Caractérisation différentielle de forums de discussion sur le VIH en vietnamien et en français : Éléments pour la fouille comportementale du web social

Océane Hô Dinh

▶ **To cite this version:**

Océane Hô Dinh. Caractérisation différentielle de forums de discussion sur le VIH en vietnamien et en français : Éléments pour la fouille comportementale du web social. Linguistique. Université Sorbonne Paris Cité, 2017. Français. NNT : 2017USPCF022 . tel-01964671

HAL Id: tel-01964671

https://theses.hal.science/tel-01964671

Submitted on 23 Dec 2018

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

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

INSTITUT NATIONAL DES LANGUES ET CIVILISATIONS ORIENTALES

ÉCOLE DOCTORALE N°265

LANGUES, LITTÉRATURES ET SOCIÉTÉS DU MONDE

THÈSE

présentée par

Océane Hồ Đình

Soutenue le 22 décembre 2017

pour obtenir le grade de **Docteure de l'INALCO**

Discipline : *Sciences du langage : linguistique et didactique des langues*

Caractérisation différentielle de forums de discussion sur le VIH en vietnamien et en français

Éléments pour la fouille comportementale du web social

Thèse dirigée par :	
M. Mathieu Valette	Professeur des universités (Sciences du langage) INALCO
Rapporteuses :	
Mme Agata Jackiewicz	Professeure des universités (Sciences du langage) Université Paul Valéry - Montpellier III
Mme Claire Oger	Professeure des universités (Sciences de l'information et de la communication) - Université Paris Est Créteil
Membres du jury :	
M. Mathieu Valette	Professeur des universités (Sciences du langage) INALCO
Mme Agata Jackiewicz	Professeure des universités (Sciences du langage) Université Paul Valéry - Montpellier III
Mme Claire Oger	Professeure des universités (Sciences de l'information et de la communication) - Université Paris Est Créteil
M. Danh Thành Dô-Hurinville	Professeur des universités (Sciences du langage) Université de Franche-Comté
Mme Evelyne Micollier	Chargée de recherche (Anthropologie) Institut de Recherche pour le Développement

inalco
Institut national
des langues
et civilisations orientales

inalco
ERTIM
Equipe de recherche
textes, informatique,
multilinguisme

Caractérisation différentielle de forums de discussion sur le VIH en vietnamien et en français

Éléments pour la fouille comportementale du web social

THÈSE

présentée par

Océane Hồ Đình

Remerciements

Je tiens à remercier en tout premier lieu Agatha Jackiewicz, Claire Oger, Evelyne Micollier et Danh Thành Dô-Hurinville d'avoir accepté de constituer le jury de cette thèse. Leur travaux dans des domaines respectifs ont guidé mes pas tout au long de cette recherche et c'est un grand honneur de voir cette dernière évaluée par elleux.

Je remercie tout particulièrement Mathieu Valette pour avoir nourri le feu de mon intérêt pour notre domaine de recherche et pour m'avoir permis de mener cette thèse à son aboutissement, par son intérêt infaillible pour mes recherches, l'acuité de son regard et ses encouragements sans cesse renouvelés. C'est aussi une rencontre humaine à laquelle j'attribue une valeur au delà de l'encadrement de thèse.

Je veux évidemment remercier toute l'ERTIM de créer au quotidien des conditions de travail aussi agréables, aspect auquel j'accorde une grande importance. Pouvoir considérer ses collègues comme des amis devrait être systématique, mais c'est assez peu le cas pour que je les en remercie. Je réserve également une pensée admirative pour Monique Slodzian et sa force entrainante qui a permis la création de cette unité de recherche. Je remercie Marie-Anne Moreaux de m'avoir transmis la passion de l'algorithmique et la rigueur de réflexion.

Je remercie Philippe Lambert de m'avoir poussée à partir, dès mes 19 ans, explorer sur le terrain le Vietnam et la langue vietnamienne, et surtout permis d'oser considérer ces sujets avec un intérêt scientifique et professionnel.

Merci Christian Jean d'avoir été là, à ta propre initiative, après que tout soit parti en fumée, pour m'aider à trouver la force de reconstruire les données et clarifier les idées.

Merci à Driss Sadoun pour les relectures constructives, amicales et impliquées, ainsi que pour les discussions ouvertes à tous les sujets qui comptent dans la vie.

Merci à Marguerite Leenhardt pour les conseils à la volée, merci à l'équipe de choc des chasseu·r/se·s de coquilles et à Capucine et Manon pour l'aide à l'impression.

Merci à Cecilia pour les respirations hebdomadaires vitales. De l'importance d'aider le corps à soutenir l'effort de l'esprit.

Merci à Ly, Capucine, Anne-Cha, Dorothée, Yasmine, Camilla, Emma et tant d'autres pour votre patience et l'assurance de votre amitié indéfectible.

Merci Manon pour ton immense patience au quotidien et tout le reste.

Merci ma soeur Estelle et mon frère Donovan pour votre amour et votre soutien.

Enfin, merci à mes parents de m'avoir insufflé cette confiance inébranlable dans la vie. Hùng, sans que tu l'aies cherché, c'est par toi que ce chemin vietnamien a commencé. Enfin, Annisa, même si tu vis à chaque instant dans chacun de mes atomes, comme tu m'as manqué pendant cette période : tu aurais été la meilleure personne pour me soutenir. Je me dis que tu peux être fière de ce que tu m'as transmis.

Sommaire

Sommaire v

Introduction 1

Chapitre I Le Viêt Nam comme axe de recherche 7

Chapitre II État de l'art 33

Chapitre III Constitution des données 73

Chapitre IV Cartographie d'un forum de discussion 117

Chapitre V Analyse sémantique des discours informels 147

Conclusion générale et perspectives 195

Annexes 201

Glossaire 293

Acronymes 301

Table des matières 303

Liste des tableaux 309

Table des figures 311

Bibliographie 313

Introduction

Contexte général

Outiller les SHS pour l'analyse des discours de l'intime

Jusqu'à l'arrivée d'Internet et du web social, l'analyse outillée des corpus textuels s'est principalement développée dans des domaines où existaient des volumes de données la justifiant, tels que la littérature et la philologie. Aujourd'hui, les possibilités se sont élargies à d'autres domaines des sciences humaines et sociales (SHS) dans le cadre des programmes d'humanités numériques. Internet a notamment favorisé la naissance de discours natifs du web qui nécessitent des méthodes d'analyse adaptées.

Les SHS ont une longue tradition de collecte de données. Par exemple, l'anthropologie, l'ethnologie ou la sociologie étudient les sociétés humaines en les observant sur le terrain, en recueillant la parole des sujets étudiés, puis en analysant les comportements. La sociologie combine le plus souvent des méthodes qualitatives et quantitatives (Jick, 1979 ; Creswell, 2003 ; Tashakkori et Teddlie, 2003). Les premières viennent en complément des secondes, pour vérifier des hypothèses, améliorer l'interprétation. Les méthodes quantitatives se basent sur des outils mathématiques et statistiques pour analyser des éléments mesurables, les méthodes qualitatives se basent sur l'interpré-

tation des données collectées. Les méthodes de collecte de données ont une influence plus ou moins forte sur lesdites données. Ainsi, les entretiens non- ou semi- directifs auront moins d'influence sur les données recueillies que l'observation d'un groupe social isolé pour une expérience, ou le recours à des questionnaires. Pourtant, lorsqu'il s'agit de traiter des sujets qui touchent à l'intime, même le recours aux entretiens semi-directifs ne garantit pas des données exploitables. Par exemple, après avoir travaillé sur des questions de sexualité en utilisant des données quantitatives ou de seconde main, Garcia (2016) s'est intéressée à l'extraconjugalité. Pour ce travail il était compliqué d'avoir recours aux entretiens enregistrés, en face à face avec la sociologue, malgré toutes les garanties de confidentialité et de déontologie. Elle s'est donc intéressée aux réseaux sociaux sur Internet, notamment les forums de discussion, où les internautes racontent leur vie extraconjugale. Garcia explique[1] qu'il est possible d'utiliser comme matériau d'analyse ces discours sur les forums, différents de ceux obtenus par entretiens.

Difficile à analyser, le web social a encore peu été exploité comme terrain d'étude sociologique. Les évolutions rapides et incessantes des outils de communication et de leurs usages, les spécificités langagières qui s'y développent, la prolifération des contenus[2], l'accélération du temps perçu, etc. sont autant d'obstacles nouveaux auxquels sont confrontées les SHS. Née (2017, dir.) montrent néanmoins que les méthodes existantes de l'analyse du discours (AD) peuvent s'adapter et se nourrir des analyses outillées par l'informatique.

1. Dans l'émission Les Nouvelles Vagues du 11 janvier 2017 sur France Culture (https://www.franceculture.fr/emissions/les-nouvelles-vagues/le-couple-15-nouvelles-alliances), où elle était invitée pour parler de son livre Amours Clandestines. Sociologie de l'extraconjugalité durable (Garcia, 2016).

2. selon IBM, en 2016 2,5 trillions d'octets de données étaient générés chaque jour dans le monde et 90% du total des données avaient été générées au cours des deux dernières années. (https://www-01.ibm.com/software/fr/data/bigdata/)

Les analyses traditionnelles en SHS nécessitent de lier les discours à des indicateurs contextuels. La sociologie s'appuie sur les données personnelles des locuteurs, telles que par exemple l'âge, la catégorie socio-professionnelle ou l'identité de genre. Lorsque les discours étudiés portent sur des sujets intimes, cette identification peut poser des problèmes de confidentialité, ou inhiber les discours spontanés.

Avec le web social, l'accès à ces discours est facilité, mais celui aux indicateurs contextuels est malaisé. Il existe pourtant des possibilités d'exploiter ces données sans y avoir recours. Il s'agit ainsi de transformer en avantage l'inconvénient que posent de prime abord les données collectées sur le web social : leur intérêt n'est plus basé sur leur traçabilité mais sur leur massivité, laquelle contribue à anonymiser les discours et ainsi conserver la confidentialité des données. De cette manière il est possible de mener les analyses sans attenter à la vie privée des personnes, tout en bénéficiant de la spontanéité des discours.

PROBLÉMATIQUE

L'autorité de la connaissance à l'aune du web social

Oger et Ollivier-Yaniv (2003, 2006) ; Sarfati (2008, 2014) ont souligné la forte stabilité des discours institutionnels : toute communauté, quelle que soit sa taille, nécessite la mise en commun de normes de sens, laquelle est prise en charge par les discours institutionnels. Toutefois, avec l'accélération des transformations sociales, les institutions, stables par définition, sont rapidement en décalage avec la société. Les discours normés que produisent les institutions sont par conséquent concurrencés par les discours informels issus du web social. Les connaissances développées au sein des échanges entre inter-

nautes s'affranchissent des hiérarchies conventionnelles et le partage d'information s'y fait de manière plus horizontale que sur les canaux traditionnels. Ainsi, la démocratisation de la prise de parole redistribue l'autorité en matière de connaissance et modifie les processus de construction des savoirs.

Cette thèse propose une méthodologie de description des discours informels du web social pour mieux appréhender les connaissances qui s'y élaborent. L'étude de ces discours para-institutionnels, contribue notamment à montrer en quelle mesure ils peuvent compléter les discours institutionnels en palliant leur lenteur d'adaptation. Cette méthodologie prend en compte la volatilité des échanges informels tout en faisant émerger des formes alternatives de stabilité. Parmi les discours du web social, nous nous intéressons plus particulièrement aux forums de discussion comme environnements où s'élaborent des connaissances spontanées. Nous montrons également en quoi au sein des forums les frontières sont poreuses entre les discours institutionnels et les discours informels.

APPLICATION

Le VIH dans un double contexte social

Notre champ d'application porte sur la problématique du virus d'immunodéficience humaine (VIH). Ce terrain applicatif recouvre plusieurs enjeux de société : sanitaire et social, évolutions des mœurs, concurrence des discours, etc. Quel que soit le contexte national, les pouvoirs publics sont tenus d'apporter une réponse face à cette épidémie mondiale contre laquelle la médecine échoue à trouver un traitement. Les institutions prennent en charge le travail de prévention, d'information et d'accompagnement auprès des populations confrontées aux risques de contamination. Par ailleurs, les échanges

informels produits sur les forums de discussion consacrés au VIH couvrent d'autres aspects de la problématique. La comparaison de ces discours avec ceux des institutions permet de faire émerger les facettes du problème que ces dernières laissent sous silence.

Même dématérialisés les discours doivent être analysés en tenant compte de leur contexte socioculturel de production. Nous nous intéressons à deux situations distinctes : l'une européenne (en France), l'autre asiatique (au Viêt Nam). Cette approche différentielle, qui met en vis-à-vis les deux contextes socioculturels, vise à révéler d'autres spécificités des discours étudiés.

PLAN DE LA THÈSE

Le premier chapitre est consacré à la description détaillée du contexte vietnamien. En plus de décrire le cadre socioculturel et historique, nous y exposons les tensions et les enjeux auxquels est confrontée la société vietnamienne contemporaine, afin de dépeindre le contexte dans lequel se développent les discours sur l'épidémie du VIH.

Le chapitre suivant consiste en un état de l'art pluridisciplinaire dans lequel nous étudions les spécificités des discours du web social, en particulier des forums, et les défis qu'ils posent autant à l'analyse des discours, qu'à la sociologie, à la textométrie et à la fouille contrastive, ou au traitement automatique des langues (TAL) en général. D'autre part, nous présentons les spécificités de la langue vietnamienne et leurs conséquences pour son traitement automatique.

Nous consacrons le troisième chapitre à la constitution du corpus d'analyse. Nous expliquons nos choix des sources, avant de décrire leur structuration, pour ensuite détailler la méthodologie de récupération et d'organisation

des données en corpus. Une volonté de reproductibilité guide l'ensemble de la rédaction du chapitre, elle se retrouve donc également dans la description de la préparation des données pour les traitements textométriques et les différentes analyses.

Le quatrième chapitre s'attache à la description des spécificités des forums, ainsi que des méthodes développées pour les analyser. Nous montrons en quoi la production nativement numérique de ces discours impose d'adapter les méthodes d'analyse à leur structuration particulière. Puis nous proposons des critères d'analyse (formels ou lexicaux) pour distinguer le discours institutionnel rapporté, présent en masse dans les échanges sur les forums, du discours informel spontané, où se situe le point aveugle des discours institutionnels.

Dans le dernier chapitre, nous proposons une analyse sémantique des discours informels. Nous étudions les niveaux lexical, stylistique et de genre textuel, et nous exposons les résultats qu'une analyse différentielle permet de faire émerger. Elle met en évidence les lacunes des discours institutionnels et la complétion apportée par les discours informels.

En annexe sont présentées les parties techniques et algorithmiques de la chaîne de traitement des corpus, depuis la collecte des données jusqu'aux analyses, en passant par la mise en forme des corpus pour les différents logiciels exploratoires. Y figurent également les lexiques constitués au cours de ce travail.

CHAPITRE I

Le Viêt Nam comme axe de recherche

1 Le contexte socio-culturel vietnamien

1.1 *Contexte historique*

Après trente ans de conflits successifs dans un contexte de sortie de Guerre Mondiale puis de Guerre Froide, le Viêt Nam n'a véritablement exercé une souveraineté en tant que nation unifiée qu'à partir de 1975. La lutte fratricide qui avait opposé le Nord-Viêt Nam au Sud-Viêt Nam – attisée par la participation des forces extérieures, opposées à l'échelle mondiale en deux blocs : au nord, la *République démocratique du Viêt Nam* (Việt Nam Dân Chủ Cộng Hòa), soutenue par le bloc communiste ; et au sud, la *République du Viêt Nam* (Việt Nam Cộng Hòa), soutenue par les Etats-Unis d'Amérique – s'est terminée par la victoire du Nord sur le Sud, le retrait des forces extérieures et l'établissement de la *République Socialiste du Viêt Nam* (Cộng hoà Xã hội Chủ nghĩa Việt Nam) pour l'ensemble du territoire. A cette date, le pays est exsangue mais indépendant, et la reconstruction s'opère dans un

contexte de fermeture, de gouvernance autoritaire, dans la ligne d'une politique communiste puissamment dirigée par un gouvernement constitué des vainqueurs, qui applique une économie planifiée à l'échelle nationale ayant des conséquences sur toutes les couches de la population. Il ne fait aucun doute que la paix est plus difficile à gérer que la guerre et la période qui suit immédiatement la guerre est une période de crise économique importante. Ainsi, le développement du pays a du mal à décoller et il faut attendre 1986 pour que l'échec de l'économie planifiée soit officialisé, et qu'un changement d'orientation politique soit amorcé, avec l'application du concept de *Đổi Mới* (« Renouveau »), spécifiquement utilisé pour caractériser le mouvement de réformes politiques qui a orienté le pays vers une reconnexion avec le reste du monde, tout en sauvant le régime. Après une génération qui n'a connu que la guerre, de 1945 à 1975, puis dix années de difficile reconstruction, aujourd'hui une génération entière est née dans un pays en plein développement et a soif de trouver sa place dans la course à la mondialisation.

1.2 *Situation politique : l'abus des libertés démocratiques est dangereux pour la santé*

Depuis 1976, année de la proclamation de la République Socialiste du Viêt Nam (RSVN), le pays a tracé son chemin en suivant avec plus ou moins de distance la voie de ses grands frères en politique que sont la Chine et la Russie. De manière spécifique, chacune de ces républiques socialistes (et aujourd'hui Cuba suit à son tour le mouvement) a opéré dans son histoire une adaptation de sa ligne politique vers une libéralisation économique, que ce soit la Chine à la fin des années soixante-dix, ou l'URSS avec la perestroïka à la fin des années quatre-vingt. Cependant au Viêt Nam, la libéralisation économique ne s'est pas accompagnée d'une libéralisation politique. Le Parti

communiste vietnamien (fondée par Hồ Chí Minh en 1930) est la seule formation politique autorisée, depuis la proclamation de la RSVN en 1976. Le Parti conserve donc, aujourd'hui encore, la mainmise politique. Côté économique en revanche, le pays est lancé sur la voie de la modernisation, avec l'adoption en 1986 d'une « économie de marché à orientation socialiste » (*kinh tế thị trường theo định hướng xã hội chủ nghĩa*), qui, après trente ans de guerre et dix ans de fermeture, voit le pays s'insérer sur le marché mondial depuis une trentaine d'années.

De cette situation résulte une société que l'on pourrait qualifier de clivée, entre une caste de dirigeants politiques, descendants des généraux nordistes victorieux qui ont pris les commandes du pays en 1975[1], et le reste de la population, qui s'occupe de ses affaires personnelles, en naviguant avec les lois et les autorités, dans un contexte de corruption importante[2], sans pour autant chercher à renverser ce pouvoir. La propagande d'état appelle pourtant régulièrement la population à aller voter, mais comment considérer les élections autrement que comme un simulacre, lorsque le seul parti autorisé est le Parti Communiste, à la tête du pays depuis la victoire du Nord sur le Sud il y a quarante ans ?

La politique est en fait l'un des principaux sujets tabous, que personne n'ose aborder à l'exception d'une poignée d'activistes indépendants, journalistes et aujourd'hui principalement blogueurs, contraints d'en faire un choix de vie, confrontés à des actes de harcèlement répétés sur eux ou leur famille,

1. Mais aussi de nombreux relais du pouvoir, dans un système aux multiples divisions administratives.

2. Dans la grande enquête sur la corruption publiée par Transparency International en 2013, 72% des vietnamiens interrogés estimaient que la police était corrompue. En 2015, l'ONG plaçait le Viêt Nam au 123ème rang sur 176 dans son classement des pays les plus corrompus (http://www.transparency.org/cpi2012/results) Voir aussi http://redtac.org/asiedusudest/2013/04/01/l%e2%80%99influence-du-parti-communiste-vietnamien-sur-la-corruption-dans-pays/comment-page-1/.

voire des violences policières graves[3]. Pourtant, en 2013, le Viêt Nam est entré au Conseil des droits de l'homme à l'ONU. Mais la même année, le gouvernement a publié le Décret 72, interdisant l'usage des blogs et réseaux sociaux pour partager des informations sur l'actualité, limitant leur usage à l'échange d'informations personnelles[4] . Ce décret est dénoncé par les organisations de défense de la presse libre et indépendante comme une « grave atteinte au droit d'informer et d'être informé »[5]. Il existe un article du code pénal vietnamien, l'article 258, encore très utilisé aujourd'hui pour justifier une ligne répressive à l'égard des dissidents, considérant comme un délit l'« abus des libertés démocratiques », ce qui permet à peu près tout, sur toute situation à même de remettre en cause les autorités[6].

C'est pour ces raisons que Reporters Sans Frontières (RSF) classe le Viêt Nam dans les pays à « situation très grave », situation qui s'est encore empirée en 2015, plaçant le Viêt Nam au 175ème rang mondial sur 180[7], seulement dépassé par quelques pays comme la Corée du Nord, la Syrie ou la Chine. RSF a rapporté régulièrement des atteintes aux droits de l'homme dans ce domaine, notamment autour du quarantième anniversaire de la fin de la guerre, le 30 avril 2015, où les voix indépendantes ont été particulièrement muselées.

3. Voir par exemple : l'agression de Trương Minh Đức en novembre 2014 (http://www.danchimviet.info/archives/91766/nha-bao-truong-minh-duc-gui-don-to-cao/2014/11) ou celle de Nguyễn Văn Đài en décembre 2015 (https://rsf.org/fr/actualites/letat-vietnamien-sacharne-contre-le-journalisme-citoyen)

4. Décret 72 sur la gestion, la mise à disposition et l'utilisation d'internet et des informations en ligne : http://datafile.chinhphu.vn/file-remote-v2/DownloadServlet?filePath=vbpq/2013/07/72-2013-nd.pdf

5. rsf.org/fr/actualites/le-vietnam-veut-interdire-aux-internautes-de-parler-dactualite

6. « *C'est à partir de l'année 2004 que les autorités vietnamiennes ont [commencé] à utiliser l'article 258 pour réprimer les délits d'expression. Depuis cette date, cette disposition du Code pénal a été utilisée pour arrêter et condamner au moins dix militants des droits de l'homme et quatre blogueurs.* » (http ://eglasie.mepasie.org/asie-du-sud-est/vietnam/2014-12-12-deux-nouveaux-blogueurs-arretes-pour-avoir-abuse-de-ab-la-liberte-democratique-bb)

7. Classement effectué par RSF : http://index.rsf.org/#!/index-details/VNM

1.3 *Situation démographique : une population jeune et dense*

La population vietnamienne a quasiment doublé en 40 ans, dépassant les 90 millions de personnes en 2014, pour un territoire moitié moins étendu que celui de la France, et composé aux deux tiers de montagnes et de hauts plateaux, où se répartissent les 53 ethnies minoritaires représentant 14% de la population. Les 86% de Kinh, l'ethnie majoritaire, se regroupent donc sur le tiers restant du territoire, composé de fines côtes et de deux grandes plaines deltaïques intensivement exploitées pour la riziculture. Les deux principales villes du pays, au nord Hà Nội - la capitale, et au sud Hồ Chí Minh Ville - la mégapole économique, sont donc très densément peuplées et concentrent l'essentiel de l'activité, suivies par d'autres villes moyennes en plein développement.

Le Viêt Nam est considéré comme ayant réalisé sa transition démographique particulièrement rapidement, cela sans avoir recours à des directives aussi coercitives que la politique de l'enfant unique en Chine. Jusque dans les années soixante-dix, le taux de fécondité vietnamien était encore de six naissances par femme. Le gouvernement nord-vietnamien s'en est alarmé dès le recensement de 1960 et a mis en place une planification familiale à partir de 1963, ce qui a fait du Nord-Viêt Nam l'un des premiers pays en voie de développement (et en guerre) à adopter une politique de limitation des naissances, à savoir : deux ou trois enfants maximum, espacés chacun de cinq à six années. Cela s'est accompagné à partir de 1970 de mesures d'incitation à la contraception (production et importation de stérilets) et d'incitation à la pratique de l'avortement. À la réunification du pays en 1976, cette politique a été généralisée à l'ensemble du pays, à l'exception des minorités[8].

8. Reportage sur la régulation des naissances au Viêt Nam : http://www.paperblog.fr/1567214/la-regulation-des-naissances-au-VietNam/

Dans les décennies qui ont suivi, les recensements ont révélé une croissance démographique foudroyante malgré les volontés de limiter les naissances, étant donné que la mortalité a fortement baissé à partir de la fin de la guerre. Par conséquent, de grandes campagnes d'information ont été réalisées, d'abord dans les villes, puis aussi dans les campagnes, et les mesures se sont faites de plus en plus incitatives : gratuité de tous les moyens de contraception (préservatifs, pose du stérilet, ligature des trompes, vasectomie), avortement remboursé, pose du stérilet obligatoire pour les femmes accouchant dans les services de santé publics, même au premier enfant, etc.[9]

Fig. 1.1 : Panneau de propagande. A gauche : « Promotion de la pilule contraceptive » Au centre : « Bien suivre la planification familiale pour un peuple riche, un pays fort, une société équitable et civilisée » A droite : « Peu de naissances , espacées dans le temps, assurent une bonne santé à la mère, des enfants sages, et le bonheur de la famille ». Photographie prise le 3.12.2006 par vn555333 (www.flickr.com/photos/vietnam555-333/312981752, sous licence Creative Commons ©).

pour en venir en 2003 à restreindre officiellement le nombre d'enfants par

9. Reportage sur la pression démographique au Viêt Nam : http://www.vninfos.com/selection/articles/pression_demogra_exode_rural.html

famille à deux (à l'exception des familles recomposées et des minorités ethniques), restriction toujours en vigueur. Dix ans après, le Viêt Nam affichait un taux de fécondité de 2,1 enfants par femme, avec cependant une disparité encore forte entre les villes et les campagnes, entre l'ethnie majoritaire et les minorités ethniques.

Malgré ce ralentissement de l'accroissement naturel, la population reste très jeune, avec un âge médian de moins de trente ans (il est de plus de quarante ans en France), c'est-à-dire que plus de la moitié de la population a moins de trente ans : 52,7% en 2010. Ce chiffre est en constante baisse depuis 1980, où il était de 69,3%, cependant la population vietnamienne reste une population très jeune encore aujourd'hui. Parmi cette majorité, la part des quinze-trente ans, qui nous intéresse principalement, représente 30% de la population totale.

D'autre part, la population urbaine représente 30% de la population totale et la dichotomie économique entre monde rural et monde urbain se creuse. Les deux principales villes du pays, qui dominent chacune une des deux régions de plaine deltaïque rizicole, au nord Hà Nội, la capitale, et au sud Hồ Chí Minh-Ville (HCMV), le cœur économique du pays, sont très densément peuplées et concentrent l'essentiel de l'activité (75% de l'activité industrielle notamment, dont 55% à HCMV). Suivent ensuite quelques villes moyennes, notamment la troisième ville du pays, qui domine la région centrale, Đà Nãng, en pleine industrialisation. Hà Nội compte 7 millions d'habitants, ayant absorbé les régions rurales en périphéries en 2008, et pour HCMV le chiffre est difficile à estimer car la ville compte plusieurs millions de clandestins, il oscille entre 9 et 11 millions. Par comparaison, Paris compte 2,2 millions d'habitants, l'Île-de-France 11,8 millions.

