Peut-on déterminer la prévalence d'une maladie en fonction de la fréquence de son nom sur le web ? Le cas du cancer au Japon

Raoul Blin

Peut-on déterminer la prévalence d'une maladie en fonction de la fréquence de son nom sur le web ? Le cas du cancer au Japon

2015/10/11

Blin Raoul

CNRS-CRLAO

blin@ehess.fr

Résumé : Nous nous sommes interrogé sur la possible corrélation entre l'ordre d'occurrence des noms de maladies dans un corpus de textes, et l'ordre de prévalence des maladies dans la population parlant la langue de ces textes et à la date de leur publication. A titre expérimental, nous avons comparé l'ordre du nombre d'occurrences des noms de cancer dans un corpus en japonais extrait du web et les données épidémiologiques au Japon. La correspondance est très grossière, limitant ainsi l'intérêt des données du web. Néanmoins, elles sont acceptables pour donner une indication lorsque les études épidémiologiques sont inexistantes ou parcellaires.

Mots clefs : corpus, prévalence des maladies, japonais, cancer,

1 Introduction

Connaître l'ordre de prévalence des maladies est important pour de nombreux acteurs de la santé et pour les populations. Mais établir un tel ordre suppose l'existence d'un dispositif d'observation (recueil, centralisation et traitement des données) très lourd, en général géré par les états et reposant sur la contribution des praticiens. Une stratégie parallèle a vu le jour. Elle consiste à collecter des informations à partir d'une contribution volontaire ((Pellerin 2008) , (Goulinet 2014)). Toutes ces initiatives impliquent une organisation lourde et centralisée. Les deux approches ont pour premier défaut de limiter les informations à celles fournies par les contributeurs. La qualité dépend de leurs compétences et de leur assiduité. D'autre part, du fait de la centralisation des données, l'accès aux résultats n'est pas systématiquement garanti : il faut être en mesure de les localiser et d'avoir un droit de consultation, lequel dépend du bon vouloir des détenteurs des informations. Ne peut-on concevoir un dispositif permettant à tout un chacun de constituer sa propre ressource et de l'analyser ?

Aujourd'hui, l'accès à l'internet ou à la téléphonie mobile se développe et la plupart des pays disposent d'une presse et de réseaux sociaux numériques en ligne, même les pays les plus démunis. Nous nous sommes donc demandé si cela ne pouvait constituer une ressource pour établir un ordre de prévalence. Si oui, il suffirait alors de disposer d'un outil d'analyse pour exploiter cette ressource, ce qui ne représente pas de difficulté compte tenu de la profusion d'outils librement accessibles.

Dans cet article, nous proposons une première solution dans ce sens. Il s'agit d'une procédure légère, consistant à collecter sur le web un corpus d'actualité, d'analyser le nombre d'occurences des noms de maladies et d'en déduire un ordre de prévalence « virtuel ». L'ensemble du dispositif repose sur des logiciels libres. Notre principale interrogation est la fiabilité du classement obtenu par cette procédure. Pour cela, nous l'avons comparé au classement fourni par des études épidémiologiques. Par opportunisme, notre intérêt s'est porté sur le cancer, le web et les études épidémiologiques japonaises. Mais la démarche et les outils sont complètement adaptables à toutes autres langues. La seule contrainte réside dans les modalités d'accès au web. Nous présentons ici l'expérience et les résultats.

2 Expérience

2.1 Constitution du corpus

Nous avons exploité un sous-corpus du corpus monolingue japonais corefjp.0.003.150528 (Blin 2015). Ce sous corpus est composé de journaux, livres blancs gouvernementaux, minutes du parlement et site de questions réponses en ligne, ouvert au grand public. Sauf les minutes du Parlement qui couvrent tout l'après guerre, les données sont récentes puisque publiées de 2008 à 2011. Au total, nous disposons d'un corpus de plus de 4 milliards de caractères.

Tableau 1: Constitution du corpus (avec nombre de caractères japonais).

2G	Journaux (Contient le corpus d'articles de P.Marchal, ERTIM-INALCO)
12M	Livres blancs
2G	Minutes Diète
5M	Questions au gouvernement
7M	Questions Réponses; site de partage d'informations
7M	*Tchats*
4,031 G caractères	

Pour des questions pratiques, nous avons travaillé sur un corpus existant mais la constitution d'un tel corpus n'est pas une difficulté en soi, dès lors que l'on dispose d'un accès à l'internet. Les outils libres de collecte étant foison, la principale difficulté est de localiser les sources (journaux, blogs). Cela ne prend toutefois que quelques heures.

Le corpus contient des segments répétés susceptibles de fausser le dénombrage des occurrences de noms. Nous avons donc constitué un corpus « intermédiaire » pour éliminer les segments répétés. Nous avons procédé comme suit. Nous savons que les termes désignant en japonais les différents types de cancer sont composés du morphe *gan* (« cancer »). A l'aide de Sagace-v4.2 (Blin 2012) nous avons relevé tous les segments du corpus qui contient la chaîne de deux hiragana がん ou le sinogramme 癌. Il s'agit des deux transcriptions les plus probables du morphe *gan* « cancer ». Nous estimons que la probabilité que le morphe soit écrit en katakana (ガン) est extrêmement faible et que les risques d'erreurs d'analyse sont élevés. C'est la raison pour laquelle nous ne l'avons pas pris en compte. Nous avons limité la longueur des segments à 21 octets (soit 7 caractères japonais en UTF8) avant et 21 octets après l'occurrence du morphe. Puis nous avons éliminé toutes les répétitions de segments. Nous avons obtenu un corpus de 173 779 segments. A ce stade, il n'est pas garanti que l'occurrence de *gan* relevée dans un segment soit effectivement celle du morphe « cancer ». En particulier pour la première graphie, il peut très bien s'agir d'une partie de mot, comme la tête du verbe がんばる, *gan ba ru*, « faire des efforts », sans rapport avec la maladie. Il faudra procéder à une analyse morphologique.

