Relever des critères pour la distinction automatique entre les documents médicaux scientifiques et vulgarisés en russe et en japonais

Sonia Krivine[1], Masaru Tomimitsu[1,2], Natalia Grabar[3,4,5], Monique Slodzian[1]

[1]INaLCO, CRIM
[2]Université de Nantes, LINA, FRE CNRS 2729
[3]Université René Descartes - Paris 5, Faculté de Médecine
[4]Inserm, U729
[5]SPIM

Résumé

Dans cet article, nous cherchons à affiner la notion de comparabilité des corpus. Nous étudions en particulier la distinction entre les documents scientifiques et vulgarisés dans le domaine médical. Nous supposons que cette distinction peut apporter des informations importantes, par exemple en recherche d'information. Nous supposons par là même que les documents, étant le reflet de leur contexte de production, fournissent des critères nécessaires à cette distinction. Nous étudions plusieurs critères linguistiques, typographiques, lexicaux et autres pour la caractérisation des documents médicaux scientifiques et vulgarisés. Les résultats présentés sont acquis sur les données en russe et en japonais. Certains des critères étudiés s'avèrent effectivement pertinents. Nous faisons également quelques réflexions et propositions quant à la distinction des catégories scientifique et vulgarisée et aux questionnements théoriques.

Mots-clés : recherche d'information translangue, corpus comparables, typologie de documents, catégorisation, document scientifique, document vulgarisé.

Abstract

In this paper, we aim to ripen the notion of corpora comparability. We study especially the distinction between scientific and popularized documents in the medical domain. We suppose that this distinction can give important informations, for instance in information retrieval. In the same time, we suppose that documents reflect the context of their production and provide features necessary for this distinction. We study and present several features, linguistic, typographic, lexical and other, for the characterization of medical documents as scientific or popularized. The results presented are acquired on data in Russian and Japanese. Some of analyzed features turn out to be relevant. We give then some remarks and suggestions as for the distinction of scientific and popularized documents and their theoretical issues.

Keywords: translingual information retrieval, comparable corpora, document typology, categorisation, scientific document, popularized document.

1. Introduction

Avec la recherche d'information translangue, un utilisateur peut formuler sa requête dans une langue et recevoir des réponses qui se présentent éventuellement en d'autres langues qu'il maîtrise ou qu'il peut faire traiter. Dans ce cadre, la recherche d'information fait abstraction

de la langue de l'utilisateur et permet d'assurer ainsi une meilleure couverture des informations. Elle repose souvent sur des corpus parallèles ou comparables. Il existe ainsi des méthodes pour collecter des corpus parallèles sur le Web, par exemple celle de (Resnik, 1999). Mais ces corpus deviennent plus difficiles à collecter lorsque l'on travaille dans un domaine de spécialité ou avec des langues « rares ». On se tourne alors vers des corpus dits « comparables », mais la notion de comparabilité des corpus reste vague. De manière générale, on considère que les documents doivent être similaires, souvent de par leur thématique (Sinclair, 1994), celle-ci étant quantifiable et qualifiable à travers le lexique (Déjean & Gaussier, 2002) ou à travers les cognats, dates et noms propres (Matsumoto & Tanaka, 2002). Dans notre travail, nous faisons l'hypothèse que le niveau lexical, tout en permettant d'accéder à la thématique des documents, n'est pas suffisant pour prendre en compte leurs dimensions culturelle et sociolinguistique, qui exercent également une influence sur la comparabilité des documents. Nous nous intéressons donc aux principes et aux critères de comparabilité des documents qui permettraient un meilleur accès aux informations multilingues et multiculturelles du Web. Notre travail s'inscrit dans le cadre du projet DECO[1]. Nous avons choisi de travailler avec des documents du domaine médical concentrés autour de la thématique *diabète et alimentation* et de proposer des critères de distinction entre les documents à visée scientifique et de vulgarisation. Cette distinction nous paraît pertinente dans le domaine médical, car elle peut fournir par exemple des indications sur le degré de fiabilité de l'information. On accordera ainsi plus de confiance aux informations trouvées dans les documents à visée scientifique. C'est d'ailleurs une des activités des portails médicaux, comme CISMeF[2] ou HON[3]. CISMeF type ainsi les documents en supports de cours, guides de bonnes pratiques, articles, etc. Tandis que HON propose un référencement des pages en tenant compte d'un indice sur la fiabilité de l'information qu'elles contiennent, se basant pour ceci sur les outils de TAL (Gaudinat & Boyer, 2003). Par ailleurs, la réorganisation des résultats par catégories (Karlgren, 2000) permet à l'utilisateur d'accéder plus rapidement aux documents qui lui semblent les plus pertinents pour sa recherche. Un scientifique manifestera peut-être plus d'intérêt pour les travaux de recherche et un visiteur occasionnel se contentera des recommandations générales susceptibles de paraître sur un site grand public.