1.4 *Mondialisation et VIH*

La propagation du VIH est concomitante avec l'ouverture des frontières, le développement du tourisme, la multiplication des échanges avec le monde extérieur. L'anthropologue médical Wolffers (2001) explique que la propagation virale du VIH procède par vagues, et se fait par trois voies principales : celle intraveineuse parmi la population des usag·er/ère·s de drogues injectables (UDI), celle sexuelle, touchant en premier les professionnel·le·s du sexe (TS), puis leurs clients, puis les partenaires de ces derniers, et enfin, la transmission dite verticale, autrement dit, de la mère à l'enfant, pendant la grossesse, l'accouchement ou l'allaitement.

Au Viêt Nam, Wolffers (2001, p.156) indique que « le premier cas d'infection par le VIH officiellement reconnu remonte à décembre 1990, à HCMV », c'est-à-dire quatre ans après le changement officiel d'orientation politique vers plus d'ouverture. La décennie 90 est celle qui a vu naître et se développer les épidémies de VIH en Asie. Au Viêt Nam il est bien entendu impossible d'attester qu'aucun cas n'ait existé avant la politique d'ouverture, mais en 1992 « on ne rapportait encore que onze cas »[10] et c'est à partir de 1993 qu'une poussée est observée parmi les personnes toxicomanes.

1.4.1 *Drogues injectables et VIH*

La population des UDI est la première touchée par le VIH dans un premier temps. Or, cette population a longtemps été stigmatisée et même criminalisée par les autorités, constat valable jusqu'à récemment, puisque ce n'est que depuis 2009 que le gouvernement vietnamien a cessé cette criminalisation en mettant progressivement en place des centres de substitution à la méthadone

10. Wolffers (2001, p.156)

pour les héroïnomanes dépendants.[11] En 1998, les UDI représentaient deux tiers des infections dénombrées. Le nombre de nouvelles infections par an parmi cette population s'est stabilisé à partir de 2000.

1.4.2 *Prostitution et VIH*

Concernant la population des travailleu·se/r·s sexuel·le·s (TS), ce n'est qu'à partir de 2003 que le nombre d'infections parmi celle-ci est devenu significatif et il n'a cessé d'augmenter depuis, augmentation qui se reflète par un discours, de plus en plus ancré, associant prostitution et VIH. Le travail de prévention s'adressant aux TS semble relativement plus efficace qu'envers les UDI. Cependant, le rapport 2010 de l'organisme international ONUSIDA[12] indique que « l'utilisation de préservatifs pendant des rapports sexuels tarifés est peu fréquente. » et que « Le commerce du sexe joue un rôle central dans les épidémies de la région. » C'est ainsi que peu à peu, la population devenue majoritaire parmi les cas de nouvelles infections au VIH est celle des clients des TS, population masculine, et très majoritairement hétérosexuelle.

Soulignons que les échanges sexuels tarifés font partie intégrante de la société traditionnelle : avant l'arrivée des Européens et la colonisation de la région, le commerce de services sexuels était toléré, car vu comme un moyen d'ascension sociale pour les femmes et leurs familles. Il a été montré (Andaya, 1998 ; Micollier, 2004a) à quel point la prostitution féminine en Asie du Sud-Est était préexistante à la mondialisation et à l'apparition d'un tourisme sexuel. Par exemple, ce dernier ne représente que 20% des échanges sexuels

11. C'est ce que note l'organisme *Sida Info Services* en 2014 : https://www.sida-info-service.org/?AIDS-2014-Le-sida-est-en-train-de

12. L'*ONUSIDA* est un programme de l'Organisation des Nations Unies créé le 1er décembre 1995 pour coordonner les efforts de lutte contre la pandémie de VIH/syndrome d'immunodéficience acquise (SIDA) à l'échelle mondiale. Le sigle anglophone est UNAIDS. Ce programme produit un rapport annuel recensant les données chiffrées et estimations pour chaque région du monde et une analyse de l'état de la pandémie.

commerciaux en Thaïlande, qui connaît pourtant un tourisme sexuel international très développé. La prostitution s'adresse dans une large majorité à la population locale. Au Viêt Nam, dans les années 1990, pendant la première décennie d'ouverture du pays au monde (après un siècle de colonisation, 30 ans de guerre et 10 ans de fermeture des frontières), la prostitution était pratiquée dans « presque tous les hôtels, restaurants, clubs, salons de beauté et de massage, bars, cafés, parcs, trottoirs, arrêts de bus, gares, et d'autres lieux tels que les quais des ports ou la plage. » (Lê, 1993b). A cette liste, Nguyễn (1997) ajoutait « les salons de coiffure, les salons de manucure et de beauté, [...] et les établissements à karaoké », Ngo (1993) parlait également de prostituées de luxe dans les villas et les établissements à 4-5 étoiles. Par conséquent, l'estimation était portée à 500 000 prostitué·e·s en 1998[13] (les chiffres officiels sont largement en deçà, mais aucune étude sérieuse n'avait pu établir de chiffres fiables[14]). Pourtant, la prostitution est officiellement illégale et le gouvernement entend lutter contre ce qu'il a qualifié de fléau social. Dans les faits, elle représente une manne financière pour de nombreux intermédiaires ainsi que pour la police, sous forme de corruption, très installée (par exemple, Walters, 2004 a mené une enquête auprès de mini-hôtels louant des chambres à des prostituées : la police ne fait pas de descente tant qu'ils paient à chaque fois que la police le requiert, ce qui aurait rapporté 42 000 $ par policier par an, selon ses estimations établies pour une circonscription).

1.4.3 *Évolution de l'épidémie du VIH dans la population vietnamienne*

A propos de l'épidémie de VIH, l'ONUSIDA fournit des chiffres sur le Viêt Nam à partir de 2001. Pour cette année-là, le nombre total de personnes

13. Khuât, 1998
14. Walters, 2004, p.82

vivant avec le VIH était porté à 140 000. En 2009 le nombre de nouvelles infections était estimé entre 15 et 37 000, portant le nombre de porteurs du VIH à 280 000, ce qui signifie que ce nombre a doublé en 8 ans. L'arrivée de l'épidémie de VIH tarde à être suivie par des mesures de protections, telles que l'utilisation de préservatifs, et surtout l'évolution des comportements pour que de telles mesures soient adoptées par la population. En 2005[15], seulement 68% des jeunes vietnamiens (de 15 à 24 ans, et de sexe masculin) ayant eu un rapport sexuel avec un·e partenaire non officiel·le au cours des douze derniers mois déclaraient avoir utilisé un préservatif lors de leur dernier rapport sexuel. (Les jeunes femmes étaient 0% à déclarer avoir eu un tel rapport.) La moyenne pour les 15-49 ans était de 58%, donc le discours de prévention semble avoir touché, malgré tout, une part plus importante de la population de moins de vingt-cinq ans. Pourtant cela reste encore très insuffisant pour endiguer la transmission d'infections par voie sexuelle. De même, seulement 2% des 15-24 ans avaient effectué un dépistage du VIH dans les douze derniers mois. Également 2% des 15-49 ans avaient fait ce dépistage.

1.4.4 *Accès aux traitements*

Blanc (2010) a montré que la médecine traditionnelle occupe une grande place dans les pratiques vietnamiennes, et même, jusqu'à l'arrivée des antirétroviraux, elle apportait une réponse efficace dans le traitement de la dépendance aux drogues injectables[16], voire dans la séroconversion de patients atteints du VIH. Mais l'interférence de l'État dans les pratiques de recours au médicament est de plus en plus importante et « les autorités décidèrent en 2003, d'une part, de garantir l'approvisionnement (dans sa continuité) en

15. Enquête démographique et de santé fournie par l'ONUSIDA
16. La pratique de partage de seringues entre usagers de ce type de drogues en fait une pratique des plus à risque de transmission du VIH

médicaments antirétroviraux (ARV), leur contrôle qualité, et, d'autre part, de mieux encadrer leurs prescriptions en s'orientant vers le développement de la production locale de médicaments antirétroviraux génériques. » Aujourd'hui, les patients vietnamiens, d'une part par le crédit accordé à la science occidentale dans les imaginaires, d'autre part encouragés institutionnellement par une volonté politique influencée par des enjeux de construction de l'identité nationale et de conquête du marché global, et enfin par l'accès facilité aux antirétroviraux, se tournent de manière quasiment systématique vers des tritérapies ARV, ainsi que des traitements post-exposition. Dans les campagnes et les régions éloignées des structures institutionnelles pourtant, la médecine traditionnelle continue à jouer un rôle majeur, et même « l'Organisation Mondiale de la Santé (OMS) encourage les pratiques de recours à la médecine traditionnelle pour suppléer l'absence de médecin ou de médicaments pour les soins de santé primaire. » Cela amène des contradictions du fait que l'État n'encourage pas ces pratiques, voire souhaite les éradiquer, du moins les standardiser.

2 UNE SOCIÉTÉ EN TENSION

Depuis l'ouverture du pays dans les années quatre-vingt-dix, le Viêt Nam fait face à des tensions entre le désir de modernisation et le maintien des traditions. La moitié de la population a grandi durant cette politique d'ouverture et a soif de développement, mais l'attachement aux traditions reste fort, et joue un rôle essentiel dans la construction de la société vietnamienne.
L'accélération des évolutions sociétales à une échelle globale est d'autant plus remarquable dans le contexte d'un pays en voie de développement, qui voit sa société se moderniser à très grande vitesse. Les tensions entre tradition et

modernité sont par conséquent exacerbées et il est intéressant d'étudier plus en détail comment la société et les individus composent avec ces deux extrêmes en se les appropriant.

2.1 *Traditions et tabous : un discours lacunaire sur la sexualité*

La société vietnamienne est historiquement constituée d'un cadre d'interactions entre les individus très normé. Fortement empreinte de confucianisme, la culture sino-vietnamienne implique un contrôle fort de la jeunesse. Ce contrôle est garanti par les grandes institutions que sont l'État, l'école et la famille. D'autre part, l'émancipation sexuelle est arrivée en même temps que l'épidémie de VIH/SIDA, avec l'ouverture du pays, ce qui implique la nécessité de la production de discours à propos de la sexualité. Ce discours est donc fortement empreint de la vigilance à maintenir face à l'épidémie qui a été qualifié de fléau social, au même titre que la prostitution. Du côté des institutions évoquées, l'école ne présente que quelques faits biologiques, et laisse la famille assumer seule l'inculcation de valeurs morales. Avec un fonctionnement hérité du confucianisme, celle-ci joue un rôle fondamental dans la structuration de la société ainsi que le contrôle de la vie privée. L'appareil d'État quant à lui, avec son système de parti unique, sa propagande nationaliste et son contrôle quasi total des médias classiques, continue d'imposer un maintien fort des traditions en se chargeant de rappeler quotidiennement à tout un chacun de veiller à se tenir loin des fléaux sociaux. Restent les organisations de jeunesse, également très chapeautées mais qui sont un cadre ouvrant la possibilité d'une vraie transmission entre pairs.

2.2 *Modernisation et accroissement de l'écart entre les générations*

Nous nous intéressons ici plus particulièrement à la population au sein de laquelle ce phénomène de tensions se fait le plus ressentir : les adolescents et les jeunes adultes. Autrement dit, la population en âge de s'imposer en tant qu'individu, de prendre des décisions pour elle-même et à qui s'offrent tous les possibles. Population qui est, par ailleurs, la plus utilisatrice des nouveaux modes de communication introduits avec Internet. Population qui, enfin, représente une large part de la population totale du pays, sachant qu'un tiers de la population a entre 15 et 30 ans, ce qui représente de l'ordre de 30 millions d'individus. Leurs sources principales d'information ne sont ni la propagande, ni le carcan familial, ni l'école. Parler de sexualité à l'école par exemple met les professeurs dans une posture d'inconfort social vis-à-vis de leurs élèves, et plus généralement les adultes vis-à-vis des adolescents, ce qui est inacceptable dans les sociétés confucéennes.

2.2.1 *L'éducation par les pairs*

La défiance vis-à-vis d'un discours trop policé pousse à se tourner vers d'autres sources d'information. Un type de structure alternatif existe au Viêt Nam : les clubs de jeunesse où se pratique une éducation par les pairs, une transmission horizontale des savoirs. La parole y est facilitée par le fait que les personnes référentes n'ont que quelques années d'expérience de plus, sont parcourues par les mêmes interrogations, et présentent une proximité créant la confiance. L'avantage de l'éducation par les pairs est qu'elle s'affranchit des barrières hiérarchiques entre celui qui transmet l'information et celui qui la reçoit.

Au niveau international, le concept de « peer education » est très utilisé dans le domaine de la prévention du VIH/SIDA à partir des années 2000, en

suivant l'idée que le message est plus à même d'être transmis si ce sont des personnes qui ont le même âge et la même culture qui le transmettent. Par exemple, le réseau international Y-PEER[17] (Youth Peer Education Electronic Ressource) lancé par l'UNFPA (United Nations Population Fund) est un réseau de jeunes autour des questions sur les droits et la santé sexuelle et reproductive, qui a pour objectif d'être au plus proche des jeunes pour les aider à s'approprier les connaissances et les capacités nécessaires pour faire des choix sains. Au Viêt Nam, l'éducation par les pairs fait partie depuis des décennies du mode d'action des clubs de jeunesse. Ces *organisations de masse* jouent un rôle important dans la société vietnamienne en ce qui concerne la prise en charge de la jeunesse. Elles permettent une socialisation autour d'une communauté de partage. Elles sont vouées à la transmission de connaissances, notamment sur les sujets qui touchent à l'intime, en échappant au problème de l'écart générationnel et en maintenant la discrétion socialement requise. Ces structures ont été étudiées par Blanc à propos de l'accès à l'information dans le domaine de la sexualité :

> « *In our survey about sexuality and AIDS in a questionnaire, we asked 407 young men an women aged between fifteen and twenty-nine living in Ho Chi Minh City about their sources of information about sexuality. They told us that neither school nor their parents were their main sources. […] Books and printed matter are more convenient for an individual use and appealed to a perceived need for privacy and intimacy. There is no difference between boys and girls.* » (Blanc, 2004)

Plutôt que vers l'école ou leur parents, les jeunes citadins interrogés (de 15 à 29 ans) préfèrent se tourner vers les livres et les documents papiers[18], en raison de l'intimité et la discrétion que ces médias permettent.

17. http://www.ypeerinaction.org/index.php?option=com_content&view=article&id=1&Itemid=2

18. 20% des interrogés disent que les documents imprimés constituent leur principale source d'information concernant la prévention du VIH, 12 % placent les organisations de jeunesse en source principale.

2.3 *Utilisation d'Internet et des TIC*

Les données de l'étude précitée ont été recueillies de 1999 à 2002, ce qui explique qu'il n'est pas mention d'Internet, mais les moyens de communication par le web sont dans la directe continuité de ce qui y a été présenté. D'abord, sans parler de l'interactivité permise par les formes plus récentes du web, Internet ne fait que démultiplier les possibilités d'accès à des informations écrites dont la consultation, si elle ne passe plus par le papier, offre la même intimité que les documents imprimés. Ensuite, les échanges rendus possibles par le Web 2.0 conservent le paramètre de discrétion, du fait de l'anonymat qu'un avatar virtuel confère à l'internaute.

En 2002, le pourcentage d'internautes dans la population au Viêt Nam restait anecdotique (moins de 2% de la population , la France en était à 30%) mais à partir de cette même année, il a augmenté de 100% par an pour atteindre 12% de la population en 2005 et n'a cessé de grandir depuis (20% en 2007, 30% en 2010, 50% en 2015).

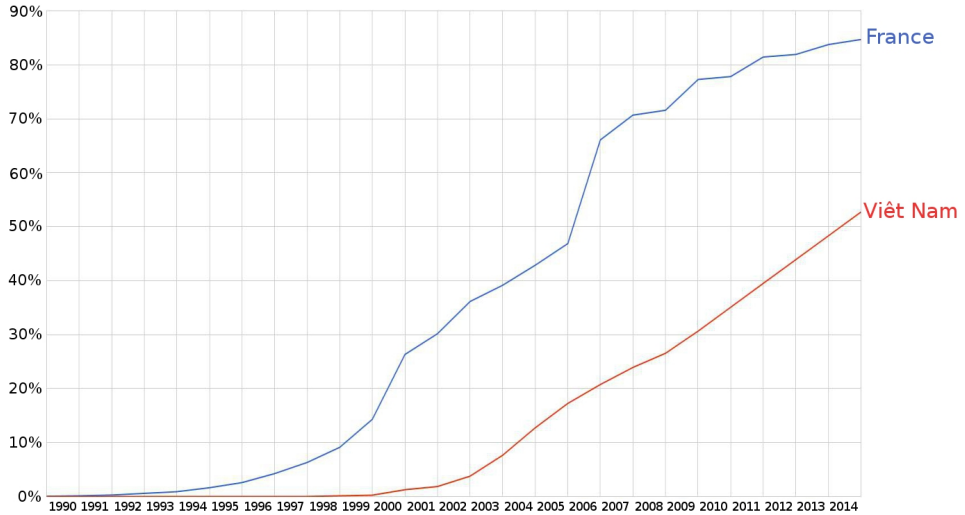


Fig. 1.2 : Pourcentage d'utilisateurs d'Internet parmi la population nationale (Source : Banque Mondiale)

D'autre part, il est important de prendre en compte l'utilisation de la téléphonie mobile, dans les pays en développement où elle a rapidement et massivement été adoptée, par comparaison aux pays qui avaient développé des lourdes infrastructures de télécommunications avant l'avènement de la téléphonie sans fil. Au Viêt Nam, le nombre de possesseurs de téléphones mobiles a été multiplié par 25 en 7 ans : de 2004 à 2011 il est passé de 5 millions à 125 millions d'abonnements[19], pour moins de 88 millions d'habitants, ce qui porte le nombre d'abonnements à plus d'un par habitant (cf. la figure I.3).

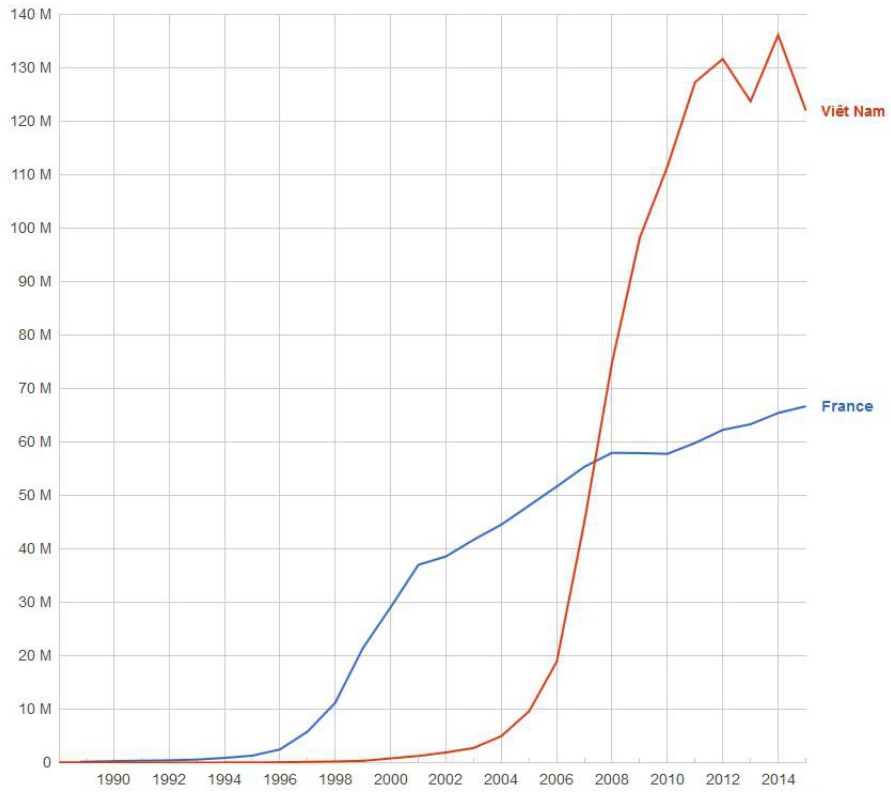


FIG. I.3 : Abonnements à la téléphonie mobile cellulaire (Source : Banque Mondiale)

19. Source des données : Banque Mondiale (https://www.google.fr/publicdata/explore?ds=d5bncppjof8f9_&ctype=l&met_y=it_cel_sets&idim=country:VNM:FRA)

L'engouement pour des accès à l'information alternatifs (ou du moins supplémentaires) aux canaux contrôlés par les autorités traditionnelles est donc perceptible comme une tendance de fond. Ces alternatives sont en fait dans la continuité des canaux existants avant Internet, comme l'éducation par les pairs[20] qui permet de s'affranchir des barrières hiérarchiques entre celui qui transmet l'information et celui qui la reçoit. C'est ce qui se retrouve dans les forums. L'accès aux technologies de l'information et de la communication (TIC) a accéléré cette évolution dans la transmission de l'information concrète et l'accès à l'information quel que soit le sujet, même les plus délicats à aborder au moyen des autres biais.

Contrairement aux canaux traditionnels de diffusion de l'information[21], il reste une divergence d'accès à Internet entre campagnes et villes. Après une décennie de boom des cybercafés dans les zones urbanisées pendant les années 2000, ils ont presque disparu, du fait de l'arrivée des abonnements privés dans les foyers et du déferlement d'appareils personnels sur le marché. De même que pour la téléphonie sans fil, l'adoption de la transmission de données sans fil se fait à grande vitesse. La mise à disposition gratuite d'Internet par wifi se généralise dans les lieux publics (par exemple les gares, et bientôt les trains), tous les bars, cafés et restaurants fournissent un accès libre, le plus petit hôtel de la moindre bourgade reculée en zone montagneuse proposera lui aussi un accès au wifi. Mais cette offre s'adresse aux touristes et à la classe moyenne urbaine en majorité. Les minorités ethniques et les populations rurales pauvres ne sont donc pas encore concernées par les données que nous

20. cf. le paragraphe 2.2.1.
21. Voir par exemple le paragraphe **??**.

étudierons.[22]

3 LE DISCOURS INSTITUTIONNEL VIETNAMIEN EN QUESTION

3.1 *L'état tout-puissant*

3.1.1 *La politique du parti unique*

L'organisation même du découpage et de l'administration territoriale influe sur les caractéristiques du discours officiel. De même que le système politique : avec un seul parti autorisé, l'unité du discours ne peut que faiblement être remise en question. Ainsi, le discours émis par le Comité Central est retransmis à tous les échelons d'un système très hiérarchisé : du gouvernement central au conseil de quartier (ou de commune pour les zones rurales), en passant par le comité de la province ou de la ville et celui du district[23], chaque niveau transmettant les directives de l'échelon supérieur, tout en lui rendant des comptes. Le discours est donc unifié à tous les échelons et vise à toucher tous les individus de la société. La volonté d'amener cette forme d'éducation à l'intégralité de la population est héritée de la doctrine communiste. Elle vaut pour tous les domaines et a recours à tous les leviers traditionnels de ladite doctrine, notamment la propagande (cf. le paragraphe 3.2.2).

Le discours est donc unifié et constant, même si une lente évolution est observable : les instances dirigeantes s'efforcent de prendre en compte les phénomènes de société, lorsque ceux-ci prennent de l'ampleur et menacent

22. A noter cependant : le réseau 3G est qualifié de très bon en 2015 sur un forum de voyageurs (https://www.ozbargain.com.au/node/204018#comment-2923008), et les cartes prépayées peu chères. Reste toutefois à posséder un appareil, ce qui n'est pas le cas majoritaire des populations rurales pauvres (qui en outre ne maîtrisent pas toujours le vietnamien écrit), mais la situation devrait évoluer en ce sens.

23. http://www.ambafrance-vn.org/L-organisation-territoriale-du

d'opérer une brèche dans l'organisation de ladite société, voire d'ébranler ses fondements. Le fait que le discours institutionnel évolue en réaction aux évolutions de la société, lui confère toujours un décalage, plus ou moins sensible, par rapport à celle-ci. Ce phénomène est intrinsèquement caractéristique de tout discours institutionnel et n'est pas propre au contexte vietnamien, cependant en ce qui concerne le discours officiel vietnamien ce décalage est particulièrement remarquable, ce que relèvent Papin et Passicousset (2010) en expliquant que « *le moindre changement* [prend au Viêt Nam] *l'allure d'un bouleversement.* »

3.1.2 *Le verrouillage des médias*

Nous l'avons vu au paragraphe 1.2, la presse indépendante a beaucoup de mal à se faire une place dans le paysage médiatique vietnamien. L'État exerce l'équivalent d'un monopole sur les médias : chaînes de télévision, journaux, radio sont sous tutelle de l'État. Si le discours est laissé relativement libre en ce qui concerne les sujets de société, d'autres sujets sont plus sensibles, notamment tout ce qui touche à la politique, les avis émis à propos du gouvernement, la critique des instances étatiques. Et même pour ce qui est des sujets de société, le phénomène d'auto-censure puissamment ancré n'est pas à négliger. En novembre 2014, la ligne politique au plus haut niveau, en la personne du Vice-Premier Ministre, a semblé vouloir présenter les gages d'une ouverture journalistique, en parlant d'adapter la loi sur le journalisme, âgée de 15 ans, aux évolutions entraînées par le numérique, mais selon Horizons Médiatiques[24], cette évolution servirait surtout à ce que le webjournalisme

24. Horizons Médiatiques est un site de travaux d'étudiants de Lyon-2 analysant les évolutions du journalisme dans le monde.

n'échappe pas à la mainmise de l'État sur la presse[25].

Dans un tel contexte, il n'est pas évident pour la population d'avoir accès à une information complète sur des sujets pour peu que ces derniers s'éloignent de ceux que la ligne politique cherche à mettre en avant, toujours dans l'idée de sauvegarder la bonne marche de la société.

3.2 *La lutte contre les fléaux sociaux*

La bonne marche de la société est mise en péril par un certain nombre de menaces, qualifiées de « fléaux sociaux » (*tai họa xã hội*), et contre lesquels l'institution entend lutter, en se posant en gardienne des valeurs et des fondements de la société.

3.2.1 *Les fléaux sociaux*

Les deux fléaux sociaux faisant l'objet du plus d'attention actuellement sont le VIH/SIDA et la prostitution. Mais la corruption, le proxénétisme, la toxicomanie, les jeux d'argent, les vols, également l'homosexualité sont visés par cette qualification. La liste, bien qu'elle n'existe pas de manière formelle, s'agrandit ou se réduit en suivant les évolutions de la société. Lorsque, pour l'une des problématiques incluses dans la liste, l'aspect discriminatoire d'une telle classification devient trop saillant, cela pousse les autorités à cesser de catégoriser en fléau social ladite problématique, et donc à adapter la ligne politique suivie pour y répondre. Par exemple, depuis 2002 existait le *Comité national de prévention et de lutte contre les fléaux de la drogue, la prostitution et le SIDA*. En 2007 celui-ci est devenu le *Comité national de prévention et de lutte contre le SIDA et de prévention et de lutte contre les fléaux de la drogue et de la pros-*

25. http ://horizonsmediatiques.fr/2015/03/la-douce-illusion-dune-ouverture-journalistique-vietnamienne/

titution[26]. Le SIDA est donc sorti des fléaux sociaux dans le discours officiel, à des fins de réduction de la discrimination. La qualification en fléau social revient à criminaliser une catégorie de population, et orientera la réponse vers la répression plutôt que vers un soutien qui permette à la population concernée de trouver des solutions pour en sortir. Les directives de lutte contre le VIH/SIDA ont effectivement évolué à partir de ce changement au niveau institutionnel (cf. le paragraphe 3.2.3).

