipses

Toute omission qui se trouve dans les SMS doit aussi figurer au niveau normalisé (intermédiaire). Nous n'ajoutons rien.

(9) a. [*Te*] [*raconterai*] [.]

ne reçoit pas de sujet : on notera

b. [*te*] [*raconterai*] [.]

Motivation

Toute sorte d'ellipse est potentiellement intéressante pour la recherche. Ce sont donc des informations qui ne doivent pas se perdre à ce niveau.

5. Variantes graphiques

En principe, les variantes graphiques sont normalisées.

Exemples :

(10) [*Mé*] [*fiche*] [*de*] [*Geo*] [*son*] *ché*] [*toi*] [*?*] → [*Mes*] [*fiches*] [*de*] [*géo*] [*sont*] [*chez*] [*toi*] [*?*]

(NB. : *géo* sera marqué comme abréviation et comme *unclear*, car il peut s'agir de *géographie*, *géologie*, *géométrie* etc.)

(11) [*ca*] [*yé*] → [*ça*] [*y est*]

Quand le son rendu par la graphie correspond à la prononciation réelle, on ne marquera pas ce *token* comme abréviation ! (Par contre, on le marquera quand la graphie correspond à la prononciation de la lettre comme en citant l'alphabet, voir 9.4 écriture rébus.)

Quelques règles spéciales qui divergent de ce principe sont les suivantes :

- Pour les onomatopées nous appliquons d'autres règles. Pour plus d'informations voir le chapitre correspondant (chap. 11).
- Si l'on reconnaît une variante graphique dans une relation syntagmatique, celle-ci est rétablie.

(12) [*pré*] [*ou*] [*postposé*] → [*pré-*] [*ou* [*postposé*]

6. Verlan

Les expressions en verlan sont rapportées au français normalisé :

(13) […] *Gro zibou bonne soirée jtd* (14830) → *gros bisou*

7. Clitiques

Les pronoms personnels sont normalisés, c'est-à-dire que les formes apostrophées sont seulement maintenues où permis par la norme et les éléments agglutinés sont séparées.

(14) [*t'*] [*es*] → [*tu*] [*es*]

(15) [*jtaime*] → [*je t'aime*]

Les clitiques apostrophés (14) ou agglutinés (15) ne sont pas marqués comme abréviations. Les éléments fusionnés (16) par contre le sont (écriture rébus, voir 9.4).

(16) [*g*] → [*j'ai*]

Quand il y a une forme inversée ou un impératif qui n'a pas de trait d'union mais qui est séparé par un espace, les deux tokens sont traités séparément. Par contre, le verbe

reçoit un trait d'union de sorte à ce qu'on puisse traiter ces entités de même manière que les autres inversions.

(17) [*peux*] [*tu*] → [*peux*-] [*tu*]

(18) [*donne*] [*moi*] → [*donne*-] [*moi*]

Si par contre les éléments sont agglutinés et apparaissent comme un seul token, ils seront séparées par un trait d'union

(19) [*donnemoi*] → [*donne-moi*]

Voir aussi chapitre tokenisation (1.3).

8. Composés

Les composés qui consistent de plusieurs tokens (« mots ») sont reconstitués en les écrivant entièrement dans la **première case** qui apparaît sous les tokens et **en laissant vides les successives** :

(20) [*pomme*] [*de*] [*terre*] → [*pomme de terre*] [ ] [ ]

Attention : nous considérons comme composés uniquement les lexèmes qui sont répertoriés comme **entrée principale** dans le PR !

9. Formes abrégées

9.1. Points dans l'abréviation

Quand une abréviation contient des points (s.v.p., etc.), ils doivent être 'rattachés' à celle-ci. Cela veut dire que souvent il faut indiquer qu'il faut retokeniser la séquence :

(21) [*etc*] [.]→ [*etc.*]

Exception : quand le point se trouve en fin de phrase, il n'est pas rattaché à l'abréviation.