L'objectif de notre travail consiste à définir des critères de caractérisation de la comparabilité des corpus. Pour ce faire, nous supposons que les textes sont le reflet de l'environnement sociolinguistique de leur production. Pour une même thématique, ils peuvent donc offrir des variations de genres, discours, supports médiatiques ou autre. D'autres travaux se sont déjà intéressés à distinguer automatiquement des types de documents. Les moyens et critères mis en oeuvre pour cette distinction dépendent des objectifs visés et de la nature des documents traités. Ainsi (Biber, 1988) adopte une démarche inductive consistant à faire émerger, grâce à un traitement statistique multidimensionnel, des traits linguistiques (ou plutôt des configurations de traits linguistiques) permettant de regrouper les textes en différents « types » et se distinguant essentiellement par leur caractère oral ou écrit. Les 67 traits linguistiques sur lesquels s'appuie ce travail sont identifiés automatiquement et mettent en jeu les marqueurs de temps et d'aspect, les pronoms, les questions, les passifs, la coordination, etc. S'inspirant de ce travail, (Habert *et al.*, 2000) effectuent également un typage de textes en cherchant à faire émerger les spécificités non lexicales des discours radio-télévisés de Mitterrand et De Gaulle.

D'autres travaux cherchent à faire émerger les genres des documents. (Kwasnik *et al.*, 2000)

[1] Le projet DECO (*Découverte et exploitation des corpus comparables pour l'accès à l'information*), débuté en 2004, s'inscrit dans le programme CNRS TCAN. Il est piloté par Béatrice Daille, LINA/Nantes.

[2] Catalogue et index des sites médicaux francophones, à l'adresse http://www.chu-rouen.fr/cismef/

[3] Health on the Net, à l'adresse http//www.hon.ch/Project/Intro_projects_f.html

envisagent ainsi l'identification automatique des genres du Web en se basant sur la longueur des mots ou les URL (tilde ~ dans une URL indique potentiellement une page personnelle). Également inspirés des travaux de (Biber, 1988), les expériences menées par (Karlgren, 2000) mettent au point une liste de variables linguistiques intuitives pour la catégorisation des genres textuels : fréquences relatives de mots ou catégories, longueur des mots, des phrases, etc. Testées sur le corpus Brown pré-catégorisé en genres[4], ces variables produisent de bons résultats pour les catégories générales (Fiction, Informative text), mais donnent des résultats mitigés pour les sous-catégories (Science-Fiction, Mystery, Romance). (Kessler *et al.*, 1997) envisagent également une détection automatique des genres en introduisant la notion de « facette générique ». Les résultats des expériences menées sur le corpus Brown sont également mitigés : Fiction, Reportage et Scientifique et technique sont bien reconnus, mais pas les textes juridiques. Notons que la notion de genre reste, dans ces travaux, mal définie.

Et un dernier travail que nous citons se place au niveau des discours. Il cherche à détecter des documents racistes et xénophobes sur le Web (Valette, 2004). L'approche théorique s'inspire des propositions de la sémantique interprétative de (Rastier *et al.*, 1994) en dégageant des critères sémantiques à plusieurs niveaux d'analyse du texte.

Dans la suite de cet article, nous présentons les méthodes d'acquisition des critères requis (section 2) et notre matériel (section 3). Nous présentons et discutons ensuite les résultats obtenus (section 4). Le travail est fait sur trois langues : français, russe et japonais. Nous présentons essentiellement les résultats pour le russe et mettons le japonais en parallèle lorsque c'est possible. Nous terminons l'article avec des conclusions et perspectives (section 5).

2. Démarche d'identification des critères

L'objectif que nous nous fixons dans ce travail consiste à définir les critères distinctifs entre les documents scientifiques et vulgarisés. Nous nous inspirons des travaux précédents en typologie de documents. Nous prenons aussi en compte la remarque de (Malrieu & Rastier, 2001) : "*aucune typologie des textes fondée sur des critères définis indépendamment des genres (comme oral vs écrit, public vs privé, etc.) n'a permis d'isoler des genres*", que nous généralisons de manière suivante : puisque nous voulons aboutir à des critères discriminants entre les documents scientifiques et vulgarisés, nous devons partir des données qui illustrent cette distinction. Donc, à partir des documents distribués en deux catégories, scientifique et vulgarisée, nous cherchons à faire émerger les critères qui permettraient ensuite d'effectuer cette distinction de manière automatique, avec des algorithmes de catégorisation ou d'apprentissage.