Un travail équivalent reste à faire en ce qui concerne les TS et les UDI. Ces derniers étant officiellement criminalisés jusqu'en 2009. Depuis des années, de nombreuses organisations internationales exigeaient la fermeture de centres de réhabilitation où étaient détenus TS et UDI. En juillet 2012 le pouvoir a cessé l'envoi de TS dans de tels centres et en a libérées 9000 l'année suivante. Pour autant la prostitution est toujours illégale, et les personnes interpelées doivent encore payer une amende. Les centres ont évolué en centres de désintoxication à la méthadone, destinés à accueillir les héroïnomanes.

Concernant l'homosexualité, elle n'a cessé d'être considérée comme une maladie mentale qu'en 2011. Le 1er janvier 2015, la loi interdisant les unions entre personnes de même sexe a été abolie (ce qui permet au gouvernement vietnamien de moderniser son image auprès de la communauté internationale), sans pour autant qu'aucune reconnaissance légale ne soit accordée aux mariages de ces couples.[27]. De plus, la discrimination n'est pas interdite. Les déclarations officielles considérant l'homosexualité comme une maladie sont encore régulières, et il n'est pas rare d'entendre dans les médias qu'elle est un comportement déviant incompatible avec la bonne moralité et les coutumes ancestrales du Viêt Nam.

26. http://www.vaac.gov.vn/Desktop.aspx/Noi-dung/He-thong-to-chuc/Uy_ban_quoc_gia_phong_chong_AIDS_va_phong_chong_te_nan_ma_tuy_mai_dam/
27. http ://360.ch/blog/magazine/2015/01/le-vietnam-entrouvre-la-porte-aux-mariages-gay/

3.2.2 *La propagande*

Héritée de la doctrine communiste, la propagande est très implantée dans le pays, passant par tous les canaux accessibles : des journaux aux fresques, banderoles et panneaux – dignes des campagnes publicitaires en Occident, seulement avec d'autres codes graphiques – en passant par les hauts-parleurs (hérités de la guerre, lorsqu'ils prévenaient la population d'un bombardement imminent), diffusant quotidiennement les dernières directives du Comité Central ou les dernières consignes du quartier, entre 5h00 et 7h00 et entre 17h et 19h chaque jour sans exception, dans l'ensemble du pays, aussi bien dans les grandes villes scintillantes de modernité que dans les campagnes les plus reculées, y compris les îles les plus excentrées des organes du pouvoir[28].

Les fléaux sociaux font bien évidemment partie des objets du discours de la propagande, mais elle sert également aux autorités pour communiquer sur d'autres sujets, tels quel les bons comportements à adopter pour la sauvegarde de l'environnement, le rappel de l'obligation du port du casque aux conducteurs de deux-roues, les bonnes pratiques d'usage des antibiotiques, l'hygiène, la limitation des naissances, le don du sang, etc. Elle propose (ou impose) des séances de sport, de la musique patriotique, communique sur les dernières actions des services de l'État, etc. Le tout dans une vision d'éducation des masses, telle qu'elle fut théorisée par les régimes communistes.

3.2.3 *Les efforts pour diversifier les moyens de lutte*

Les autorités prennent de plus en plus conscience que le tout répressif n'est pas une bonne solution pour éradiquer les fléaux sociaux. C'est pour-

28. Pour une vision actualisée de leur perception, lire l'article traduit du journal Thanh Niên par Courrier International : http://www.courrierinternational.com/article/2010/11/25/cessez-de-nous-casser-les-oreilles

quoi elles s'efforcent de construire d'autres approches. En ce qui concerne le VIH/SIDA, la prostitution et la toxicomanie, avant 2002 existaient :

- un comité national de prévention et de lutte contre le SIDA,
- un comité national de prévention et de lutte contre la drogue, et
- un département directeur du Gouvernement sur la prévention et la lutte contre les fléaux sociaux.

Ces trois organes ont fusionné en 2002 en *Comité national de prévention et de lutte contre les fléaux de la drogue, la prostitution et le SIDA*, et cinq ans plus tard, celui-ci a été rebaptisé *Comité national de prévention et de lutte contre le SIDA et de prévention et de lutte contre les fléaux de la drogue et de la prostitution*, toujours en place en 2010[29].

Cette même année, les autorités ont organisé un mois d'action nationale pour la prévention et la lutte contre le SIDA : dans tout le pays, à tous les échelons, des films, des reportages et des spots, ainsi que journaux papiers et électroniques ont été diffusés, des discussions, conférences ont été organisées, des banderoles, affiches et brochures ont été produites. De plus, des programmes d'activités de clubs de prévention et de lutte ont été organisés, ainsi que de la propagande ambulante (hauts-parleurs et cinéma), et des concours de prévention et de lutte. Ces actions ont comptabilisé 1 900 000 participants.

D'autre part sont organisées des distributions gratuites de préservatifs (mais seulement à certains publics considérés en grand risque de contamination), de cadeaux aux personnes atteintes de la maladie, et aux personnes travaillant dans la prévention et la lutte.

Il existe aussi un *Comité National de Prévention et de Lutte contre le SIDA*, qui fournit sur son site internet[30] des informations très actualisées, par exemple

29. http://www.vaac.gov.vn/Desktop.aspx/Noi-dung/He-thong-to-chuc/Uy_ban_quoc_gia_phong_chong_AIDS_va_phong_chong_te_nan_ma_tuy_mai_dam/
30. http://tiengchuong.vn

à propos des aides (financières et autres : stabilité, emploi, insertion sociale) proposées aux séropositi·f/ve en situation précaire.

Tous les échelons sont mobilisés pour atteindre également les régions essentiellement peuplées d'ethnies minoritaires. Les instances centrales ont demandé aux localités d'introduire la lutte contre ces trois fléaux dans leur programme socioéconomique annuel.

En matière de stupéfiants, il s'agit toujours de lutte contre ce qui est considéré comme de la criminalité, mais des actions sont menées en dehors de la répression : ramassage et destruction de seringues usagées, mise à disposition de seringues stériles, mise en place de centres de désintoxication par la méthadone, par exemple.

Le gouvernement communique sur sa volonté de multiplier les actions, les formes d'action et l'évaluation des résultats, de réduire la discrimination, tout en maintenant une ligne volontariste de lutte (incarnée par les termes guerriers *phòng / chống* (« se défendre contre un danger, un ennemi »)). Pourtant l'épidémie continue de progresser, ce qui laisse à penser que des avancées doivent encore être faites concernant les réponses à apporter.

4 CONCLUSION

Nous l'avons vu, le contexte vietnamien favorise la production d'un discours institutionnel lissé et homogène. Vis-à-vis de l'épidémie de VIH, ce discours suit une ligne résumée par l'expression « en lutte contre le SIDA » (*phòng chống AIDS*), discours diffusé par tous les canaux de la propagande. D'autre part, l'accès à Internet, de plus en plus généralisé dans tout le pays, permet à la population d'accéder à un discours médical international, immédiatement actualisé concernant les dernières avancées prophylactiques dans

le domaine du VIH/SIDA. Discours médical qui suit également une normalisation à l'échelle globale. Enfin, la prise de conscience de la complexité des problèmes posés par l'épidémie à la société a fait émerger une volonté de diversification des méthodes de lutte. L'information sur les risques liés à l'épidémie atteint donc l'ensemble de la population. Cependant, la question de l'efficacité n'est pas résolue, car nous avons affaire à des processus sociaux complexes : « L'éducation est un processus cognitif, le SIDA est un comportement »[31]. Selon Blanc (2004), l'éducation sexuelle est plus à même de modifier les connaissances et les attitudes, mais moins efficace pour modifier les comportements. Elle ajoute en outre qu'il est difficile de mesurer l'impact de l'éducation sexuelle sur les comportements des adolescents[32]. Encore faut-il avoir accès aux comportements des adolescents et des jeunes adultes, a fortiori lorsqu'il s'agit des domaines les plus intimes. C'est en visant cet objectif que nous estimons utile d'explorer les forums de discussion sur Internet voués à ces sujets. En partant de l'analyse des discours institutionnels, nous souhaitons décaler le point de vue afin de mettre en regard les discours institutionnels et ceux nés avec le développement des TIC. Ces dernières ayant été adoptées dans un contexte de mondialisation, leur confrontation avec les discours institutionnels est valable autant dans des pays en développement que dans les pays développés. Avant d'entériner des distinctions en fonction de contextes sociaux ou linguistiques, des bases communes peuvent être établies.

31. Hochhauser et Rothenberger, 1992, p.103
32. Blanc, 2004, p. 253

CHAPITRE II

État de l'art

1 Points de vue sur les discours institutionnels

Comment définir le discours institutionnel ? Pour commencer, soulignons qu'il n'y a pas un seul discours institutionnel, mais des discours spécifiques à un domaine. Dans chaque domaine s'établit une institution, dont le champ d'application se restreint à celui-ci.

1.1 Le rôle de l'institution dans la construction des connaissances

L'institution a pour vocation d'établir des normes qui constituent le cadre au sein duquel se joueront les interactions entre les différents acteurs du domaine. Ce cadre prend la forme d'une structure hiérarchisée, et définit des critères déterminant les détenteurs de l'autorité. Dans le processus de construction des connaissances, l'autorité s'établit par la constitution d'une expertise. Cette expertise est ensuite reconnue par les autres acteurs du domaine, qui se soumettent à cette influence de pensée, lorsque la relation de confiance est en place. L'activité institutionnelle consiste donc également à établir et maintenir une relation de confiance vis-à-vis des règles qu'elle élabore. La confiance

est construite par l'activité institutionnelle en garantissant une maîtrise des connaissances, ainsi qu'en établissant des normes nécessaires à une formalisation du savoir. Nous sommes donc en présence de deux pré-requis à la circulation de l'information : la structure hiérarchisée entre les acteurs, et la structure de formalisation des connaissances. L'évolution perpétuelle de ces dernières est en contradiction avec la fixation de normes et oblige l'institution à ne jamais cesser son activité d'évaluation et d'adaptation, sous peine de voir son autorité remise en question.

Des processus de reconfiguration sont à l'œuvre parmi les détenteurs de cette autorité, qu'il nous importe de mettre en lumière.

1.2 *Analyse du discours institutionnel*

L'AD a examiné le discours institutionnel tantôt sous l'angle discursif, tantôt sous l'angle sociologique, tantôt en analysant les rapprochements possibles entre les deux, et ce que chacun peut apporter à l'autre.

1.2.1 *Une fonction de production de sens commun*

En s'appuyant sur le postulat que le concept de discours est synonyme de lien social, Sarfati (2014) estime que l'AD classique, en ne traitant pas le versant discursif de l'activité institutionnelle, a laissé de côté sa dimension sociologique, or l'activité énonciative se trouve configurée par des paramètres socio-discursifs, autant qu'elle les configure. Selon lui, c'est la problématique du sens commun qui permet de relier l'analyse sociologique et l'analyse discursive. L'activité institutionnelle, en tant qu'elle est un élément qui *fait lien entre les individus*, nous intéresse par sa production de sens commun. Selon la définition donnée par Sarfati, les institutions sont des « dispositifs énoncia-

tifs collectifs » et « d'instanciation de normes ». Les individus impliqués dans l'activité institutionnelle voient donc leur conduite modalisée, à la fois dans la production discursive et dans le respect des normes. Ils occupent ainsi la place de *sujets-acteurs* de l'institution. Étant donné que la production de sens est un processus fait d'instanciation, de reformulation, d'innovation, et de remaniement, « le partage du sens, dans le cadre et à partir des institutions de sens, constitue un éclairage rigoureux sur les mécanismes de circulation des énoncés, la formation des évidences collectives, compte tenu des contraintes que font peser sur les sujets-acteurs des dispositifs de régulation et de répartition de socialisation du sens. » Ainsi, les sujets-acteurs des dispositifs institutionnels sont pris dans un rapport de domination, qu'elle soit consensuelle (les sujets nouant des liens de subordination) ou coercitive (les sujets nouant des liens de subversion), avec l'institution.

Pour Sarfati, par la production de mécanismes discursifs normant les rapports entre les individus et l'institution, ou des individus entre eux, l'activité institutionnelle entend regrouper des individus autour d'une communauté de sens, spécifique à chaque domaine de pratique.

1.2.2 *Une fonction de figement des connaissances et d'autorité prescriptive*

Pour Oger et Ollivier-Yaniv (2003), le discours institutionnel est défini au sens strict comme « le discours produit officiellement par un énonciateur singulier ou collectif qui occupe une position juridiquement inscrite dans l'appareil d'État, qu'il soit fonctionnaire ou représentant politique ». Est ensuite introduit un élargissement des discours concernés par l'analyse, pour comprendre également « le discours produit par les mêmes énonciateurs en dehors des contextes officiels. » Le corpus à l'étude dans ce travail, ainsi que dans d'autres (Oger et Ollivier-Yaniv, 2006, notamment) porte sur le discours

du Ministère de la Défense. Des problématiques spécifiques à ce type de discours institutionnel sont donc plus particulièrement étudiées : comment produire un discours cohérent, lissé, gommer les dissensions internes, avoir réponse à tout face aux discours adverses, etc. Le discours institutionnel dans un tel contexte fait face à une méfiance, voire une hostilité de la part des récipiendaires (journalistes, opinion publique, opposants politiques).

L'analyse de la fabrication des discours institutionnels a permis d'introduire la différence entre le discours instituant et les discours institutionnels. Plutôt que de déclarer que les discours institutionnels auraient un caractère meta-prescriptif dans le maintien de leur propre cohérence, Oger et Ollivier-Yaniv (2006) font la distinction entre le discours instituant, qui joue ce rôle de cadrage, de lissage, de « gommage des formes de diversités et d'hétérogénéité » et qui est trop souvent considéré comme le seul discours institutionnel, et les productions discursives des acteurs de l'institution, qui correspondent à une pluralité de situations. Le discours instituant a donc une fonction de regroupement d'une communauté autour d'une identité énonciative, construite à base de prescriptions internes de contraintes discursives, ainsi que de prescriptions externes adressées au public récipiendaire. Un discours qui se veut donc stable et par conséquent peu renouvelé, cadre d'autorité prescriptive dans lequel une variété de discours institutionnels sont produits.

Ces travaux ont donc permis un relâchement définitoire du concept de discours institutionnel en embrassant une variété de pratiques langagières qui ressortissent à divers appareils (partis politiques, syndicats, etc.) et non plus seulement à l'État (Krieg-Planque et Oger, 2010). D'une approche donnant le primat à l'énonciateur et aux conditions d'énonciation, les recherches se sont acheminées vers des acceptions déliées du cadre juridique et davantage déterminées par les notions d'autorité prescriptive et de sanction. Le dis-

cours institutionnel apparaît dès lors élargi en termes de pratiques sociales et en conséquence, plus varié en termes de genres textuels.

Les observations faites par Oger et Ollivier-Yaniv (2003, 2006) étaient menées dans le domaine militaire. En ce qui concerne le domaine médico-sanitaire, qui nous intéresse plus particulièrement ici, la contrainte de production d'un discours homogénéisé est moins forte ; il n'en reste pas moins que les discours institutionnels doivent répondre aux attentes d'un public toujours plus exigent et qu'ils conservent un rôle d'autorité prescriptive. Nous retrouvons les phénomènes de stabilisation des énoncés et d'obéissance à des régularités réduisant la diversité des énoncés possibles, même si la soumission à un discours instituant est moins contrainte.

Il est important de noter dès à présent que ces phénomènes de figements, d'expressions stabilisées sont reproduits dans les discours non institutionnels, lorsque les publics à qui s'adressent les discours institutionnels prennent à leur tour la parole, par des phénomènes d'appropriation. La diversification des types de prise de parole dans l'espace public entraîne un renouvellement des problématiques dans la légitimité du discours institutionnel. La mise en question de cette légitimité est évoquée par Krieg-Planque et Oger (2010), et fait l'objet des développements de la suite de ce chapitre.

L'AD a examiné le discours institutionnel sous sa facette discursive, sociologique ou encore juridique, en s'intéressant tantôt aux procédés de production de sens par l'activité institutionnelle, ou bien aux rapports des individus aux normes, ou encore à l'évolution des lois. Sans avoir la prétention de remettre en cause l'un ou l'autre de ces angles d'analyse du discours institutionnel, nous nous proposons de l'éclairer sous un autre angle encore, celui des marges, voire de l'extériorité, en déplaçant le point de vue au niveau des discours qui se font en dehors de l'institution, des discours pourtant produits

par des parties prenantes du domaine, mais qui ne détiennent a priori pas l'autorité sur la production du discours, et qui malgré cela influent sur son évolution.

2 Reconfigurations de l'autorité dans la construction des connaissances

Dans le but de questionner la construction des connaissances à l'heure du numérique, nous souhaitons étudier comment les discours institutionnels sont aujourd'hui mis en confrontation, comment Internet contribue à reconfigurer la notion d'expertise et donc d'autorité dans la construction des connaissances, que nous appelons autorité gnoséologique. Pour cela, nous souhaitons nous appuyer sur plusieurs angles d'analyse des phénomènes sociaux et humains. Nous étudierons ce qu'en disent respectivement l'AD et la sociologie.

2.1 *Autorité gnoséologique selon l'analyse du discours*

En dehors des domaines politiques ou militaire, cette reconfiguration de l'autorité prescriptive est perceptible dans tous les domaines, que ce soit par exemple dans la culture (Saunier et al., 2014), ou la santé, par la concurrence et la complémentarité entre une prescription verticale traditionnelle et une prescription horizontale des médias sociaux, dans le contexte de porosité de plus en plus forte entre les professionnels et un public ayant développé une expertise. La généralisation des usages d'Internet a rénové la notion d'autorité pour y inclure notamment, des conditions formelles de production (genre du document) ou de qualité des contenus (sources, éditeurs, etc.) (Broudoux, 2007).

Nous entendons le concept d'autorité telle qu'il est défini par Wilson (1983), inventeur de la notion d'autorité cognitive, entendue comme « une relation d'influence de pensée, impliquant au minimum deux personnes, l'une accordant à l'autre sa confiance parce qu'elle maîtrise un domaine spécifique de compétences ». Cette définition prend en compte le fait qu'une institution de sens coïncide avec un domaine de pratique spécifique, et qu'au sein de la communauté de sens concernée, la partie qui détient l'autorité est détentrice d'une expertise, reconnue par toutes les parties, à laquelle celles-ci se soumettent.

Avec les mutations entraînées par la généralisation d'Internet, la sanction éditoriale se voit complexifiée et n'est plus réservée aux seules autorités traditionnelles d'un domaine. Les textes institutionnels que ces autorités produisent sont concurrencés, et parfois disqualifiés, par les textes non institutionnels, que nous qualifierons d'informels, produits sur le web social. Ce dernier opère une remise en cause radicale des anciennes expertises, des institutions et de toutes les « positions acquises », et permet l'émergence, la montée en puissance de nouvelles expertises (Serres, 2010). Certains forums de discussion, notamment, parce qu'ils sont régis par des procédures de modération exigeantes, ne sont pas les avatars modernes des conversations dites du café du commerce auxquels on voudrait souvent les réduire ; ils sont des lieux d'élaboration et de co-construction de savoirs experts. Les forums reposant sur le partage d'expériences individuelles, la mise en commun de paroles singulières relèverait d'une *sagesse des foules* (*Wisdom of Crowds*, Surowiecki, 2004), qui permettrait d'accorder une valeur fiable à la connaissance construite par la multitude. Ainsi, l'émergence de ces moyens d'expression entraîne par corollaire l'émergence de nouvelles autorités.

Concernant l'évaluation de l'information, notons que l'évolution des rapports d'autorité sous l'influence d'Internet et du web social amène à constater

une inversion entre l'autorité et la notoriété[1]. C'est ce que remarque Serres :

> « dans l'ancien modèle, c'était d'abord l'autorité, la compétence, l'expertise... qui conféraient une certaine notoriété, le nouveau régime d'autorité cognitive, en émergence sur le web, inverse le rapport et c'est la notoriété qui semble conférer désormais l'autorité, la compétence venant ensuite. La circulation massive des énoncés, la notoriété, le buzz, deviennent de plus en plus les sources de l'autorité, de la crédibilité. Ce qui pose évidemment problème pour l'évaluation de l'information. »

Nous souscrivons à cette mise en garde et le propos présent n'est pas de retirer le rôle d'autorité aux institutions traditionnelles pour accorder un crédit aveugle à la multitude d'expressions d'individualités sur Internet. Nous conservons donc une posture critique à l'égard de cette expression, mais postulons qu'il est indispensable d'exploiter ces ressources nouvelles, en vue de développer l'expertise nécessaire à l'appréhension de la construction, la transmission et l'appropriation des connaissances dans leurs formes actuelles.

Dans l'objectif de mieux saisir les reconfigurations de la prise de parole institutionnelle et les mutations que peut entraîner le numérique sur les discours, nous proposons de reprendre les trois variations discursives du sens commun définies par Sarfati (2008) :

- le canon désignant le discours instituant,
- la vulgate désignant le discours transmis par les acteurs de l'institution dans un cadre moins formel, et
- la doxa désignant le discours communément répandu et qui ne fait que rarement référence au canon.

C'est ce que fait Longhi (2014) en les appliquant au contexte de Twitter. Longhi démontre que certains acteurs de la doxa « qui bénéficient d'une légiti-

1. A propos des technologies d'évaluation de l'autorité des sources du web, voir Lauf (2014).

mité et d'une audience conséquentes, peuvent servir, pour certains lecteurs, de « transmetteurs » du discours canonique, et acquérir le statut de vulgate. » Ajouté à ce phénomène le fait que les trois états de discours sont concomitants, en raison du caractère instantané du web, cela entraîne « parfois le brouillage de la légitimité des énonciateurs face à l'institution. »

2.1.1 *Application au domaine médico-sanitaire*

Nous appliquons les trois variations discursives de Sarfati au domaine médico-sanitaire, moyennant quelques aménagements. Tout d'abord, nous devons introduire la dimension mondiale inhérente au domaine médico-sanitaire. En effet, les crises sanitaires majeures ne peuvent être analysées qu'à l'échelle mondiale, et la recherche scientifique constitue un savoir accessible à l'ensemble des sociétés, dans le monde connecté d'aujourd'hui – en mettant de côté les logiques de marché de l'industrie pharmaceutique et les inégalités d'accès à Internet – et la santé concerne tout un chacun dans son individualité. C'est donc au niveau mondial que nous devons penser le premier niveau discursif qu'est le canon. C'est le discours académique qui en est le détenteur, par le biais de la publication d'articles scientifiques, qui rendent compte des progrès de la science et de la médecine. La communauté de sens – ici entendue comme tous les sujets-acteurs intéressés d'une manière ou d'une autre par les progrès de la médecine – lui accorde un rôle de détenteur de la vérité. Ce discours est ensuite investi par les acteurs de l'institution médicale, aussi bien les acteurs des politiques publiques de santé, que les acteurs sur le terrain de la prévention et du soin. Ceux-ci produisent le deuxième niveau discursif qu'est la vulgate. Ces acteurs, en plongeant ce discours dans la société, y intègrent des valeurs culturelles, par conséquent les discours seront variables d'une société à une autre, d'une région du monde à une autre. C'est ce qui

explique que le discours institutionnel vietnamien sera différent des discours institutionnels français, chinois ou américain. Chaque vulgate appréhendera le canon à l'aune de valeurs morales propres à sa culture. Cela sous-entend que différents tabous véhiculés par les valeurs morales vont entrer en jeu. Par conséquent les discours institutionnels seront caractérisés par le non-dit, des zones d'ombres qui varieront d'une culture à l'autre. C'est ce qu'il nous importe ici d'étudier, et c'est ici que le troisième niveau discursif qu'est la doxa prend toute son importance, comme l'expliquent Oger et Ollivier-Yaniv (2003) : « le sens à restituer réside aussi (surtout ?) dans le non-dit ». Nous cherchons à comprendre quelles sont les omissions des discours institutionnels, en les comparant aux discours de la doxa. En effet, la mise en regard des uns par rapport aux autres permet de révéler les lacunes qui resteraient imperceptibles sans ce mécanisme.

En conclusion, le canon est porteur de la vérité scientifique, la vulgate s'en empare et y adjoint des valeurs morales pour produire son discours et la doxa s'empare de ce dernier en comblant ses omissions pour produire le sien. Autrement dit, chacune des trois variations du discours couvre des aspects spécifiques de la communauté de sens, de telle sorte que, selon nous, s'opère une complétion, par un niveau, des deux autres, dans une relation que nous choisissons de décrire comme un *triangle discursif* (voir la figure ii.1). Finalement, c'est par la confrontation entre les différents niveaux que nous saisirons mieux les spécificités de chacun d'eux.

Au-delà de ce schéma théorique, la réalisation de ces trois niveaux est plus complexe et dans les faits, les frontières présentent une certaine porosité. Une bonne illustration de cette porosité est soulignée par Longhi avec le phénomène des internautes de la doxa acquérant le statut de vulgate. Dans tous les domaines, ce phénomène est relevé par les observateurs, avec parfois une cer-

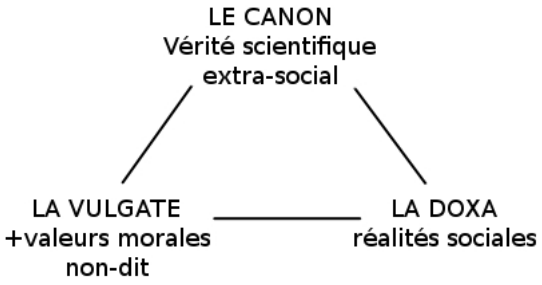


FIG. II.1 : Le triangle discursif

taine inquiétude, parfois de l'admiration, toujours avec intérêt, par exemple dans le domaine de la mode et la beauté, un article du journal Les Échos en mars 2015, intitulé *La petite blogueuse qui bouscule les grands de la beauté*, constatait que « Cette vénération, ce pouvoir d'influence n'ont pas échappé aux ténors du make-up, qui voient leur autorité prescriptive remise en question par des collégiennes immatures. »[2]

2.1.2 *Prescription vs description des connaissances*

Le rapport de force qui se dessine entre le canon et la vulgate d'une part et la doxa d'autre part est celui qu'établissent Slodzian et Valette (2009) entre connaissances prescrites et connaissances décrites. Les approches prescriptives, dans leur volonté d'objectivation, s'appuient sur une analyse conceptualisante, partant du principe qu'il n'y a qu'une seule vérité à expliciter. La recherche de prescription des connaissances se retrouve dans le mouvement de construction d'ontologies dans le domaine des sciences de l'information. Les ontologies se donnent pour objectif de représenter la connaissance dans un champ d'information, sous la forme d'ensembles structurés de concepts et de relations. Dans le contexte de l'informatique, ces représentations ont pour objectif d'améliorer les systèmes d'intelligence artificielle, afin de permettre

2. http://www.lesechos.fr/24/03/2015/lesechos.fr/0204250876008_la-petite-blogueuse-qui-bouscule-les-grands-de-la-beaute.htm

aux machines une meilleure appréhension du monde réel. En interaction avec le contexte d'Internet, les ontologies ont permis l'émergence du Web sémantique, selon l'expression inventée par Berners-Lee (2001), qui a parfois été jusqu'à être appelé le *Web 3.0* mais qui a aussi beaucoup été critiqué (par exemple Doctorow, 2001, et Shirky, 2005). En effet, comment prétendre saisir le monde uniquement avec des concepts définis a priori, d'autant plus dans un monde dont l'évolution et les mutations s'accélèrent de manière exponentielle ? Le principe fondamental des ontologies repose sur une volonté d'objectivation, de représentation de la vérité vue comme un tout unique. Il en résulte des méthodes d'extraction de connaissances déliées des textes et des interprétations possibles. Puisqu'il existerait une et une seule vérité à représenter, il suffirait de trouver les bons concepts pour y parvenir. Il s'agit donc d'une mise à distance avec les situations réelles de production des discours, qui souffrent d'un cadre trop contraint dans lequel ces approches prescriptives voudraient les faire tenir. A l'inverse, les discours ayant le plus de proximité avec des situations réelles, comme par exemple les discours informels, tels que ceux des forums ou des réseaux sociaux, permettent de se rendre compte à quel point la connaissance est composite, changeante, voire constituée de vérités contradictoires, avec un intérêt porté vers l'expérience plutôt que la transcription de la réalité.