9.2. Abréviations phraséologiques

Dans les SMS on utilise souvent des abréviations. Il s'agit d'abréviations bien connues comme *q* pour *que* ou *a+* pour *à plus*, mais aussi d'abréviations qui sont reconnues uniquement par l'auteur et l'adressé.

Dans le corpus, nous voulons marquer toutes les formes abrégées comme telles afin qu'on puisse les rechercher plus tard. Par conséquent, chaque token qui contient une forme abrégée reçoit l'attribut *Abbreviation*. Les acronymes (comme *mdr* pour *mort de rire*) et les abréviations (*s.v.p.*) sont catégorisés de la même manière pour questions de simplicité. Si un chercheur s'intéresse aux formes abrégées, il peut facilement les retrouver et les classifier d'après son propre schéma.

Les formes abrégées non standard (donc pas dans le PR) qui peuvent être devinées facilement sont écrites en toutes lettres au niveau normalisé. Toutes les formes abrégées qui sont écrites en toutes lettres **reçoivent aussi** l'attribut *Abbreviation*.

9.3. Noms propres abrégés, acronymes, formes abrégées lexicalisées

A côté des formes abrégées mentionnées ci-dessus, il existe encore d'autres formes abrégées. Y appartiennent *IBM* pour *International Business Machines, Laser* pour *Light Amplification by Stimulated Emission of Radiation*, mais aussi *NE* pour *Neuchâtel*, *di* pour *dimanche*.
Pour décider quelle forme abrégée sera écrite en toutes lettres, nous nous posons les questions suivantes : 1) qu'est-ce qu'on peut attendre des collaborateurs du projet et 2) qu'est-ce qui pourrait être intéressant pour des futures recherches.
De là ressortent les règles suivantes :

a. Les formes qui se trouvent abrégées comme **entrée principales dans le dictionnaire** ne sont pas écrites en toutes lettres (*Laser* reste *Laser, resto* reste *resto, HLM* devient *H.L.M.* selon le PR), mais reçoivent l'attribut abréviation.
b. Les **noms propres (sauf toponymes)** ne sont pas analysés (*IBM* reste *IBM*). Cela vaut aussi pour les noms de clubs : *HCFG* (hockey club Fribourg-Gottéron) reste *HCFG*.
c. Les autres formes abrégées qui ne posent **pas de problèmes d'interprétation** sont écrites en toutes lettres (*NE* pour *Neuchâtel* devient *Neuchâtel, lu* pour *lundi* devient *lundi* etc.).

(22) *Non t. beau demain et encore lu → non temps beau demain et encore lundi*

d. Quand il y a un **doute** quant à l'interprétation ou que la forme abrégée ne peut pas être analysée (*Hey sgt c!*) elle est reprise telle quelle au niveau normalisé et on la marquera par *unclear*.
e. Les **noms d'individus**, p.ex. *bisous E.* ne reçoivent pas l'attribut *Abbreviation*, car ils ne sont pas intéressants pour la recherche.
f. Les abréviations qui consistent de **signes** seront marquées comme abréviations et écrites en toutes lettres.

(23) *+ → plus*
(24) *x → fois*

Exemples

Dans le SMS il y a la forme abrégée *s.v.p.* On indiquera *Abbreviation* et – comme *S.V.P.* existe comme entrée dans le dictionnaire – on adaptera l'abréviation au standard, donc la forme normalisée sera *S.V.P*.

(25) *s.v.p. → S.V.P.*

(26) *Christine organise les **rdv**. → Christine organise les **rendez-vous**.*

On choisit à nouveau l'attribut *Abbreviation* et on écrira en toutes lettres *rendez-vous*.

