Citer ce document :
Falaise, A. (2013). Manuel du *Corpus de français tchaté getalp_org*. In Banque de corpus CoMeRe, Chanier, T. (ed.). ORTOLANG/CoMeRe. [cmr-getalp_org-tei-v1]

CODAGE DU CORPUS

Ce dont ce document ne parle pas

Pour des informations sur le tchat en général, ou sur ce corpus, mais ne concernant pas directement le codage, se reporter aux documents suivants :

- A. Falaise, « Constitution d'un corpus de français tchaté », *actes de RECITAL 2005*, Dourdan, 2005
  - http://www-clips.imag.fr/geta/User/achille.falaise/publis/AF-Article_RECITAL_2005.pdf
  - [lien alternatif] http://taln.limsi.fr/site/talnRecital05/actes-articles.htm
  - [transparents] http://taln.limsi.fr/site/talnRecital05/session4R/falaise.ppt

- A. Falaise, *Premier pas vers une TA interactive pour le tchat*, rapport de stage de M2R, INPG, Grenoble, 2004
  - http://www-clips.imag.fr/geta/User/achille.falaise/publis/AF-Rapport_de_stage_M2R.pdf

Ces documents décrivent des versions précédentes du corpus, il ne faut donc pas se fier aux indications de codage qui y sont données. Par contre, les informations qu'ils donnent à propos du tchat, de la méthode de collecte de ce corpus, des questions légales, etc... sont toujours d'actualité.

A chaque canal de tchat (= « salon de discussion »), enregistré durant environ trois mois (cette durée peut être plus courte pour quelques canaux), correspond trois fichiers.

1. Les fichiers périphériques

Les 2 premiers sont des objets Java *Properties* sauvegardés par la méthode *store*(), et peuvent donc être directement chargés en Java dans un tel objet par la méthode *load()* (cf. Javadoc).

Il n'est toutefois pas indispensable de passer par Java pour utiliser ces fichiers. Il s'agit de fichiers « texte ». Ils débutent par deux lignes de commentaires (débutant par des symboles #), indiquant la date de génération du fichier. Ils comportent ensuite une succession de lignes formatées sur le modèle *proriété*=*valeur*. Les caractères spéciaux sont codés de manière conventionnelle : `\u`*`Code`*, où *`Code`* est le code UTF-16 du caractère en hexadécimal. Par exemple, le caractère « é » est noté « `\u00E9` ».

1.1 Les fichiers .statformes.db

Ces fichiers, nommés *NomCanal*.statsformes.db d'après le nom du canal, donnent la liste (non ordonnée) des formes lexicales relevées dans le corpus, ainsi que le nombre de leurs occurrences.

```
#
#Tue Oct 25 16:05:42 CEST 2005
W2156496=2
acceptez=2
vache=6
graine=2
rouiller=6
PostQuitMessage=2
After=1
d\u00E9clarations=2
```

Exemple du début d'un fichier .statsformes.db.

1.2 Les fichiers .statsparticipants.db

Ces fichiers, nommés *NomCanal*.statsparticipants.db d'après le nom du canal, donnent la liste (non ordonnée) des intervenants (pseudos) du canal (humains ou robots), relevés dans le corpus, ainsi que le nombre de leurs interventions.

```
#
#Tue Oct 25 16:05:42 CEST 2005
Fallen-Angel[OQP]=222
Anonyme6447699=2
Gui-Gui{}lycee{}=2
SysTeM|Failure[afk]=10
KRAKRAKRAKRAKRAKRASEUHHHHHh=22
Pro=2
Kralhen=18
Fallen-Angel[OQP][OQP]=16
```

Exemple du début d'un fichier .statsparticipants.db.

2. Le corpus proprement dit

L'ensemble des interventions d'un canal donné est regroupé dans un fichier XML (*NomCanal*.xml). Ce fichier est encodé en UTF-8.

2.1 L'élément <canal>

L'élément racine <canal> porte les attributs suivants :

- *nom* : le nom du canal
- *nbconnectes* : le nombre d'utilisateurs uniques se connectant ou se déconnectant, déterminés d'après les logins de connexion (et non les pseudos); c'est à dire à peu de choses près le nombre d'utilisateurs connectés à un moment ou un autre sur le canal
- *nbparticipants* : le nombre d'intervenants uniques, déterminés d'après les pseudos (un même utilisateur peut avoir plusieurs pseudos, on comptera alors plusieurs intervenants); c'est à dire le nombre d'utilisateurs envoyant des messages et/ou des commandes.
- *nbmessages* : le nombre de balises <message> (voir cette balise)
- *nbevenements* : le nombre de balises <evenement> (voir cette balise)
- *nbcommandes* : le nombre de balises <commande> (voir cette balise)
- *nbMotsMessages* : le nombre de mots[1] dans les interventions des balises <message>, paratexte (date, heure, pseudo de l'auteur) non compris.
- *nbmots* : le nombre de mots dans toutes les interventions, paratexte (date, heure, pseudo de l'auteur) non compris.
- *nbformes1* : le nombre de formes uniques apparaissant au moins une fois.
- *nbformes2* : le nombre de formes uniques apparaissant au moins deux fois.

```
<canal
      nom="c++"
      nbconnectes="2094"
      nbparticipants="441"
      nbmessages="55135"
      nbevenements="27547"
      nbcommandes="185"
      nbMotsMessages="146319"
      nbmots="273330"
      nbformes1="9416"
      nbformes2="8402"
/>
```

Exemple d'éléments racine (canal c++).