La constitution des corpus scientifiques et vulgarisés est décrite dans la section suivante. Les critères distinctifs entre les documents scientifiques et vulgarisés sont, selon notre hypothèse, le reflet de l'environnement sociolinguistique de leur production. Ils sont donc liés à la situation dans laquelle ils sont fabriqués, transmis, utilisés, etc. (Habert *et al.*, 2001) distinguent des critères « externes » et « internes ». Les premiers sont relatifs au contexte de production des documents (auteurs, support, date de création et/ou de parution, taille de l'article en occurrences ou octets, cadre de production, mode de transmission, type de destinataires, objectifs de la parution, etc.). Ils peuvent être observés dans les documents ou à travers les sites où les documents se trouvent. Dans les documents, le contexte sociolinguistique de leur production est souvent reflété à travers leur structure et contenu, qui peuvent apparaître à travers les critères « internes ». Ces critères sont observables uniquement dans les textes. Ils caractérisent les doc-

[4] http://khnt.hit.uib.no/icame/manuals/brown/index.htm

uments par leur contenu et concernent, par exemple, le niveau de style, la personnalisation, la technicité, etc. Les critères internes donnent des indications sur le contexte de production des documents.

Pour l'exploration des corpus et la détection des critères nous utilisons différents outils, qui varient en fonction des langues. Notons tout de suite que, pour les langues et écritures traitées, très peu d'outils sont disponibles. Ces outils permettent le plus souvent d'accéder à des données linguistiques élémentaires qu'il convient ensuite d'interpréter. Pour le russe, nous utilisons Unitex[5] et Lexico 3[6]. Le fonctionnement d'Unitex est basé sur les transducteurs à états finis. Il permet de reconnaître les motifs linguistiques et d'afficher les concordances autour de ces motifs. Lexico 3 est un outil de statistiques textuelles. Il nous a permis de récupérer, pour les deux corpus, la liste des fréquences absolues des mots simples ainsi que de calculer les segments répétés (suite de formes dont la fréquence est supérieure ou égale à 2 dans le corpus) de chaque corpus. Pour le japonais, nous avons utilisé la chaîne de traitement, y compris l'étiqueteur morphosyntaxique, ChaSen (Matsumoto *et al.*, 2001). Pour les deux langues, nous avons écrit des programmes Perl ou Java pour l'accès à tel ou tel critère. Une fois les critères extraits et isolés, nous les contrastons entre les deux corpus.

3. Préparation et présentation des données

La thématique des corpus concerne le diabète et l'alimentation, en particulier la prise de poids (obésité) due au diabète. Le thème du diabète, pandémie des pays fortement industrialisés, laisse supposer une importante production langagière dans les trois langues traitées. Pour constituer les corpus, nous avons établi une liste de mots-clés, complétée par la consultation de la terminologie médicale UMLS (NLM, 2005) et surtout par des équivalents textuels rencontrés dans les documents. Par exemple, en russe parmi les requêtes les plus productives nous avons : диабет + питание (diabète + alimentation) et диабет + ожирение (diabète + obésité). Ces mots-clés ont été utilisés essentiellement dans le moteur de recherche généraliste http://www.google.ru, mais aussi dans quelques autres moteurs « russophones » (http://www.yandex.ru, http://www.rambler.ru et http://www.aport.ru). En japonais, nous avons utilisé le moteur http://www.google.co.jp sur la requête :

糖尿病　　食事療法　*(diabète régime alimentaire)*

La taille visée de chaque corpus, scientifique et vulgarisé, dans chaque langue était d'environ 200 000 occ. Par contre, nous avons rencontré des difficultés à recueillir les documents scientifiques dans les trois langues. Nous avons donc pu collecter des corpus vulgarisés d'environ 200 000 occ. et des corpus scientifiques d'environ 100 000 occ. Nous avons 150 documents vulgarisés et 45 scientifiques en russe, et 426 et 199 en japonais. La distribution des documents entre les deux catégories a été faite selon l'intuition des auteurs, locuteurs des langues considérées et habitués du domaine étudié. Nous verrons dans la section 4 que cette distinction propose des critères qui semblent être pertinents. Mais nous verrons également qu'il n'existe pas de frontière « stricte » entre les documents scientifiques et vulgarisés. Les documents en russe se sont présentés en trois encodages (win1251, iso-8859-5 et koi-8r), ils ont été convertis en koi-8r et utf-16 pour pouvoir être traités par les outils. Les documents japonais étaient en 4 encodages (shift-jis, euc-jp, iso-2022-jp, utf-8), ils étaient convertis en utf-8, également en fonction des exigences de l'outil ChaSen. Certains des documents n'ont pas pu être traités faute

[5] Disponible gratuitement à http://www-igm.univ-mlv.fr/\~~unitex (version 1.2beta, qui intègre des fonctionnalités de transcodage de fichiers).