2.2 Constitution du vocabulaire, liste des maladies

Les noms de cancer en japonais sont composés sur le schémas < nom commun d'organe + *gan* >. Par exemple *i-gan* (« estomac-cancer »). Pour dénombrer les noms de cancer, il nous suffit donc de compter les noms communs d'organes précédant *gan*. Nous avons le choix entre deux stratégies. La première consiste à ne chercher que les noms communs désignant un organe, ou même plus spécifiquement ceux que l'on sait susceptibles d'être touchés par la maladie. L'intérêt est que nous sommes sûrs de ne pas avoir de noms communs « parasites ». La contrepartie est qu'il faut disposer d'une bonne connaissance en anatomie pour établir la liste et d'une bonne expérience lexicale pour à la fois couvrir la terminologie spécialisée mais aussi

la terminologie commune. Il n'y a pas de garanti que l'ensemble du vocabulaire soit parfaitement couvert. L'autre stratégie est de collecter tous les noms communs qui apparaissent dans cette position, sans aucune restriction sémantique a priori. Plus le lexique est exhaustif, mieux c'est. L'intérêt est que nous pourrions voir émerger des noms auxquels nous ne nous attendions pas. La contrepartie est le risque de lister des noms indésirables, notamment à cause d'un mauvais découpage du texte. Malgré tout, partant de l'intuition que dans cette distribution les erreurs seraient en nombre limité, nous avons opté pour la deuxième stratégie, qui promettait des résultats plus riches.

Nous avons utilisé le lexique japonais libre au format Sagace, JLFS.1.2.2015-07-02_v150920[1]. Il contient 92 870 noms communs et 10 particules. Il n'est pas aisé de constituer un lexique d'une telle taille. Si nous n'en avions pas disposé, nous aurions opté pour la première stratégie et constitué manuellement un petit lexique des noms d'organes les plus susceptibles d'être affectés par la maladie. Heureusement, il existe en japonais de nombreux lexiques libres du même type que le JLFS et facilement formatables pour Sagace. On trouve par ailleurs de nombreuses listes de vocabulaire spécialisé sur l'Internet.

2.3 Relevé des occurrences

A l'aide du logiciel Sagace (Blin 2014), nous avons compté le nombre d'occurrences des noms communs dans la chaîne <nom commun + *gan*>, par exemple 胃癌, *i-gan*, litt. « estomac-cancer », « cancer de l'estomac », 肺臓癌, *haizou-gan*, « poumon-cancer ». Nous avons pris en compte les deux écritures (hiragana et sinogramme) du morphe *gan*. Le morphe fonctionne aussi bien comme un nom commun que comme un suffixe. Nous nous en sommes toutefois tenu à la position de suffixe car c'est celle qui est la moins susceptibe d'entraîner des erreurs d'analyse avec Sagace (mais aussi les autres analyseurs morphologiques). La description de la chaîne à dénombrer est présentée en annexe (1 et 2).

Nous avons testé l'autre schémas possible <nom commun + $no_{\text{particule de détermination}}$ + *gan* > (« cancer de NC ») mais le nombre de mots indésirables grandit tandis que le nombre d'occurrences diminue très sensiblement (seulement quatre occurrences dans le cas le plus fréquent relevé dans cette position, à savoir 血液, *ketsu eki*). Il n'y a donc pas d'intérêt à nos yeux de s'embarrasser ici de l'étude de ce schémas qui présente plus d'inconvénients que d'avantages. Cette remarque est néanmoins propre aux noms construits avec le morphe *gan* (« cancer »). Nous ne nous prononçons pas pour d'autres termes.

3 Résultats

Nous obtenons une liste de 40 noms communs désignant des organes ou parties du corps (table 2). Les noms étant classés par ordre décroissant, nous désignerons désormais cette liste par « classement corpus ». Les noms présents par erreur étaient en petit nombre et facilement repérables. Ils ont été éliminés à la main. C'est le cas par exemple de *mondai* (« problème »), *kyonen* (« l'année dernière »). La liste obtenue contient des synonymes, comme par exemple *zin* (腎, « rein ») et zin*zou* (腎臓, « rein-organe »). Cela confirme l'intérêt de travailler avec un grand lexique qui permet de récupérer tous les termes désignant un même organe et ne cantonne pas à un registre particulier de langue. Certains termes relevés sont liés par des relations partie-tout : comme *shôchô* (小腸, « intestin grêle ») et *naizô* (内臓, « entrailles »). Comme à l'accoutumé, nous obtenons une courbe de répartition des nombres d'occurrences conforme à la loi de Zipf.

Table 2 : Liste brute des noms communs d'organes relevés dans le corpus, suffixés de *gan* (« cancer »), et les occurrences.

