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

Can the prevalence of a disease be predicted by using the frequency of its name on the WEB ? The case of cancer in Japan

2016/06/08
R.Blin
CNRS-CRLAO

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-clés : corpus, prévalence des maladies, japonais, cancer,

Résumé : We wondered if there is a correlation between the frequency of the names of diseases on the web, and the prevalence of these diseases within the population who use this web. As an experiment, we compared the number of occurrences of the names of cancers in a Japanese corpus from the web, and epidemiological data in Japan. We founded a rough correlation between the both. In this paper, we discuss the bias in the corpus.

Mots-clés : corpus, prevalence, Japan, oncology

1 Introduction

Connaître l'ordre de prévalence des maladies est important pour les acteurs de la santé et pour les populations. Mais établir un tel ordre suppose l'existence d'un dispositif d'observation et de traitement des données à grande échelle. Il en existe mais ils sont lourds (réseaux de contributeurs, volontaires ou désignés), très centralisés (pilotés par des états ou des entreprises (Goulinet 2014). Les informations sont limitées à celles fournies par les contributeurs. La qualité dépend de leurs compétences et de leur assiduité. 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 de disposer d'un droit de consultation.

L'Internet pourrait permettre de contourner ces difficultés. 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 en ligne, même les pays les plus démunis. Nous nous sommes donc demandé si cela ne pouvait pas constituer une ressource pour établir un ordre de prévalence à moindre frais et par tout un chacun. Plusieurs travaux laissent l'espérer. Ceux-ci (voir par exemple White et al. (2016) qui en cite d'autres) ont compté les occurrences de noms de maladies dans les requêtes sur les moteurs de recherches. Ils ont montré la corrélation entre nombre d'occurence des noms et l'évolution d'une épidémie ou encore fait apparaître les effets secondaires de médicaments, qui n'avaient pas encore été détectés par les moyens traditionnels de surveillance. Le résultat est encourageant mais la méthode suppose une fois encore une centralisation des données. En effet, la collecte suppose de disposer d'un moteur de recherche d'une audience très importante. Or il n'en existe qu'une poignée, appartenant à des groupes privés. La diffusion des informations en leur possession reste à la discrétion de ces groupes qui ne sont pas engagés à fournir l'ensemble de leurs données. Les moyens techniques mis en oeuvre ne sont pas non plus à la portée de tout un chacun. Ne peut-on concevoir un dispositif léger, permettant à tout un chacun de constituer ses propres ressources et de les analyser ?

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'actualités, d'analyser le nombre d'occurences des noms de maladies et d'en déduire un ordre de prévalence « virtuel ». Le dispositif repose sur des logiciels libres dont l'accès et la mise en oeuvre sont faciles. La principale interrogation est la fiabilité du classement obtenu par cette procédure. A titre expérimental, nous avons utilisé ce dispositif pour analyser un corpus réel et nous avons comparé les résultats à ceux d'études épidémiologiques « classiques ». Par opportunisme, notre intérêt s'est porté sur le cancer, le web et les études épidémiologiques japonaises. Mais la démarche et les outils sont adaptables à toute autre langue. Nous présentons ici l'expérience et les résultats.

2 Expérience

2.2 Ressources

Nous avons exploité un corpus de textes bruts non balisés et non segmentés, en japonais[1]. Il est composé de journaux, livres blancs gouvernementaux, minutes du parlement et sites 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. Lorsqu'un segment (correspond ici à une phrase) est répété, seule une occurrence de ce segment est conservée.

Nous avons ensuite constitué un lexique. Les noms de cancer en japonais sont composés sur le schéma < 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 d'organes précédant *gan*. Nous avons le choix entre deux stratégies. La première consiste à se doter d'un petit lexique spécialisé et à ne chercher que les noms désignant un organe, ou même plus spécifiquement ceux que l'on sait susceptibles d'être touchés par la maladie. L'intérêt de cette approche est que nous sommes sûrs de ne pas avoir de noms « 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. 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

[1] Blin, Raoul. 2015. “Corefjp-0.003.150528, (Another) Corpus for Written Contemporary Japanese.” http://goo.gl/p0Tx7h.

seraient en nombre limité, nous avons opté pour la deuxième stratégie, qui promettait des résultats plus riches. Nous avons utilisé le lexique existant JaLexGram (Blin 2015).

2.3 Résultats

Le logiciel Sagace (Blin 2014) a effectué le comptage et produit une liste de 40 noms communs désignant des organes ou parties du corps (table 1). Nous désignerons par « classement corpus » cette liste ordonnée par ordre décroissant. Les noms présents par erreur (ex. *mondai* « problème », *kyonen* « l'année dernière ») étaient en petit nombre et facilement repérables. Ils ont été éliminés. La liste contient des synonymes (par ex. *zin* « rein » et zin*zou* « rein-organe »). Leurs résultats ont été regroupés dans des rubriques chapeau. Certains termes relevés sont liés par des relations partie-tout : comme *côlon* et *gros intestin*. Cette relation est bien prise en compte par les travaux scientifiques : le nombre d'occurrences du tout égale les nombres d'occurrences des parties. Par contre, dans le corpus, la relation entre parties et tout est inconnue. Concrètement, le tout ne vaut pas la somme des parties. Nous avons donc préféré étudier séparément les résultats pour les parties et le tout. Comme à l'accoutumé, nous obtenons une courbe de répartition des nombres d'occurrences conforme à la loi de Zipf.

