HAL open science

A lexicometric analysis of the Japanese novel kokoro (Natsume Soseki)

Raoul Blin

To cite this version:

Raoul Blin. A lexicometric analysis of the Japanese novel kokoro (Natsume Soseki). 2017. hal-01473996v1

HAL Id: hal-01473996

https://hal.science/hal-01473996v1

Preprint submitted on 22 Feb 2017 (v1), last revised 28 Feb 2017 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

Analyse lexicométrique du roman japonais *kokoro* (« Le pauvre cœur des hommes ») de Natsume Soseki

2017-02-22
R.Blin
CNRS-CRLAO

English title: A lexicometric analysis of the Japanese novel kokoro (Natsume Soseki)

Summarize : This text provides the lexicometric analysis of kokoro, performed by the automatic system Mecabenligne-v0.1

Ce texte présente les résultats de l'analyse lexicométrique du roman japonais *kokoro* à l'aide du dispositif Mecabenligne-0.1 [1] .

1 Le roman..... 1
2 Outil d'analyse..... 1
3 Evaluation..... 2
4 Bibliographie..... 2

1 Le roman

Le roman *kokoro* (« Le pauvre coeur des hommes ») de Natsume Sôseki (1867-1917) est paru en 1914. C'est un classique de l'époque moderne (fin du XIXè, début du XXè siècle), époque qui voit la langue évoluer vers sa forme contemporaine : évolution du système morphologique verbal, bouleversement du lexique. Le roman (en japonais) est désormais librement accessible à l'adresse :

http://www.aozora.gr.jp/cards/000148/files/773_14560.html

2 Outil d'analyse

Mecabenligne-v0.1 est un outil en ligne pour l'analyse lexicométrique de sites web en japonais. Il récupère la page d'une adresse web donnée par l'utilisateur. La page est débarrassée de son entête (<head>...</head>) ainsi que des éventuels hurigana[2] (<ruby>...</ruby>). Les retours chariot (<br/>) sont transcrits en retours chariots '\n'. Enfin les balises html sont éliminées. Un traitement particulier est réservés aux pages du site Aozora. Elles sont débarrassées des éventuelles méta-données bibliographiques.

Le texte est ensuite lemmatisé à l'aide de Mecab (Kudo 2006), analyseur morphologique très utilisé en traitement automatique du japonais. Deux dictionnaires sont mis à disposition pour l'analyse. Il est recommandé de sélectionner le dictionnaire Kindai Bungo UniDic (小木曽, 小町, et 松本 2013) pour traiter les textes de l'époque moderne (fin XIXè - début XXè S.). Les textes contemporains (à partir de la deuxième moitié du XXè S.) sont traités à l'aide du dictionnaire Ipadic-v102 (Asahara et Matsumoto 2003)

L'analyse produit cinq fichiers :

1 http://5.39.94.64/divers/mecabenligne/mecabenligne.php
ou http://rkappa.fr/divers/mecabenligne/mecabenligne.php

2 *hurigana* : *kana* inscrit à coté d'un *kanji* pour en indiquer la lecture.

1. Le fichier texte produit à partir de la page web.
2. L'analyse par mecab.
3. La liste des morphèmes avec leurs occurrences.
4. Une analyse linguistique de synthèse comportant les informations :

    Nombre de morphes

    Nombre total des occurrences de morphes

    Informations quantitatives sur les catégories morpholexicales

    Informations quantitatives sur les strates lexicales toutes catégories confondues

    Informations quantitatives sur les strates lexicales en fonction de la catégorie morpholexicale

3 Évaluation

Nous n'avons pas effectué d'analyse sur le résultat produit par le dispositif Mecabenligne. Une configuration quasiment identique a été évaluée sur un échantillon de textes modernes et a obtenu un taux de succès de plus de 94 %[3] (小木曽, 小椋, et 近藤 2008). Nous ne disposons d'aucun résultat pour les textes contemporain mais on constate que l'analyse de sites d'actualité contiennent de nombreuses sous-chaînes inconnues ou traitées de façon erronnée. Il est de ce fait difficile de croire que le niveau soit aussi bon.