(27) *Je peu invité **qqn** cet **apr** juska 5**h**? → Je peux inviter **quelqu'un** cet **après-midi** jusqu'à 5**h** ?*

Toutes les trois formes recevront l'attribut *Abbreviation*. Comme *qqn* – bien qu'il s'agisse d'une abréviation relativement commune – n'apparaît pas dans le Petit Robert, on écrira en toutes lettres *quelqu'un*. La même chose vaut pour *apr → après*. *h* finalement se trouve comme entrée dans le Petit Robert (voir aussi chapitre 15 horaire). On le laissera tel quel quand il indique l'heure, tout de même en le marquant comme abréviation s'il ne fait pas partie d'un token « NUM » (p.ex. [5h]) reconnu par le tokenizer . Normalement, le tokenizer traite *5h* comme un token.

Attention : Quand h est employé dans un autre contexte, il est écrit en toutes lettres :

(28) *Tu viens pour une petit h ? → tu viens pour une petite heure ?*

Pour la forme abrégée

(29) *cu ou cya → see you*

il faut ajouter l'information qu'il s'agit d'une abréviation anglaise. On marque donc en premier *anglais*, puis on normalise la forme à *see you* par moyen de *edit gloss*. Il ne faut pas oublier de le marquer comme abréviation.

9.4. Ecriture rébus

Aussi les rébus sont des formes abrégées

(30) *bonne **n8** → bonne **nuit***

(31) ***C** → c'est,* [se > sɛ] et non pas [s] ou [k]

(32) ***G** → j'ai,* [ʒe] et non pas [g]

Elles sont normalisées et indiquées comme abréviations.

Il s'agit de rébus quand la lettre prend plus que le son habituel dans la chaîne phonique, donc quand elle est prononcée comme si on citait l'alphabet.

10. Noms

Le corpus est d'ores et déjà anonymisé. Les noms propres ne devraient donc plus pouvoir être attribués aux auteurs et leurs proches. Les noms de famille ont été remplacés par *<Lastname>* et les prénoms ont été tournés, c'est-à-dire qu'ils ont été remplacés par d'autre prénoms. Cependant, cette anonymisation n'est pas complète. D'une part, parce que certains noms ne peuvent pas être anonymisés comme ils sont homographes avec des mots fonctionnels (*Pierre* s'écrit de la même manière que *pierre* 'rocher'). D'autre part, il est possible que certains noms aient échappés au processus. Il faut donc contrôler si l'anonymisation est complète.

10.1. A quoi reconnaît-on si un nom a été anonymisé?

Les noms anonymisés sont marqués comme tels. Ils sont déjà copiés dans le niveau normalisé. Si au niveau normalisé il y a déjà le même nom qu'au niveau SMS, tout est correct. S'il n'y a pas encore de nom, il faut marquer le SMS comme *missing anonymisation*. Ceci vaut aussi pour les noms homographes avec d'autres mots. Ces cas seront remaniés à la main.
Quand on indique *missing anonymization*, le casier restera vide (pour des raisons pratiques).

10.2. Noms abrégés

Les noms abrégés comme p.ex. *bisous E.* ne sont pas marqués comme abréviations (voir chapitre 9.3 formes abrégées). Ils ne reçoivent pas non plus l'attribut *unclear*.

10.3. Noms de produits

Les noms de produits, p.ex. *Facebook* ne sont pas anonymisés et pas non plus marqués. Ils sont simplement normalisés graphiquement mais autrement repris tels quels.

(33) *faceebbookk → Facebook*

NB : Si un nom est employé de manière créative comme verbe (34) le verbe prendra une minuscule (*facebook*).

(34) *Yeah! Cool alors on se **Facebook** pour dire quand on se croise et pour savoir combien ça fait :-D merci ;-)* (9305) → *facebook*

10.4. Noms propres étrangers

Les noms propres de chansons, groups etc. étrangers sont repris tels quels mais reçoivent l'attribut de la langue correspondante (ici anglais) :

(35) *Hello!ca va?Esk t'as les **bac&love** suivants(je viens de finir le 7)?tu pourra mles preter un dc jour stp?jsuis tro pressé dlire la suite!à demain bonne soirée:-)* (10320)

10.5. Toponymes

Dans les toponymes, chaque mot correspond à un token. [*Lac*] [*Léman*] constitue alors deux tokens. Ils sont repris tels quels, à part s'il y a des variantes orthographiques