乳 (sein) 339 ; 胃 (estomac) 293 ; 肺 (poumon) 240 ; 大腸 (gros intestin) 191 ; 胆管 (voies biliaires) 178 ; 食道 (tube digestif) 149 ; 前立腺 (prostate) 147 ; 肝臓 (foie) 82 ; 膵臓 (pancréas) 82 ; 肝 (foie) 80 ; 甲状腺 (thyroïde) 75 ; 卵巣 (ovaires) 60 ; 直腸 (rectum) 59 ; 膵 (pancréas) 54 ; 子宮 (utérus) 36 ; 喉頭 (larynx)

1 Accessible sur : https://sharedocs.huma-num.fr/wl/?id=S7rkT7cv3msgg5qKi3kt9UeeykcVYZ1s

32 ; 腎臓 (rein) 29 ; 結腸 (côlon) 28 ; 皮膚 (peau) 27 ;膀胱 (vessie) 18 ; 血液 (sang) 16 ; 舌 (langue) 15 ; 腎 (rein) 13 ; 全身 (corps entier) 13 ; 胆道 (voies biliaires) 12 ; 腺 (glande) 11 ; 咽頭 (pharynx) 4 ; 乳頭 (mamelon) 3 ; 十二指腸 (duodénum) 3 ; 副腎 (glande surrénale) 2 ; 小腸 (intestin grêle) 2 ; 膣 (vagin) 2 ; 頸 (col de l'utérus) 2 ; 肺臓 (poumon) 1 ; 度肝 (épatique) 1 ; 内臓 (entrailles) 1 ; 骨髄 (moelle osseuse) 1 ; 口腔 (bouche) 1 ; 唾液腺 (glande salivaire) 1 ; 下腹部 (bas ventre) 1 ;

4 Confrontation aux données épidémiologiques

Pour évaluer la corrélation entre l'ordre établi à partir du corpus et l'ordre de prévalence des cancers dans la réalité, nous comparons les résultats du corpus à ceux d'une étude épidémiologique (désormais EE) détaillée de la prévalence des cancers au Japon (Matsuda et al. 2014), disponible sur l'internet. Nous nous concentrerons plus particulièrement sur le nombre de personnes atteintes d'un cancer et non pas décédées de cette maladie. Nous exploitons les données de l'année 2011 qui correspond à la dernière année couverte par le corpus. L'EE distingue homme et femme. Une synthèse est disponible mais pour les seules maladies aux deux sexes (nous parlerons de maladie « genrées »). Elle n'inclut pas par exemple les cancers de la prostate et du sein. Nous lui avons préféré notre propre synthèse, qui inclu les données communes et genrées. Sur la base de l'EE, nous disposons donc d'un classement de prévalence des cancers chez les hommes (classement « EE-homme »), chez les femmes (« EE-femme ») et mixte (« EE-mixte »).

Pour effectuer une comparaison, nous avons procédé à une unification des noms du classement du corpus avec ceux des classements EE. Par exemple dans le corpus, le poumon est désigné par 肝 (*hai*, « poumon ») ou 肝臓 (*hai*, « poumon-organe »). Nous avons additionné les occurrences des deux termes et regroupé le tout sous le terme utilisé dans l'EE (肝, *hai*). Il en va de même pour 肝臓 et 肝. Nous avons englobé dans la catégorie plus large de l'EE des noms isolés du corpus : par exemple 胆管 devient la catégorie 胆嚢・胆管 de l'EE, quand bien même 胆嚢 n'apparaît pas dans le corpus. Nous avons aussi rassemblé des noms du corpus lorsqu'ils l'étaient dans l'EE. Par exemple : 腎・尿路. Dans l'EE-femmes, nous avons retiré les noms organes déjà comptabilisées dans un plus grand ensemble (子宮 26741 = (子宮体部 ;14763)+ (子宮頸部 ;11378)). Les occurrences de maladies dans l'EE n'obéissent pas à une courbes de type zipf, contrairement aux données linguistiques. Il faut donc s'attendre à de possible difficultés lors de la comparaison avec les résultats du corpus.

Nous avons tout d'abord comparé les classements sur la base de l'écart des fréquence des maladies. Nous n'avons travaillé pour cela que sur les 15 maladies communes aux quatre classements. Dans chaque classement, nous avons repéré la maladie d'occurrences la plus élevées occmax. A chaque maladie, nous avons attribué la valeur (occurence/occmax). Puis nous avons comparé la différences de « fréquence » entre items identiques, entre classements. Selon le résultat, le corpus est plus proche du corpus EE-mixte.

Tableau 3 : Différence moyenne de fréquences entre chaînes ((EE)h(omme), (EE)f(emme), (EE)m(ixte), c(orpus)).