Finalement, nous estimons que, parmi ces différentes conceptions de la construction de la connaissance, la confrontation se joue sur une opposition entre mise à distance et proximité, entre concept a priori et expérience vécue.

2.2 *Autorité gnoséologique selon la sociologie*

Dans le domaine de la recherche en sciences humaines et notamment en sociologie, nous retrouvons la confrontation entre différentes conceptions de la construction de la connaissance. Tout d'abord, le choix des sources étudiées est soumis à leur degré de stabilité. Par exemple, Chateauraynaud et Debaz (2009) constatent que « la tendance du chercheur est de suspendre la circulation et la prise en compte de multiples sources hétérogènes, pour se concentrer sur des sources plus stables, dotées de références ou de garanties de fiabilité » alors que c'est justement la prolifération actuelle qui permet de dépasser « [le] silence et [les] rapports de pouvoir hiérarchique de la période précédente ». Ils tentent donc d'outiller les chercheurs en sociologie dans les tâches de veille de l'information adaptées au numérique. Il s'agit de développer des méthodes adaptées aux nouvelles formes de production du discours, pour être en mesure de détecter leurs apports spécifiques dans la construction des connaissances, notamment dans le travail de détection et d'interprétation des signaux faibles. Cette tâche existait avant l'utilisation en masse d'Internet, mais elle nécessite d'être mieux formalisée en vue de sa généralisation. Chateauraynaud (2007) définit les signaux faibles comme « des événements mineurs, a priori négligeables, qui, pour tout analyste attentif, rendent manifeste une modification dont les conséquences, bien qu'incertaines au moment de leur détection, peuvent être décisives, et irréversibles, pour une entité ou un ensemble d'entités ». A l'heure où Internet est un média adopté par tous les niveaux d'énonciation, allant des plus hautes sphères institutionnelles jusqu'aux masses de citoyens lambda, « tout signe précurseur parvenant à un degré suffisant de visibilité publique produit une vive agitation » qui entraîne des évolutions dans les prises de paroles institutionnelles. Par conséquent, si l'importance accordée aux signaux faibles est de plus en plus grande, ceux-

ci provenant de sources variées, hétérogènes et difficilement définissables a priori, « la question de la validité des expertises et des formes de régulation associées se pose de manière cruciale » (Chateauraynaud, 2007). La confrontation apparaît entre des savoirs vérifiés diffusés dans des discours stables et l'émergence de nouvelles sources de construction de la connaissance, dans un contexte de redéfinition de la fiabilité des expertises.

D'autres limites apparaissent sur lesquelles buttent les méthodes traditionnellement employées en sociologie pour analyser les discours. Oger et Ollivier-Yaniv (2003) indiquent par exemple que « bon nombre de sociologues se sont penchés sur le caractère potentiellement artéfactuel ou biaisant de la nature même du dispositif de recherche par entretiens. » En effet, « ce dispositif méthodologique est une sollicitation qui vise à susciter, voire provoquer une parole et un discours de la part d'acteurs sociaux. » Le discours n'est donc pas une transposition transparente de la pensée de l'enquêté, et par conséquent il n'est pas observé complètement naïvement par les chercheurs qui, conscients de ces limites, appliquent des ajustements pour corriger ces bais. Mais les limites ne peuvent être totalement dépassées, notamment sur les catégories de sincérité/mensonge.

La rivalité qui se joue ici est celle entre d'une part des énoncés accompagnés d'informations attestées sur les énonciateurs eux-mêmes, mais qui restent, par la nature même des entretiens, policés, lissés, consciemment adaptés par leurs énonciateurs en fonction de leur récepteur et d'autre part des énoncés produits sous le régime de l'anonymat que permettent les forums notamment, pour l'analyse desquels une redéfinition de la notion de fiabilité des sources est nécessaire, mais qui permettent d'accéder à des strates de la connaissance sociologique non saisissables par des méthodes telles que les entretiens ou les questionnaires traditionnellement employés.

3 L'ANALYSE DES DISCOURS INFORMELS

L'analyse des discours numériques (que Paveau écrit ADN) a un rôle à jouer dans le dépassement des limites rencontrées par la recherche traditionnelle en sciences humaines. Des espaces de recherche s'ouvrent pour explorer ce champ, par exemple la revue scientifique *Recherches En Sciences sociales sur InternET (RESET)*[3] considère Internet comme « un terrain d'enquête utile et nécessaire à la compréhension de certains phénomènes sociaux », et a pour objectif de combler le vide laissé dans le paysage éditorial francophone.

Paveau (2012a) a étudié les énoncés natifs du web qui désignent les discours « produits directement dans les environnements du web », par exemple les discussions sur les forums, les échanges sur Twitter ou les commentaires à propos d'une publication. Elle regrette que ces discours soient trop souvent analysés par l'AD « d'une manière traditionnelle, c'est-à-dire logocentrée : on extrait les énoncés des écrans, et on les présente dépouillés de leur matérialité technologique […] Autrement dit, on enlève le *techno-* de ces technodiscours ». Paveau (2013b) enjoint les chercheurs à « Penser le contexte comme un écosystème où s'élabore le discours et non comme un arrière-plan. » Cela sous-entend qu'il est nécessaire d'accorder de l'importance non seulement au texte mais aussi aux éléments non langagiers. C'est ce qu'assuraient déjà Slodzian et Valette (2009) : l'objectivation à laquelle se prête la matérialité des textes du web n'est possible qu'en prenant en compte les « conditions de production et d'interprétation des documents », à l'inverse des approches prescriptives qui partent du principe qu'il n'y a qu'une seule vérité à expliciter. Comme nous avons déjà commencé à l'examiner, l'éventail des discours pour appréhender et expliciter le monde et la société s'est complexifié avec la popularisation d'Internet, qui, dans son évolution vers l'interactivité, dans

3. http://www.journal-reset.org/index.php/RESET

sa forme qui a émergé à partir de 2005 avec le web 2.0 - le web des réseaux sociaux, devient de plus en plus un reflet du monde réel.

Cardon (2008, 2009, 2010) s'est attaché à établir une sociologie du web 2.0. Cardon (2009) observe que le modèle du web 2.0 réussit mieux à des groupes peu institutionnalisés, horizontaux, souvent internationaux, groupes sociaux qui vont des mouvements altermondialistes aux groupuscules d'extrême droite, en passant par les réseaux d'experts environnementaux du GIEC par exemple. Nous retrouvons l'opposition et la concurrence entre un mouvement vertical et un mouvement horizontal dans la transmission du savoir et l'établissement de l'autorité. Cardon (2010) explique qu'Internet n'a pas changé les modèles de compétition politique, mais a permis une autonomisation de la société civile, libérée du cadre institutionnel de production de discours. Internet bénéficie souvent d'une présupposition d'égalité car c'est un univers peu hiérarchisé, où a priori tout énoncé en vaut un autre, sans que la position sociale, le parcours d'études ou le niveau d'influence de l'énonciateur ne contribue à la valeur de l'énoncé.

Toutefois, une hiérarchisation existe dans l'évaluation des informations, puisque, si elle n'est pas définie a priori, elle l'est a posteriori, sur des critères de réputation de l'énonciateur, ou de notoriété de l'énoncé fondée sur la fréquentation. Le crédit est accordé par la multitude. Il en résulte que nous nous trouvons dans un système de méritocratie, dans lequel nous assistons à la reproduction d'inégalités sociales dans la prise de parole : le web 2.0 valorise ceux qui occupent le terrain, qui sont les plus réactifs, dans l'écriture, mais aussi dans le fait de savoir rebondir, circuler, élargir leur réseau. Cette expression tout azimuts couvre les voix silencieuses. Le système entraîne la recherche de la visibilité car « les propos « légitimes » sont ceux qui apparaissent « en haut » des hiérarchies (des moteurs de recherche, des classe-

ments des blogs, des fils d'actualité, des portails d'information, des agrégateurs de news, etc.) » (Cardon, 2009).

3.1 *Spécificités des forums*

Les forums occupent une place particulière dans le web centré sur l'interactivité. Dans le cas des réseaux sociaux, l'inclusion du contexte technodiscursif tel que défini par Paveau nous apparaît primordiale dans l'analyse du discours, car il influence fortement la production. Les réseaux sociaux reposent principalement sur les relations entre les utilisateurs, ce qui, dans une certaine mesure, est moins fondamental dans le cas des forums, cela sera explicité dans la suite de cette sous-partie. Cette différence entre forums et réseaux sociaux est l'une des raisons pour lesquelles il existe un débat pour déterminer si les forums font partie du web 2.0, une des raisons étant qu'ils préexistaient à cette évolution d'Internet. Nous considérons qu'ils en font partie dans le sens où leurs utilisateurs sont les créateurs des contenus. Les forums s'enrichissent au fur et à mesure de leur utilisation. C'est la production et les échanges discursi·f/ve·s qu'ils permettent aux internautes qui construisent la connaissance, l'expertise, l'actualisation des savoirs.

Peut-on par ailleurs parler d'un genre spécifique aux forums ? Paveau (2012b) fait tout d'abord la distinction entre la notion de genre de discours et celle de genre textuel. Pour Paveau, les analyses textuelles ont trop tendance à mettre de côté le contexte de production des textes alors qu'elle souligne qu'il n'y a pas que du langage dans le langage et que « l'écriture numérique native possède des traits particuliers qui lui sont donnés par le dispositif technologique ». Le genre de discours serait donc plus adapté que le genre textuel pour veiller à inclure dans l'analyse les éléments non textuels participant à la production de sens, et non seulement le texte. L'inclusion du contexte de

production dans l'analyse des énoncés natifs du web nous permet de préciser notre position sur l'existence d'un genre spécifique aux forums. Existe-t-il dans le cas des forums la communauté de sens dont parle Sarfati, ou encore la communauté discursive définie par Paveau (2012b) comme « un ensemble de locuteurs qui partagent des usages [...] et des rapports au langage et au discours » ? L'originalité des réseaux sociaux réside selon Paveau dans le fait qu'ils « s'accompagnent généralement de leur guide d'usage ou de « savoir vivre », contrairement aux réseaux sociaux IRL[4], pour lesquels la maîtrise des usages est plutôt du ressort du capital culturel et social ». En cela, les forums confirment leur appartenance au web 2.0, puisqu'il existe bel et bien des normes de production du discours partagées par les internautes sur les forums et cela pourrait venir confirmer l'existence d'un genre. Seulement ce dernier ne serait pas valable uniquement pour les forums, il s'agirait plutôt d'un genre incluant l'ensemble des énoncés natifs du web, dans lequel nous incluons à part entière les forums.

Pour en revenir aux spécificités des forums, par rapport aux plateformes de réseaux sociaux à proprement parler, la relation entre les utilisateurs n'occupe pas la même importance dans l'édification de l'information. En effet, dans le cas des réseaux sociaux, c'est l'interaction qui crée la valeur ; tout comme cela peut être le cas dans les forums fermés tels que définis par Sidir et al. (2006) : une communauté fermée et formée pour une durée donnée. Il nous faut à ce point de la discussion établir la distinction entre les forums tels que ceux auxquels se sont intéressés Sidir et al. et les forums ouverts, c'est-à-dire des discussions consultables par tout internaute, où tout internaute peut devenir membre et où tout membre peut intervenir dans la conversation, ajouter sa voix au concert, s'exprimer personnellement. Des modes de

4. IRL est l'abréviation de « in real life » (dans la vraie vie) qui signifie que l'on parle des interactions sociales non médiées par Internet.

censure, de modération des discours par les administrateurs peuvent intervenir dans ces forums ouverts, mais cela relève le plus souvent de la veille de mise en conformité avec les règles d'usage du medium. Les forums sur lesquels nous portons notre attention sont de ce type. Contrairement aux réseaux sociaux ou aux forums fermés, les interactions existent entre les intervenants et sont évidement constitutives de l'élaboration des données mais le niveau de proximité entre les intervenants n'importe pas autant. Les liens qui relient les membres peuvent être moins forts et moins durables. Des liens qualifiables de forts peuvent se créer au fil des participations, mais ce n'est pas la ressource première recherchée dans ce medium. De plus, une hiérarchie plus marquée apparaît entre différents types d'intervenants en fonction de l'importance de leur expérience sur un sujet donné. Expérience qui sera reconnue par la communauté au fur et à mesure des interventions et de la qualité des informations proposées par un membre. C'est ce partage de savoir qui est spécifique aux forums et qui est recherché par les internautes s'y manifestant (de manière plus ou moins visible). Pour faire la lumière sur ces spécificités, nous faisons à nouveau appel à la cartographie du web 2.0 établie par Cardon (2008)[5] et notamment la notion de clair-obscur qu'il y a développée. Au sein des plateformes classées dans la zone du *clair-obscur*, les informations que les internautes révèlent d'eux-mêmes s'adressent de manière privilégiée à un cercle de proches, tout en sachant que ces informations restent accessibles à un lectorat anonyme. Le positionnement et la stratégie de chaque réseau social au sujet de la visibilité (des contenus et des membres) évolue avec le temps, et cette typologie établie en 2008 voit se déplacer les différents réseaux sociaux sur le graphique, en fonction de ces évolutions. Par exemple, Facebook a modifié la possibilité de réglage de confidentialité, ce

5. http ://www.internetactu.net/2008/02/01/le-design-de-la-visibilite-un-essai-de-typologie-du-web-20/

qui rend quasi inaccessible les informations à l'extérieur du cercle de proches choisis, même Twitter propose de rendre privés des tweets. Sur les forums, la politique diffère de l'un à l'autre. Par exemple dans un des forums sur lesquels nous avons travaillé, celui de l'association Sida Info Service, la consultation de certaines rubriques est publique mais d'autres (notamment la rubrique principale, qui représente 60% du forum) réservent leur accès aux membres. Alors que le forum vietnamien de HIV Online est extrêmement consulté par toute une foule de lecteurs qui n'interviennent pas. Cette information peut être constatée grâce à l'affichage statistique présenté pour chaque conversation : le nombre de *vues*[6] d'une conversation est 25 à 25 000 fois plus élevé que le nombre d'interventions publiées dans cette même conversation. De même, sur 187 internautes ayant consulté le forum sur une période de 15 minutes[7], 2 sont membres et 185 ne sont pas inscrits (et donc ne sont pas en mesure de publier eux-même une intervention. Cela montre à quel point les informations produites par les interactions entre membres inscrits sur le forum constituent en même temps un ensemble de connaissances accessibles à un cercle de personnes beaucoup plus étendu et consultées à grande échelle. Nous considérons que cela constitue une différence majeure en l'état actuel entre ces forums et les réseaux sociaux, ces derniers cultivant l'entre-soi plus que la mise à disposition de connaissances, alors que dans un cas comme dans l'autre les données sont produites par les utilisateurs.

3.1.1 Formalisation de la structuration des forums

Les méthodes d'analyse conversationnelle ont été développées pour des conversations non médiées par ordinateur et enregistrées par l'analyste. Mar-

6. i.e. le nombre de fois que la conversation a été consultée par un internaute.
7. Chiffres collectés à titre d'exemple, tels qu'affichés sur le forum le 27 août 2016.

coccia (2004) a montré les défis que posent les spécificités des forums à ces méthodes. Afin de pouvoir décrire les forums avec les termes de l'analyse conversationnelle, Marcoccia opère un relâchement définitoire des concepts traditionnels, que nous reprenons à notre compte. Il explique qu'il faut admettre « une définition rendant compte du caractère souple et continu de l'interaction » pour pouvoir considérer que les discussions sur un forum constituent des conversations, au sens où l'analyse conversationnelle les entend. En effet, les conversations sur les forums sont caractérisées par :

- un nombre non borné d'intervenants
- un entremêlement possible de plusieurs thématiques
- le fait de n'être potentiellement jamais terminées, et par conséquent
- la possibilité pour deux de ses interventions qui se suivent d'être séparées d'un long écart temporel
- et le fait qu'un nombre illimité de conversations puissent être en cours de manière synchrone.

En raison de tous ces points, le maintien d'une lisibilité impose une structuration forte, garantie par l'interface du forum. « L'objectif de l'interface est de permettre la structuration progressive du forum afin d'assurer une bonne lisibilité de la dynamique de l'interaction qui s'y déroule, quels que soient les procédés de visualisation de la conversation utilisés. » (Donath et al., 1999) D'un forum à l'autre des variations existent dans leur présentation, mais la forte structuration, nécessaire au maintien intemporel de la lisibilité, constitue une régularité avantageuse pour l'analyse des forums.

Nous reviendrons sur la description des niveaux de structuration des forums de manière détaillée au chapitre IV (partie 2), en l'appliquant à nos corpus. Pour l'heure, à des fins d'unification terminologique, établissons simplement trois niveaux de structuration fondamentaux et trois termes sélectionnés pour les désigner : (i) *rubrique* plutôt que *catégorie*, (ii) *conversation*

plutôt que *fil de discussion* et (iii) *intervention* plutôt que *message*. Ces choix résultent d'une volonté de désigner avec précision les concepts afin de ne pas confondre ceux-ci avec le sens large de mots tels que *catégorie*, *réponse* ou *message*. Dans le cas de la *conversation*, le terme a été retenu au détriment du *fil de discussion* à des fins de cohérence syntaxique. A partir de cette base, nous empruntons à Marcoccia son travail de typologie plus fine des interventions : les interventions initiatives désignent les interventions qui initient une nouvelle conversation, les interventions réactives celles qui sont publiées à la suite d'autres au sein d'une même conversation. A ces types d'interventions nous ajoutons les interventions uniques comme celles qui ne reçoivent aucune réponse, mais nous y reviendrons en détails dans les chapitres iv et v.

Paveau (2013a) recense les éléments technodiscursifs qui sont abandonnés par une analyse logocentrée dans les interfaces de Tweeter ou Facebook. En ce qui concerne les forums, la contextualisation est également inséparable du texte. Tout d'abord, l'interface de saisie contraint le discours. Par exemple, la liste des émoticônes qu'il est possible d'utiliser est fermée. Ensuite, le texte est produit dans le cadre d'une conversation, c'est-à-dire une série d'interventions qui se présentent généralement sous la forme de réactions aux interventions précédentes. Le positionnement temporel, la chronologie sont donc primordiaux, ainsi que le cadre dans lequel l'énoncé est produit – cadre qui indique à quel autres interventions l'énoncé réagit – de même que l'identification de l'auteur de l'énoncé (qui donnera des informations telles que le statut de l'auteur au sein du forum), ou encore l'évaluation du texte par les membres du forum participant à la conversation. Par conséquent, nous avons veillé dès le début à conserver un maximum de métadonnées sur le texte. Nous détaillerons l'application de ce choix dans le chapitre iii, relatif à la constitution des données.

3.1.2 *Le TAL confronté à l'écriture sur les forums*

Quels que soient les traitements ultérieurs auxquels un texte sera soumis pour son analyse, et quelle que soit la langue (avec des variations), certaines étapes primordiales sont indispensables avant toute chaîne de traitement. La particularité des données produites au sein de forums est qu'elles sont difficilement traitables par des outils développés pour traiter des textes normalisés, bien orthographiés, sans abréviations, etc. Un point positif nous aide malgré tout : les forums auxquels nous nous intéressons, appliquent une veille sur les bonnes pratiques d'écriture, que ce soit par les gestionnaires ou entre membres, se rappelant à l'ordre lorsqu'il y a un trop grand relâchement sur ce plan. Par conséquent, cela permet d'éviter qu'une trop grande quantité de données ne soient pas traitables par des outils automatiques. Cela n'empêche pas que des pratiques spécifiques aux forums s'y développent, qui posent des défis aux outils, heureusement relevables. L'avantage de ces données étant qu'elles sont potentiellement disponibles en grande quantité, cela explique que notre attention se porte plus sur les performances de précision que de rappel.

4 Méthodes et outils de corpus

4.1 *Segmentation du corpus pour la fouille contrastive*

« *Dans la langue, il n'y a que des différences.* » (Saussure, 2002)

Concevoir le corpus avec la possibilité de le segmenter en sous-corpus, et réfléchir aux unités de segmentation pertinentes, permet d'avoir recours au contraste pour faire émerger les singularités de chaque sous-partie par rapport aux autres (notamment avec le calcul des spécificités). C'est cette fonction différentielle que nous aurons à l'esprit en organisant en corpus homogène des données composites (sources, langues, rubriques, etc.). De plus, multiplier les possibilités de segmentation du corpus (rubriques, conversations, interventions, cf. le chapitre iii, sous-partie 2.3.2, et le chapitre iv, partie 2) permet de multiplier les angles d'approche pour analyser les données : plutôt que d'aborder le corpus uniquement par le lexique, il est alors possible d'exploiter des catégories méta-textuelles et philologiques (le genre, l'auteur, la chronologie, etc.)[8]

4.2 *Analyse des données textuelles*

La communauté scientifique s'intéressant à l'analyse statistique des données textuelles (ADT) a développé des logiciels, se revendiquant tantôt de la lexicométrie (Lexico), tantôt de la textométrie (TXM), tantôt de la logométrie (Hyperbase), mais appliquant tous des méthodes de calcul, statistique (spécificités, cooccurrences, etc.) ou non (segments répétés, concordances, etc.), sur des données textuelles, en vue d'analyser les productions langagières avec

8. cf. Pincemin (2012).

l'appui de données chiffrées objectives[9] et des représentations visuelles (carte des sections, histogrammes de ventilation, courbes d'accroissement du vocabulaire, nuages de mots, etc.). Cependant ces données quantitatives ne prennent un sens que par l'interprétation humaine qui en est faite. En cela, l'ADT s'oppose au TAL, dont l'objectif principal est l'automatisation des processus et l'évaluation des performances obtenues (Eensoo et Valette, 2015). Avec l'ADT, nous nous situons au contraire dans une perspective de construction des connaissances, par l'interprétation humaine des résultats obtenus grâce à des outils informatiques de calcul et de visualisation. La puissance informatique vient donc en assistance de l'exploration et la fouille des données. Cette différence fondamentale permet de produire des connaissances qualitatives sur les données et non seulement quantitatives (ici encore, voir Eensoo et Valette, 2015). Pour autant, nous travaillons avec une volonté de reproductibilité des méthodes, c'est pourquoi nous entrons de manière détaillée dans la description des aspects techniques, ainsi que nous le ferons pour les étapes de traitement dans le chapitre III.

Nous prendrons l'exemple de deux logiciels d'ADT, ceux auxquels nous avons eu recours pour analyser nos corpus, en pointant leurs spécificités.

4.2.1 Lexico (5)

4.2.1.1 Lexico et le vietnamien

Pour l'encodage de l'écriture vietnamienne, le standard Unicode s'est imposé dès les années 2000[10], et depuis lors, la plupart des données disponibles pour la constitution de corpus sont codées selon cette norme. Pourtant ce n'est

9. Le concept d'objectivité existe ici par opposition à une lecture humaine, par exemple aucune occurrence ne sera oubliée par le calcul, ou le contexte d'une forme sera considéré de la même manière, que ce soit en début de lecture ou en fin de lecture.
10. Sur l'encodage du quốc ngữ, voir le paragraphe 5.2.2, page 63.

que dans sa version 5 (disponible depuis 2016) que Lexico a permis un affichage des caractères Unicode. Le logiciel est ainsi devenu accessible aux langues comme le vietnamien, le chinois, le japonais ou l'arabe. Des traitements étaient possibles dans les versions précédentes de Lexico, mais l'intégration d'Unicode lève le frein majeur posé au traitement de corpus en langues non occidentales.

4.2.1.2 *Segments répétés et carte des sections*

Lexico propose une fonctionnalité dont ne disposent pas les autres logiciels d'ADT : le calcul des segments répétés, autrement dit le repérage des suites de mots qui se trouvent répétées plusieurs fois dans le corpus. Étudier des segments de texte plutôt que des mots, permet d'éviter le caractère ambigu que présentent les unités minimales. Les segments constituent plus facilement des unités linguistiques porteuses de sens (Mayaffre, 2007). De plus, Mayaffre (ibid, p.9) explique en quoi les segments répétés offrent une alternative à la lemmatisation : « [Leur étude] permet de désambiguïser les termes de manière formelle et surtout de manière endogène, en corpus et non en référence (arbitraire) au dictionnaire ou à la langue. » Cette désambiguïsation est particulièrement pertinente dans le cas de la langue vietnamienne, où les unités minimales sont invariables donc ne sont de toute façon pas lemmatisables, présentent une forte ambiguïté, et fonctionnent la plupart du temps par segments constitués de plusieurs unités.

Le calcul des segments répétés de Lexico génère une liste de segments allant de 2 à n unités minimales. Les paramètres sont réglables, mais l'intérêt se portera vers les segments les plus longs trouvés par le calcul. Plus les segments comprennent de mots, moins ils sont ambigus et plus ils sont caractéristiques du corpus, et pertinents à observer.

Un autre point fort de Lexico est la représentation graphique du corpus en carte des sections[11]. Elle permet de visualiser sous la forme d'un carré chaque unité de texte, tel que définie par la segmentation du corpus, et l'ensemble du corpus comme une suite de carrés, et d'utiliser une colorisation graduée de ceux-ci, selon que l'unité de texte qu'ils représentent contient une proportion plus ou moins forte de la forme (ou le groupe de formes) recherchée. Elle permet également de visualiser simultanément la répartition de deux formes (ou groupes de formes) dans le corpus, grâce à deux couleurs distinctes.

4.2.2 *TXM*

TXM[12] permet d'analyser des corpus encodés en Unicode et sous format XML principalement[13]. TXM intègre l'étiqueteur morpho-syntaxique Treetagger[14], ainsi que le logiciel de calcul statistique R[15]. TXM est développé en Java et est multiplateforme (Windows, Linux et Mac). L'équipe propose régulièrement des formations, et répond de manière réactive aux problèmes que peuvent rencontrer ses utilisateurs.