Table 1 : 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 (côlon) 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) 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 ;

2.4 Confrontation aux données épidémiologiques

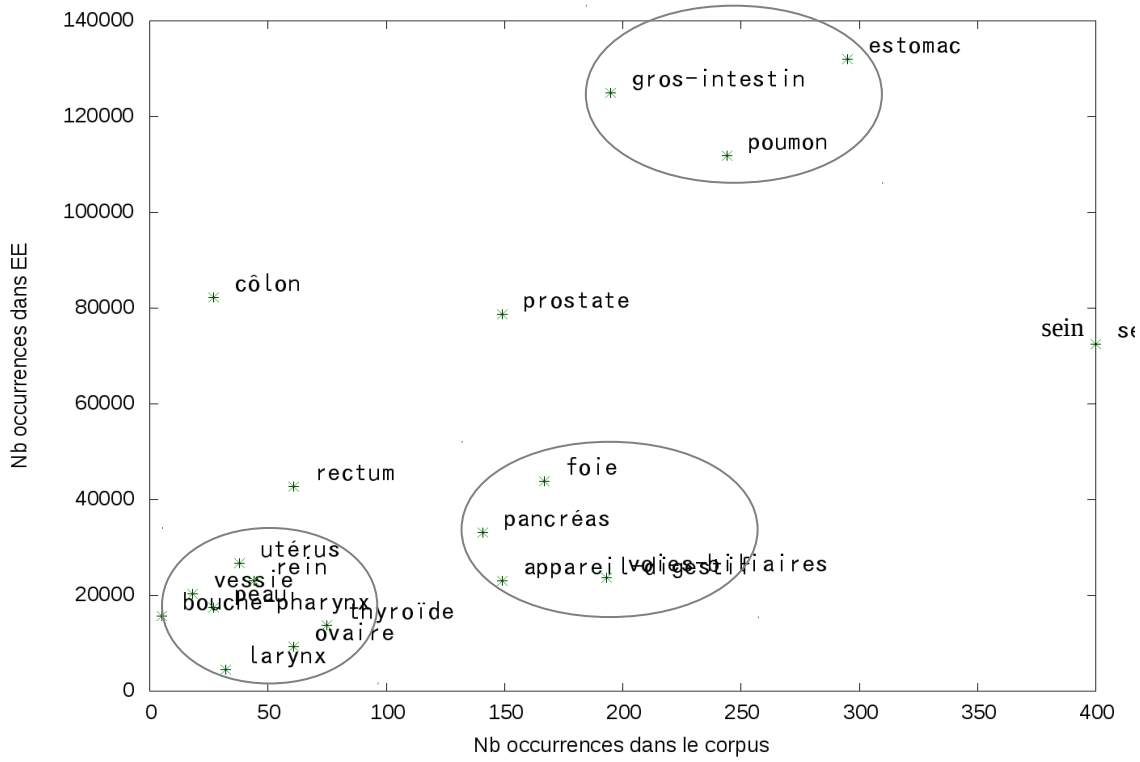
Pour évaluer la corrélation entre le classement corpus et l'ordre de prévalence des cancers dans la réalité, nous comparons le classement corpus aux classements fournis par une étude épidémiologique (désormais EE) des cancers au Japon (Matsuda et al. 2014). Nous nous concentrons sur le nombre de personnes atteintes d'un cancer et non pas sur celles 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 hommes et femmes et propose une synthèse pour les seules maladies communes aux deux sexes. Nous lui avons préféré notre propre synthèse, qui inclu les données communes et celles spécifiques aux hommes et aux femmes (nous les qualifierons de « genrées »). Les données du corpus et de l'EE ont été traitées de sorte à obtenir les catégories identiques dans l'EE et le corpus.

Nous avons tout d'abord évalué la corrélation entre classements sur la base de l'écart entre chaque maladie et la maladie de fréquence plus élevée (Tableau 2). La procédure est la suivante. Dans chaque classement, nous avons repéré la maladie dont le nombre d'occurrences est le plus élevé (occmax) puis à chaque maladie, nous avons attribué la valeur : occurrence/occmax. Du seul point de vue de la moyenne, le classement corpus serait plus proche des classements EE-femme et mixte. Par contre, l'écart type penche en faveur d'une plus grande proximité avec EE-mixte. De ce fait, et pour pouvoir discuter aussi des maladies genrées, nous nous sommes concentré sur la comparaison avec le classement EE-mixte. Nous avons tenté de comparer les classements bruts (indépendamment de la fréquence) avec différentes méthodes mais les résultats obtenus étaient incohérents d'une méthode à l'autre et les résultats inexploitables. Nous ne les présentons pas ici. Calculé sur la base des occurrences, l'indice de corrélation entre le classement EE-mixte et le classement corpus vaut 0,64.