Corpus	Différence moyenne	Variance	Ecart moyen
EE-h,EE-f	0.196707	0.067173	0.259178
EE-h,EE-hf	0.212708	0.078449	0.280087
EE-f,EE-hf	0.198081	0.052508	0.229146
EE-f,c	0.174629	0.021675	0.147226
EE-h,c	0.199686	0.019725	0.140444
EE-hf,c	**0.179698**	**0.013963**	**0.118164**

Nous avons quantifié la proximité des classements homme, femme et mixte de l'EE avec le classement du corpus. Pour cela, nous avons transformé chaque classement en une chaîne (string) en concaténant les noms dans leur ordre d'apparition dans le classement (EE-hommes, EE-femmes, EE-mixte, Corpus). Par exemple, la chaîne correspondant au classement du corpus est : <乳 胃 肺 大腸 胆嚢・胆管 肝臓 食道 前立腺 膵臓 甲状腺 卵巣 直腸 腎・尿路 子宮 喉頭 結腸 皮膚 膀胱 血液 舌 全身 胆道 腺 口腔・咽頭 乳頭 十二指腸 副腎 導管 小腸 腟 頸 度肝 内臓 唾液腺 骨髄>. Nous avons ensuite attribué à chaque nom un numéro unique à deux chiffres : 胃 est remplacé par 01, 胆嚢・胆管 est remplacé par 04. L'encodage des noms est le même pour toutes les chaînes. Le classement EE-mixte <胃 大腸 肺 結腸 ...> ainsi encodé commence par : <02 04 03 17...>. Enfin, nous avons comparé la proximité entre les quatre chaînes, à l'aide des calculs de distance de Levenshtein (Levenshteĭn 1965) et de Jaro-Winkler[2]. Les deux modes de mesure produisent des résultats très différents (tableau 4). Cela est peut-être dû, entre autres, au fait que les chaînes ne sont pas constituées des mêmes éléments[3]. Nous avons tenté une deuxième mesure en ne retenant que les noms qui sont communs aux quatre chaînes (Tableau 5). Il apparaît encore plusieurs différences entre les deux modes de mesures. Ces modes de mesure quantitative de la distance ne fournissent donc pas des indices fiables.

Tableau 4 : Distance entre chaînes ((EE)h(omme), (EE)f(emme), (EE)m(ixte), c(orpus)).

Distance de Levenshtein	Distance de Jaro-winkler
fm 24	hm .931
hm 25	fm .923
hf 26	hf .882
mc, hc **58**	fc .837
fc 61	**mc .821**

2 Winkler, W. E., « The state of record linkage and current research problems », Statistics of Income Division, Internal Revenue Service Publication R99/04, 1999. http://www.census.gov/srd/papers/pdf/rr99-04.pdf

3 Les mesures de proximités pourraient être encore plus incohérentes en présence de mots ex aequo dans les classements. En effet, la transformation d'un classement sous forme de chaîne linéaire oblige à imposer un ordre (arbitraire) aux éléments ex aequo.

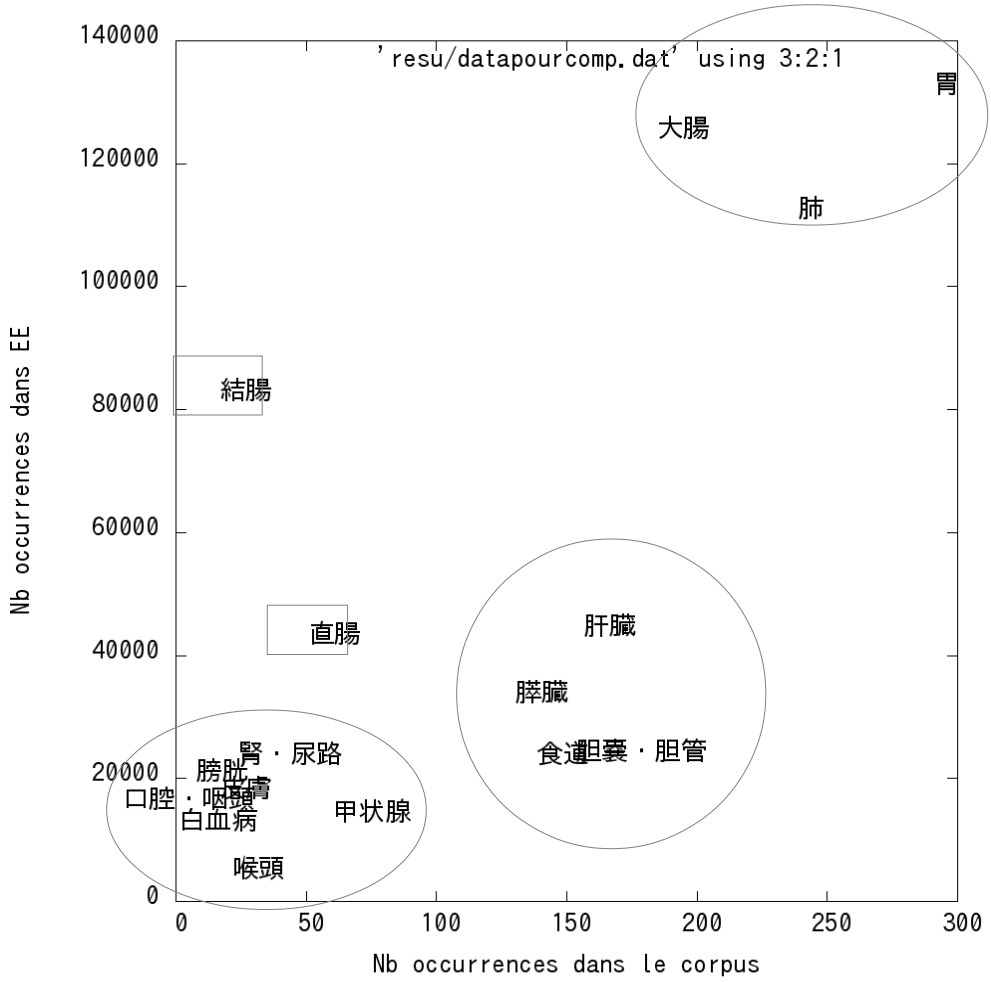
	hc .771

Tableau 5 : Distance entre chaînes (EE : Etudes Epidémiologiques, h(omme), f(emme), m(ixte), c(orpus mixte)), ne contenant que les termes communs aux quatre chaînes.