[6] Disponible gratuitement à http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/lexico3.htm

de reconnaître leur encodage ou bien de le convertir.

4. Résultats d'identification des critères

Lors de l'analyse des corpus, nous avons examiné plusieurs critères (de très nombreuses balises HTML, longueur des mots, des phrases et des documents, modalité, pronoms, ponctuation, adverbes, subordination, catégories syntaxiques, connecteurs logiques, etc.). Tous ces critères ne se sont pas avérés discriminants entre les documents scientifiques et vulgarisés de notre échantillon. Nous présentons ici les critères qui nous semblent être les plus intéressants.

Images. La présence d'images peut donner des indications sur la catégorie des documents, les textes scientifiques étant par définition plus sobres. Nous étudions donc la balise HTML <*img*>, car c'est cette balise qui introduit les images. Nous trouvons ainsi que les documents vulgarisés en russe contiennent beaucoup plus d'images que les documents scientifiques : une moyenne de 44,0 images par document dans le corpus vulgarisé contre seulement 15,5 dans le corpus scientifique. Par contre, dans les documents en japonais, les deux catégories présentent le même comportement : une moyenne de 21,54 images par document dans les deux corpus. Nous pouvons donc considérer que le nombre moyen d'images par document est un critère discriminant dans les corpus en russe, mais pas en japonais. Toutefois, nous ne disposons pas d'outils permettant de préciser davantage les caractéristiques de l'image (caractère animé/statique, nombre de couleurs, etc.). Quant à la fonction sémiotique des images dans les pages HTML, elle reste à analyser (valeur illustrative, fonction de navigation, caractère publicitaire, etc.).

Tableaux. Les tableaux peuvent être plus présents dans les documents scientifiques, où ils servent alors pour présenter les données chiffrées des études statistiques ou autres. C'est la balise <*table*> qui permet d'accéder à cette information. Travaillant avec les documents HTML, nous nous sommes vite aperçus que les tableaux, en plus de leur première vocation, sont également utilisés comme un moyen de mise en page. Ils sont alors vus comme une alternative aux cadres ou <*frame*>, et aux listes. Ils servent aussi pour le positionnement des éléments graphiques et de l'information paratextuelle (numéros de chapitres, titre du chapitre en cours, etc.). Si la balise <*caption*>, permettant d'associer un commentaire à un tableau, pourrait constituer un marqueur supplémentaire des tableaux, elle s'avère absente des corpus russes. Même étant détournée, la balise <*table*> correspond à un critère discriminant dans les corpus en russe. Mais, contrairement à nos attentes, cette balise est plus fréquente dans le corpus vulgarisé que dans le corpus scientifique : une moyenne de 13 vs. 8,5. Cette balise semble être ainsi le reflet de la complexité de la mise en page : dans un document vulgarisé l'information est « habillée », tandis que dans un document scientifique sa présentation est plus « crue ».

Listes. On peut s'attendre à ce que les listes (balises <*ul*> et <*ol*>) servent à introduire les données dans les documents scientifiques. Nous constatons effectivement une présence importante de listes dans les documents scientifiques en russe : une moyenne de 1,55 par document pour la balise <*ol*> et de 2,26 pour la balises <*ul*> dans le corpus scientifique contre 0,40 et 0,50 respectivement dans le corpus vulgarisé.

Typographie. Sous la typographie nous regroupons la mise en forme italique ou grasse de caractères. Ce sont les balises <*i*>, <*b*> et <*strong*> qui permettent de le faire. Pour le russe, les données recueillies attestent d'un usage des balises <*b*> et <*i*> largement plus répandu dans le corpus scientifique que dans le vulgarisé : 13,18 balises <*i*> et 30,97 balises <*b*> par document pour le corpus scientifique contre respectivement 2,61 et 8,01 dans le corpus vulgarisé. Pour le japonais, on observe la tendance inverse : les balises <*b*> et <*strong*> sont plus répandues dans

le corpus vulgarisé : 13,05 balises par document dans le corpus vulgarisé contre 8,87 dans le scientifique. Quant aux balises <*i*>, elles sont trop peu utilisées (0,31 balises <*i*> par document dans le corpus scientifique contre 0,33 dans le vulgarisé) pour permettre une comparaison significative. Leur sous-emploi peut cependant constituer un indice caractéristique des pratiques culturelles du Web japonais.