4.2.2.1 *Cooccurrences, expressions régulières et tri des données*

TXM met en avant une fonction en particulier : le calcul des cooccurrences, autrement dit les formes qui apparaissent plus fréquemment dans le contexte d'une forme ou d'une expression donnée. La distance maximale entre le pa-

11. cf. par exemple p. 165
12. Heiden et al. (2010)
13. De nombreux autres encodages sont possibles, ainsi que d'autres formats qu'XML. Le programme d'import accepte par exemple le format des logiciels Hyperbase, Alceste, et également des corpus sous format texte, en les transformant au format XML.
14. développé par Helmut Schmid en Allemagne (et par Achim Stein depuis 2003 pour ce qui est des ressources permettant de traiter le français. http://www.cis.uni-muenchen.de/~schmid/
15. développé au départ par Robert Gentleman and Ross Ihaka, mais aujourd'hui par une communauté mondiale. https://www.r-project.org/contributors.html

tron sur lequel est effectuée la recherche et ses cooccurrents est paramétrable dans TXM, alors que dans Lexico elle est calculée au sein des sections définies par l'utilisateur (par exemple dans la limite d'une intervention)[16].

TXM propose une fonction de recherche par expressions régulières, qui peut porter sur les formes pleines, mais qui inclue également les étiquettes de parties du discours ainsi que les lemmes, les différents niveaux de recherche étant combinable dans une même expression. Cette fonction donne une grande souplesse à la recherche de patron, par exemple pour une concordance, mais aussi pour une recherche de cooccurrences. D'autre part, tous les résultats de recherche sont présentés dans leur contexte, avec la possibilité d'accéder à chaque texte par un clic sur la ligne correspondante dans la liste des résultats. Cette liste peut être triée selon tout type de métadonnée qui a été associée aux textes. Nous pourrons ainsi trier les résultats par auteur, par date, ou par toute autre information stockée au préalable.

5 LE TRAITEMENT AUTOMATIQUE DU VIETNAMIEN

La langue vietnamienne est considérée comme une langue « peu dotée ». Cette qualification porte sur le niveau de formalisation de la langue pour son traitement par des procédures automatiques, et donc par conséquent en termes d'outils de TAL, ainsi que sur la quantité de ressources linguistiquement annotées, qui permettraient d'y appliquer des procédures d'apprentissage automatique.

Pourtant, que ce soit au Viêt Nam ou à l'étranger, plusieurs laboratoires de recherche produisent des travaux depuis les années 2000, progressant peu à

16. Lexico propose en effet une manière différente de calculer les cooccurrences : le contexte de la forme n'est pas limité par une distance fixe définie pour le calcul, mais par les limites du segment de texte défini par la segmentation du corpus. Par conséquent, si les segments de texte sont de longueur très différentes, les calculs en seront influencés.

peu dans cette tâche, s'inspirant des dernières technologies développées pour des langues mieux dotées, tout en faisant face au défi du manque de données, et en proposant aussi des pistes innovantes, s'adaptant aux spécificités de la langue vietnamienne.

5.1 *Les acteurs du traitement automatique du vietnamien*

En France, deux laboratoires ont été particulièrement actifs dans ces recherches, ce sont le Laboratoire d'Informatique de Grenoble (LIG, anciennement CLIPS-IMAG), et le LORIA à Nancy. Au Viêt Nam, toutes les grandes villes (Hà Nội, Hồ Chí Minh Ville, Đà Nãng, Huế etc.) comptent des laboratoires universitaires d'informatique (le VNLP à Hà Nội, par exemple), ainsi que des centres de recherche comme le Centre Lexicographique du Viêt Nam (VietLex) ou le Centre de Recherche International (MICA), ou encore l'Institut Français d'Informatique, à Hà Nội, par exemple. Au Japon, à Ishikawa, l'Institut de Science et de Technologie (JAIST) compte deux éminents chercheurs d'origine vietnamienne qui travaillent en collaboration avec les centres vietnamiens sur les progrès les plus avancés du domaine : les Professeurs Hồ Tú Bảo et Nguyên Lê Minh. D'autres universités en Europe, comme l'Université Catholique de Louvain en Belgique ont aussi publié des recherches en ce sens.

Depuis 2003, est organisée au Viêt Nam la Rencontre en Informatique Viêt Nam-France (RIVF), qui rassemble des chercheurs du monde entier dans le domaine des technologies de l'information et de la communication. Au départ tenue en français, elle s'est toujours voulue d'envergure internationale et compte désormais sur la scène de la recherche en informatique et communication. Elle est soutenue par l'organisation internationale des ingénieurs en informatique, l'IEEE. La douzième conférence (RIVF2016) a eu lieu en novembre 2016 à Hà Nội. Ce fut également l'occasion de la tenue du

quatrième atelier international Vietnamese Language and Speech Processing (VLSP 2016)

5.2 *Les spécificités du vietnamien et ses conséquences en TAL*

5.2.1 *L'écriture de la langue vietnamienne*

Parmi les nombreuses langues d'Asie, l'écriture de la langue vietnamienne est la seule à se baser sur l'alphabet latin. Cela est dû à son passé colonial, et même antérieurement, à sa position géographique sur l'une des routes explorées par les Européens dans leur quête vers la Chine. La route de la soie par les terres étant dominée par les Anglais, d'autres puissances ont privilégié les voies maritimes, telles que les Portugais, les Italiens ou plus tard les Français. Pour atteindre la Chine par le sud, il était donc nécessaire de s'établir en Asie du Sud-Est. Ce sont les missions d'évangélisation au catholicisme qui, dès le XVIème siècle, ont les premières poussé les Européens à s'intéresser de près aux langues parlées sur ces territoires. Afin de convertir de nouvelles populations il était nécessaire de se faire comprendre de celles-ci dans leurs langues. Les missionnaires portugais ont établi un système d'écriture à partir de l'alphabet latin, principalement dans sa version portugaise, pour transcrire la langue vietnamienne. Cette langue avait auparavant fait l'objet de transcriptions à partir de sinogrammes, mais la nécessité de la maîtrise de l'écriture chinoise avait constitué un frein à l'adoption de ce système par la majeure partie de la population. Au XVIIème siècle, le jésuite français Alexandre de Rhodes a systématisé cette transcription de la langue vietnamienne à l'aide de l'alphabet latin, en ajoutant au dictionnaire vietnamien-portugais une version en latin, ainsi qu'une grammaire. Ce système de transcription fut répandu dans un premier temps par les catholiques vietnamiens, puis par l'adminis-

tration coloniale française pendant la première moitié du XXème siècle, mais également par les nationalistes vietnamiens, qui lui donnèrent la dénomination de chữ quốc ngữ (« écriture de la langue nationale », souvent abrégé en quốc ngữ), et enfin à la réunification du pays en 1975 par le gouvernement vietnamien, avec une visée d'unification de l'ensemble des habitants du territoire. Le quốc ngữ est l'écriture officielle des administrations vietnamiennes depuis 1954.

La langue vietnamienne est une langue tonale : elle dispose de 6 tons, marqués à l'écrit chacun par un diacritique réservé (ou l'absence de diacritique pour le ton *ngang*). D'autre part, elle dispose de 12 voyelles et 27 consonnes, digrammes et trigramme consonantiques.

5.2.2 *L'encodage du quốc ngữ*

Avant que le standard Unicode ne s'impose à l'échelle mondiale, l'encodage des caractères hors ASCII[17] ne bénéficiait pas d'une normalisation. En ce qui concerne le quốc ngữ, plusieurs tentatives d'encodage coexistaient. Cette hétérogénéité représentait un frein à la constitution d'un corpus de données textuelles numériques homogènes, analysables par des traitements automatisés. En 1991, Unicode, dès sa version 1.1, a alloué une plage de codes pour représenter les caractères spécifiques au quốc ngữ[18], en fait des combinaisons de voyelles et de diacritiques. S'en est suivie une phase de développement

17. American Standard Code for Information Interchange. Table des 256 premiers codes de représentation de caractères, développée par les américains, donc majoritairement des lettres de l'alphabet latin, n'incluant aucun caractère accentués – puisqu'ils sont inexistants en anglais – les chiffres, quelques symboles (mathématiques, monétaires ou de ponctuation) et des caractères de contrôle. Dans la deuxième partie de cette table, quelques caractères propres aux écritures européennes sont malgré tout inclus.

18. Au sein de la plage *Latin étendu additionnel*, les codes allant de 1EA0 à 1EF9, soit seulement 90 caractères, suffisants pour inclure toutes les combinaisons du quốc ngữ (outre les caractères ASCII). L'encodage sur un octet est par conséquent suffisant pour écrire du vietnamien, contrairement au chinois, par exemple.

de programmes permettant de convertir des textes écrits avec les normes précédentes (VISCII, VNI, VPS, Windows-1258, etc.) vers l'UTF-8 d'Unicode. Puis, à partir des années 2000, la norme UTF-8 a été adoptée unanimement. A présent, la quasi totalité des productions textuelles numériques en quốc ngữ étant codées selon cette norme, le traitement automatique des corpus est grandement facilité.

5.2.3 *La saisie du quốc ngữ au clavier*

Le nombre de combinaisons possibles entre les 11 voyelles (a, â, ă, e, ê, o, ô, ơ, u, ư et y) et les 5 diacritiques de tons (̀ , ́ , ̉ , ̃ et ̣) rend improbable l'existence de claviers proposant un nombre suffisant de touches. En revanche, l'adaptation d'un clavier classique (que ce soit de type Qwerty ou Azerty) est très simple, du fait que certaines lettres de l'ASCII présentes par défaut sur les touches de ces claviers n'ont pas d'utilité en quốc ngữ (il s'agit des lettres suivantes : f, j, r, w, x et z). Ces touches peuvent donc être réemployées pour ajouter des diacritiques spécifiques au quốc ngữ. Des logiciels légers et téléchargeables librement permettent cette opération, sans avoir à se procurer un clavier supplémentaire. Ainsi, passer d'un système d'écriture à un autre au cours de la frappe est très simple. Parmi ces logiciels, citons VietKeys[19] et UniKey[20]. Plusieurs normes de saisie sont disponibles (également personnalisables). La norme VNI utilise les touches des chiffres situées au-dessus des touches alphabétiques sur les claviers standards. La norme Telex

19. Développé par Đặng Minh Tuấn depuis 1994, sous licence publique GNU. Depuis 2004 les normes d'encodage autres qu'Unicode ne sont plus maintenues. Il est téléchargeable sur http://www.vietkey.com.vn

20. Développé par Phạm Kim Long depuis 1994, sous licence publique GNU depuis 2000. Ce logiciel intègre 17 normes d'encodage, dont celle d'Unicode. Il est téléchargeable sur http://www.unikey.org.

utilise les touches f, j, r, s et x[21] pour noter les 5 diacritiques de tons (cf. le tableau II.1), le w combiné aux touches a, o et u pour marquer respectivement les voyelles ă, ơ et ư, et le z combiné à la touche Alt pour passer d'un système d'écriture à l'autre (par exemple du français au quốc ngữ). Pour les lettres â, ê, ô, et đ, il suffit de taper deux fois successives sur les touches a, e, o et d. Avec ce petit nombre de combinaisons de touches, l'intégralité des combinaisons de diacritiques et lettres utilisées en quốc ngữ est couverte.

Tab. II.1 : Touches du clavier pour marquer les tons

Touche	Diacritique
f	̀
s	́
r	̉
x	̃
j	̣

5.2.4 *La segmentation*

Une première approche simplistes serait de considérer que, le vietnamien étant une langue dite monosyllabique, la syllabe et l'unité de sens se recoupent (Cao (1985) : « la plupart du temps, le mot, le morphème et la syllabe coïncident presque entièrement »). Mais les différents auteurs s'accordent à dire que malgré le fait que le vietnamien s'écrive avec des caractères du latin étendu, la segmentation dans cette langue est très difficile car il ne s'agit en réalité pas d'une langue véritablement monosyllabique (Tuan et al., 2004). Cette tentation d'assimiler la syllabe au mot est sans nul doute motivée par le fait que graphiquement, les syllabes sont séparées par des espaces. Mais en réalité on ne peut se baser sur l'espace pour segmenter les unités lexicales, car la compo-

21. Les lettres r, s et x sont pourtant utilisées en quốc ngữ, mais uniquement en position initiale de syllabe, donc saisie après une voyelle, elle marquera sans ambiguïté le ton qui lui est dévolu par Telex.

sition est très caractéristique de la formation de mots en vietnamien, comme le remarquent Hô et al. (2003). La segmentation est donc une étape fondamentale et délicate pour le traitement automatique du vietnamien en général et pour la recherche d'information (Hô et al., 2003) ou pour la catégorisation des textes (Nguyen et al., 2006).

L'unanimité se fait s'agissant de dire que c'est le contexte dans lequel les unités apparaissent qui permet de déterminer les frontières de mots (en même temps que leur catégorie (Hô et al., 2003) ou leur désambiguïsation sémantique (Tuan et al., 2004)).

Par ailleurs, Hô et al. (2003) considèrent que, malgré l'invariabilité morphologique, il existe bel et bien des suffixes et des préfixes. Cette analyse nous semble plus poussée linguistiquement que ce qu'en disent Nguyen et al. (2006). Il est à préciser cependant que ces affixes ne sont pas soudés graphiquement.

Hô et al. (2002) affirment que la plupart des unités lexicales sont composées de deux mot graphiques. Ils expliquent en quoi un découpage en bigrammes, pris en partant de la gauche et en allant vers la droite, validé par un lexique, est la meilleure méthode pour une segmentation satisfaisante. Nous validons cette approche comme principe général à partir duquel penser la segmentation du vietnamien, pourtant il n'est pas rare de rencontrer des constructions de quatre syllabes pour désigner un concept.

Le principe de la segmentation en vietnamienest donc de repérer les ngrammes (majoritairement des bigrammes), afin de reconstruire à partir d'unités monosyllabiques les unités de sens. Du fait du système d'écriture actuel du vietnamien (le quốc ngữ), le texte se présente sous la forme de syllabes alphabétiques, toutes séparées par des espaces (ou des caractères de ponctuation). Concrètement, la tâche nommée segmentation consiste donc à faire perdre à

certains espaces leur statut de frontière de mots, en les transformant en un lien unissant deux syllabes, et en laissant les autres espaces comme frontières réelles de mots (au même titre que les caractères de ponctuation). Autrement dit, marquer graphiquement la différence entre deux types d'espaces. Prenons un exemple avec l'énoncé suivant :

> « *Suy giảm miễn dịch : Suy giảm chức năng bảo vệ cơ thể chống lại sự tấn công của các mầm bệnh.* »

Suy giảm	/	**miễn dịch :**	/	**Suy giảm**	/	**chức năng**	/	**bảo vệ**	/ **cơ thể** ...
diminuer	/	immuniser	/	diminuer	/	fonction	/	protéger	/ organisme
Immunodéficience :				*Diminution des défenses immunitaires ...*					

... **chống**	/	**lại**	/	**sự**	/	**tấn công** ...			
lutter contre	/	(marq.) idée d'arrêt	/	(marq.) substantivation	/	agresser			
... qui protègent contre				*les assauts ...*					

... **của**	/	**các**			/	**mầm**	/	**bệnh.**	
(marq.) possession	/	(marq.) pluriel défini			/	germe	/	maladie	
... des	/	*germes pathogènes.*							

Le segmenteur vnTokenizer a été développé par Lê et al. (2008). Il repose sur une approche hybride combinant des automates à état finis[22] et des expressions régulières pour un découpage en bigrammes, suivi d'un modèle statistique pour la désambiguïsation et le choix de la segmentation la plus probable (Lê et al., 2008).

L'opération de différentiation des types d'espaces (lien entre 2 syllabes / séparateur d'unité de sens) est effectuée par vnTokenizer en remplaçant les

22. FSA pour Finite State Automata

espaces-*liens* par des tirets bas (underscore ou low line) entre les unités graphiques formant une seule unité de sens, et en maintenant les espaces comme séparateurs d'unités signifiantes. Ainsi, le résultat pour l'énoncé pris en exemple est le suivant :

> « *Suy_giảm miễn_dịch : Suy_giảm chức_năng bảo_vệ cơ_thể chống lại sự tấn_công của các mầm bệnh* »

vnTokenizer [23] peut s'utiliser seul, mais qui fait aussi partie du vnToolkit[24], ensemble d'outils pour le traitement automatique du vietnamien.

Une autre suite d'outils de traitement de la langue vietnamienne, JVnTextPro, a été développée au sein du groupe de recherche VNLP (Vietnamese Natural Language Processing) de l'Université de Hà Nội. Cette suite intègre un outil de segmentation, JVnSegmenter, dont la dernière version a été mise en ligne date de 2007. La suite est également open source et basée en Java. Le segmenteur a été entraîné sur un corpus de 8 000 phrases, et nécessite d'être entraîné sur d'autres corpus pour de bonnes performances.

5.2.5 *L'étiquetage morpho-syntaxique*

Nguyen (2000) faisait la remarque qu'il n'existait pas encore de classification standard des mots vietnamiens. Elle n'a cessé de le répéter dans les publications de ses différents travaux et, 6 ans après, ses propos sont repris par Lê et al. (2006) qui précisent : la principale difficulté vient du fait de l'ambiguïté des rôles grammaticaux de nombreuses unités lexicales. Les mutations de catégorie (sans variation morphologique) sont très fréquentes.

23. La dernière version (3.0) date de 2007.
24. L'ensemble vnToolkit est développé par le laboratoire VietLex (à Hanoi : http://www.vietlex.com/) depuis 2006. La dernière version (2.0.1) date de janvier 2008.

Un ensemble d'étiquettes morpho-syntaxiques a été construit par Nguyễn TMH et al., reprenant la norme MULTEXT (dont le principe est rappelé par Nguyen et al. (2005)), composé de 14 catégories de premier niveau et de nombreuses spécifications propres à la langue vietnamienne dans le deuxième niveau. Nguyen et al. (2003) disaient que MulText comptabilisait 11 catégories communes, et que traditionnellement 9 catégories étaient généralement admises pour le vietnamien. Certaines catégories de MulText n'ont pas été conservées (Déterminant, Adposition) du fait des particularités de la grammaire vietnamienne, d'autres ont été rajoutées aux 9 catégories traditionnelles (Adverbe, Numéral), pendant que la catégorie Adjonction disparaissait et que la catégorie Modalité a été conservée en plus des catégories de MulText. Cela faisait donc 10 catégories.

Lê (2005) précise que MulText a en fait 14 catégories de premier niveau, mais que 11 seulement ont des spécifications dans le 2e niveau (toutes sauf les 3 dernières). En 2006 les deux s'accordent en disant conserver 11 catégories de premier niveau pour le vietnamien : Nom (N), Verbe (V), Adjectif (A), Pronom (P), Adverbe (R), Adposition (O), Conjonction (C), Déterminant (D), Numéral (M), Interjection (I), Particule Modale (T) (Lê et al., 2006), ce qui contredit l'article de 2003 car les catégories Déterminant et Adposition sont toujours là. De plus nous supposons que la catégorie Résidu (X) ne doit pas avoir disparu totalement, mais elle ne figure plus dans la liste. Pour finir, rien n'est dit de ce qu'il est advenu des catégories Eléments non autonomes (U) et Abréviation (Y).

L'étiqueteur des parties du discours (POS-tagger) de Dinh et Hoang (2003) utilise lui un étiqueteur des parties du discours déjà disponible pour l'anglais, puis fait correspondre ces parties du discours étiquetées à leurs vis-à-vis vietnamien dans un corpus automatiquement aligné.

VnTokenizer est à la base d'un ensemble d'outils qui s'appuient dessus pour des traitements plus avancés, comme par exemple l'étiquetage syntaxique[25], mais nous avons finalement fait le choix de n'utiliser que le segmenteur. Avant de prendre cette décision, qui n'est pas sans conséquence puisque nous n'avons de fait plus la possibilité de travailler au niveau des parties du discours, nous avons effectué un travail de comparaison entre deux outils d'étiquetage syntaxique du vietnamien : vnTagger, de la suite vnToolkit du LORIA, et jvnTagger de la suite JVnTextPro du VNLP . Une comparaison des résultats n'a pas conclu à nos yeux une analyse satisfaisante de la langue, trop de formes présentant une ambiguïté. En effet, l'idée de calquer un jeu d'étiquettes morpho-syntaxiques développé pour une langue comme l'anglais trouve vite ses limites, avec un langue où la morphologie n'est d'aucune aide puisque les syllabes sont invariables, et où chacune de ces syllabes pourra tantôt jouer le rôle de sujet, tantôt de verbe, sans qu'aucun indice graphique ne vienne les distinguer, ce n'est que par le contexte et la syntaxe que l'étiquette de partie du discours pourra être déterminée. Il n'est en tout cas pas possible de se baser sur un modèle de dictionnaire et règles morphologiques, comme c'est le cas des étiqueteurs développés sur les langues occidentales.

Pour illustrer ce propos, reprenons notre énoncé d'exemple page 68.

> « *Suy_giảm miễn_dịch : Suy_giảm chức_năng bảo_vệ cơ_thể chống lại sự tấn_công của các mầm bệnh* »

Chaque unité de sens peut se voir affecter plusieurs étiquettes POS tag[26], en fonction du contexte et de l'ordre des mots dans la phrase. Voici les POS tags possibles pour chaque unité de sens de l'exemple prises hors contexte.

25. étiquetage syntaxique et non morpho-syntaxique, les unités de la langue vietnamienne étant invariables
26. Part-Of-Speech tag ou étiquette de partie du discours

Suy giảm	**V** : diminuer, s'amoindrir **N** : baisse, chute
miễn dịch	**V** : immuniser **N** : immunité(s) **Adj** : immunitaire
chức năng	**N** : fonction, rôle **Adj** : fonctionnel.le(s)
bảo vệ	**V** : protéger **N** : protection(s), gardien(s) **Adj** : protecteur
cơ thể	**N** : corps, organisme(s)
chống	**V** : lutter contre **Adv** : contre, en opposition à
lại	**marqueur** de l'idée d'arrêt
sự	**marqueur** de substantivation
tấn công	**V** : attaquer, agresser **Adj** : offensi.f/ve(s)
của	**marqueur** de possession
các	**marqueur** de pluriel défini
mầm	**N** : germe(s)
bệnh	**N** : maladie

La tâche d'étiquetage s'apparente donc à une désambiguïsation en fonction de règles syntaxiques et de règles de combinaisons entre des mots-outils et des mots pleins (par exemple *sự* + V => N), ce qui n'est pas impossible et fait l'objet des recherches les plus récentes dans le domaine du traitement automatique de la langue vietnamienne. Elle nécessite simplement des modélisations à un niveau plus complexe, voire qui prennent en compte plusieurs niveaux d'analyse de la langue. Nous avons bon espoir de constater dans un avenir proche le développement d'un outil fournissant des résultats satisfaisants.

CHAPITRE III

Constitution des données

Après un préambule à propos du choix des sources des données que nous nous proposons d'analyser, nous décrirons dans le détail les méthodes développées pour passer des données en ligne au corpus de travail (partie 2), ainsi que les différents traitements appliqués auxdites données en vue du recours à des logiciels d'ADT pour les analyser (partie 3). En fin de chapitre (parties 4 et 5) nous récapitulerons les étapes de la chaîne de traitement et présenterons les données quantitatives du corpus obtenu.

1 Préambule : le choix des sources

Nous souhaitons étudier le recours aux TIC – en particulier le web social et notamment les forums – comme moyen d'accès et d'échange d'information. L'objectif est d'apporter aux SHS un point de vue complémentaire aux méthodes traditionnelles de collecte de données. Pour cela nous avons élaboré une méthode adaptée à ce type particulier de données élaborées au sein des forums, type que nous avons décrit au paragraphe 3.1.1 du chapitre II. Pour élaborer cette méthode et l'appliquer en contexte réel, nous avons porté notre choix sur la problématique de la transmission du VIH au Viêt Nam, problé-

matique décrite au chapitre 1. Pour sélectionner nos sources, nous avons ainsi porté notre attention sur des forums consacrés majoritairement au sujet du VIH/SIDA, recouvrant à la fois la maladie, les risques de contamination et sa prévention, les antirétroviraux et le quotidien des personnes touchées mais aussi les sexualités en général.

L'un des aspects de la méthodologie s'appuyant sur le contraste, nous avons également collecté des données hors des forums, dans le but de faire émerger les spécificités des discours sur les forums par rapport au discours institutionnel. Dans la même optique, nous avons également choisi de collecter des données en dehors du contexte vietnamien, toujours dans le but de faire émerger des spécificités par le contraste. Toutes ces collectes ont été menées avec le souci de construire un corpus homogène, condition indispensable à une mise en comparaison.

1.1 *Choix des sources vietnamiennes*

Le forum sélectionné pour être la source principale des textes analysés est hébergé sur le site web intitulé HIV Online (`hiv.com.vn`). Nous l'avons déjà évoqué au paragraphe 3.2.3 du chapitre 1, la lutte contre l'épidémie de VIH est une préoccupation qui se reflète au sein de l'organisation même du gouvernement depuis la fin des années 1990. Mais c'est en 2004 que la lutte prend le tournant d'Internet. Cette année-là, le ministère de la santé vietnamien estimait que dès l'année suivante le nombre de ménages ayant un membre atteint du VIH s'élèverait à 1 million, soit autant de personnes touchées directement ou indirectement par ce sujet. Parallèlement, le ministère des télécommunications établissait que le nombre d'internautes était alors de 5 millions et augmentait de 100% par an. Le gouvernement a dès lors énoncé la volonté de permettre à tous les résidents urbains et à 80% de résidents ruraux l'accès à

des informations sur la prévention du VIH, en l'espace de 5 ans. Dans cet objectif, le programme intitulé « Lutte contre le VIH/SIDA en ligne » a été lancé, et qualifié de projet novateur dans la lutte contre le VIH en mai 2004, juste avant la mise en activité du site HIV Online. Dès la fin 2004, ce site a été étendu aux questions de la jeunesse et la sexualité[1]. Les informations présentes sur le site peuvent donc être considérées comme représentatives du discours institutionnel vietnamien sur le VIH (il serait difficile de l'être plus). Toutefois nous avons établi la constatation suivante : la plupart des articles publiés dans cette vitrine institutionnelle sont tirés d'autres sources web. Le site n'est donc pas à proprement parler producteur d'information, mais essentiellement agrégateur et transmetteur d'information existante (en ce sens, il correspond bien à l'application de la vulgate définie au paragraphe 2.1.1 du chapitre II, page 41), ce qui ne lui ôte rien de sa qualification de vitrine institutionnelle.

1. Source : page de présentation du site : http://hiv.com.vn/LienHe/GioiThieu/

Fig. iii.1 : Capture d'écran de la page d'accueil du site HIV Online (16 août 2016)

A tout moment de la navigation sur le site HIV Online, il est possible d'accéder au forum de discussion, dont l'URL est `forum.hiv.com.vn`. La langue utilisée sur le forum est pour la quasi totalité le vietnamien. Le recours à l'anglais est présent lorsque des articles écrits en anglais sont reproduits sans traduction, mais cela reste marginal. L'activité sur le forum est beaucoup plus manifeste que la publication d'articles sur le site lui-même. Nous pouvons

noter que, dès le titre et l'URL, le site comme le forum s'adressent explicitement aux personnes vietnamophones intéressées par les questions qui ont trait au VIH, et par extension les questions qui ont trait à la sexualité – ce que confirme le sous-titre du site : *Tuổi trẻ, Giới tính & HIV* (« Jeunesse, sexualité et VIH »).