Tableau 2 : Comparaison des fréquences entre classements (EE-h(omme), EE-f(emme), EE-h(ommes et)f(emmes), c(orpus)).

Classements	Différence moyenne	Variance	Ecart type
EE-h,EE-f	0.1967	0.0671	0.2591
EE-h,EE-hf	0.2127	0.0784	0.2800
EE-f,EE-hf	0.1980	0.0525	0.2291
EE-f,c	0.1746	0.0216	0.1472
EE-h,c	0.1996	0.0197	0.1404
EE-hf,c	0.1796	0.0139	0.1181

Graphe 1 : Croisement des occurrences dans le corpus et dans EE-mixte



3 Discussion

Le traitement du corpus et l'analyse des données ne posent aucune difficulté technique. Nous avons utilisé des outils et ressources faciles à se procurer ou concevoir. Notre intérêt se porte donc dans cette section sur la fidélité du classement corpus par rapport au classement épidémiologique. En effet, le classement à base de corpus n'a d'intérêt que si il rend compte de la réalité.

L'indice de corrélation entre les données du corpus et les données EE-mixte est honorable et montre que statistiquement, il existe une corrélation. Le classement corpus est proche des classements mixte et féminin. Cette proximité n'est pas étonnante car le corpus n'est pas marqué par le genre. Au Japon, l'accès au Web est égal pour les deux sexes, ce qui fait un lectorat mixte. Il n'est pas possible de prédire la proportion homme/femme du coté de la rédaction mais rien ne permet de dire qu'il est marqué pour le genre. Une proportion

importante du corpus est constituée de textes journalistiques. 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[2]), mais la part d'interventions féminines n'est pas prédictible 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 rédactions serait nécessaire. Il est impossible d'avancer 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 connue, elle ne présumerait pas du contenu : un rédacteur peut s'intéresser à des thèmes associés au sexe opposé[3].

La corrélation, même si elle est statistiquement bonne, doit cependant être relativisée. On peut parler d'une corrélation mais avec une granularité grossière. Si l'on excepte quatre cas isolés, le graphique de dispersion fait apparaître trois groupes de noms : un premier est constitué des noms de bas de classement, un second regroupe les noms de milieu et un troisième groupe les noms de haut de classement. Plus on monte dans le classement, plus le nombre de composant de ces groupes diminue, de même que l'écart augment entre les composants.

L'observation détaillée du graphique, notamment à l'intérieur des trois groupes, montre de nombreuses incohérences entre le classement corpus et le classement EE. Le classement corpus surclasse tout aussi bien qu'il déclasse par rapport au second. Les différences sont parfois significatives et concerne des maladies non minoritaires. 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), le cancer de la prostate est déclassé en huitième position tandis que celui du sein est propulsé à la première place. Les incohérences pour les maladies genrées sont de bons phénomènes pour réfléchir aux possibles sources de parasitages dans le corpus.

Nous voyons au moins quatre sources possibles de parasitages. Tout d'abord, 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 données corpus ne vont pas dans ce sens. Les deux cancers genrés ont une létalité comparable, voire supérieure pour le sein (femmes : 12 731 et hommes : 10 828 [4]). Cancers du sein et de la prostate devraient donc être évoqués un même nombre de fois, voire moins souvent pour le sein. Or c'est le contraire qui est observé. Par ailleurs, le cancer du pancréas, qui est quasiment toujours fatal et dans un délai bref (quelques mois) n'est pas retrogradé dans le corpus par rapport au classement EE, alors qu'il laisse aux personnes touchées peu de temps pour s'exprimer.

Une autre explication aux incohérences entre les deux classements serait la taille des populations non seulement atteintes, mais aussi susceptibles de l'être. Selon cette hypothèse, la supériorité du nombre d'occurrences du cancer du sein dans le corpus devrait s'expliquer par une population féminine supérieure 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 femmes en plus sur un total de 120 millions de Japonais) [5]. 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.

Une troisième 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

---

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

[3]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.

[4]Voir annexe.

[5]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 .

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. Les deux semblent aussi pauvres en thématiques associées.

Une dernière explication tiendrait à l'image des maladies : certaines maladies pourraient être évoquées plus difficilement que d'autres. Spéculons. 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 positive. 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.

4 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. La légèreté du dispositif le distingue de tous les lourds 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[6] 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 psycho-sociologie et semblent difficiles à transcrire de façon objective sous forme de pondérations quantitatives. En conséquence, dans une 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.

5 Bibliographie

Blin, Raoul. 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. “Metadonnées Du Lexique-Grammaire Du Japonais jalexGram-0.010.” http://goo.gl/3BvzGU.

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.

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.

[6]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.

White, Ryen W., Sheng Wang, Apurv Pant, Rave Harpaz, Pushpraj Shukla, Walter Sun, William DuMouchel, and Eric Horvitz. 2016. “Early Identification of Adverse Drug Reactions from Search Log Data.” *Journal of Biomedical Informatics* 59 (February): 42–48. doi:10.1016/j.jbi.2015.11.005.