Pronoms personnels. Nous nous attendons à ce que les pronoms personnels soient plus présents dans les documents vulgarisés, les documents scientifiques étant par définition impersonnels. Nous présentons les résultats obtenus à partir des documents en russe. (1) Le pronom personnel de première personne du singulier, я (*je*), apparaît ainsi comme un critère fortement discriminant entre les documents vulgarisés et scientifiques. À l'exception des cas d'homonymie, car ce pronom correspond aussi à la dernière lettre de l'alphabet cyrillique et aux initiales de patronymes, tous les deux étant courants dans les documents, nous repéront 11 occurrences de я dans le corpus scientifique, soit une moyenne de 0,29 par document. Ce pronom apparaît lorsque les thérapeutes citent les paroles de leurs patients au style direct. Tandis que dans le corpus vulgarisé, nous en dénombrons 1 046 occurrences, soit une moyenne de 7,12 par document. (2) Quant au pronom de deuxième personne du singulier, ты (*tu*), il semble également être un critère discriminant. Nous en dénombrons une moyenne 0,37 par document dans le corpus scientifique contre 0,92 dans le vulgarisé. Les 14 occurrences de ты du corpus scientifique sont toutes localisées dans un même document, Fig. 1, analysé plus loin. (3) Le pronom de première personne du pluriel мы (*nous*) est plus fréquent dans les documents vulgarisés, sans être totalement absent du corpus scientifique : une moyenne de 2,25 occurrences dans le vulgarisé contre 1 dans le scientifique. Comme le remarque (Kinn, 2005), l'emploi de мы (*nous*) dépend étroitement du genre textuel et, à l'intérieur du genre de l'article scientifique, de la discipline et des tendances des différentes traditions académiques. Le sous-emploi de мы (*nous*) dans notre corpus scientifique peut constituer un critère discriminant. Nous pensons par ailleurs que son sous-emploi dans le corpus scientifique est lié aux multiples possibilités laissées par la langue russe à l'auteur du texte d'éviter de se mentionner lui-même, par exemple les nombreuses tournures impersonnelles. (4) Nous remarquons un écart dans l'utilisation du pronom de deuxième personne du pluriel вы (*vous*). Nous en trouvons en moyenne 2,55 dans le corpus scientifique contre 4,45 dans le vulgarisé. Dans les documents scientifiques, ce pronom apparaît lors de l'échange entre thérapeutes, sous forme de citations au style direct, de recommandations sur la manière d'aborder les questions de poids avec leurs patients. Les proportions que ce pronom montre dans les deux corpus peuvent être suffisantes pour le considérer comme un critère. (5) Et enfin, les pronoms de troisième personne он (*il*), она (*elle*), оно (*neutre*) et они (*ils*, tous genres confondus). Ces pronoms ne montrent pas une différence dans les deux corpus et ne nous semblent donc pas être des critères discriminants.

Critère « poli »/« neutre ». En japonais, les formules de politesse peuvent constituer un critère discriminant entre les deux catégories de documents étudiées. La politesse est exprimée dans le prédicat (verbe, adjectif ou nom), qui se trouve à la fin des phrases. (Doi, 1999) a ainsi étudié les fins de phrases du japonais, mais sans pourtant aller jusqu'à distinguer les types discursifs de documents, et (Ishida et al., 2004) ont mené une étude comparative des fins de phrases sur trois corpus différents. Dans notre corpus, nous trouvons ainsi des expressions caractéristiques des documents scientifiques, comme :

思われる *(être amené à considérer que)*

報告する *(rapporter)*

行った *(effectuer une enquête, une expérimentation, etc.)*

Quant aux documents vulgarisés, nous y découvrons des expressions très différentes, comme par exemple :

「し下さい。 *(vous êtes priés de ..., faites ... s'il vous plaît)*

Cette expression représente une façon d'introduire l'impératif. Elle ne pourrait pas apparaître dans les textes scientifiques.

Les textes scientifiques montrent ainsi un souci de transmettre des informations de « vérité absolue » et se caractérisent par l'absence de rapport entre l'auteur et ses lecteurs tandis que dans les documents vulgarisés, nous l'avons vu, ce rapport peut être exprimé de différentes façons : conseils, interdictions, recommandations, etc. La présence de ces tournures lexicales, qui témoignent des styles « poli » et « neutre », constitue un critère discriminant entre les deux catégories de documents étudiées. Nous trouvons ainsi que le style « poli » est plus fréquent dans le corpus vulgarisé, avec une moyenne de 32,52 phrases par document, que dans le corpus scientifique (une moyenne de 13,62 phrases « seulement »). À l'inverse, les tournures « neutres » sont plus fréquentes dans le corpus scientifique (23,46 phrases en moyenne) que dans le corpus vulgarisé (1,98 phrases).