FIG. III.2 : Logo, titre et sous-titre du site HIV Online

Le premier élément du sous-titre annonce que la cible du site est en priorité les jeunes. Plus encore, ce n'est pas le VIH qui est mis en avant dans la présentation du site (une fois passé le titre), mais plus globalement la jeunesse. De même pour le forum, celui-ci est intitulé « Le forum de la jeunesse, de la sexualité et du VIH ». Le site comme le forum veulent donc attirer les jeunes qui s'interrogent à propos du VIH, mais se veulent plus généralistes et ouvrent la possibilité aux jeunes de s'exprimer sur tous les sujets qui les préoccupent. C'est aussi pour cette raison que ce forum a été choisi comme source principale des données.

1.2 *Choix des sources françaises*

Nous avons constitué notre corpus vietnamien avec d'une part les échanges sur le forum et d'autre part les articles publiés sur le site. A des fins de comparaisons multilingues et culturelles, nous avons produit un autre corpus, similaire en genre, domaine et discours, correspondant au contexte français. Le même schéma a pu être conservé : un site émanant d'une volonté des ins-

tances publiques de moderniser les moyens de lutte contre l'épidémie de VIH, ainsi que les échanges discursifs produits sur le forum attenant. Notre choix s'est porté sur l'association Sida Info Service (SIS). Un autre site fonctionne sur le même schéma : il s'agit de Seronet[2]. Le premier (SIS) s'adresse au grand public et le second est destiné « aux personnes séropositives au VIH et aux personnes porteuses d'une hépatite virale mais aussi à leurs proches ou à tous ceux qui les soutiennent ». Notre choix s'est donc porté sur le premier car – tout comme dans le cas de HIV Online – nous souhaitons porter notre attention sur un public aussi large que possible, considérant que la problématique du VIH concerne la population en âge d'avoir une sexualité dans son ensemble.

Parmi les services proposés par SIS, la ligne d'écoute téléphonique reste primordiale, comme nous pouvons le constater sur le site en ligne[3] (cf. la figure iii.3) : le bandeau-titre affiche le numéro en gros caractères juste sous le titre de l'organisme, auquel a été accolé le nom de domaine « .org » ; mais cela ne suffit pas à indiquer l'adresse électronique exacte, puisque celle-ci contient des traits d'union. De plus, la police des caractères du numéro téléphonique est de plus grande taille et de couleur verte. Nous pourrons aussi constater dans les discussions du forum que le renvoi des internautes vers la ligne téléphonique est fréquent, les deux medias sont véritablement complémentaires.

De même que dans le cas du site HIV Online et de son forum, l'accès au forum de Sida Info Service[4] est possible à tout moment de la navigation sur le

2. D'autres associations françaises œuvrent dans le domaine de l'information sur le VIH/SIDA. AIDES est la plus ancienne et la plus instituée, Act-Up mène de nombreuses actions de visibilité, par exemple. Cependant ces associations ne proposent pas de plateforme offrant la possibilité à tout un chacun de s'exprimer. En revanche, c'est sous l'impulsion de AIDES que les services publics (le ministère de la Santé, via l'Agence française de lutte contre le sida) ont financé la mise en place d'une part de Sida Info Service (en 1990 pour la ligne téléphonique, en 2003 pour le site internet) et d'autre part de Seronet (en 2008).

3. sida-info-service.org

4. forum.sida-info-service.org

FIG. III.3 : Capture d'écran de la page d'accueil du site Sida Info Service (16 août 2016)

site d'information. La justification du choix de cette instance parmi le paysage associatif français est renforcée par son organisation similaire à celle de notre source vietnamienne.

2 DU WEB AU CORPUS

Les étapes de collecte de données sur Internet (parfois nommée aspiration) d'une part et de choix dans la structuration du corpus d'autre part sont interdépendantes. Elles sont, chacune à sa manière, liées à la structuration des forums en ligne. La première y sera liée au niveau microstructurel (architecture des pages html, pagination, présentation des métadonnées, etc.), la seconde au niveau macrostructurel (organisation des conversations en rubriques, présentation des interactions entre participants, etc.). Si des différences secondaires peuvent apparaître d'un forum à l'autre (ce qui entraîne des variations au niveau microstructurel), la structure générale des échanges est partagée par l'ensemble de ces systèmes (i.e. au niveau macrostructurel).

La tâche de collecte des données étant dépendante des spécificités de structuration du forum-source, une étude précise de cette structuration est indispensable pour l'implémentation de la collecte, car elle détermine les détails techniques de cette dernière. Dans le cas de l'organisation du corpus, celle-ci sera pour partie fonction des généralités de structuration des forums (organisation des échanges sous forme de conversations, ordonnancement chronologique, échanges sous forme de réponses, etc.), mais les choix seront également fonction des analyses prévues sur les données.

2.1 *Étude des données en ligne*

2.1.1 *Deux genres textuels à organiser en un corpus homogène*

Une analyse de productions discursives dont la méthodologie s'appuie sur le contraste implique que les contextes de production de ces discours sont eux-même dissemblables. Nous comparons ici des forums et des sites d'information institutionnels. Leur point commun – outre le thème traité –

est leur medium d'affichage, constitué de pages internet, celles-ci présentant l'intérêt d'une accessibilité dans une temporalité asynchrone par rapport à leur moment de production. Tout le reste les distingue. Leur organisation en corpus homogène impose l'établissement d'une structure commune, qui nécessite une étude préalable spécifique à chacun.

2.1.1.1 *Analyse technodiscursive des forums*

Pour analyser les données textuelles, leur contextualisation est capitale. C'est pourquoi la conservation d'un maximum de métadonnées qui leurs sont associées est intégrée à la méthodologie développée, dès la collecte des données. Pour cela, il est nécessaire de prendre connaissance des choix éditoriaux des architectes du forum en partant du plus bas niveau, jusqu'aux détails les plus précis.

- Les pages sont-elles construites dynamiquement ou sont-elles statiques ?
- Combien y a-t-il de messages par page, et s'il y a plusieurs pages pour une discussion, comment accède-t-on aux pages suivantes ?
- A partir de quelle logique sont formées les URL : l'arborescence est-elle transparente, peut-on reproduire l'URL à partir d'informations telles que l'identifiant de la rubrique, celui de la conversation, ou celui de l'intervention, le nombre de messages par page, etc. ?
- Les sujets de discussion sont-ils organisés par thèmes ? par rubriques ?
- Quelles sont les informations présentées pour identifier et qualifier les intervenant·e·s ? Ont-iels :
  - un avatar (i.e. l'image identifiant un membre à chaque intervention) ?
  - une signature (i.e. la phrase de présentation choisie par un membre) ?
  - une évaluation (établie par leurs pairs, par le volume de leurs interventions, ou par un autre critère) ?

- Une prise de parole peut-elle également être évaluée ?
- existe-il une hiérarchie entre les prises de parole au sein d'une même discussion ?
- etc.

Tous ces paramètres entreront en compte pour élaborer la méthodologie de récupération des données, ainsi que celle de structuration des corpus, afin de pouvoir par la suite accéder à l'ensemble de ces informations.

2.1.1.2 *Analyse des sites d'information institutionnels*

Contrairement aux productions discursives des forums, le paramètre de l'interactivité n'entre pas en jeu, donc toute forme d'évaluation par les lecteurs est absente (nous traitons ici des sites qui ne proposent pas aux internautes de commenter les articles), le texte est produit par un auteur unique et a une date unique de production (celle de l'ajout de l'article sur le site web). Le texte présente ainsi un aspect linéaire, beaucoup plus simple à traiter. En revanche, du fait que ces informations sont considérées dans ce genre discursif comme secondaires, elles ne sont pas toujours aussi systématiquement disponibles que sur les forums. Toujours avec le même souci d'homogénéisation, nous stockerons pourtant ces textes avec la même structuration que les productions plus complexes des forums. Certains champs resterons vides et des analyses sur l'interaction ou le profil d'auteur par exemple ne pourront être menées, ce qui est normal puisque celles-ci sont non-pertinentes, mais la mise en compatibilité des structures est nécessaire pour tout le travail de fouille contrastive.

2.1.2 *Etude de l'organisation des forums en rubriques thématiques*

La première étape avant d'envisager la collecte des données consiste en un parcours attentif de la structure du forum en vue de déterminer comment

y sont classées les rubriques, dans l'objectif d'établir la liste des conversations que nous souhaitons analyser, et que nous choisirons de récupérer. Nous verrons que cette tâche n'est pas aussi triviale qu'elle pourrait apparaître.

Dans les deux cas qui nous intéressent (vietnamien et français), les forums ne sont pas structurés exactement de la même manière, mais nous trouvons des similitudes dans l'existence d'une classification thématique en deux niveaux. Un premier niveau propose plusieurs « forums », un deuxième niveau propose des rubriques au sein de chaque « forum ».

Fig. III.4 : Forums sur *forum.hiv.com.vn* (capture du 16 août 2016)

Fig. iii.5 : Rubriques sur *forum.hiv.com.vn* (capture du 16 août 2016)

Fig. iii.6 : Forums et rubriques sur *forum.sida-info-service.org* (capture du 16 août 2016)

Tab. III.1 : Premier niveau de structuration sur le site vietnamien

titre du forum	id	traduction
Hướng Dẫn Thành Viên Mới	1	Guide pour les nouveaux membres
Nội dung quan trọng	10	Contenu important
HIV - Điều không mong muốn	3	VIH – L'indésirable
Tuổi trẻ & lối sống	4	Jeunesse et mode de vie
Tình yêu & Giới Tính	5	Amour & Sexualité
Vòng tay cộng đồng !	8	Communauté, unissez-vous !
Hoạt động Offline	13	Activité Hors ligne
Hỏi và đáp liên quan đến HIV/AIDS	14	Questions et réponses relatives au VIH/SIDA
Hành trang vào đời - Kỹ năng sống	18	Bagages, compétences pour affronter la vie
[...] xử lý lỗi kỹ thuật Diễn đàn [...]	11	Gestion des erreurs techniques sur le forum etc.
CLB Những người điều hành diễn đàn	12	Club des gestionnaires du forum

Tab. III.2 : Premier niveau de structuration sur le site français

id	titre du forum
1	Forum VIH/SIDA, IST
12	Forum Hépatites
27	Forum Séropos
13	SIS Association
9	Santé sexuelle
18	Détente

Dès le premier niveau, nous pouvons constater de grandes disparités dans la répartition des thématiques entre le site français et le site vietnamien. Ce

dernier multiplie dès le premier niveau les possibilités de répartition des conversations – reste à déterminer si cela facilite ou rend plus difficile le parcours d'un internaute venu chercher un type d'information en particulier. Le premier niveau en français reste très général et l'affinement de la recherche se fera au deuxième niveau.

Au deuxième niveau, cette dissemblance entre français et vietnamien s'accentue, puisque l'ensemble des rubriques proposées tous forums confondus en français s'élève à 13[5], lorsque le site vietnamien totalise une cinquantaine de rubriques, ainsi que des clubs fermés (réservés aux administrateurs, modérateurs, ou un groupe restreint de membres). Pour la constitution du corpus, les rubriques ne seront donc pas traitées exactement de la même manière dans les deux contextes.

Tab. iii.3 : Deuxième niveau de structuration sur le site français

id	titre du forum	id	titre de la rubrique
10	Forum VIH/SIDA, IST	1	Forum général sur le VIH/SIDA
		24	Forum Infections Sexuellement Transmissibles (IST)
		34	TPE - Traitement d'urgence contre le VIH
		36	PrEP, TASP, etc.
12	Forum Hépatites	29	Forum général sur les hépatites
		14	Suivi et traitement des hépatites et co-infections
27	Forum Séropos	8	Traitement VIH & Suivi Médical
		2	Entraide entre Séropositi·f/ve·s
		6	Forum Sérodifférence & Proches
13	SIS Association	4	Sida Info Service - Hépatites Info Service - Ligne Azur
		35	Le comptoir de SIS
9	Santé sexuelle	17	Santé sexuelle Lesbiennes Gays Bi Trans
18	Détente	5	Détente :-)

5. 12 hors connexion, plus la rubrique Détente, réservée aux membres connectés.

Tab. III.4 : Deuxième niveau de structuration sur le site vietnamien

id	titre	traduction
1	Hướng Dẫn Thành Viên Mới	Guide pour les nouveaux membres
78	Hướng Dẫn Sử Dụng Forum	Guide d'utilisation du forum
81	Xét Nghiệm HIV và Xử trí Phơi Nhiễm	Dépistage et TPE
82	Thông Tin cảnh Giác	Alertes
10	Nội dung quan trọng	Contenu important
1	Thông báo	Remarques
7	Góp ý của bạn	Votre avis
70	Họ đã sống như thế !	Ils vivent ainsi !
3	HIV - Điều không mong muốn	VIH – L'indésirable
6	Thông tin về HIV	Informations sur le VIH
21	Tủ thuốc cho bạn	Des médicaments pour vous
36	Cùng chia sẻ tâm sự	Partager son état d'esprit
49	Địa chỉ cần biết	Adresse utiles
64	Mang thai, sinh em bé	Grossesse, accouchement
75	Vấn đề pháp lý có liên quan HIV/AIDS	Questions juridiques liées au VIH/sida
4	Tuổi trẻ & lối sống	Jeunesse et mode de vie
16	Giao lưu & Kết bạn	Echanger et se faire des amis
20	Nhịp sống giới trẻ	Mode de vie des jeunes
25	Góc tùy bút	Au fil de la plume
26	Ma thúy và Mai dâm	Drogue et prostitution
28	Hè hè... cười cái coi... hì hì...	Hi hi, pour rire...
30	Lên án & Cảnh tỉnh	Critiques & coups de gueule
38	Ước mơ của bạn và tôi	Vos rêves et les miens
69	Exchange rellations, Make friends[...]	Echanger, se faire des amis et partager

Tab. iii.4 : Deuxième niveau vietnamien (suite)

id	titre	traduction
5	Tình yêu & Giới Tính	Amour & Sexualité
17	Tình yêu là...	L'amour c'est …
19	Điều khó nói	Choses difficiles à dire
22	Quan hệ an toàn ...	Safe sex
24	Trung tâm tư vấn tình ...	Courrier du cœur
35	Chút tình gửi gió ...	Un peu d'amour porté par le vent
50	Khám phá giới tính	Explorer la sexualité
57	Là người đồng tính nam	Etre gay
58	Là người đồng tính nử	Etre lesbienne
8	Vòng tay cộng đồng !	Communauté, unissez-vous !
27	Thêm một góc nhìn	Donner son point de vue
31	Khi hai trái tim cùng nhịp...	Quand deux cœurs battent
33	Họ vẫn bên nhau	Ils sont toujours ensemble
71	Mái ấm Tình Thân	aide psychologique
13	Hoạt động Offline	Activité Hors ligne
48	Chúng ta offline	Nous, dans la vie non connectée
55	Việc tìm người - Người tìm việc	Recherche d'emploi
59	Hoạt động Từ thiện	Bénévolat
14	Hỏi và đáp liên quan đến HIV/AIDS	Questions et réponses relatives au VIH/sida
56	Kiến thức cơ bản về HIV/AIDS	Connaissances de base sur le VIH/sida
62	Quan hệ tình dục, bao cao su[...]	relations sexuelles, préservatif, lubrifiant
63	Sử dụng Ma túy, Kim tiêm[...]	drogues, seringues, objets coupants
65	Xét nghiệm HIV	dépistage du VIH
66	Phơi nhiễm HIV	Exposition au VIH

Tab. III.4 : Deuxième niveau vietnamien (suite)

id	titre	traduction
67	Khác	Autres
68	Kỳ thị và phân biệt đối xử	Stigmatisation et discrimination
72	Phòng Giải Tỏa Tâm Lý	Soutien moral
18	Hành trang vào đời - Kỹ năng sống	Bagages pour affronter la vie
53	Nghệ thuật sống	Art de vivre
74	Ứng phó nghịch cảnh	Faire face à l'adversité
11	xử lý lỗi kỹ thuật Diễn đàn [...]	Gestion des erreurs techniques [...]
45	Câu lạc bộ Spam	Club des spams
46	Giải đáp thắc mắc	Dépannage
60	Nhà kho	Stockage
12	CLB Những người điều hành diễn đàn	Club des gestionnaires du forum

Une telle dissemblance inciterait même à rapprocher le premier niveau vietnamien (11 « forums » thématiques) du deuxième niveau français (13 rubriques). En effet, nous trouvons des correspondances entre les intitulés des rubriques françaises de second niveau et les intitulés des forums thématiques vietnamiens de premier niveau :

Français (niveau 2)		Vietnamien (niveau 1)
(1) Forum VIH/SIDA, IST	⟺	(14) Questions et réponses relatives au VIH/SIDA
(2) Entraide entre Séropositi·f/ve·s	⟺	(8) Communauté, unissez-vous !
(5) Détente	⟺	(4) Jeunesse et mode de vie

Cependant nous trouvons tout de même plus de correspondances en com-

parant les rubriques de deuxième niveau entre elles :

Français (niveau 2)		Vietnamien (niveau 2)
(1) Forum général sur le VIH/SIDA	⟺	(6) Informations sur le VIH (56) Connaissances de base sur le VIH/SIDA
(8) Traitement VIH & Suivi Médical	⟺	(21) Des médicaments pour vous
(34) TPE - Traitement d'urgence contre le VIH (36) PrEP, TASP, etc.	⟺	(21) Des médicaments pour vous (66) Exposition au VIH (81) Dépistage et TPE
(2) Entraide entre Séropositi·f/ve·s	⟺	(71) aide psychologique[6]
(17) Santé sexuelle Lesbiennes Gays Bi Trans	⟺	(57) Etre gay (58) Etre lesbienne
(5) Détente :-)	⟺	(28) Hi hi, pour rire...

Comme nous pouvons le constater, l'idéal de deux forums parfaitement parallèles est loin d'être atteint, mais c'est le propre des données réelles de ne pas l'être, et c'est là l'intérêt fondamental de notre discipline : confronter et travailler à adapter des méthodologies théoriques d'analyse à des données produites en contexte réel.

Un élément vient s'ajouter à la difficulté de s'y retrouver sur le forum vietnamien : certaines rubriques ont des frontières très floues, par exemple dans le forum thématique intitulé "VIH - L'indésirable" nous trouvons aussi bien des informations pratiques (adresses utiles, médicaments, questions juridiques) que des espaces donnant la possibilité de donner libre cours à ses états d'âme. Alors que nous trouverons des rubriques s'en rapprochant à d'autres endroits du forum ("Au fil de la plume" dans le forum thématique "Jeunesse et mode

6. au sein de la communauté

de vie" ou "Connaissances de base sur le VIH/SIDA" dans le forum thématique "Questions et réponses relatives au VIH/SIDA", qui comporte même une rubrique simplement intitulée "Autres").

Parmi les grandes différences entre les deux contextes en ce qui concerne les rubriques proposées nous pouvons noter que :

- hépatites, IST et vie associative sont absentes des thèmes pré-établis par le forum vietnamien ;
- drogue, prostitution, histoires d'amour, grossesse, vie courante, etc. sont absentes des thèmes pré-établis par le forum français.

Cette constatation ne signifie pas que ces sujets ne sont pas abordés par les internautes, mais ils ne participent pas à l'organisation respective des deux forums. Ces thèmes pré-établis ne font que donner une indication sur l'idée que se font a priori les commanditaires, créateurs, administrateurs des deux forums à propos des centres d'intérêts de leurs publics, ainsi que sur le type de public visé.

2.2 *Délimitation des données à collecter*

2.2.1 *Sélection des rubriques à analyser*

Nous avons vu en quoi la structuration des données diffère d'un forumà l'autre. De plus, nous constatons que le point d'entrée principal diffère également. Avec un intitulé aussi général que « Forum général sur le VIH/SIDA » il est compréhensible que la plupart des nouvelles discussions soient ouvertes à cet endroit du forum français (cf. le diagramme III.7).

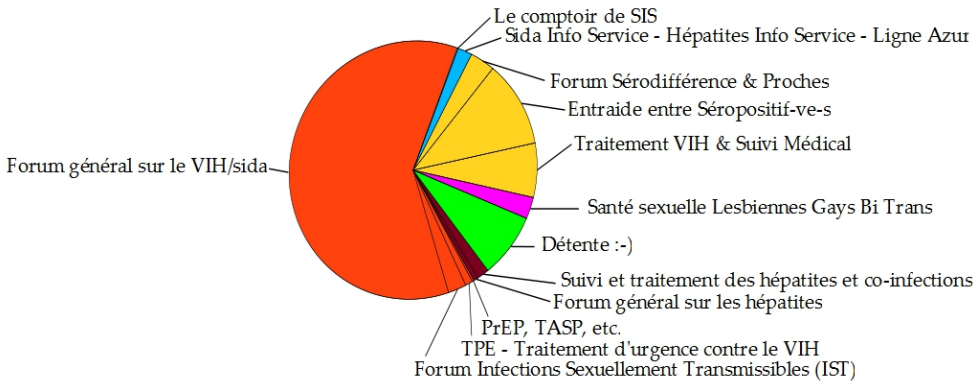


Fig. iii.7 : Répartition des interventions par rubrique sur le forum français

En revanche, dans le forum vietnamien, le foisonnement des rubriques rend moins prévisibles les lieux où se concentrera l'intensité de l'activité. [7]

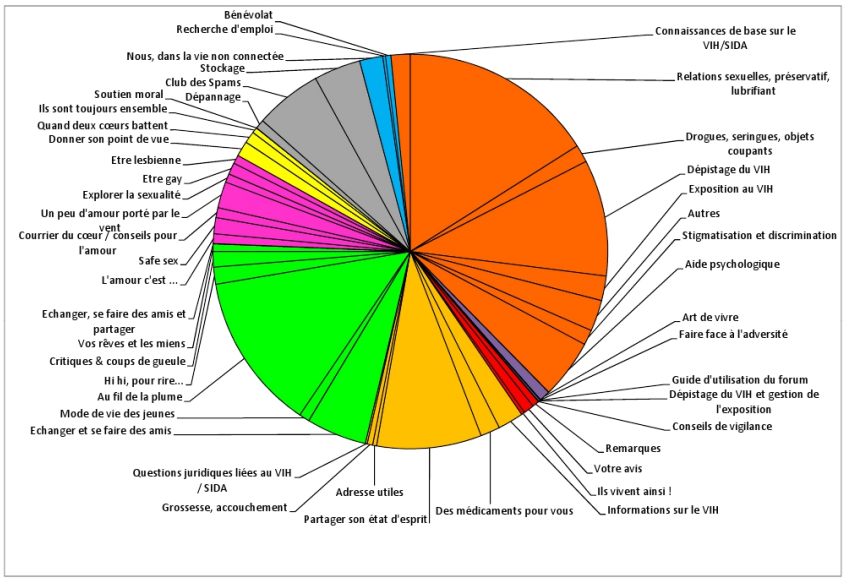


Fig. iii.8 : Répartition des interventions par rubrique sur le forum vietnamien

7. Les données ayant servi à constituer les diagrammes iii.7 à iii.10 ont été recueillies en juin 2016 pour le forum vietnamien et en juillet 2016 pour le forum français.

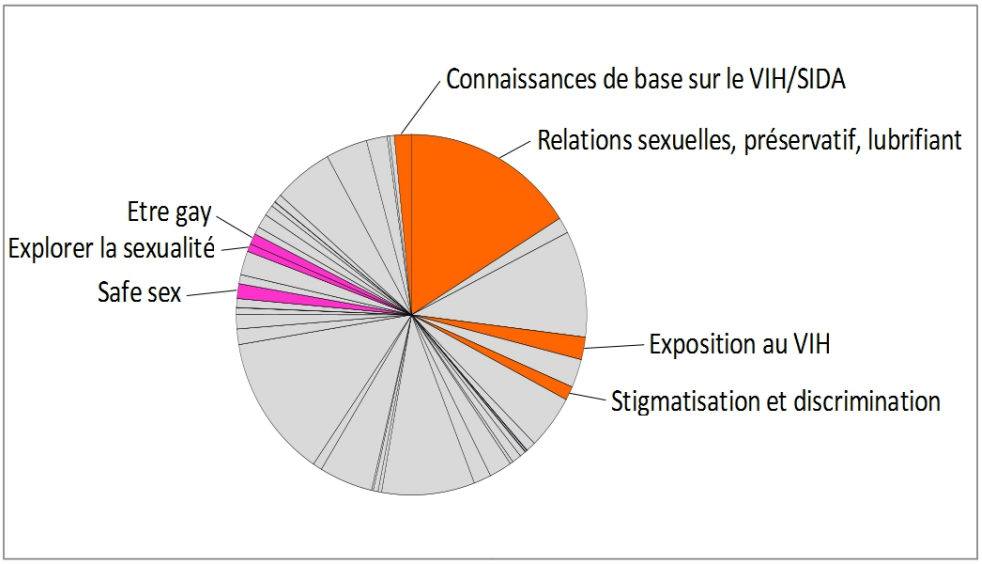
La rubrique la plus active est intitulée *Quan hệ tình dục, bao cao su, chất bôi trơn* (« Relations sexuelles, préservatif, lubrifiant »), autrement dit nous sommes immédiatement dans la dimension concrète et pratique.

Les rubriques proposées par le forum français étant plus générales que sur le forum vietnamien, nous avons retenu un plus grand nombre de rubriques en vietnamien (7 rubriques sélectionnées) qu'en français (5). Nous nous sommes concentrée sur un nombre restreint de rubriques, mais maintenant que toute la procédure est automatisée, il serait simple d'ajouter de nouvelles rubriques. Les rubriques retenues pour les analyses au sein de la thèse sont les suivantes.

Tab. III.5 : Rubriques sélectionnées sur le forum vietnamien

id	titre	traduction
5	Tình yêu & Giới Tính	Amour & Sexualité
22	Quan hệ an toàn ...	Safe sex
50	Khám phá giới tính	Explorer la sexualité
57	Là người đồng tính nam	Etre gay
14	Hỏi và đáp liên quan đến HIV/AIDS	Questions et réponses relatives au VIH/SIDA
56	Kiến thức cơ bản về HIV/AIDS	Connaissances de base sur le VIH/SIDA
62	Quan hệ tình dục, bao cao su[...]	relations sexuelles, préservatif, lubrifiant
66	Phơi nhiễm HIV	Exposition au VIH
68	Kỳ thị và phân biệt đối xử	Stigmatisation et discrimination

Fig. iii.9 : Rubriques sélectionnées sur le forum vietnamien



Nous avons concentré notre attention sur les forums thématiques les plus susceptibles de donner un aperçu des comportements à risque et des pratiques réelles des internautes vis-à-vis du VIH. Alors qu'à la lecture du forum français il semble que sa cible majoritaire soit la population homosexuelle, la rubrique "Pour les homosexuels" du forum vietnamien pourrait laisser entendre en creux que le reste du forum s'adresse avant tout aux hétérosexuels, ou du moins que les sujets s'intéressant spécifiquement aux pratiques homosexuelles sont classés dans une rubrique à part. Cependant, gardons-nous de tirer des conclusions trop hâtives, puisque le forum français propose également une rubrique intitulée "Santé sexuelle LGBT". Les deux forums ont donc créé une rubrique s'adressant spécifiquement à cette communauté. Le forum français entier comptant une prise de parole majoritairement homosexuelle, il pourrait en être de même pour le forum vietnamien. C'est par l'analyse détaillée de son contenu que nous constaterons qu'en effet l'évocation des pratiques homosexuelles ne se cantonne pas à la rubrique qui leur est consacrée.