(36) [*Lac*] [*Lemand*] → [*Lac*] [*Léman*]

Ils ne sont pas adaptés à la langue cible. Ils reçoivent l'attribut de la langue correspondante uniquement s'il s'agit d'un toponyme qui connaît une traduction dans la langue cible et dont l'orthographe indique clairement la langue d'origine (p.ex. le *Umlaut* en allemand)

(37) *Zürich* → *Zürich* même en français [attribut langue : German]

11. Onomatopées

Comme dans les BDs, dans les SMS on utilise des formes onomatopéiques telles que *whoaa*, *grrrr, boah* etc. Ces formes seront marquées.

11.1. Délimitation

En linguistique, le terme *onomatopée* est très vaste. Il couvre aussi des mots comme *coucou* ou *couiner*. Mais nous ne voulons pas marquer ce genre de mots. Les limites entre les onomatopées au sens étroit et les mots lexicaux comme *aboyer* d'une part et les interjections comme *ah* d'autre part sont floues. Pour cette raison, nous nous orienterons à des règles simples qui ont un sens pour les futures recherches.

Le marquage de ces tokens a deux buts. D'une part, ces tokens peuvent être ignorés lors du Part-of-Speech-Tagging et d'autre part, ils doivent être retrouvables comme groupe par les futurs chercheurs. Nous définirons donc *onomatopée* de la manière suivante :

- Des tokens qui ne sont pas répertoriés dans le Petit Robert
- Des tokens qui ont un caractère onomatopéique

Cette définition ne correspond pas à la définition linguistique habituelle du terme. Mais de cette manière nous atteignons nos buts :

- Tous les tokens marqués de cette manière peuvent être exclus lors d'analyses de linguistique computationnelle.
- Les tokens qui ne posent pas de problèmes aux taggeurs PoS (parce qu'ils sont répertoriés (p.ex. *ah*, *grrr*)) ne demandent pas d'efforts supplémentaires aux collaborateurs.

- Tous les tokens marqués de cette manière peuvent être regroupés avec d'autres onomatopées comme *ah* si cela est requis, car les formes qui correspondent à la norme se trouvent dans les dictionnaires et peuvent être répertoriées auparavant.

Avec cette catégorie d'éléments nous ne respectons pas un autre principe de normalisation, c'est-à-dire que nous considérons exceptionnellement aussi la graphie pour des raisons de simplicité et cohérence entre collaborateurs.

11.2. Procédé concret

Concrètement, les tokens avec caractère onomatopéique sont recherchés dans le Petit Robert. S'ils s'y trouvent (p.ex. *ah*, *grrr*), ils sont repris tels quels. S'ils ne s'y trouvent pas (p.ex. *whoaa*, *boah*, mais aussi *ahhhh*) ils sont également repris tels quels et en plus marqués comme onomatopées.

12. Orthographe

L'orthographe de référence est celui du Petit Robert (http://pr.bvdep.com/).

Pour les lexèmes qui ont plusieurs variantes normatives on se décidera pour une seule :

- *ciao* pour *ciao/tchao* (et toutes les variantes *tchô, tcho* etc.)

A ce sujet consulter le Wiki sur ILIAS.

13. Erreurs de frappe

Quand il s'agit d'une erreur de frappe évidente, on corrigera l'orthographe :

(38) *Coucou bien recu tes messages. **Par** de clés malheureusement. A++* (7499)

La seule forme qui a du sens dans ce contexte est un *pas* à la place de *par*.

14. Langues

Les SMS sont déjà annotés pour les langues, c'est-à-dire que pour chaque SMS il a été déterminé quelle était la langue principale et quels emprunts et emprunts ad-hoc s'y trouvaient. Cette identification a été effectuée au niveau du SMS, les tokens-mêmes n'ont pas encore de marque pour la langue. Cette tâche reste à faire pour les emprunts **ad-hoc**. Les tokens comme *cu* ou *night* doivent être annotés comme anglais.

