Constitution d'un dictionnaire de fréquences du français à l'aide de SAGACE

Raoul Blin

Constitution d'un dictionnaire de fréquences du français à l'aide de SAGACE

FrWoFr-v1

2014/12/30

R.Blin
CNRS-CRLAO

Résumé : Nous montrons comment relever la fréquence des mots dans un corpus en français à l'aide du logiciel SAGACE.

Summary : We show how to count the frequency of the words in a french corpus with the software SAGACE.

Le logiciel SAGACE (Blin 2012) est un analyseur de corpus conçu pour produire un concordancier et compter les fréquences de *patterns* de mots dans une corpus, étant donné un lexique. Ne pouvant faire d'analyses morphologiques, il a été essentiellement utilisé pour des langues très faiblement flexionnelles, et en particulier le japonais (entre autres (Blin 2013)). Cependant, l'existence de lexiques exhaustifs de formes fléchies comme le Lefff (Sagot 2010)change complètement la donne. Avec un tel lexique, Il est désormais possible d'utiliser SAGACE pour travailler aussi sur le français. On peut même s'attendre à ce que la qualité des résultats pour le français soit meilleure que pour une langue comme le japonais car les effets négatifs induits par l'absence de séparateurs de mots en japonais (voir discussion dans (Blin 2014)n'existent pas en français. Une étude reste cependant à faire sur les performances en qualité, notamment en comparaison avec d'autres logiciels. Ceci n'est pas notre propos ici.

Nous avions besoin, dans le cadre d'un projet de traducteur hybride du japonais vers le français de disposer d'une liste des fréquences approximatives des mots (et non des morphes) du français écrit. Nous avons pour cela utilisé le logiciel SAGACE, le Lefff et wikipedia comme corpus.

Le corpus est le frwiki-20130216-pages-articles-multistream.xml[1] et compte plus de 190 millions d'occurrences de mots selon le comptage de la présente étude. Il contient de nombreuses répétitions dont l'effet sur les résultats n'a pas été mesuré.

Le choix du lexique s'est porté sur le Lefff, pour son exhaustivité et sa license. Nous avons utilisé Lefff-ext-3.2[2] qui a été reformaté pour être exploitable par SAGACE et simplifié de sorte à ne contenir que des informations pertinentes pour la présente étude. Le lexique ainsi obtenu[3] comporte toutes les formes de chaque mot. Une particularité est que les différentes formes d'un mot reçoivent un identifiant commun. Par exemple les deux formes, singulier et pluriel du mot *attaquable*, reçoivent un même trait « mot :"attaquable" », réservé à ce mot, :

attaquable [[cat:adj & mot:"attaquable"]]
attaquables [[cat:adj & mot:"attaquable"]]

SAGACE atribue en plus, automatiquement, un trait « ref » unique propre à chaque entrée.

SAGACE cherche des patterns constitués par les entrées du lexique. Dans cette étude exploratoire, nous nous en sommes tenu à chercher des mots séparés par un blanc graphique. La description du pattern cherché est :

1 https://dumps.wikimedia.org/frwiki/20140822/
2 http://gforge.inria.fr/frs/download.php/32626/lefff-ext-3.2.tgz
3 Lefff-ext.3.2_prSagace : http://sharedocs.huma-num.fr/wl/?id=qK&filename=Lefff-ext.3.2_prSAGACE.dic.tar.gz

```
>0 EspaceSimple
=0 cat:motfrancais /-affich:trait:lemme /-compte
=0 EspaceSimple
```

'Espacesimple' est un mot réservé pour décrire l'espace (un octet). 'motfrancais' correspond aux entrées du lexique, y compris les signes de ponctuations. Avec le trait 'lemme', on s'assure de compter les mots et non les morphes.