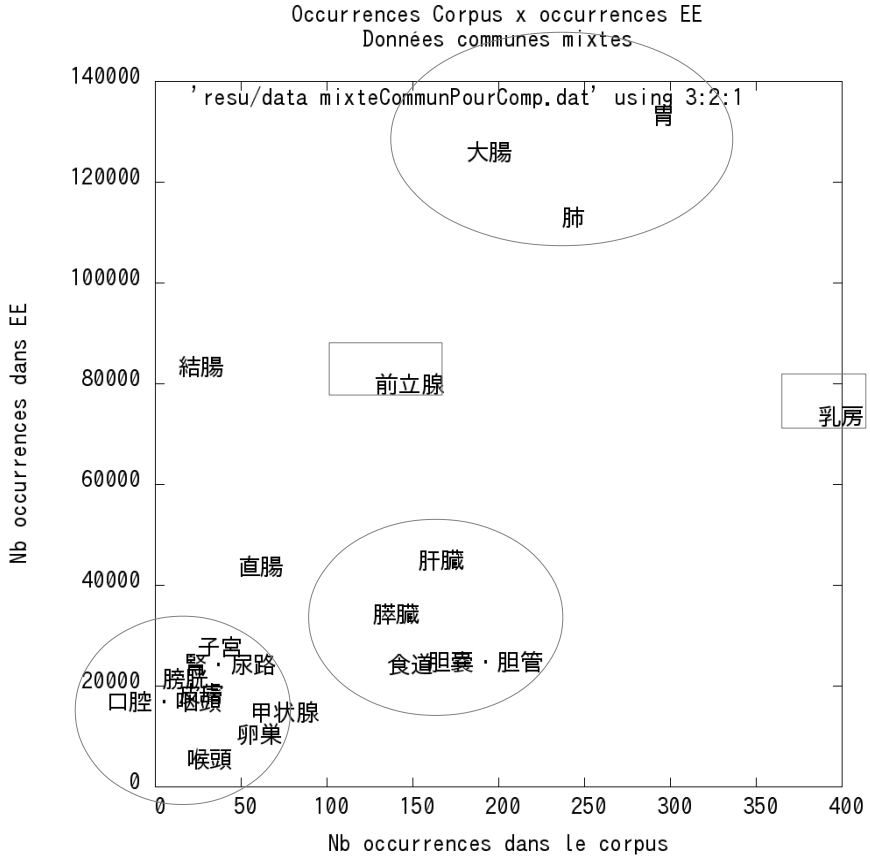
Distance de Levenshtein	Distance de Jaro-winkler
hm 12	hm .971
mc 14	fm .946
hc 16	hc .919
fm 19	**mc** .915
fc 21	hf .904
hf 22	fc .893

Nous avons procédé à une troisième et dernière analyse consistant à croiser les occurrences des noms dans le corpus et les occurrences dans l'EE-mixte, en ne prenant en compte que les noms communs aux quatre classements (Graphe 1). Nous voyons alors se dégager trois groupes de noms : les noms les plus fréquents (dans les deux classements), moyennement fréquents et le reste. On relève des noms isolés (結腸 et 直腸) témoignant de l'incohérences entre le classement du corpus et celui de l'EE. On pourrait donc parler d'une cohérence mais à une échelle très grossière. Les noms isolés sont plus nombreux encore si l'on élargit les classements aux noms genrés communs aux deux classements (corpus et EE-mixte) (Graphe 2). Ainsi, deux noms isolés supplémentaires apparaissent (prostate et sein). Pour autant, on ne peut pas dire que les noms genrés soient systématiquement isolés puisque deux d'entre eux (vagin et utérus) entrent dans un des groupes existant (celui des noms de plus petite occurrence).

Graphe1 : Croisement des occurrences dans le corpus et dans EE, limité aux noms présents dans les quatre classements.



Graphe2 : Croisement des occurrences dans le corpus et dans EE, limité aux noms présents dans le corpus et dans EE mixte.



5 Discussion

Le traitement du corpus (collecte et analyse) ne pose aucune difficulté technique. Nous avons utilisé des outils banals faciles à se procurer. Notre intérêt se porte donc ici sur la fidélité du classement du corpus par rapport aux classements des données épidémiologiques. En effet, le classement à base de corpus n'a d'intérêt que si il rend compte de la réalité. Malheureusement, les données obtenues dans cette étude ne permettent pas de mesurer quantitativement cette fidélité. Il apparaît seulement une cohérence grossière, qui laisse entièrement ouverte la question de l'intérêt de recourir au corpus pour évaluer la prévalence des maladies. Nous revenons sur les résultats.