Dans les analyses à l'aide du dictionnaire contemporain, des erreurs sont susceptibles d'être provoquées par les termes inconnus. Lorsqu'une chaîne de katakana n'est pas reconnue, Mecab la traite (éventuellement à tort) comme un nom commun :

オキナワ　　名詞,普通名詞,*,*,*,*

Lorsqu'une sous-chaîne de kanji n'est pas reconnue, mecab traite chaque kanji comme un nom commun (éventuellement à tort). Le résultat est alors compté comme « inconnu » :

M　　名詞,普通名詞,*,*,M,lecture supposée,漢字読み:音 代表表記:M

Mecab traite les chaînes de caractères latins non reconnues comme des noms communs (éventuellement à tort) :

JAPANESE, *, 名詞, 組織名

La version utilisée dans la présente analyse est Mecabenligne-v0.1, mecab 0.996, Kindai Bungo UniDic Ver.1.4 , ipadic-v102.

4 Bibliographie

Asahara, Masayuki, et Yuji Matsumoto. 2003. « IPADIC User Manual ». Nara Institute of Science and Technology.

Kudo, Taku. 2006. « MeCab: yet another part-of-speech and morphological analyzer. » http://mecab.sourceforge.net.

小木曽智信，小椋秀樹，et 近藤明日子．2008．« 近代文語文を対象とした形態素解析辞書の開発 ». In 言語処理学会第 14 回年次大会発表論文集，225-28．言語処理学会．https://dl.dropboxusercontent.com/u/73297026/report/unidic-MLJ_report2009.pdf.

小木曽智信，小町守，et 松本裕治．2013. « Morphological Analysis of Historical Japanese Text ». 自然言語処理 = Journal of natural language processing 20 (5): 727-48.

3 La F-mesure pour la segmentation est supérieure à 0,97, pour la catégorisation elle se situe entre 0,94 et 0,96, pour la lemmatisation entre 0,93 et 0,98.

5 Résultats

Voici les résultats de l'analyse du roman « kokoro », tel que transcrit sur le site :

http://www.aozora.gr.jp/cards/000148/files/773_14560.html

Dans les résultats qui suivent,
- les fréquences sont calculées par rapport au nombre total d'occurrences de morphes (ponctuations non comprises), soit 97 865.
- les proportions sont calculées par rapport au nombre total d'occurrences de morphes (ponctuations non comprises), soit 5 809.

5.1 Synthèse

Nombre de morphes : 5 826
Nombre de morphes (sans les ponctuations): 5 809
Nombre de signes de ponctuations : 17

Nombre d'occurrences de morphes : 107 669
Nombre d'occurrences de morphes (sans les ponctuations): 97 865
Nombre d'occurrences de signes de ponctuation : 9 804

Nombre de phrases : 4 654

5.1.1 Comptage par catégories (nombre d'items différents et proportion, nb d'occurrences et fréquence)

助詞	:	69	.01188	32 616	.33328
名詞	:	3 181	.54760	21 286	.21750
動詞	:	1 478	.25443	15 278	.15611
助動詞	:	42	.00723	13 747	.14047
補助記号	:	17	.00293	9 804	.10018
代名詞	:	41	.00706	4 121	.04211
副詞	:	323	.05560	2 959	.03024
形容詞	:	215	.03701	1 975	.02018
接尾辞	:	160	.02754	1 578	.01612
連体詞	:	26	.00448	1 501	.01534
形状詞	:	193	.03322	1 253	.01280
接続詞	:	11	.00189	551	.00563
接頭辞	:	35	.00603	452	.00462
記号	:	3	.00052	413	.00422
感動詞	:	32	.00551	135	.00138