14.1. Langues retenues

Lors de la détermination de langue du SMS, plusieurs langues ont été annotées. Partant de l'idée que la recherche dans le corpus sera faite pour les langues nationales et l'anglais, nous nous limiterons à ces langues-là. Les emprunts ad-hoc d'autres langues seront marqués comme *other*.

Les pseudo-emprunts comme *no problemo* (qui est ni la forme correcte italienne, ni espagnole) sont rangés sous *other*.

14.2. Procédé concret

Quand un token est marqué pour une langue, le système le copie directement dans le niveau normalisé. Il est donc recommandable de marquer en premier la langue.

14.3. Adaptation à l'orthographe

Les SMS montrent parfois une orthographe près du parlé, c'est-à-dire qu'ils sont rédigés d'une manière pseudo-phonétique. Ainsi nous trouvons des formes comme *Tchuss!* pour *tschüss!* Où possible, ces formes sont normalisées d'après la langue dont elles proviennent. On écrira donc *tschüss* d'après le Duden pour l'allemand.

(39) *Tchuss → tschüss*

14.4. Adaptation morphologique

Lorsqu'un élément d'une langue étrangère montre une flexion française dans le SMS, ces traits morphologiques seront retenus. Si un auteur écrit

(40) *il m'a phoné → il m'a phoné*

phoné sera repris tel quel. Si par contre il n'y a aucune flexion au départ, une telle n'est pas ajoutée :

(41) *des choses cool → des choses cool (pas cooles)*

Si la flexion correspond à la langue de départ, on la maintient dans sa forme originale (ou adaptée à la norme de la langue correspondante) :

(42) *il m'a phoned → il m'a phoned*

(43) *phond → phoned*

14.5. Homophonie avec des mots du français

Il peut arriver qu'une variante graphique corresponde à l'orthographe du même mot dans une autre langue (p.ex. *rich* pour *riche* correspond à l'anglais). S'agit-il d'un emprunt ad-hoc ou d'une variante graphique ? De manière générale on optera pour une variante graphique qu'on normalisera (donc pas de marquage d'emprunt et forme normalisée *riche*). Font exception à cette règle les graphies homophones qui se trouvent dans un syntagme ou une séquence entièrement dans la langue étrangère :

(44) *Il est un homme very rich.*

Là on marquera *very* et *rich* comme emprunts ad-hoc.

14.6. Homophonie entre deux langues étrangères

Quand il y a un doute à propos de l'origine d'un mot, comme il peut être le cas entre le suisse allemand et l'allemand standard, le contexte immédiat est décisif. Si le mot en question se trouve dans une séquence qui est clairement attribuable à l'une des deux langues, le mot douteux est marqué comme emprunt de cette langue. Quand il s'agit de mots isolés, dans le doute on optera pour l'allemand standard.

(45) *i bi am bahnhof xi → gsw*

(46) *bahnhof → g*

14.7. Toponymes et noms propres

Les toponymes et noms propres de langues étrangères (p.ex. *Ticino, New York, IBM*) seront repris tels quels. On ne les marquera pas non plus comme emprunts ad-hoc, car il est impossible de trouver une délimitation des langues. *Berlin* par exemple correspond à la forme française, allemande et anglaise de la ville. Aussi faudrait-il marquer l'acronyme *IBM* comme emprunt anglais, car il dénomme *International Business Machines*. Ce procédé n'est pas praticable pour les collaborateurs, ni semble-t-il sensé pour des éventuelles recherches.

15. Horaire

En français, la manière habituelle d'indiquer l'heure est la forme suivante : *21h00*. Celle-ci est reconnue comme un token et nous la laisserons telle quelle. Si nous avons une forme comme *21h* (où le *h* suit le chiffre) le *21* et le *h* apparaissent également comme un seul token. On ne le marquera pas comme abréviation dans ce contexte-là.

Quand *heure(s)* est écrit en toutes lettres dans le SMS, on l'écrira également en toutes lettres au niveau normalisé.