Modalité de l'incertitude. Nous nous attendons à ce que les marqueurs d'incertitude soient plus fréquents dans le corpus vulgarisé. En russe, le marqueur privilégié de l'incertitude est la particule invariable бы (à prononcer */by/*), grammaticalement proche des terminaisons en *-rai-* du mode conditionnel en français. La particule бы apparaît effectivement comme un critère discriminant entre les documents scientifiques et vulgarisés : une moyenne de 0,74 occurrences par document dans le corpus scientifique contre 1,44 dans le vulgarisé.

Ponctuation. En ce qui concerne la ponctuation, elle peut refléter la complexité des phrases (virgules, point virgule, deux points, parenthèses, etc.), apporter des informations sur l'émotivité (points d'interrogation et d'exclamation), introduire les citations (guillemets), etc. Dans le corpus en russe, les documents scientifiques semblent ainsi se caractériser par un sous-emploi des signes de ponctuation à valeur émotive : une moyenne de 4,0 points d'interrogation par document dans le corpus scientifique contre 6,23 dans le vulgarisé et une moyenne de 1,76 point d'exclamation par document dans le corpus scientifique contre 5,63 dans le vulgarisé. Par ailleurs, les guillemets (plus fréquents dans le corpus vulgarisé que dans le scientifique) semblent également distinguer les deux corpus : une moyenne de 3,76 dans le corpus scientifique contre 17,32 dans le vulgarisé. Il faut cependant noter que les guillemets remplissent différentes fonctions dans ces deux corpus. Dans le corpus scientifique, les citations apparaissent entre guillemets lorsqu'elles sont extraites de documents vulgarisés, ce qui permet de leur limiter leur fiabilité : "продукты-друзья" (*produits-amis*), "продукты-враги" (*produits-ennemis*), Будьте осторожны с "диабетическими" продуктами (*Soyez prudents avec les produits "pour diabétiques"*). Tandis que dans le corpus vulgarisé, ce sont les noms de marques et de produits alimentaires qui apparaissent entre guillemets, mais aussi des néologismes "углеводистость" (*glucidité*), des citations (caution solidaire) d'autres ouvrages de vulgarisation, ou des citations à valeur proverbiale, comme "Ножом и вилкой копаем мы могилу себе" (*C'est avec un couteau et une fourchette que nous creusons notre tombe*). Même si nous ne constatons pas un emploi homogène de la ponctuation, en particulier des guillemets, elle peut être considérée comme un critère discriminant entre les deux corpus.

medi.ru »» Фармацевтические компании »» Ново Нордиск »» Сахарный диабет I типа »» Новости диабетологии

Информация для профессионалов здравоохранения ! Соглашение об использован

Управление диабетом детей и подростков (руководство для обучающих управлению диабетом)

ОБУЧАЮЩЕЕ РУКОВОДСТВО ДЛЯ ДЕТЕЙ И ИХ РОДИТЕЛЕЙ ДЕТСКОЕ ОТДЕЛЕНИЕ, КЛИНИКА УНИВЕРСИТЕТА ГЛОСТРУП ДАНИЯ, 1999

Бирте С. Ольсен, консультант-педиатр;
Хенрик Мортенсен, главный врач, старший детский эндокринолог;
Медицинские сестры по диабету Лене Повлсен и Кристен Дюрлов

[Содержание]

ДИАБЕТ И АЛКОГОЛЬ

ИНФОРМАЦИЯ ДЛЯ СПЕЦИАЛИСТОВ

Люди с диабетом должны принимать меры предосторожности во время приема алкоголя. Без специальных знаний и аккуратного планирования могут возникнуть опасные ситуации. Симптомы низкого сахара крови часто ошибочно принимаются за опьянение и поэтому не распознаются и не лечатся. По этой причине важно не пить слишком много. Друзья должны знать об особой опасности алкоголя для людей с диабетом. Люди с диабетом должны всегда носить диабетическую идентификационную карточку.

Сахар крови

Важно проверять сахар крови перед сном после приема алкоголя. Прием алкоголя повышает риск гипогликемии в последующие 24 часа. Этот риск повышается из-за тенденции заменять алкоголем еду и напитки.

ИНФОРМАЦИЯ ДЛЯ СПЕЦИАЛИСТОВ

физическая активность может также повышаться во время приема алкоголя. Например, прием алкоголя часто сопровождается танцами или бодрствованием позже обычного. Кроме того, печень занята тем, что разрушает алкоголь, и ее функция повышать сахар крови не выполняется. Поэтому инъекции глюкагона будут неэффективны, когда в крови имеется алкоголь.