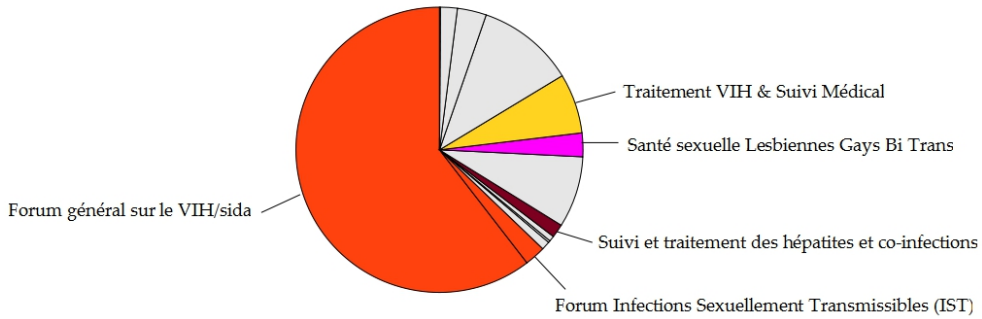
Cet exemple prouve que les rubriques créées a priori peuvent induire en erreur et mal orienter les déductions sans une analyse plus approfondie. Ce qui reste commun aux deux forums dans ce cas, se trouve dans leur organisation en rubriques a priori.

Pour les raisons que nous avons évoquées, il n'est pas aisé de trouver une correspondance parfaite entre les rubriques françaises et vietnamiennes. Nous avons sélectionné les rubriques suivantes au sein du forum français :

Tab. III.6 : Rubriques sélectionnées sur le forum français

id	titre de la rubrique
1	Forum général sur le VIH/SIDA
8	Traitements VIH & Suivi Médical
17	Santé sexuelle LGBT
24	Forum Infections Sexuellement Transmissibles (IST)
14	Suivi et traitements des hépatites et co-infections

Fig. III.10 : Rubriques sélectionnées sur le forum français



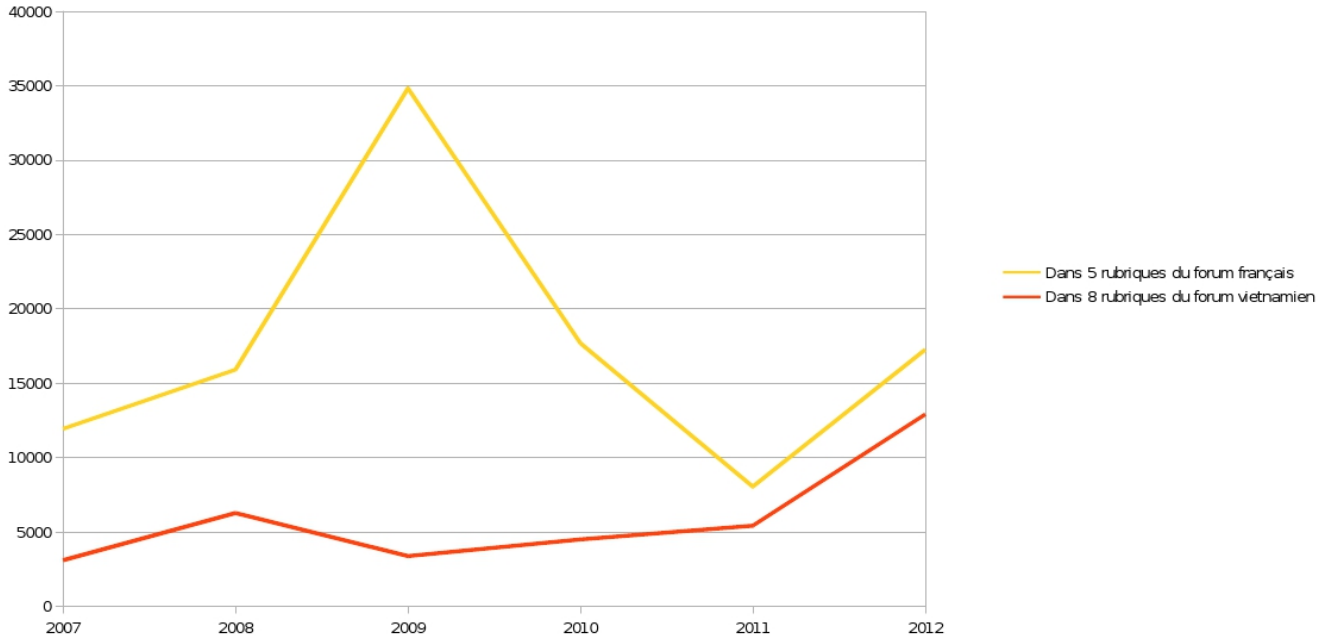
Une fois ces rubriques choisies, notre programme (exposé dans ce chapitre dans la partie 4, notamment le paragraphe 4) nous permet d'obtenir l'ensemble des conversations qui ont été ouvertes dans chacune de celles-ci, depuis leur création dans le forum, jusqu'au moment présent – celui de l'exé-

cution du programme.

2.2.2 *Choix de l'empan temporel*

La récupération des conversations est possible depuis leur ouverture, jusqu'au moment de la collecte. Cependant, dans le cadre de notre étude, l'aspect temporel (la diachronie autant que la chronologie) est une donnée essentielle. C'est pourquoi il est nécessaire de borner temporellement le corpus. Nous avons travaillé sur les échanges produits sur les forums au cours d'une durée de 6 années : de 2007 à 2012. Le forum vietnamien (i) a été créé en 2004, (ii) a connu un taux d'activité en perpétuelle croissance, notable à partir de 2007 et (iii) est toujours actif en 2016. Dans le contexte français, les réponses aux interrogations se faisaient depuis 1990 par téléphone. Sur Internet, le forum français (i) a été créé en 2003 et (ii) c'est à partir de 2006 que la diversification des moyens d'information s'est généralisée et que le forum a vu son activité croître considérablement, (iii) activité toujours forte jusqu'à mi-2016. Si des solutions étaient proposées dans le contexte français depuis déjà plusieurs décennies (avec la ligne d'écoute téléphonique), les deux forums suivent, malgré les contextes distincts, un cheminement général parallèle, à seulement un an près, et suivant la même tendance de croissance progressive. Cela dit, dans le détail des rubriques étudiées nous constatons une courbe du taux d'activité assez différente entre le forum français et le forum vietnamien (cf. le graphique iii.11).

FIG. III.11 : Nombre d'interventions publiées par année

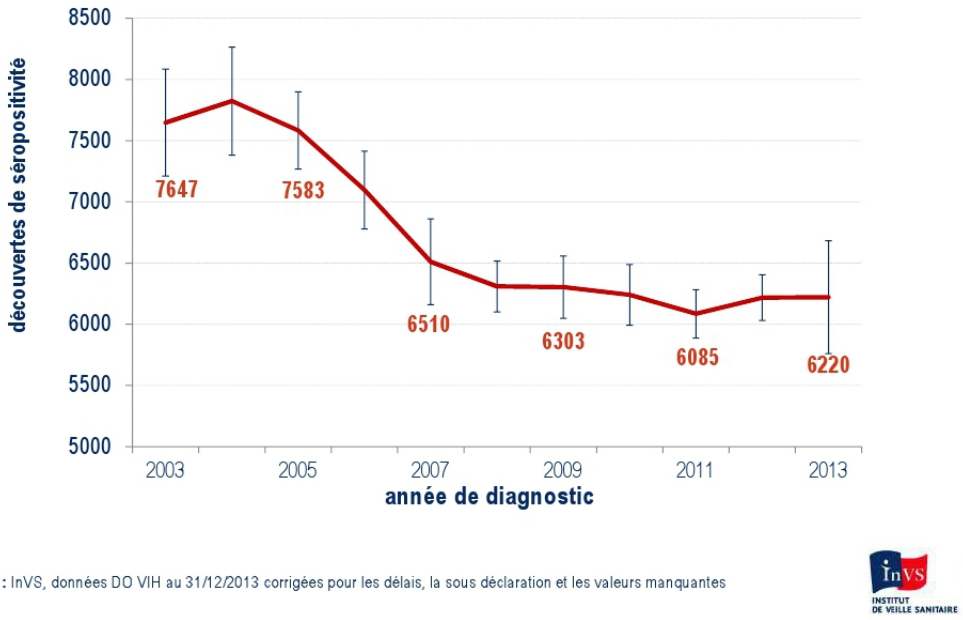


D'une part, en chiffres absolus, le nombre d'interventions est plus conséquent dans le corpus français que dans le corpus vietnamien alors que nous avons retenu moins de rubriques. Mais les rubriques retenues dans le corpus français représentent plus de 70% des interventions du forum français, lorsque les rubriques retenues dans le corpus vietnamien représentent moins de 25% des interventions du forum vietnamien[8]. Donc cela n'indique pas une activité plus importante en français qu'en vietnamien, c'est même le contraire si l'on compare le nombre total d'interventions publiées sur chacun des deux forums (230 000 sur 13 ans pour le français, contre 350 000 sur 12 ans pour le vietnamien). Différence plus saillante : les rubriques étudiées (qui représentent 70%) du forum français ont connu un pic d'activité en 2009, puis un affaiblissement progressif de l'intérêt jusqu'en 2011, et ont retrouvé une activité similaire à 2006 depuis 2012, alors que le nombre d'interventions publiées

8. Ces proportions peuvent être visualisées respectivement dans les diagrammes III.10 et III.9.

par les vietnamiens dans les rubriques étudiées (25% du forum) a suivi une tendance ascendante d'année en année. Nous n'avons pas trouvé d'explication à cette différence. Ce n'est pas du côté de la courbe du nombre de nouvelles infections qu'il faut chercher :

Fig. iii.12 : Nombre de découvertes de séropositivité par année.



Source : InVS, données DO VIH au 31/12/2013 corrigées pour les délais, la sous déclaration et les valeurs manquantes

L'année 2009 ne présente pas de caractère inattendu dans la tendance générale du nombre de sérologies positives. Peut-être que l'explication est à rechercher du côté de la visibilité du site ou du forum, mais nous n'avons trouvé aucune donnée attestant une telle hypothèse.

2.2.3 *Choix concernant le corpus institutionnel*

Que ce soit en français ou en vietnamien, dans un souci de cohérence, nous avons choisi de constituer le corpus institutionnel à partir des textes publiés sur les sites-*vitrines* auxquels sont associés les forums. En effet, les deux forums que nous avons choisis, *forum.hiv.com.vn* et *forum.sida-info-service.org*,

sont chacun rattaché à un sites présentant des articles d'actualité ou d'information dans le domaine concerné (i.e. le VIH), respectivement *hiv.com.vn* et *sida-info-service.org*. Nous les qualifions de « vitrines » (i) en raison de leur existence plus ancienne que celle de leur forum, (ii) parce que les discussions informelles ayant lieu sur les forums viennent en complément des informations présentées par les sites, et aussi (iii) parce qu'ils représentent la réalisation émergeant de la volonté des pouvoirs publics de proposer un espace de présentation et de transmission des connaissances dans le domaine.

2.2.3.1 *Collecte des données sur le site d'information institutionnel HIV Online*

La méthodologie de récupération des données sur le site vietnamien est présentée dans l'annexe A. La sélection des rubriques a été faite dans l'objectif de rendre possible une comparaison avec le forum attenant, donc en repérant des similitudes sur le plan thématique. Les 5 rubriques retenues sont exposées dans le tableau III.7[9]

Tab. III.7 : Rubriques institutionnelles sélectionnées sur le site vietnamien

rubrique	id	traduction
Kiến thức HIV	01	FAQ sur le VIH
HIV / Thông Tin	02	Informations sur le VIH
HIV / Thuốc và Sức khỏe	03	Médicaments et santé
Giới Tính / Tư Vấn	04	Conseils sur la sexualité
Giới Tính / Chuyện hai người	05	Amour et Sexualité

9. Au sein du corpus, contrairement aux rubriques du forum, les identifiants des rubriques institutionnelles ne correspondent pas à leurs identifiants sur le site, car au moment de la constitution du corpus ces derniers n'étaient pas affichés de manière visible, c'est pourquoi nous avons attribué aux rubriques des identifiants arbitraires. Depuis 2014 l'intégralité de l'architecture du site a été modifiée, ce qui a affecté la formation des URL, qui incluent maintenant de manière transparente les identifiants des rubriques. La procédure de constitution de corpus par aspiration de données étant automatisée pour ce site (et adaptée à la nouvelle architecture de 2014), nous pourrions facilement la relancer avec les identifiants réels du site, mais dans un souci de cohérence dans la rédaction du manuscrit, nous choisissons de conserver les identifiants arbitraires de 01 à 05 pour faire référence au corpus de travail.

Autre élément à noter : tout comme pour le forum (cf. le paragraphe iv.2.1.2 Créativité des rubriques , page 120), il est rare que les titres des rubriques, leurs URL ou leur positionnement dans l'organisation générale du site subissent des modifications, pourtant cela est arrivé au cours de cette thèse. Le contenu en revanche n'a pas été touché, il faut donc uniquement corriger les titres et les chemins d'accès dans l'arborescence du site, autrement dit les URL internes. (Dans le tableau suivant (iii.8), les URL indiquées sont relatives, la racine commune du forum étant `http://forum.hiv.com.vn`.)

Tab. iii.8 : Modifications sur le site vietnamien

	ancienne architecture	nouvelle architecture
01	Khái quát */khai-quat-hiv*	Kiến thức HIV */Information/QuestionHIV*
02	Thời Sự HIV > > Thông Tin */hiv/thong-tin*	(pas de changement) */Thong-Tin-7*
03	Thời Sự HIV > > Thuốc và Sức khỏe */hiv/thuoc-suc-khoe*	Thời Sự HIV > > Thuốc điều trị HIV */thuoc-dieu-tri-hiv-9*
04	Giới Tính > > Tư Vấn */gioi-tinh/tu-van*	Giới Tính > > Chuyện ba người */chuyen-ba-nguoi-29*
05	Giới Tính > > Chuyện hai người */gioi-tinh/hai-nguoi*	(pas de changement) */chuyen-hai-nguoi-28*

2.2.3.2 *Collecte des données sur le site d'information institutionnel Sida Info Service*

La construction du corpus institutionnel français n'a pas fait l'objet d'une procédure automatisée. La procédure de sa construction n'apporterait pas d'informations supplémentaires sur la méthodologie, puisque l'une des raisons pour lesquelles ce site a été choisi est que son organisation suit le même schéma que le site vietnamien. Nous avons sélectionné 783 articles sur le site, ayant été publiés entre 2007 et 2012 , ainsi que 47 articles de 2013. Nous avons reproduit ces 830 articles dans autant de fichiers texte. Nous n'avons donc

conservé que le texte, sans plus de sémiotique. A cet ensemble de fichiers textes, est associé un fichier de métadonnées, dans lequel sont consignées les informations relatives à chaque article, dans l'optique d'une conformité avec les autres corpus. Il n'est pas question dans ce corpus de rubrique, donc un identifiant commun a été arbitrairement attribué à l'ensemble des articles : le nombre 99. Il n'est pas possible de reconstruire leur URL à partir des données conservées, tout au plus est-il possible de retrouver les articles en ligne grâce à leur titre.

2.3 *Choix de structuration des données en corpus*

Au cours de ce travail, nous avons veillé dès le début à conserver le maximum de métadonnées associées au texte, dans l'optique d'une analyse non logocentrée (mais pas multimedia pour autant), qui prenne en compte le contexte de production des énoncés, car nous considérons que les textes étudiés sont inséparables de leur contexte technologique, et que leur interprétation ne peut se faire sans celui-ci. Le texte reste cela dit le matériau d'étude principal.

2.3.1 *Gestion des différents types d'éléments sémiotiques*

La méthodologie présentée dans cette thèse ayant pour objet central le texte, nous n'avons pas traité les images. Nous devons pourtant admettre que les images publiées à l'intérieur même de l'intervention participent à part entière de la signification de ladite intervention, qui peut aller jusqu'à perdre totalement son sens une fois privée de l'image qui l'accompagne, voire n'être composée que d'image(s). Cependant, l'analyse sémiotique des images n'étant pas l'objet de la présente recherche, nous ne disposons pas des outils adaptés et nous avons écarté ce genre dans le présent travail. Nous avons

toutefois eu le souci de conserver pour chacune de ces images – dans tous les cas où ce fut possible – dans la version la plus brute du corpus (c'est-à-dire dans sa version antérieure à toute adaptation en vue d'une analyse), une adresse URL, dans le but de pouvoir y accéder ultérieurement. A nouveau, nous reconnaissons les limites de ce choix dans l'optique d'une véritable analyse ultérieure, du fait de la pérennité, assez rapidement précaire, de ces liens internet.

Nous avons également décidé de ne pas inclure la sémiotique des couleurs.

En revanche, les seuls éléments hybrides – entre l'image et le signe sémantique – que nous avons tenu à conserver sont les émoticônes, que nous avons codés en chaînes textuelles[10], car nous considérons qu'elles participent intensivement à la sémantique de l'énoncé, contrairement à l'avatar de l'auteur, ou d'éventuelles autres images d'illustration. Comme l'explique Halté (2013), l'émoticône est un signe qui imite une émotion, afin de la rendre perceptible au cours de l'énonciation d'un contenu, dans le cadre d'une communication médiée par un ordinateur. Halté a travaillé sur les conversations synchrones que sont les tchats, mais son étude des émoticônes vaut pour leur pendant asynchrone que sont les conversations sur les forums. A la suite de Halté, nous considérons que les émoticônes participent au sens des productions discursives, c'est pourquoi nous avons tenu à les inclure dans les données à analyser. Même s'il serait vain d'établir un dictionnaire de significations des émoticônes car leur interprétation dépend trop du contexte[11], nous pouvons en établir une liste finie pour chaque forum, puisque celle-ci est définie par les concepteurs des forums. Pour la même raison d'interprétabilité,

10. Voir en annexe M la liste complète des émoticônes utilisées sur les forums analysés.
11. Cf. http://tempsreel.nouvelobs.com/les-internets/20150512.OBS8809/ce-que-voulez-vraiment-dire-quand-vous-utilisez-des-emojis.html.

et également en raison de la grande créativité dans l'ensemble d'émoticônes proposées par chaque forum, ces listes ne sont pas facilement mises en parallèle d'un forum à l'autre, mis à part un nombre restreint de base, que constitue l'intersection de ces différents ensembles. Une liste d'émoticônes existe donc pour chaque forum.

Un autre élément, textuel celui-ci, a également été écarté : il s'agit de la signature. Cet élément est un énoncé choisi par chaque membre du forum, configuré pour être inséré à la fin de chacune des interventions qu'il publie. Ce texte se répète donc à chaque intervention d'un membre, tout comme son avatar ou son pseudonyme. Ce sont donc des éléments à écarter d'une analyse lexicométrique des échanges, car leur caractère redondant fausserait les calculs. (En revanche, ces éléments seront utiles à une étude portant sur les intervenant·e·s plus que sur les échanges).

2.3.2 *Profondeur de segmentation du corpus : la conversation ou l'intervention comme unité minimale ?*

Appréhender des données collectées en masse sur le Web impose de se poser la question de leur segmentation en unités de base, comparables entre elles, qui conservent une cohérence autant sur le plan thématique que structurel. Dans le cas des forums, la tâche sus-dite est aiguillée par leur construction même, dès leur conception : une hiérarchie établie qui va du forum à la rubrique, de la rubrique à la conversation et de la conversation à l'intervention. Pourtant, il reste à trancher la sélection du niveau structurel le plus pertinent pour saisir au mieux les données : est-ce réellement l'unité minimale que constitue l'intervention, ou ne serait-il pas plus pertinent de considérer la conversation dans son ensemble ?

Après avoir étudié les deux pistes, a émergé la nécessité de travailler au

niveau de la conversation. Structure spécifique au genre discursif développé dans les forums, la conversation (aussi appelée fil de discussion) est ouverte au sein d'une rubrique, en parallèle de celles déjà existantes. La conversation est initiée par une intervention initiative et comprend toutes les intervention publiées au sein de celle-ci en réaction aux précédentes, triées par ordre chronologique. Cette structure implique une unité thématique assez bien délimitée, ainsi que l'appartenance à une thématique plus globale, donnée par le niveau supérieur : la rubrique, qui regroupe un ensemble de conversations et produit par là même une cohérence thématique.

Nous pouvons ajouter qu'à cette unité structurelle et thématique appartiennent un titre et une intervention initiative.

- Le titre contribue à la cohérence et à l'unité thématique de l'ensemble des interventions publiées au sein de la conversation, puisque le plus souvent les membres du forum accèdent à la conversation par celui-ci. Ce comportement est dû à la présentation des conversations sur la page du forum : ces dernières sont présentées par leur titre, sous la forme d'une liste, à la manière d'un sommaire. On peut donc considérer que les conversations sont indexées par leur titre, d'où son importance capitale, intégrée par les internautes, qu'iels soient lect·eur/rice·s, contribut·eur/rice·s, ou surtout initiat·eur/rice·s d'une nouvelle conversation.
- L'intervention initiative possède un statut particulier, la distinguant des autres interventions qui suivront dans la conversation, puisque c'est par cette intervention qu'est lancée la nouvelle conversation. Elle comporte donc des informations méritant un traitement spécifique (cf par exemple l'étude des intervention-témoignage, p.159).

Par conséquent nous devons également prendre en compte un niveau plus fin de la structure du forum : l'intervention (tout en considérant celle-ci dans

son contexte).

Dans le cas d'une conversation concernant une demande d'information ou d'encouragement, nous pouvons considérer que les interventions suivant l'intervention initiative sont des réponses à celle-ci. Ainsi nous serons en présence d'une question dans le titre ou l'intervention initiative, avec de fortes probabilités de trouver la réponse dans le reste de la conversation. Nous pouvons donc scinder ce type de conversation en deux parties aux propriétés distinctes : le titre et l'intervention initiative d'une part, et l'ensemble des autres interventions d'autre part. Il en est de même concernant l'initiat·eur/rice, qui a la propriété d'avoir ouvert la conversation, d'être celui/celle qui pose la question et attend une réponse des autres membres du forum. Ainsi son statut restera distinct des autres intervenant·e·s durant toute la durée de la conversation, ses interventions suivantes correspondant à des précisions, des compléments d'informations à propos de ce pour quoi il cherche une réponse.

Dans le cas d'un récit de longue durée (parfois plusieurs années), le titre et l'intervention initiative n'ont pas un statut aussi distincts du reste, en revanche l'initiat·eur/rice le restera tout au long de le conversation, et c'est la totalité de ses interventions dans leur ensemble qui est à distinguer du reste. Les interventions d'autres membres se limiteront bien souvent à des réactions au récit principal, des encouragements. Ce type de conversation s'apparente aux productions discursives observables dans les blogs.

Dans le cas d'un débat cette fois, même le statut de l'initiat·eur/rice ne sera pas systématiquement à distinguer du reste de la conversation.

Comme nous l'avons évoqué, la segmentation en conversation n'est pas suffisante et, si celle-ci constitue la structure principale à laquelle nous nous intéresserons, il nous sera parfois nécessaire de pouvoir accéder, au sein d'une conversation, au niveau inférieur de segmentation des données, à savoir l'intervention,

que nous pourrions qualifier d'unité minimale. Ce niveau de structuration apporte des informations distinctes et complémentaires. Par exemple, nous l'avons vu, il est souvent utile de considérer à part le statut de l'intervention initiative pour ses particularités, ou encore, il est utile de pouvoir comptabiliser le nombre d'interventions dans une conversation afin de mesurer son attractivité. En effet, se contenter de compter le nombre de mots serait trompeur, puisque la taille des interventions peut fortement varier, et une conversation totalisant 2 longs articles scientifiques peut afficher un nombre de mots égal à une conversation constituée de courts mais nombreux échanges.

Nous conserverons donc la possibilité des deux segmentations, la conversation dans son ensemble restant l'unité sur laquelle se portera le plus notre attention. En outre, nous avons évalué qu'il est peu pertinent de considérer l'intervention hors de son contexte de publication, autrement dit hors de la conversation dans laquelle elle a été publiée, et sans pouvoir accéder aux interventions qui la précèdent, l'aspect chronologique étant un paramètre essentiel dans l'étude des productions textuelles des forums. Pour conclure, nous considérons donc que la segmentation en interventions est un complément d'information à la segmentation principale en conversations.

3 TRAITEMENT DES DONNÉES POUR L'ANALYSE

3.1 *Choix des outils de traitement automatique des langues*

3.1.1 *Choix des outils de segmentation du vietnamien*

Nous avons sélectionné l'outil vnTokenizer pour effectuer la tâche de segmentation[12]. L'opération qui, en vietnamien, consiste en l'union de syllabes pour marquer les unités de sens est effectuée par vnTokenizer en insérant des tirets bas (underscore ou low line) entre les unités graphiques formant une seule unité de sens. Il n'est pas parfait, d'autant plus pour traiter des données produites au sein d'un forum, mais il présente le double avantage de ne pas nécessiter un apprentissage[13] et d'être en accès libre et en open source, ce qui nous a permis de l'adapter aux spécificités de nos données, en l'enrichissant de nouvelles fonctionnalités, comme la gestion de la structure XML et de l'arborescence par rubriques, la normalisation des émoticônes selon un standard établi pour ce travail (voir l'annexe M), ou la gestion des citations : elles sont conservées dans le corpus brut, comme information annexe, afin de garder la possibilité d'en faire une analyse, mais écartées des corpus soumis aux ADTs, pour éviter la répétition factice d'énoncés.

3.1.2 *Choix des outils d'étiquetage morpho-syntaxique*

La tâche d'étiquetage en parties du discours du vietnamien a été discutée à la fin du paragraphe II.5.2.5. En l'état de la recherche dans ce domaine lors des analyses suivies dans cette thèse, nous avons choisi de nous passer d'outils d'étiquetage en partie du discours (pour ce qui est du vietnamien), faute d'y trouver un apport significatif.

12. Voir p.5.2.4
13. cf. l'outil JVnSegmenter, p.5.2.4

Le corpus français est, lui, étiqueté par Treetagger, au moment de son import dans le logiciel TXM (cf. la sous-partie suivante). Le français ne fait pas partie des langues peu dotées. Les outils développés pour traiter cette langue, s'ils sont moins nombreux que ceux traitant l'anglais, ont été éprouvés et intégrés dans les travaux développés par les chercheurs en sciences humaines. Par exemple, l'étiqueteur morpho-syntaxique Treetagger est intégré dans des plateformes comme TXM ou SATO[14].

3.1.3 *Positionnement quant au nettoyage des données*

Les corpus étant aspirés de manière automatisée sur internet, le travail de nettoyage de données dépend des analyses ultérieures et des résultats que l'on souhaite obtenir. Pour notre part, nous avons souhaité réduire ce nettoyage au minimum, d'une part pour rester au plus près des discours tels qu'ils sont produits et perdre le moins d'informations possible spécifiques à ce média, et d'autre part car notre but est de confronter les outils d'analyse tels que les logiciels d'ADT à ce type de données.