La combinaison de ce pattern et du lexique a plusieurs défauts avec SAGACE. Le premier est que les occurrences commençant par une majuscule, en début de phrase ou de citation, ne sont pas comptabilisées. Cela affecte plus particulièrement les déterminants. Pour y remédier, la solution la plus simple, à défaut d'être la plus élégante, serait de dédoubler les entrées du lexique pour que chaque morphe figure avec et sans majuscule en initiale. Le deuxième défaut est de ne pas comptabiliser les occurrences de morphes immédiatement suivies d'une ponctuation : mots devant un point, une virgule etc. Pour ce faire, il faudra se donner une liste de signes de ponctuations (catégorisés comme tels) et modifier la description du troisième composant du pattern :

```
=0 cat:EspaceSimple | ponctuation
```

Pour l'instant, SAGACE ne permet pas de traiter l'espace comme une entrée lexicale ordinaire et cette description n'est pas encore possible.
L'usage des résultats nous autorisait à nous en tenir à un comptage frustre. Les « manques » étant compensés par la taille du corpus.

Bibliographie

Blin, Raoul. 2012. “SAGACE v4.2.0.” http://crlao.ehess.fr/japonais-coreen/corpus/sagace/manuel/Manuel.pdf.

———. 2013. *DIctionnaire de Fréquence Du Japonais Contemporain - 16.000 Noms-*. Librairie You Feng. Paris.

———. 2014. “Comparaison de deux outils d’analyse de corpus japonais pour l’aide au linguiste, Sagace et MeCab.” In *Actes TALN-RECITAL 2014*, 497. Marseilles, France. http://hal.archives-ouvertes.fr/hal-01054370.

Sagot, Benoît. 2010. “The Lefff, a Freely Available and Large-Coverage Morphological and Syntactic Lexicon for French.” In *Proceedings of the 7th International Conference on Language Resources and Evaluation*. Istanbul, Turkey.

Annexe

Les mots les plus fréquents obtenus (la liste complète est téléchargeable[4]).

de 23099868	- 1336226	premier 533652
la 9309482	a 1275496	nom 515863
et 7322741	avec 1129460	mais 486218
= 5657608	plus 1029810	deux 479266
: 4035227	son 1007587	aussi 454073
est 2937445	avoir 972117	commun 450381
être 2751044	se 970877	entrer 418011
un 2464553	faire 818830	site 370881
dans 2348487	pas 765756	» 365555
par 2272629	« 728283	En 344143
une 2104957	ne 643236	grand 342465
pour 1885880	pouvoir 630597	même 337085
sur 1803534	sa 621965	mettre 307916
qui 1607789	tout 575503	sou 307648

4 http://sharedocs.huma-num.fr/wl/?id=2K&filename=FrWofr-v1.txt.tar.gz

titre 307414
français 294321
ver 293351
nouveau 292416
population 287822
France 278086
national 277928
ville 275299
où 272409
autre 271115
alors 261038
monde 254294
très 247449
année 247019
externe 246569
ainsi 246229
devenir 244247
partie 240221
après 237467
également 232418
depuis 232084
voir 230543
prendre 223990
sans 223889
in 218714
référence 217065
permettre 213347
lieu 212826
puis 206622
contre 205927
bien 205702
janvier 203603
Elle 200664
mar 198499
général 197905
utiliser 196827
maire 194561
trois 192978
situé 192844
groupe 190018
dernier 189816
je 189668
image 185330
an 185083
plusieurs 183909
lors 181221
fin 181195
certain 180957
se trouver 180219
avant 178189
page 177162
si 176431
dire 172980
région 172033
jour 165597
politique 165204
nombreux 162336
date 161685
partir 159612
petit 159115
mai 158986
pays 158384
légende 157661
mort 157536
début 157089
septembre 155497
connaître 154819
film 153632
donc 153427
naître 153297
non 152158
quelque 150243
devoir 149903
juin 149663
famille 149043
peu 148658
fois 148556
encore 144877
octobre 144576
cour 142661
moins 142146
décembre 141987
football 141971
leurs 140953
ligne 140234
droit 139569
forme 138776
juillet 138228
novembre 137006
nombre 136342
point 135198
/ 133918
De 133688
place 132683
seul 131674
avril 131442
août 129628
ancien 127978
février 127772
temps 127351
sortir 126410
membre 125777
notamment 124758
genre 122664
jeu 120918
produire 120459
naissance 120447
principal 120198
tour 120181
type 119090
série 118070
ils 117572
On 116965
personne 115827
club 115384
tel 114296
pendant 113835
suite 111877
Un 111754
meilleur 111536
article 110951
langue 110513
roi 110283
vie 109102
lien 108958
web 108900
réaliser 107964
consulter 107888
? 105644
nommer 104364