2.2 Les éléments <message>, <commande>, <evenement> et <commentaire>

Cet élément <canal> porte quatre types de sous-éléments:

- un élément <commentaire> : brève description du canal (généraliste, technique, thématique, etc.)
- un nombre indéterminé d'éléments <message>, qui correspondent aux interventions des utilisateurs (humains ou robots) à destination des autres utilisateurs (humains) du canal
- un nombre indéterminé d'éléments <commande>, qui correspondent aux interventions identifiées comme étant destinées aux robots
- un nombre indéterminé d'éléments <evenement>, qui correspondent aux notifications d'événements affichées sur le canal

La classification automatique entre <message> et <commande> n'est pas parfaite; c'est pourquoi il arrive

1Afin de tenir compte des spécificités du tchat, un mot est défini par l'expression régulière suivante
`[\s.:/\\\-'\"+;!,?(){}\]\[]([^\s.:/\\\-'\"+;!,?(){}\]\[]+)[\s.:/\\\-'\"+;!,?(){}\]\[]`
c'est à dire « n'importe quoi compris entre deux blancs ou caractères .:/\'"+;!,?(){}[] ».

(rarement) qu'un message soit identifié à tort comme une <commande>, et inversement.

2.3 Les éléments <message>, <commande>, <evenement>

Les éléments <message>, <commande> et <evenement> possèdent les attributs suivants :

- numero : le numéro d'ordre (unique dans un canal donné) de l'intervention.
- annee, mois, jour, heure, min : la date et l'heure de l'intervention
- idauteur : l'identifiant numérique unique de l'auteur de l'intervention (le cas échéant), basé sur le login; cet identifiant ne varie pas, même si l'utilisateur change de pseudo; par contre, il varie si un utilisateur se reconnecte avec un login différent

Ils possèdent en outre deux sous-éléments :

- <auteur> qui contient (le cas échéant) le pseudo de l'auteur de l'intervention
- <contenu> qui contient l'intervention effectivement affichée sur le canal

2.4 Spécificités de l'élément <evenement>

Cet élément contient en outre un attribut *type*, qui indique le type d'événement :

- "connexion" : un utilisateur vient de se connecter
- "deconnexion" : un utilisateur vient de se déconnecter
- "changementpseudo" : un utilisateur vient de changer de pseudo; au lieu d'un sous-élément <auteur>, on alors deux sous-éléments <pseudosource> et <pseudocible>
- "action" : « action » (en principe purement verbale) d'un utilisateur, et explicitement annoncée en tant que telle
- "changementmode" : un utilisateur change de statut
- "kick" : un utilisateur (disposant du statut approprié) vient de kicker (« éjecter ») un autre utilisateur
- "autre" : c'est suffisamment explicite non ?

```
<evenement
      numero="1"
      type="deconnexion"
      annee="2004"
      mois="2"
      jour="7"
      heure="0"
      min="0"
      idauteur="41914"
>
      <auteur>SysTeM|Failure</auteur>
      <contenu>
      SysTeM|Failure(~70ce2a25@fc73c398a4bef9b4) s'est déconnecté: Ping timeout
      </contenu>
</evenement>
```

Exemple d'élément <evenement>

Lors de la connexion et de la déconnexion, le login de l'utilisateur est donné en clair. Dans le corpus, il est brouillé par un procédé de hachage, qui empêche la reconstitution du login original, tout en conservant le caractère distinctif du login.

LICENCE D'UTILISATION DU CORPUS

Ce corpus de tchat, réalisé par Achille Falaise (GETA-CLIPS) est distribué avec l'aimable autorisation de Kévin Labécot, responsable du service Botstats (http://www.botstats.com), et dépositaire des droits d'auteurs relatifs aux canaux de tchat qu'il a reçu la charge de publier.

L'utilisation, la reproduction et la modification de ce corpus sont autorisées exclusivement dans le cadre d'activités scientifiques non-commerciales.

Ce corpus, en l'état ou modifié, ne saurait être diffusé ou utilisé à l'extérieur des milieux scientifiques, même partiellement.

Toute reproduction ou diffusion, effectuée dans le contexte d'utilisation ci-dessus, devra être accompagnée de la présente licence, ainsi que du manuel d'utilisation joint (manuel.pdf).

Le 16/01/2005 à Grenoble,

Achille Falaise, achille.falaise@imag.fr