Еда и напитки

Figure 1. Exemple d'un document à la catégorisation « ambigüe » (medi.ru/doc/051406.htm)

Bilan. Au terme de cette étude, nous obtenons un ensemble de critères qui semblent être discriminants pour distinguer les documents scientifiques et vulgarisés : certaines balises HTML, pronoms personnels, ponctuation, expressions de politesse et marqueurs d'incertitude. Dans la majorité des cas, les critères sont différents dans les deux langues présentées ici, le russe et le japonais, et la comparaison avec le français, qui reste à faire, risque de révéler des différences plus grandes encore. Néanmoins, si nous suivons le travail de (Kessler *et al*., 1997) et tâchons d'hiérarchiser nos critères, nous voyons qu'ils sont assez cohérents et se positionnent sur l'axe objectif / subjectif. Dans cette optique, les documents scientifiques sont caractérisés d'objectifs et les documents vulgarisés de subjectifs. Il est intéressant de voir que cette distinction, qui apparaît d'ailleurs comme assez intuitive, est traduisible par des critères élémentaires de type linguistique, typographique et autre. Ainsi, lorsqu'un document contient les tournures « polies » et/ou les marqueurs de l'incertitude, et/ou dénombre plusieurs indices d'images, pronoms personnels ou guillemets, ce document sera considéré comme vulgarisé. Par contre, lorsque nous rencontrons les tournures de type « neutre » et pas de marqueur d'incertitude ou bien lorsque les images sont absentes et lorsque les pronoms personnels et les guillemets sont rares, le document pourra être considéré comme scientifique. Une autre remarque vient à ce stade de réflexion : l'axe subjectif / objectif se présente comme un continuum sur lequel les documents peuvent venir se positionner avec un certain degré de subjectivité et/ou d'objectivité, en fonction de critères qui pourront y être distingués. Il va de soi qu'à côté des documents qui sont facilement catégorisables, d'autres le sont beaucoup moins. Par exemple, le document de la Fig. 1 retient notre attention par la mixité des critères qu'il contient. Il relève ainsi par plusieurs aspects du

scientifique : mise en page, titres, localisation dans un portail médical sous la rubrique *Information destinée aux professionnels de la santé* contenant plusieurs autres articles scientifiques, indication des auteurs et d'organismes médicaux d'affiliation en tête de l'article, un filigrane en fond du texte *Information destinée aux spécialistes*. Ce document présente en même temps plusieurs critères de vulgarisation : dessin coloré, emploi du pronom ты (*tu*), dont nous avons parlé plus haut, et des impératifs. L'emploi des pronoms ты vient ainsi du fait qu'il s'agit en réalité d'un guide de conduite destiné aux adolescents diabétiques intitulé *Diabète et alcool chez l'adolescent*, qui contient des conseils formulés à la deuxième personne du singulier. De par cette analyse, il est difficile de catégoriser ce document vers le pôle scientifique ou vulgarisé. Il se situerait plus vraisemblablement vers le milieu de l'axe subjectif / objectif. Cette étude permet aussi de dégager un autre point important, sur lequel nous ne nous sommes pas prononcé jusqu'ici : la nature des catégories scientifique et vulgarisée. Il est en effet difficile de statuer sur cette distinction établie intuitivement, mais nous considérons, en nous basant en particulier sur les travaux de (Rastier *et al*., 1994), que le discours des documents que nous avons analysés est médical. Nous distinguons également, à l'intérieur de chaque corpus analysé, scientifique et vulgarisé, plusieurs genres : article, manuel, recette de cuisine, message, etc. Par contre, il nous est difficile de définir à quel niveau se fait la distinction entre les catégories scientifique et vulgarisée. Correspond-elle aux « champs génériques » introduits par (Rastier *et al*., 1994), qui représentent une couche intermédiaire entre les discours et les genres ? Ou peut-être se fait-elle plutôt à un niveau supérieur aux discours, car de nombreux discours (médical, juridique, technique, religieux, ...) peuvent produire des documents vulgarisés et scientifiques ? Ou peut-être s'agit-il d'une distinction purement pragmatique ? Mais la question reste alors la même : à quel niveau cette distinction se fait-elle ? Nous pensons que nous sommes ici en face d'un manque théorique et que la jonction entre la linguistique et la linguistique de corpus pourrait apporter certaines réponses à ces questions.