5.1.2 Comptage par strates lexicales (nombre d'items différents et proportion, nb d'occurrences et fréquence)

和	:	3 321	.57170	84 958	.86811
漢	:	2 199	.37855	11 253	.11498
記号	:	20	.00344	10 217	.10440

混	:	180	.03099	921	.00941
固	:	65	.01119	247	.00252
外	:	39	.00671	66	.00067

5.1.3 Comptage par strates lexicales par catégories (nombre d'items différents, nb occurrences)

連体詞-混	:	4	18
連体詞-和	:	22	1 483
記号-記号	:	3	413
補助記号-記号	:	17	9 804
接頭辞-漢	:	27	136
接頭辞-和	:	8	316
接続詞-和	:	11	551
接尾辞-漢	:	96	420
接尾辞-和	:	64	1 158
感動詞-和	:	32	135
形状詞-漢	:	112	807
形状詞-混	:	2	26
形状詞-和	:	79	420
形容詞-混	:	4	70
形容詞-和	:	211	1 905
名詞-漢	:	1 920	9 531
名詞-混	:	88	240
名詞-外	:	39	66
名詞-固	:	65	247
名詞-和	:	1 067	11 195
名詞-	:	2	7
動詞-混	:	70	486
動詞-和	:	1 408	14 792
助詞-和	:	69	32 616
助動詞-和	:	42	13 747
副詞-漢	:	42	354
副詞-混	:	9	66
副詞-和	:	272	2 539
代名詞-漢	:	2	5
代名詞-混	:	3	15
代名詞-和	:	36	4 101

5.2 Extraits des résultats de la lemmatisation

5056, た, タ, 助動詞, *, 助動詞-タ, 和
4654, 。, , 補助記号, 句点, *, 記号
4509, の, ノ, 助詞, 格助詞, *, 和
4177, は, ハ, 助詞, 係助詞, *, 和
4055, て, テ, 助詞, 接続助詞, *, 和
3685, に, ニ, 助詞, 格助詞, *, 和
3605, 、, , 補助記号, 読点, *, 記号
3219, を, ヲ, 助詞, 格助詞, *, 和
2586, と, ト, 助詞, 格助詞, *, 和
2439, だ, ダ, 助動詞, *, 助動詞-ダ, 和
2367, 為る, スル, 動詞, 非自立可能, 文語サ行変格, 和
1992, が, ガ, 助詞, 格助詞, *, 和
1991, 私, ワタシ, 代名詞, *, *, 和
1745, も, モ, 助詞, 係助詞, *, 和
1413, の, ノ, 助詞, 準体助詞, *, 和
1294, です, デス, 助動詞, *, 助動詞-デス, 和
1053, なり-断定, ナリ, 助動詞, *, 文語助動詞-ナリ-断定, 和
929, ない, ナイ, 助動詞, *, 助動詞-ナイ, 和
909, まし, マシ, 助動詞, *, 文語助動詞-マシ, 和
837, 言う, イウ, 動詞, 一般, 五段-ワア行, 和
777, 有る, アル, 動詞, 非自立可能, 文語ラ行変格, 和
760, で, デ, 助詞, 格助詞, *, 和
743, か, カ, 助詞, 係助詞, *, 和
732, 其の, ソノ, 連体詞, *, *, 和
713, 「, , 補助記号, 括弧開, *, 記号
685, , , 空白, *, *, 記号
658, 」, , 補助記号, 括弧閉, *, 記号
645, 居る, イル, 動詞, 非自立可能, 文語上一段-ワ行, 和
636, 其れ, ソレ, 代名詞, *, *, 和
634, さん, サン, 接尾辞, 名詞的, *, 和
616, 私, ワタクシ, 名詞, 普通名詞, *, 和
613, から, カラ, 助詞, 格助詞, *, 和
597, 先生, センセイ, 名詞, 普通名詞, *, 漢
581, 事, コト, 名詞, 普通名詞, *, 和

