# SAGACE-v3.3 ; Analyseur de corpus pour langues non flexionnelles

Blin Raoul

CRLAO - CNRS , 54 Raspail 75006 Paris
blin@ehess.fr

**Résumé** Nous présentons la dernière version du logiciel SAGACE, analyseur de corpus pour langues faiblement flexionnelles (par exemple japonais ou chinois). Ce logiciel est distribué avec un lexique où les catégories sont exprimées à l'aide de systèmes de traits.

**Abstract** We present a software program named SAGACE, designed to search for and extract word strings from a large corpus. It has been conceived for poor flexional languages, such as Japanese or Chinese. It is associated with a lexicon where categories are expressed with feature systems.

**Mots-clés :** corpus , lexique , analyseur, japonais , chinois
**Keywords:** corpus, lexicon , analyzer , japanese , chinese

SAGACE est un logiciel d'analyse de corpus destiné plus particulièrement à l'étude de langues faiblement flexionnelles. Il est actuellement exploité entre autres pour la recherche sur le japonais, sur un corpus de grande taille (26 millions de phrases) non balisé. Il est distribué[1] gratuitement sous licence CECILL.

Nous présentons ici ses principales fonctionnalités et caractéristiques.

## 1 Conceptions et fonctionnalités de base du logiciel

Le logiciel est constitué d'un moteur de recherche, qui associé à un lexique manipule des corpus tagués ou non. Le moteur assure deux fonctions :

1) Concordancier

Extraction d'une chaîne et de son contexte, affichage dans différents formats, dont KWIC. Les segments de textes extraits sont de toutes natures et maîtrisables par l'utilisateur : phrase, paragraphe, autres. Ces segments sont définis grâce à des marques d'arrêt comme par exemple

[1] ljp.homelinux.net

un rond (marque de fin de phrase en japonais), un saut de ligne, tout autre marque dont éventuellement une balise insérée par l'utilisateur dans le corpus.

2) Extraction et comptage de collocations

# 2 Lexique

Les lexiques exploités par SAGACE sont des listes d'entrées dotées d'une représentation syntaxique exprimée à l'aide d'une formule propositionnelle de traits.

ex : 私 [[ pronom & lecture:"watasi" & lecture:origine:"japonais" ...]]

SAGACE dispose de plusieurs fonctions de manipulations du lexique : listage du contenu (recherche d'un lemme particulier, listage du contenu d'une catégorie etc.), insertion d'entrée. La fonction d'insertion évite une ouverture manuelle du lexique. Elle permet aussi de gérer les droits d'accès au lexique.

Il existe actuellement deux lexiques immédiatement exploitables par SAGACE. Le premier, LEXS-J-β3.0.0 comporte 380.000 entrées pour le japonais et le second LEXS-CHS- β1.0.0 comporte 40.000 entrées pour le chinois simplifié. Ils sont librement distribués[2].

# 3 Nature des chaînes recherchées et description

Le logiciel cherche des chaînes de lemmes. Il ne travaille pas sur des syntagmes et se démarque en cela de logiciels comme UNITEX ou NOOJ, entre autres.

Une chaîne à chercher est décrite composant par composant, en indiquant la valeur du composant et différentes options de recherches.

Chaque composant de la chaîne à chercher est soit un lemme donné (exemple : 私 (*watasi*, je)), soit n'importe quel lemme d'une catégorie donnée (exemple : pronom) décrite à l'aide d'une formule propositionnelle de traits. Avec ce langage, il est aussi possible de composer de nouvelles catégories à partir des catégories existantes. Par exemple : combiner deux ou plusieurs catégories en une seule (par disjonction), exclure une sous catégorie (usage de la négation etc.), créer un sous-ensemble (par conjonction). Le langage offre ainsi la possibilité à l'utilisateur de décrire des nouvelles catégories sans avoir à modifier le contenu du lexique.

Les chaînes manipulées peuvent être discontinues. Il est possible de rechercher des morphèmes disjoints.

Il est enfin possible de chercher des lemmes "inconnus", compris entre deux lemmes connus. Cette fonction permet par exemple de relever tous les morphèmes qui occupent une place donnée dans la chaîne (entre deux éléments connus).

[2] ljp.homelinux.net

# 4 Les corpus manipulés

Le logiciel travaille sur des textes non balisés ou comportant des balises structurelles (distinctions titre/corps de texte) au format xml. L'intérêt du texte non balisé est de pouvoir effectuer une étude sans nécessiter aucun prétraitement.

Il est possible depuis la version 3.3 de prendre en compte la structure du texte en distinguant les titres et leur niveau d'imbrication, et le corps du texte. Cette fonction permet de faire des recherches conditionnées, comme par exemple de limiter la recherche à du texte situé dans la portée d'un titre remplissant certaines conditions.

Présentement, le logiciel est utilisable en ligne et travaille avec un corpus non-tagué de 26 millions de phrases japonaises (en majorité extraites du web).

Un corpus avec balisage de la structure du texte est actuellement en cours de constitution.

# 5 Conclusion

SAGACE est distribué pour installation sur poste individuel et une version est utilisable en ligne[3] pour simplifier au maximum l'utilisation et pour mutualiser les données (lexique et corpus).

Le logiciel est avant tout destiné à l'usage scientifique (aide à l'analyse linguistique) mais il est conçu aussi pour entrer dans des chaînes de traitement plus complexes. En l'occurrence, un travail est en cours pour concevoir un prototype d'analyseur d'opinion : cet analyseur, qui repose sur l'analyse linguistique effectuée par SAGACE, permettra par exemple à un utilisateur français de connaître l'image/réputation d'un objet donné (un produit, une marque, etc.) sur le web japonais.

***Matériel demandé : tableau, panneau pour poster, connection internet***

[3] http://ljp.homelinux.net