16. Unclear

L'attribut *unclear* est réservé pour les formes pour lesquelles il est impossible d'identifier ce que l'auteur voulait dire. Il peut s'agir d'une faute de frappe qui rend la forme ininterprétable, de mots dans des langues inconnues ou de formes homophones qui ne sont pas clairement interprétables au sein du syntagme (donc de formes ambigües). Toutes ces formes seront reprises telles qu'elles apparaissent dans le SMS et marquées comme *unclear*.

Exemples :

(47) *géo*

L'abréviation *géo* est marquée comme *unclear* (et abréviation), car il est impossible de savoir s'il s'agit de *géographie*, *géologie*, *géométrie* etc.

(48) *Hey sgt c!*

Il est impossible de savoir ce que veut dire *sgt c*.

17. Chiffres et nombres

Les nombres ne sont pas changés s'ils apparaissent comme chiffres dans le SMS.

(49) *5 chiens →5 chiens*

Si les nombres sont utilisés comme rébus, ils sont écrits en lettres

(50) *bonne n8 → bonne nuit.*

Si les nombres apparaissent en combinaison avec des abréviations, les chiffres restent inaltérés, les abréviations par contre seront analysées.

(51) *2x → 2 fois.*

Les nombres ordinaux abrégés par 1^{er}, $2^{ème}$, 2^{e} etc. sont écrits en toutes lettres et sont indiqués comme abréviations :

(52) *1er → premier*

(53) *2e → deuxième*

18. Agglutination

Les mots sont séparés comme le prescrit le Petit Robert.

Les auteurs des SMS ne respectent pas toujours les normes de l'orthographe française. Très souvent, à la place des apostrophes, les mots sont agglutinés :

(54) [*jaime*] → [*j'aime*]

(55) [*juska*] → [*jusqu'à*]

(56) [*Tas*] → [*tu as*]

Dans ces cas-là, on réintroduit l'apostrophe (voir aussi le chapitre 3 apostrophe). Mais on trouve aussi d'autres mots ou des séquences entières agglutinées.

(57) [*Monamour*] → [*mon amour*]

On les séparera également.

Quand les éléments (qui ont une relation syntagmatique) sont –pas conformément à la norme- séparés par des traits d'union, on maintiendra cette orthographe pour rendre compte de la volonté de l'auteur d'explicitement marquer ce passage comme unité.

(58) [*super-bon-chaleureuse-nuit*] → [*super-bonne-chaleureuse-nuit*]

Par contre, on corrigera les variantes orthographiques dans les sous-segments.

NB : Les traits d'union sont modifiés que lors qu'ils se trouvent entre deux éléments qui ne peuvent pas constituer une unité lexicale (cf. (59)). Là, il faut marquer l'option *needs retokenizing*. Pour les pronoms personnels voir chapitre 7 clitiques.

(59) [*Salut*] [*Marie-rdv*] [*demain*] → [*salut*] [*Marie - rendez-vous*] [*demain*] (needs retokenizing)

19. Conjugaison verbale

Comme beaucoup de terminaisons verbales du français sont muettes, les auteurs des SMS ne leur prêtent pas toujours attention. Soit elle sont omises/abrégées, soit elles correspondent à une forme qui n'est pas compatible avec le sujet ou avec la construction verbale. On trouve p.ex. des formes comme :

(60) *I pleu a Maurice pour la premiere foi.*

(61) *Je peu invité qqn cet apr juska 5h?*

(62) *Peut-tu me confirmer que tu l'a bien recu?*

Mais aussi des changements de mode comme dans :

(63) *tain mon gars j'suis tombé sur un son electro de fou, un truc de dingue haha! J'te ferais écouter un de ces 4. Cya!*

Lors de la normalisation des messages (voir exemples ci-dessous), nous ne voulons pas interpréter. Pour cela, nous corrigerons seulement là où aucun doute n'est possible. Les corrections se basent uniquement sur les formes clairement identifiables comme fausses d'un point de vue purement grammatical. En outre, nous ne voulons pas perdre des informations qui pourraient témoigner d'un changement en cours comme p.ex. le remplacement du subjonctif par l'indicatif ou la perte de distinction entre futur et conditionnel.