Nous avons recourru à des mesures de proximité pour évaluer la distance entre les classements, à l'aide de deux modes de calculs (Levenshtein et Jaro-Winkler). Les deux modes de mesure de proximité produisent des résultats incohérents et n permettent pas de savoir avec quel classement (homme, femme ou mixte) des EE le classement du corpus est le plus proche. Par contre, la comparaison entre les fréquence laisse penser qu'il y a une moindre différence dans les fréquences des maladies pour les classements EE-mixte et corpus. Partant de l'a priori que le corpus est rédigé par et adressé à un public mixte, nous privilégions la comparaison du classement du corpus avec le classement mixte des EE.

Nous avons tenté de comparer les classements sous forme de graphiques, en croisant les occurrences des noms du classement EE-mixte avec les occurrences des noms du corpus. Avec les seules données communes à tous les classements (graphe 1) et donc excluant les noms « genrés », trois groupes de termes se détachent : les termes les plus fréquents dans les deux classements, un groupe intermédiaire et un groupe moins fréquent. Si l'on regroupe par paquets de trois noms ou plus, donc avec une granularité très grossière, on peut parler d'une cohérence entre classement du corpus et classement mixte des EE. La question est de savoir si à

l'usage, cette granularité est acceptable. A l'intérieur des groupes, le corpus ne rend pas fidèlement compte des réalités. Ainsi dans le plus petit groupe, correspondant aux termes/maladies les plus fréquent(s), si le premier terme (estomac) est le même pour le corpus et les EE, le corpus « se trompe » dans l'ordre des deux suivants (poumon et intestin). Il y a des erreurs plus remarquables encore sur les cas isolés. Ainsi, le cancer du côlon, à forte prévalence (proche du groupe des plus prévalents) est rétrogradé dans le corpus au rang des moins prévalents. Avec le corpus, il y a donc un risque de sous-estimation sur des cas non marginaux. L'intégration des maladies genrées fait apparaître deux nouvelles maladies isolées (cancer de la prostate et du sein). Toutefois, on voit que l'isolement n'est pas systématique pour les maladies genrées puisque le cancer de l'utérus est complètement intégré à un groupe existant (groupe de faible occurrence). Sein et prostate pourraient donc être des cas à part. Nous reviendrons dessus.

Les mesures de proximité ne permettent pas de dire si le corpus rend mieux compte des classements chez l'homme, la femme ou dans le groupe mixte. A défaut de pouvoir dire si il existe une préférence, relevons les éléments qui pourraient justifier une préférence. Il pourrait y avoir une préférence si le corpus était massivement rédigé par un des deux genres. Malheureusement, il est impossible de le dire pour le corpus dont nous disposons dans cette expérience. Celui-ci est composé pour une bonne part de textes journalistiques généralistes. A notre connaissance[4], le lectorat est complètement mixte et ne peut justifier une préférence pour un sexe. Il n'est pas possible de prédire la proportion homme/femme du coté de la rédaction. Le monde du travail au Japon est certes majoritairement structuré autour des hommes si l'on en juge par la domination en nombre des employés *statutaires* homme (double de celui des femmes en 2011[5]), mais la part d'interventions féminines n'est pas prédictibles à partir de tels chiffres puisque les femmes sont malgré tout présentes dans des emplois non statutaires. Une étude au cas par cas dans les entreprises serait nécessaire. D'autre part, il est impossible d'élaborer la moindre hypothèse sur la répartition par genre des textes « libres » du corpus (textes de questions réponses entre internautes). Il faudrait établir une statistique à partir de l'examen des signatures de chaque question/réponse puisque ceux-ci sont signés. Mais quand bien même la répartition des tâches serait connues, elle ne présumerait pas du contenu : un rédacteur peut s'intéresser à des thèmes associés au sexe opposé[6].

On observe aussi bien des sur-classements que des déclassements dans le corpus, par rapport aux EE. C'est le cas des deux maladies genrées, cancer de la prostate et du sein. Alors que les deux maladies sont à peu près aussi fréquentes avec une légère domination du sein (72 472 cancers du sein, 78 728 de la prostate), leur classement dans le corpus suit des mouvements contradictoires dans le corpus. Le cancer de la prostate est déclassé en huitième position tandis que celui du sein est propulsé à la première place (Tableau 7). Nous ne pouvons expliquer ce phénomène mais il nous servira de base de réflexion sur les possibles sources de parasitages dans le corpus. La fréquence d'évocation d'une maladie pourrait être corrélée à sa létalité. Un malade décédé ne peut plus évoquer sa maladie. Par conséquent, plus une maladie est létale, plus le nombre de locuteurs potentiels diminue. Cependant, les deux cancers en question ont une létalité comparable (femmes : 12 731 et hommes : 10 828 [7]). Il est même supérieur pour le sein et devrait donc entraîner un changement d'ordre inverse à celui observé, à savoir une rétrogradation du cancer du sein par rapport au classement des EE. Par ailleurs, le cancer du pancréas, qui est quasiment toujours fatal et dans un délais bref (quelques mois) n'est pas retrogradé alors qu'il laisse aux personnes touchées peu de temps pour s'exprimer. Une autre explication serait la taille des populations non seulement atteintes, mais aussi susceptibles de l'être. Par exemple, une femme est certainement plus sensible aux problèmes de cancer du sein qu'un homme. On pourrait donc s'attendre à ce que des femmes même non touchées évoquent la maladie plus fréquemment. Selon cette hypothèse, la supériorité du nombre d'occurrences du cancer du sein dans le corpus devrait s'expliquer par une population féminine supérieur en nombre. Mais les données démographiques invalident cette hypothèse puisqu'il n'y a pas de différence significative entre les deux populations (3 millions de

4 Nou ne disposons pas de statistiques.

5 D'après les statistiques gouvernementales : http://www.stat.go.jp/data/roudou/report/2014/dt/zuhyou/a00100d.xls (consulté en 2015).