5. Conclusions et perspectives

L'étude que nous avons menée sur les corpus en russe et en japonais nous a permis de confirmer la validité de certains critères pressentis pour la distinction entre les documents scientifiques et vulgarisés. Cette étude ne représente qu'une première étape du travail. Elle pourra être complétée par l'utilisation d'autres outils (étiqueteur morphosyntaxique du russe, outil d'analyse d'images, etc.), par l'analyse d'autres critères et elle pourra de cette manière atteindre une homogénéisation entre les critères de différentes langues. Nous pensons par ailleurs que ces critères, bien qu'acquis sur un corpus réduit, pourront être testés sur des corpus portant éventuellement sur d'autres thématiques, et généralisés. Ces critères pourront également être utilisés comme une base d'apprentissage par les algorithmes de catégorisation. Les critères recueillis pourront alors être pondérés et regroupés en des faisceaux plus fins, car c'est bien par des faisceaux de critères pondérés que l'on peut obtenir une catégorisation efficace des textes (Valette, 2004). Une collaboration entre la linguistique et l'étude des corpus peut, par ailleurs, apporter certaines réponses aux questions théoriques autour des genres, des discours et de la nature linguistique des documents scientifiques et vulgarisés.

Références

BIBER D. (1988). *Variations accross speech and writing*. Cambridge University Press.

DÉJEAN H. & GAUSSIER E. (2002). une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica*.

DOI (1999). *Enquête des pages Web autour des expressions des fins de phrases*. Rapport interne. Article en japonais.

GAUDINAT A. & BOYER C. (2003). WRAPIN (worldwide online reliable advice to patients and individuals). In *8th Annual World Congress on the Internet and Medicine (MedNet)*, Geneva.

HABERT B., GRABAR N., JACQUEMART P. & ZWEIGENBAUM P. (2001). Building a text corpus for representing the variety of medical language. In *Corpus Linguistics*, Lancaster.

HABERT B., ILLOUZ G., LAFON P., FLEURY S., FOLCH H., HEIDEN S. & PRÉVOST S. (2000). Profilage de textes : cadre de travail et expérience. In M. RAJMAN, Ed., *5èmes Journées d'Analyse des Données Textuelles (JADT)*, Lausanne.

ISHIDA ET AL. (2004). *Analyse des caractéristiques stylistiques des articles académiques*. Rapport interne. Article en japonais.

KARLGREN J. (2000). *Stylistic experiments for information retrieval*. Doctoral dissertation, department of Linguistics, Stockholm university.

KESSLER B., NUNBERG G. & SCHÜTZE H. (1997). *Automatic detection of text genre*. Palo Alto Research Center. ftp ://parcftp.xerox.com/pub/qca/papers/genre (consulté le 04/11/2005).

KINN T. (2005). *Plays of the we-hood : what do we mean by we ?* in Akademisk proza, 3-2005, Actes du séminaire.

KWASNIK B. H., CROWSTON K., NILAN M. & ROUSSINOV D. (2000). *Identifying Document Genre to Improve Web Search Effectiveness*. The Bulletin of the American Society for Information Science and Technology. http ://www.asis.org/Bulletin/Dec-01/kwasnikartic.html (consulté le 26/11/2005).

MALRIEU D. & RASTIER F. (2001). *Genres et variations morphosyntaxiques*. in Traitement automatique des langues, vol. 42, p. 548-577. http ://www.revue-texto.net/Inedits/Malrieu_Rastier/Malrieu-Rastier_Genres.html {consulté le 14/10/2005}.

MATSUMOTO K. & TANAKA H. (2002). Automatic alignment of Japanese and English newspaper articles using an MT system and a bilingual company name dictionary. In *LREC*, p. 480–484.

MATSUMOTO Y., KITAUCHI A., YAMASHITA T., HIRANO Y., MATSUDA H., TAKAOKA K. & ASAHARA M. (2001). *Morphological Analysis System ChaSen. Manual, version 2.2.8*. Rapport interne, Nara 7 Institute of Science and Technology.

NLM (2005). *UMLS Knowledge Sources Manual*. National Library of Medicine, Bethesda, Maryland. www.nlm.nih.gov/research/umls/.

RASTIER F., CAVAZZA M. & ABEILLÉ A. (1994). *Sémantique pour l'analyse - De la linguistique à l'informatique*. Paris, Masson.

RESNIK P. (1999). Mining the web for bilingual texts. In *37th Annual Meeting of the Association for Computational Linguistics (ACL)*, College Park, Maryland. Disponible à http ://www.umiacs.umd.edu/users/resnik/pubs.html. Visité le 15/09/2000.

SINCLAIR J. (1994). *EAGLES. Corpus typology*. Rapport interne, EAG-CWG-IR-2. Disponible à http ://www.ilc.pi.cnr.it/EAGLES96/. Visité le 02/03/2003.

VALETTE M. (2004). *Sémantique interprétative appliquée à la détection automatique de documents racistes et xénophobes sur Internet*. Approches sémantiques du document numérique, Actes du 7e colloque international sur le document électronique, 22-24 juin 2004, Patrice Enjalbert and Mauro Gaio ed. http ://www.revue-texto.net/Inedits/Valette/Valette_Princip.pdf {consulté le 04/10/2005}.