De cela dérivent les règles suivantes :

a. Les « faux » modes des verbes ne sont pas changés : l'indicatif reste même dans des subordonnées qui exigent le subjonctif, le conditionnel n'est pas altéré même si intuitivement on s'attendrait à un futur (cf. (67)). Par contre, les impératifs clairs doivent recevoir la forme correcte à l'impératif. En cas de doutes, on optera tout de même pour la forme de l'indicatif.
b. L'orthographe qui rend le son [e] de l'infinitif et du participe passé est corrigée selon la position de l'élément et la structure dans laquelle il apparaît (cf. (65)).
c. Les fautes de désinences personnelles (personne et nombre) sont corrigées, car elles sont clairement reconnaissables dès qu'on a un sujet dans la même phrase (cf. (68) - (71)).
d. L'accord de genre sur le participe n'est que corrigé là où il est clairement reconnaissable, c'est-à-dire à la troisième personne dont le sujet est exprimé dans le même syntagme verbal (cf. (72)). Dans le doute on optera pour le singulier du masculin. Nous ne nous orientons pas au sexe de l'auteur.

Exemples :

(64) *I pleu a Maurice → il pleut à Maurice*

(65) *Je peu invité qqn → je peux inviter quelqu'un*

(66) *Peut-tu me confirmer → peux-tu me confirmer*

(67) *J'te ferais écouter un de ces 4. → je te ferais écouter un de ces 4.*

(68) *I vont → ils vont*

(69) *I mange → il mange* (pas : *ils mangent*)

(70) *I mange, mes parents → il mange, mes parents* (pas : *ils mangent*) et de manière analogue

(71) *Mes parents, i mange → mes parents, il mange* (pas : *ils mangent*)

(72) *Je m'appelle Marie. Je suis venu hier → je suis venu hier* (pas : *venue*)

20. Accord

Les accords de nombre et genre sont ajoutés là où ils manquent sur les adjectifs, les noms et les participes si, et seulement s'ils sont exigés d'une manière non ambiguë, donc strictement grammaticale. Cela veut dire quand au sein du même syntagme il y a une disparité entre deux composants ou si un complément attributif ne s'accorde pas avec son sujet (cf. (73)). On tiendra aussi compte de la prononciation indiquée par les variantes graphiques (cf. (74)). Quand deux éléments doivent montrer un accord selon la norme française, mais seulement un des deux porte un certain trait, les deux éléments seront normalisés comme portant ce trait. Un cas particulier est la configuration dans laquelle deux éléments sont logiquement au pluriel, mais l'un de ces deux ne peut pas porter la marque morphologique. Parfois, les auteurs créent alors un pluriel inexistant. Ce dernier doit être normalisé selon la norme (donc au singulier). Par contre, l'élément qui peut porter le trait (mais qui ne le porte pas toujours !) sera au pluriel (cf. (75) et (76).

(pour l'accord verbal voir le chapitre 19 conjugaison verbale)

Exemples :

(73) *Je viens la semaine prochain. → je viens la semaine prochaine.*

(74) *Lé journé sont ensoleillé → les journées sont ensoleillées*

(75) *les jeudis matin* (913) → *les jeudi matins*

(76) *les allers-retour* (896) → *les aller-retours*

21. Adjectifs utilisés comme adverbes

Nous n'adapterons pas les adejctifs qui sont utilisées comme adverbes, sans pour autant présenter la forme adverbiale.

Dans l'exemple (77), *direct* n'est **pas** adapté à *directement*.

(77) *Marielle, en fait je voulais passer en ville mtnt pour pouvoir rentrer direct après, donc j'arriverai pas avant la pause de 4h au - Désolée j'avais oublié Biz* (13406)