6 Le cas extrême, hors champs médical, étant celui des sites pornographiques : on peut s'attendre à ce qu'un corpus de textes pornographiques fasse la part belle aux noms de parties du corps de la femme, quand bien même les rédacteurs sont certainement, en majorité, des hommes.

7 Voir annexe.

femmes en plus sur un total de 120 millions de Japonais) [8]. A moins que 2,5 % de la population soit capable de faire basculer la tendance. Cela serait plausible si l'accès au web était plus fort chez les femmes, ce qui n'est pas le cas à notre connaissance[9]. Une autre explication serait à chercher du coté des activités associées à la maladie, qui permettraient de multiplier des mentions à la maladie. Justement, le cancer du sein est l'un de ces rares cas, peut être le seul. Il est associé à la chirurgie réparatrice et plastique et souvent mentionné dans les textes relatifs à ces disciplines. Cela pourrait justifier le surclassement du cancer du sein dans le corpus. Mais cela ne justifierait pas de rétrograder le cancer de la prostate derrière celui du foie, comme on peut le constater. Tous deux semblent aussi pauvres en thématiques associées. On ne peut alors exclure une part de tabou : certaines maladies pourraient être évoquées plus difficilement que d'autres. Le cancer de la prostate s'accompagne souvent d'une gêne de l'activité sexuelle masculine. Il touche donc aux signes de virilité et pourrait de ce fait être plus souvent caché. Au contraire, le cancer du foie pourrait être moins tabou car indirectement associé à une image possitive. En effet, le cancer du foie est souvent conséquence d'une forte consommation d'alcool. Or l'image de l'alcool est plutôt valorisée dans la société japonaise (associée à la sociabilité, l'endurance, voire à la virilité). Le cancer du foie pourrait être interprété comme une conséquence d'une activité globalement « positive ». Cela justifierait qu'il figure devant le cancer de la prostate dans le corpus.

6 Conclusion

Nous avons proposé un dispositif facile à mettre en oeuvre pour classer les noms de maladie en fonction de leur fréquence dans un corpus. Ce dispositif se distingue de tous les dispositifs existants de suivis épidémiologiques, qui font intervenir de nombreux acteurs et sont en général organisés autour d'un système centralisateur. Nous avons comparé le classement ainsi obtenu pour les cancers sur le web japonais[10] avec une étude épidémiologique. Il s'avère que le classement à partir du corpus ne rend compte que grossièrement des tendances réelles. De nombreux décalages apparaissent. Il est plus particulièrement gênant de constater des décalages sur des maladies non marginales. Nous n'avons pas pu fournir d'explications solides pour justifier ces décalages. Les seules hypothèses que nous avons avancées relèvent de la psychosociologie et semblent difficiles à transcrire de façon objective sous forme de pondérations quantitatives. En conséquence, dans un région où il existe des études épidémiologiques fiables, les corpus ne peuvent être utilisés que comme appoint. Ils peuvent en effet renseigner sur des maladies absentes des statistiques parce qu'elles sont moins fréquentes ou même ignorées par les institutions. Le corpus peut par contre servir en première approximation en l'absence de toute donnée épidémiologique.

8 Hommes : 62,184M, Femmes : 65,615M ; Ministère des Affaires intérieures et des Communications Bureau de la statistique http://www.stat.go.jp/data/jinsui/2013np/img/05k25-1.gif .

9 Nous ne disposons pas de données chiffrées.

10 Les ressources, outils et un script complet sont disponibles « prêt à l'emploi » pour le japonais et le français. Il ne reste qu'à ajouter les corpus. Du texte html brut peut suffit.

7 Bibliographie

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

———. 2014. “Comparing Two Analyzers of Japanese Corpora for Helping Linguists: MeCab and Sagace (Comparaison de Deux Outils D’analyse de Corpus Japonais Pour L’aide Au Linguiste, Sagace et Mecab) [in French].” In *Proceedings of TALN 2014 (Volume 2: Short Papers)*, 491–98. Marseille, France: Association pour le Traitement Automatique des Langues. http://www.aclweb.org/anthology/F14-2018.

———. 2015. “Corefjp-0.003.150528, (Another) Corpus for Written Contemporary Japanese.” http://goo.gl/p0Tx7h.

Goulinet, Géraldine. 2014. “Rôle Socio-Culturel Des Communautés Virtuelles de Patients Dans Le Suivi Des Maladies Chroniques - Vers Un Nouveau Modèle D’éducation Thérapeutique ?” http://www.adjectif.net/spip/spip.php?article282.

Levenshteĭn, V. I. 1965. “Binary codes for correction of deletions and insertions of symbol 1.” *Problemy Peredachi Informatsii* 1 (1): 12–25.

Matsuda, Ayako, Tomohiro Matsuda, Akiko Shibata, Kota Katanoda, Tomotaka Sobue, Hiroshi Nishimoto, and Japan Cancer Surveillance Research Group. 2014. “Cancer Incidence and Incidence Rates in Japan in 2008: A Study of 25 Population-Based Cancer Registries for the Monitoring of Cancer Incidence in Japan (MCIJ) Project.” *Japanese Journal of Clinical Oncology* 44 (4): 388–96. doi:10.1093/jjco/hyu003.