512, へ, へ, 助詞, 格助詞, *, 和
431, 成る, ナル, 動詞, 非自立可能, 文語四段-ラ行, 和
418, 物, モノ, 名詞, 普通名詞, *, 和
411, Ｋ, ケー, 記号, 文字, *, 記号
404, 様, ヨウ, 形状詞, 助動詞語幹, *, 漢
401, 奥, オク, 名詞, 普通名詞, *, 和
377, 時, トキ, 名詞, 普通名詞, *, 和
377, 今, イマ, 名詞, 普通名詞, *, 和
366, 無い, ナイ, 形容詞, 非自立可能, 文語形容詞-ク, 和
352, 無い, ナイ, 形容詞, 非自立可能, 形容詞, 和
351, から, カラ, 助詞, 接続助詞, *, 和
336, む, ム, 助動詞, *, 文語助動詞-ム, 和
296, 父, チチ, 名詞, 普通名詞, *, 和
296, 何, ナニ, 代名詞, *, *, 和
281, 思う, オモウ, 動詞, 一般, 文語四段-ハ行カ行, 和
278, れる, レル, 助動詞, *, 文語下二段-ラ行, 和
276, 御, オ, 接頭辞, *, *, 和
275, ます, マス, 助動詞, *, 助動詞-マス, 和
264, 自分, ジブン, 名詞, 普通名詞, *, 漢
257, 見る, ミル, 動詞, 非自立可能, 文語上一段-マ行, 和
250, 来る, クル, 動詞, 非自立可能, 文語カ行変格, 和
246, 此の, コノ, 連体詞, *, *, 和
237, 彼, カレ, 代名詞, *, *, 和
236, ば, バ, 助詞, 接続助詞, *, 和
235, が, ガ, 助詞, 接続助詞, *, 和
219, 聞く, キク, 動詞, 一般, 文語四段-カ行, 和
214, 板, イタ, 名詞, 普通名詞, *, 和
210, 然し, シカシ, 接続詞, *, *, 和
208, 又, マタ, 接続詞, *, *, 和
187, 彼方, アナタ, 代名詞, *, *, 和
187, まで, マデ, 助詞, 副助詞, *, 和
183, 母, ハハ, 名詞, 普通名詞, *, 和
183, だけ, ダケ, 助詞, 副助詞, *, 和
180, 言う, イウ, 動詞, 一般, 文語四段-ハ行, 和
175, 人, ヒト, 名詞, 普通名詞, *, 和
172, 中, ナカ, 名詞, 普通名詞, *, 和
171, 方, ホウ, 名詞, 普通名詞, *, 漢
168, 媼, オウナ, 名詞, 普通名詞, *, 和
167, 前, マエ, 名詞, 普通名詞, *, 和
167, どう, ドウ, 副詞, *, *, 和
167, ず, ズ, 助動詞, *, 文語助動詞-ズ, 和
158, 気, キ, 名詞, 普通名詞, *, 漢
152, 所, トコロ, 名詞, 普通名詞, *, 和
151, 奏する, ソウスル, 動詞, 一般, 文語サ行変格, 混
151, 出来る, デキル, 動詞, 非自立可能, 上一段-カ行, 和
150, 返る, カエル, 動詞, 一般, 文語四段-ラ行, 和
143, 唯, タダ, 副詞, *, *, 和
139, で, デ, 助詞, 接続助詞, *, 和
138, より, ヨリ, 助詞, 格助詞, *, 和
133, 顔, カオ, 名詞, 普通名詞, *, 和
132, ず, ズ, 助動詞, *, 助動詞-ヌ, 和
130, 行く, イク, 動詞, 非自立可能, 文語四段-カ行, 和
129, 一, イチ, 名詞, 数詞, *, 漢
127, 話, ハナシ, 名詞, 普通名詞, *, 和
126, 為, タメ, 名詞, 普通名詞, *, 和