Pellerin, Cheryl. 2008. “Internet : Le Nouvel Outil de Surveillance Des Maladies Infectieuses Émergentes | IIP Digital.” http://iipdigital.usembassy.gov/st/french/article/2008/04/20080404170842lcnirellep0.3346826.html#axzz3mM9XRPKI.

8 Annexes

1) Script de requête

Requête pour construire le corpus :

```
Travail:exemples
...
Arret:"。"
...
InsereDebSegment:"<deb/>"
Convertit les caracteres latins en 1 octet
...
Sauvegarde au format KWIC sans répétition avant=21 apres=21

Reqet:
>0 cat:癌|がん
```

Requête pour l'extraction de noms des maladies (cancer) :

```
Travail:stat
...
Arret:"。"
...
Stat ordonnee:decroissant
InsereDebSegment:"<deb/>"
Convertit les caracteres latins en 1 octet

Reqet:
>0 cat:particule  | ponctuation | debsegment | markDeb | chiffre
=0 cat:nomCommun | kanji /-affich:trait:reflem /-affich:" " /-affich:trait:lemme /-compte
=0 の
=0 cat:癌|がん
=0 cat:(particule & -の) | ponctuation | copule
```

1) Mortalité des cancers

Tableau 6 : Mortalité par cancer, Japon, 2011[11]

肺 70293, 胃 49830, 大腸 45744, 肝臓 31875, 結腸 31050, 膵臓 28829, 胆嚢・胆管 18186, 直腸 14694,

11 http://ganjoho.jp/data/reg_stat/statistics/dl/cancer_mortality(1958-2013).xls

乳房 12731, 食道 11970, 子宮 10846, 前立腺 10823, 悪性リンパ腫 10390, 白血病 8156, 膀胱 7008, 口腔・咽頭 6888, 卵巣 4705, 多発性骨髄腫 4066, 子宮頸部 2737, 脳・中枢神経系 2126, 子宮体部 2034, 甲状腺 1637, 皮膚 1453, 喉頭 954

2) Chaînes ordonnées après unification des termes du corpus

EE, hommes <胃 前立腺 肺 大腸 結腸 肝臓 直腸 食道 膵臓 腎・尿路 膀胱 悪性リンパ腫 胆嚢・胆管 口腔・咽頭 皮膚 白血病 喉頭 甲状腺 多発性骨髄腫 脳・中枢神経系 >

EE, femmes <乳房 大腸 胃 結腸 肺 子宮 膵臓 子宮体部 肝臓 直腸 子宮頸部 胆嚢・胆管 悪性リンパ腫 甲状腺 卵巣 皮膚 腎・尿路 白血病 膀胱 口腔・咽頭 多発性骨髄腫 食道 脳・中枢神経系 喉頭 >

EE, mixte <胃 大腸 肺 結腸 前立腺 乳房 直腸 肝臓 直腸 膵臓 膀胱 子宮 食道 悪性リンパ腫 胆嚢・胆管 食道 腎・尿路 皮膚 膀胱 皮膚 口腔・咽頭 甲状腺 白血病 卵巣 多発性骨髄腫 脳・中枢神経系 喉頭 >

Corpus : <乳房 胃 肺 大腸 胆嚢・胆管 肝臓 前立腺 食道 膵臓 甲状腺 直腸 卵巣 腎・尿路 子宮 喉頭 皮膚 結腸 膀胱 白血病 口腔・咽頭 >

3) Détail des quatre classements

Tableau 7: alignement des classements

EE-hommes	EE-femmes	EE-mixte	Corpus
胃	乳房	胃	乳房
前立腺	大腸	大腸	胃
肺	胃	肺	肺
大腸	結腸	結腸	大腸
結腸	肺	前立腺	胆嚢・胆管
肝臓	子宮	乳房	肝臓
直腸	膵臓	肝臓	前立腺 / 食道
食道	子宮体部	直腸	膵臓
膵臓	肝臓	膵臓	甲状腺
腎・尿路	直腸	子宮	直腸 / 卵巣
膀胱	子宮頸部	悪性リンパ腫	腎・尿路
悪性リンパ腫 胆嚢・胆管	胆嚢・胆管	胆嚢・胆管	子宮
口腔・咽頭	悪性リンパ腫	食道	喉頭
皮膚	甲状腺	腎・尿路	皮膚
白血病	卵巣	膀胱	結腸
喉頭	皮膚	皮膚	膀胱
甲状腺 多発性骨髄腫 脳・中枢神経系	腎・尿路	口腔・咽頭	白血病
	白血病	甲状腺	口腔・咽頭
	膀胱	白血病	舌
	口腔・咽頭	卵巣	全身
	多発性骨髄腫	多発性骨髄腫	腺
	食道	脳・中枢神経系	口腔・咽頭
	脳・中枢神経系	喉頭	十二指腸 [occ:3]

	喉頭		導管 [occ:2] 副腎 [occ:2] 小腸 [occ:2] 頸 [occ:2] 度肝 [occ:1] 内臓 [occ:1] 骨髄 [occ:1] 唾液腺 [occ:1] 下腹部 [occ:1] 腟 [occ:1]