126, 出す, ダス, 動詞, 非自立可能, 文語四段-サ行, 和
126, 上, ウエ, 名詞, 普通名詞, *, 和
124, 言葉, コトバ, 名詞, 普通名詞, *, 和
122, 知る, シル, 動詞, 一般, 文語四段-ラ行, 和
120, 目, メ, 名詞, 普通名詞, *, 和
120, 其処, ソコ, 代名詞, *, *, 和
118, 立つ, タツ, 動詞, 一般, 文語四段-タ行, 和
117, り, リ, 助動詞, *, 文語助動詞-リ, 和
115, 此れ, コレ, 代名詞, *, *, 和
114, 持つ, モツ, 動詞, 一般, 文語四段-タ行, 和
113, 要る, イル, 動詞, 一般, 文語四段-ラ行, 和
113, 二人, フタリ, 名詞, 普通名詞, *, 和
112, 内, ウチ, 名詞, 普通名詞, *, 和
109, 未だ, マダ, 副詞, *, *, 和
109, 彼の, カノ, 連体詞, *, *, 和
108, 間, アイダ, 名詞, 普通名詞, *, 和
108, 僧, ソウ, 名詞, 普通名詞, *, 漢
107, 心, ココロ, 名詞, 普通名詞, *, 和
106, たり-完了, タリ, 助動詞, *, 文語助動詞-タリ-完了, 和
105, 考える, カンガエル, 動詞, 一般, 文語下二段-ハ行, 和
104, 直ぐ, スグ, 副詞, *, *, 和
104, 妻, ツマ, 名詞, 普通名詞, *, 和
101, 何時, イツ, 代名詞, *, *, 和
101, こう, コウ, 副詞, *, *, 和
100, 分かる, ワカル, 動詞, 一般, 文語四段-ラ行, 和
99, 同じ, オナジ, 連体詞, *, *, 和
98, 申す, モウス, 動詞, 非自立可能, 文語四段-サ行, 和
98, よ, ヨ, 助詞, 終助詞, *, 和
97, 出る, デル, 動詞, 一般, 下一段-ダ行, 和
96, 後, ノチ, 名詞, 普通名詞, *, 和
94, 少し, スコシ, 副詞, *, *, 和
94, ほど, ホド, 助詞, 副助詞, *, 和
93, 仕舞う, シマウ, 動詞, 非自立可能, 文語四段-ハ行, 和
92, 良く, ヨク, 副詞, *, *, 和
92, 向かう, ムカウ, 動詞, 一般, 文語四段-ハ行, 和
91, 通る, トオル, 動詞, 一般, 文語四段-ラ行, 和
91, られる, ラレル, 助動詞, *, 文語下二段-ラ行, 和
90, 為る, スル, 動詞, 非自立可能, サ行変格, 和
90, 口, クチ, 名詞, 普通名詞, *, 和
90, や, ヤ, 助詞, 係助詞, *, 和
89, 対する, タイスル, 動詞, 一般, 文語サ行変格, 混
87, 見える, ミエル, 動詞, 一般, 文語下二段-ヤ行, 和
86, と, ト, 助詞, 接続助詞, *, 和
85, 何処, ドコ, 代名詞, *, *, 和
85, ながら, ナガラ, 助詞, 接続助詞, *, 和
84, 人間, ニンゲン, 名詞, 普通名詞, *, 漢
83, 来たる, キタル, 動詞, 非自立可能, 文語四段-ラ行, 和
83, 取る, トル, 動詞, 一般, 文語四段-ラ行, 和
83, もう, モウ, 副詞, *, *, 和
81, 私, シ, 名詞, 普通名詞, *, 漢
81, 知れる, シレル, 動詞, 一般, 文語下二段-ラ行, 和
81, 死ぬ, シヌ, 動詞, 一般, 文語ナ行変格, 和
79, 女, オンナ, 名詞, 普通名詞, *, 和
79, 伯父, オジ, 名詞, 普通名詞, *, 和
77, けれど, ケレド, 接続詞, *, *, 和