3.2 *Mise en forme pour les logiciels d'ADT*

Après les traitements précédents, les corpus nécessitent d'être mis en forme pour qu'ils soient interprétables par les logiciels d'ADT. Nous avons développé un programme à ces fins, qui automatise la procédure en fonction du format attendu par le logiciel voulu.

14. Cf. Dupuis et al. (2010).

3.2.1 *Mise en forme des corpus pour Lexico*

Dans le cas de Lexico, le programme transforme le corpus en un seul fichier texte balisé. Pour les corpus en vietnamien, ceux-ci ayant été au préalable segmentés avec vnTokenizer, il ne faut pas oublier de retirer le tiret bas (underscore) de la liste des séparateurs (*délimiteurs de forme*) proposée par défaut dans l'interface d'import de Lexico.

Le fonctionnement de Lexico se base sur des calculs de bas niveau. Cette donnée a également certaines conséquences sur la préparation des corpus. Par exemple, il est nécessaire d'ordonner le corpus chronologiquement avant son import s'il est prévu d'analyser la courbe d'accroissement du vocabulaire. De même, il est indispensable de prévoir en amont quels seront les critères de partitionnement en sous-corpus (par exemple en année ou en mois), afin d'inclure avant l'import toutes les balises qui seront nécessaires par la suite. Ce partitionnement consistera en une répartition par le logiciel des segments de texte en fonction des critères marqués par ces balises. Encore faut-il avoir déterminé à quoi correspond un segment de texte : nous retrouvons ici la distinction, qui a été discutée au paragraphe 2.3.2, entre la conversation et l'intervention. L'interprétabilité du corpus par Lexico nécessite que, dans le fichier texte unique représentant l'ensemble du corpus, ces segments soient délimités par des caractères dédiés. A cette fin, nous avons utilisé la liste suivante de caractères :

Car.	Section délimitée
£	Corpus (informel/institutionnel)
¤	Rubrique
#	Conversation
µ	Intervention

Il faudra ajouter ces caractères à la liste de *délimiteurs de section* dans l'interface d'import de Lexico. Ce travail préalable présente par la suite l'avantage qu'une seule et même version du corpus peut être exploitée alternativement selon telle ou telle segmentation, sans avoir besoin d'un nouvel import de données dans le logiciel.

La construction de corpus par aspiration sur Internet a pour conséquence de produire des données inégales, malgré toutes les précautions qui seront prises. Il ne pourra jamais être exclu que le corpus comprenne plusieurs normes d'encodage, des textes mal orthographiés ou en *langage sms*. Ces variations pourront être à l'origine de biais plus ou moins importants, particulièrement saillants lors des tâches de classification de textes (comme l'AFC[15] dans Lexico). Afin de faire émerger de véritables spécificités plutôt que des particularités graphiques, il sera indispensable d'homogénéiser le corpus, en supprimant les textes minoritaires (ou au contraire ne conserver qu'eux si l'on cherche à les étudier – la classification aura alors permis de les faire émerger). La volonté de travailler sur des corpus ayant subi le minimum de prétraitements trouve ici une pierre d'achoppement. La nécessité d'homogénéité des données se montre prégnante et révèle la difficulté d'exploiter les logiciels d'ADT tels qu'ils ont été conçus, sur des productions discursives natif du web. Pourtant, si toutes leurs potentialités ne sont pas exploitables avec de telles données, ils gardent un intérêt certain pour explorer celles-ci, qui sera détaillé dans les chapitres d'analyse.

3.2.2 *Mise en forme des corpus pour TXM*

Plutôt que d'opter pour le format XML-TEI mis en avant par l'équipe, nous avons choisi de préparer nos corpus pour leur import sous un autre format,

15. Analyse factorielle des correspondances

également proposé (et qui sera transformé en format XML-TEI par le programme d'import de TXM) : le corpus est constitué de fichiers de texte brut (autant de fichiers qu'il comporte de textes, avec l'extention .txt), assortis d'un fichier structuré (au format CSV[16]) contenant les métadonnées associées à ces textes. Ce fichier doit respecter un format contraint afin d'être bien interprété : sa première ligne doit contenir les intitulés des métadonnées, le premier intitulé doit être "id", et le fichier doit porter le nom metadata.csv. Le choix de ce format nous permet en outre d'avoir accès en un seul fichier à une somme d'informations compilées dont nous pouvons immédiatement tirer des observations globales essentielles. Par conséquent, l'étude de ce fichier nous permet d'accéder à une vision d'ensemble du corpus avant même toute analyse textométrique. En revanche, ce format impose de faire le choix en amont du niveau de segmentation du corpus : en effet chaque texte est stocké dans un fichier distinct. Donc selon que nous travaillerons sur les conversations entières ou les interventions une à une, le nombre de fichiers ne sera pas le même. Un corpus de 5 conversations contenant chacune 10 interventions, donnera tantôt 5 fichiers tantôt 50 selon que les analyses porteront sur les conversations ou les interventions. Notre programme donne la possibilité de générer les deux formats. Un dernier point à noter à ce sujet est que TXM a été conçu pour traiter de gros fichiers plutôt que de nombreux petits fichiers et les performances d'import ne sont pas optimisées pour des corpus tels que ceux constitués de plusieurs milliers de petits fichiers textes représentant chacun une intervention.

16. Coma Separated Values, format de fichier présentant des valeurs séparées par un caractère défini, usuellement une virgule.

3.3 *Anonymisation*

Les méthodes que nous présentons s'affranchissent de la nécessité d'associer les discours à des informations personnelles concernant les internautes qui les ont produits. Pourtant nous restons consciente que nous analysons des informations qui touchent à la santé et à la sexualité. C'est pourquoi la protection des données personnelles ne peut être négligée.

L'utilisation du pseudonyme comme identification d'un membre en ligne fait déjà figure d'anonymisation des énonciateurs. Cependant nous souhaitons avoir la possibilité de ne pas afficher ces pseudonymes dans le cas où les données collectées seraient diffusées. Notamment par exemple pour l'utilisation d'outils en ligne.

Au moment de la constitution des corpus, une liste de l'ensemble des intervenant·e·s est constituée. La procédure d'anonymisation attribue à chaque intervenant·e de cette liste un code unique qui est utilisé pour l'identifier dans les corpus anonymisés. Ce code sera utilisé dans les métadonnées de chaque intervention pour désigner son aut.eur/rice, mais également dans le corps du texte, lorsqu'un·e intervenant·e s'adresse à un·e autre en l'appelant par son pseudonyme.

Il est fréquent (particulièrement dans le forum vietnamien) qu'un pseudonyme soit constitué d'une suite de caractères alphabétiques, suivie de plusieurs chiffres. Dans ce cas, lorsqu'un.e internaute désigne un.e autre par son pseudonyme, nous avons constaté qu'un usage répandu consiste à omettre les chiffres pour ne garder que la suite de caractères alphabétiques. Le remplacement des pseudonymes dans le texte inclue donc cet usage : ils sont recherchés dans leur forme extensive, ainsi que sans leur fin en chiffres.

4 DESCRIPTION DE LA CHAÎNE DE TRAITEMENT

TAB. III.9 : LANGAGES ET PROGRAMMES INFORMATIQUES UTILISÉS

	Langage		Programme	Étapes de traitement des corpus
Batch	PERL	Web :: Scraper	getTopics.pl	Collecte des URL (2007-2102)
			getPosts.pl	Collecte des conversations
	Java		vnTokenizer	Segmentation des corpus vietnamiens
	PERL		generate.pl	Génération des corpus pour les logiciels d'ADT
	SQL (MySQL Workbench)			Élimination des doublons du corpus français

L'ensemble des programmes et leurs options est détaillé en annexe A.

5 DESCRIPTION QUANTITATIVE DU CORPUS

Le corpus comprend 4 parties : deux sous-corpus en vietnamien (institutionnel et informel) et deux en français (idem). L'empan temporel s'étend de 2007 à 2012, c'est-à dire que le corpus se compose de tous les articles qui ont été publiés sur le site et de toutes les conversations qui ont été ouvertes sur les forums pendant cette période (sauf pour le corpus institutionnel français qui est statique). Nous avons expliqué que les conversations restent ouvertes et peuvent se poursuivre à l'infini, mais seules les interventions qui ont été publiées pendant la période définie sont sélectionnées. Le corpus informel vietnamien est composé à partir de 7 rubriques de *forum.hiv.com.vn*, celui français à partir de 5 rubriques de *forum.sida-info-service.org*. Le titre et le nombre de conversations collectées pour chaque rubrique est présenté dans le tableau III.10.

Tab. iii.10 : Description quantitative des corpus

Corpus		id	Titre	conv.
hiv.com.vn	Forum	56	Kiến thức cơ bản về HIV/AIDS *Connaissances de base sur le VIH/SIDA*	53
		68	Kỳ thị và phân biệt đối xử *Stigmatisation et discrimination*	127
		66	Phơi nhiễm HIV *Exposition au VIH*	263
		22	Quan hệ an toàn ... *Safe sex...*	325
		57	Là người đồng tính nam *Etre gay*	372
		50	Khám phá giới tính *Explorer la sexualité*	417
		62	Quan hệ tình dục, bao cao su, chất bôi trơn *Relations sexuelles, préservatif, lubrifiants*	769
			Total	**2 326**
	Site	02	HIV / Thông Tin *Informations sur le VIH*	6
		03	HIV / Thuốc và Sức khỏe *Médicaments et santé*	90
		04	Giới Tính / Tư Vấn *Explorer la sexualité*	790
		05	Giới Tính / Chuyện hai người *Amour et Sexualité*	798
			Total	**1 684**

	Corpus	id	Titre	conv.
sida-info-service.org	Forum	24	Forum IST	150
		14	Suivi et traitements hépatites et co-infections	279
		17	Santé sexuelle LGBT	413
		8	Traitements VIH & Suivi Médical	765
		1	Forum général sur le VIH/SIDA	5 911
			Total	**7 518**
	Site	99		830
			Total	**830**

Le nombre de mots pour chaque sous-corpus est présenté dans le tableau suivant, ainsi que la longueur moyenne des interventions. Le nombre de mots est calculé (après le traitement de vnTokenizer pour le vietnamien) avec la liste des séparateurs suivante : `. , :;!?/-"'&()[]{}<>` ainsi que les espaces simples, les tabulations et les retours à la ligne.

	Corpus	nb conversations	nb interventions	nb mots	moyenne
Vn	Forum	2 326	13 900	1,4 millions	100 mots/interv.
	Site	1 684 articles		1,13 millions	695 mots/article
Fr	Forum	7 518	96 705	6,8 millions	70 mots/interv.
	Site	830 articles		560 000	675 mots/article

CHAPITRE IV

Cartographie d'un forum de discussion

1 Introduction et définitions

A partir du corpus constitué, nous montrerons dans ce chapitre que les forums ne sont pas uniquement constitués de discours natifs du web (Paveau, 2012a), la doxa de Sarfati (2008, cf. p.43), discours que nous qualifions d'informels spontanés (DIS). Nous constaterons en effet que les discours institutionnels sont également présents dans les forums, sous la forme du discours institutionnel rapporté (DIR). Ces discours sont à rapprocher de la vulgate de Sarfati. C'est pourquoi il est plus exact d'envisager la typologie des discours des forums comme un continuum, du plus institutionnel au plus informel.

Avant d'étudier le discours informel spontané (DIS), ce que nous ferons dans le chapitre v, nous nous attacherons dans ce chapitre à établir des critères de catégorisation des différents discours qui s'élaborent sur les forums (partie 3). Pour cela nous devons dans un premier temps décrire les spécificités structurales des forums. Après la description formelle et fonctionnelle des struc-

tures (partie 2), nous proposerons des indices pour la détection du discours institutionnel rapporté (DIR) sur les forums. Dans cet objectif, nous caractériserons les structures en définissant des profils de rubriques et de conversations (sous-parties 3.1 et 3.2), puis nous étudierons le contenu des interventions, à l'aide de critères formels (nombre de mots, sous-partie 3.3), lexicaux (observations sur les phénomènes d'institutionnalisation, paragraphe 3.4.1) et discursifs (description d'un discours , paragraphe 3.4.2).

2 DESCRIPTION STRUCTURALE DES FORUMS

Dans cette section, nous allons considérer les trois niveaux de la structure des forums : les rubriques, les conversations et les interventions. La structuration particulièrement hiérarchisée des connaissances sur les forums est une caractéristique constitutive de leur élaboration, qu'il est essentiel de prendre en compte dans la perspective de leur analyse. Chaque niveau est porteur d'informations qui permettent de caractériser les données sous un angle spécifique. Tout d'abord les rubriques compartimentent le forum en grandes catégories thématiques ; puis, réparties dans les différentes rubriques, les conversations délimitent un espace d'échanges restreint, resserrant encore plus les thématiques abordées ; et, en dernier niveau de segmentation du forum, les prises de parole de chaque intervenant, à la manière des tours de parole dans un polylogue, peuvent être considérées comme unités de texte minimales.

2.1 Rubriques

La rubrique est le niveau de la structure du forum qui présente le caractère le plus figé. Les rubriques sont créées par les administrateurs du forum et non

par n'importe quel internaute. Elles sont déterminées de manière verticale et constituent le cadre dans lequel les internautes interviennent.

D'un point de vue pratique, les rubriques constituent les premières informations accessibles lors de l'affichage de la page d'accueil du forum. A la manière d'un sommaire, elles dessinent le cadre au sein duquel se fait l'accès aux informations plus détaillées. Elles sont un guide pour la navigation des internautes, que ce soit en tant que lecteur, en tant que contributeur, mais aussi pour un regard analytique sur l'organisation des connaissances ou des interactions.

Avant d'étudier le profil ou les préoccupations des *participants* du forum, penchons-nous sur les thèmes qui ont concentré l'attention de ses *concepteurs*. Le réel n'existe pas sans contexte, il est toujours perçu, interprété et organisé en étant plongé dans une culture, et ces rubriques sont pour nous autant d'indications quant à la compartimentation du réel tel qu'elle a été pensée par les concepteurs du forum dans leur contexte particulier. C'est aussi pour cela que, comme nous avons pu le constater, cette organisation varie beaucoup d'un forum à l'autre, en fonction de leur contexte d'existence. Par exemple, dans les cas qui nous intéressent, les douze rubriques proposées en français (cf. le tableau III.3, page 86) face à la cinquantaine en vietnamien (cf. le tableau III.4, page 87) nous orientent soit vers une segmentation du réel très générique, peu riche en informations de prime abord, mais qui laisse une grande ouverture possible au sein de chaque rubrique, soit vers une segmentation foisonnante, voire anecdotique (avec des intitulés tels que *Nos rêves*, *Ils sont encore ensemble*,...) qui peut a priori laisser perplexe.

Ce rubriquage du réel par les concepteurs au moment de la création du forum est la première somme d'informations interprétables par tout type d'utilisateur du forum : aux lecteurs il facilite l'accès aux informations recherchées ;

à l'internaute qui souhaite ouvrir une nouvelle conversation, il impose de se positionner en référence à une base partagée par tous ; à qui veut observer les interactions au sein du forum, il donne des indications sur la culture au sein de laquelle le forum existe.

2.1.1 *Coquille vide ou colonne vertébrale : le cadre institutionnel*

Le rubriquage du forum présente un caractère paradoxal : il constitue la base commune en dehors de laquelle aucune prise de parole ne sera possible, mais il a été défini a priori, avant que toute conversation n'ait lieu. On pourrait donc le voir comme une simple coquille vide à remplir, ou au contraire comme la colonne vertébrale du forum à partir de laquelle se construiront tous les échanges langagiers.

Selon nous, l'organisation en rubriques représente le cadre institutionnel de ces échanges informels. En effet : d'une part, la conception des forums que nous avons étudiés émerge d'une volonté des pouvoir publics de fournir une plateforme d'accès à l'information[1], d'autre part, les rubriques n'émergent pas des échanges spontanés nés sur la plateforme, mais de la mise en place du cadre de départ, qui présente un caractère figé. C'est pourquoi nous lui attribuons ce caractère institutionnel, dans un contexte d'élaboration de discours informel.

2.1.2 *Actualisation et enrichissement du rubriquage*

Bien que les rubriques puissent être modifiées ou voir leur nombre augmenter une fois le forum déjà en service, ces modifications resteront exceptionnelles. De plus, l'information de la compartimentation initiale restera dans la plupart des cas accessible malgré les modifications ultérieures, grâce aux

1. Cf. les paragraphes iii.1.1, p.74 et iii.1.2, p.77.

identifiants incrémentiels des rubriques. Par exemple, dans les cas que nous avons étudiés, des rubriques on été ajoutées par les administrateurs au fur et à mesure des progrès de la médecine : dans le forum français (pourtant très peu productif en création de nouvelles rubriques) des rubriques ont été créées pour le sujet des nouveaux traitements à prendre dans les 48 heures suivant une exposition au VIH (TPE[2]), et pour le sujet encore plus récent des nouveaux traitements à prendre avant une éventuelle prise de risque (PrEP[3]). Leurs identifiants, numériquement les plus élevés de la liste, (respectivement 34 et 36, cf. le tableau III.3, page 86) nous permet de savoir que ces rubriques ont été crées en dernier, ce que nous confirme la date des premières prises de parole au sein de ces rubriques (respectivement, décembre 2012 et décembre 2014).

Dans le contexte vietnamien, au moment de la constitution du corpus nous trouvions des rubriques spécifiques au préservatif, ou aux risques dûs au partage de seringues, mais encore aucune rubrique consacrée aux traitements les plus récents. Pourtant, nous constaterons que ce thème était déjà très présent dans le forum. Finalement, deux rubriques ont été créées a posteriori (nous le vérifions par leurs identifiants : 81 et 82 et la date des premiers messages postés en leur sein : mai 2016 et janvier 2014) et mises en avant par leur positionnement en haut de page :

id	titre de la rubrique	traduction
81	Xét Nghiệm HIV và Xử trí Phơi Nhiễm	Dépistage et TPE
82	Thông Tin cảnh Giác	Alertes

Ainsi, le sujet des nouveaux traitements s'est donc institutionnalisé en pas-

2. , Post Exposure Profilaxis (PEP) en anglais.
3. Pre Exposure Profilaxis, i.e. traitements pré-exposition

sant du statut informel de sujet souvent abordé au fil des conversations, au statut de titre de rubrique.

Cette observation commune aux deux forums confirme qu'ils suivent une évolution parallèle à peu de décalage près. La création de nouvelles rubriques est un critère à prendre en compte dans les tâches de veille sur l'évolution des comportements et des sujets préoccupant les internautes, ainsi que sur leur institutionnalisation.

Lors de l'explication de la sélection des rubriques, parmi toutes celles mises à disposition par les forums, en vue de constituer le corpus (iii.2.1.2), nous avons détaillé les disparités de leur organisation en français et en vietnamien au niveau de leur structure générale, établie a priori, en tant que réceptacle des futurs échanges. Dans ce chapitre, nous nous concentrerons sur les rubriques sélectionnées pour constituer le corpus. L'objet ici n'étant pas la comparaison d'un forum par rapport à un autre, d'un contexte socio-culturel par rapport à un autre, les exemples qui y seront pris, contrairement au chapitre suivant, se restreindront à l'un ou l'autre forum par souci de cohérence.

2.2 *Conversations et Interventions*

Contrairement aux rubriques, la création des conversations et des interventions est laissée au libre choix des internautes. Les rubriques proposent de grands thèmes, au sein desquelles les internautes ont la possibilité d'ouvrir une nouvelle conversation, ou d'intervenir dans une conversation déjà ouverte.

2.2.1 *Description d'une conversation*

Les conversations, aussi appelées fils de discussion, sont réparties dans les différentes rubriques, en fonction de la thématique définie par ces dernières. Les conversations délimitent un espace d'échanges restreint, resserrant encore plus les thématiques abordées, souvent limitées à quelques unes. Les conversations sont composées de l'ensemble des interventions qui ont été publiées en son sein. Ces interventions sont triées chronologiquement, afin que la conversation, lorsqu'elle est consultée ultérieurement, conserve sa cohérence et sa lisibilité. Une conversation est ouverte par un membre, qui en est l'initiateur. Celui-ci définit un titre, qui sera valable pour toute la durée de la conversation, et publie une première intervention. Celle-ci recevra un certain nombre de réponses, de 0 à n. Les conversations restent théoriquement toujours ouvertes, ce qui signifie que tout membre peut intervenir dans n'importe quelle conversation à tout moment, même des années après la dernière intervention. n est donc théoriquement infini. Cependant le fonctionnement majoritaire de l'activité sur les forums entraine une obsolescence rapide des conversations : le comportement adopté majoritairement (et même encouragé par les administrateurs, pour des considérations de lisibilité) sera d'ouvrir une nouvelle conversation plutôt que de faire réémerger une conversation ancienne en y publiant une nouvelle intervention. Pourtant, le comportement inverse est également encouragé par les administrateurs, pour les mêmes considérations de lisibilité : si la nouvelle intervention constitue une véritable réponse à la conversation, elle ne doit pas amener à ouvrir une nouvelle conversation, mais bien en conserver la continuité, même si elle survient après une longue inactivité. Dans tous les cas, une conversation comporte toujours un titre, un initiateur et n interventions, triées par ordre chronologique. La date attribuée à la conversation correspond à l'instant précis de son ouver-

ture par son initiateur, soit celle de la première intervention (i.e. le message d'ouverture), car une conversation ne peut être ouverte sans la publication d'une première intervention.

2.2.2 *Description d'une intervention*

Une intervention correspond à un tour de parole dans le polylogue que constitue la conversation. Elle est le résultat d'un acte de publication de la part d'un internaute. Il s'agit d'un texte produit par ledit internaute et qui constitue l'unité de texte minimale que nous étudions. Elle peut être constituée d'un seul mot - voire d'une simple image, une émoticône, une série d'émoticônes,... - aussi bien que d'un texte long de plusieurs paragraphes. L'intervention s'insère dans une conversation en possédant un rang dans la chronologie de cette dernière. La date attribuée à l'intervention correspond à l'instant précis de sa publication en ligne par son auteur.

Nous distinguons plusieurs types d'interventions : nous appelons intervention initiative[4] celle qui ouvre une conversation, nous appelons intervention réactive celle qui est publiée à la suite d'une autre intervention. Il existe également le cas particulier de la conversation qui ne contient qu'une seule intervention. Nous appelons celle-ci intervention unique, ou encore conversation à intervention unique (Conv.IU) lorsque nous parlons des conversations. Nous reviendrons sur ce cas particulier au paragraphe 3.2.

3 TYPOLOGIE DES DISCOURS

Avant de caractériser les discours nous avons besoin de caractériser les structures dans lesquelles ils adviennent, et en premier lieu les rubriques.

4. Nous empruntons l'expression à Marcoccia (2003).

La méthodologie de constitution en corpus de données extraites des forums a maintenu la structuration originelle et caractéristique des forums en rubriques, ce qui permet d'établir des sous-corpus correspondant à celles-ci, afin d'éclairer chacune en regard des autres. Nous nous attacherons à dessiner des profils de conversations, dont l'agrégation et l'organisation nous permettra de dégager des profils de rubriques. Cet angle de vue nous permettra de rapprocher certaines rubriques entre elles, pour y percevoir une homogénéité de discours. Une fois cette première catégorisation effectuée, nous entrerons au sein des conversations, pour y distinguer plus précisément les différents types de discours qui s'y développent.

3.1 *Typologie des rubriques basée sur le nombre d'interventions par conversation*

3.1.1 *Catégories de conversations en fonction de leur nombre d'interventions*

En vue d'établir une typologie des rubriques, nous nous pencherons sur le nombre d'interventions échangées entre internautes au sein d'une conversation, selon la rubrique dans laquelle celui-ci a été ouvert[5]. Ce qui importe ici est le nombre d'interventions qu'a produit l'ouverture de cette conversation, quelle qu'en soit la longueur en nombre de mots. Dans le corpus du forum vietnamien, le nombre d'interventions par conversation va de 1 à 123. Nous avons déterminé quatre catégories de conversations :

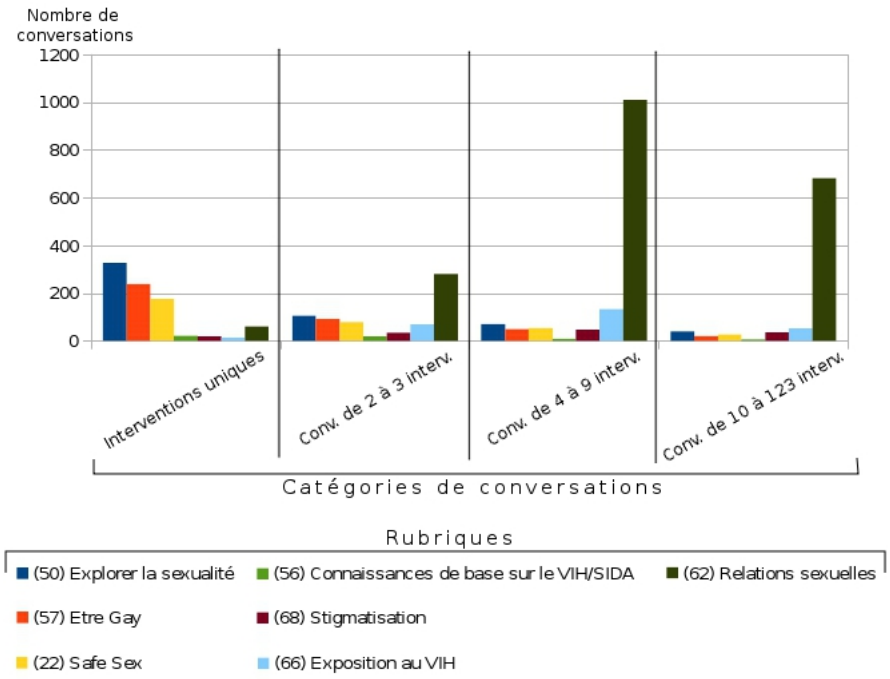
5. Nous disposons de ces données grâce au fichier de synthèse que constitue le fichier metadata.csv, produit pour chaque rubrique par notre programme (cf. le paragraphe III.3.2.2). Ce fichier permet d'avoir une vue d'ensemble sur les conversations (ou sur les interventions, selon le niveau spécifié à l'exécution du programme) produites au sein de la rubrique. Pour rappel, il s'agit d'un fichier tabulaire, avec dans chaque colonne une information particulière (le titre, le nombre d'intervenants, le nombre d'interventions ou la date d'ouverture de la conversation par exemple) et dans chaque ligne l'ensemble de ces informations pour une conversation donnée (se référer à l'annexe D). Pour établir cette typologie, nous exploitons la colonne indiquant le nombre d'interventions par conversation.

- les conversation à intervention unique (Conv.IU) ;
- les conversations de 2 à 3 interventions (Conv.2-3) ;
- les conversations de 4 à 9 interventions (Conv.4-9) ;
- les conversations de 10 à 123 interventions (Conv.10+).

Ces distinctions ont été ajustées en fonction de l'objectif visé : établir une typologie des rubriques à l'aide du critère du nombre d'interventions par conversation.

Pour chaque rubrique, les conversations ont été réparties dans ces quatre catégories (voir l'histogramme iv.1).

Fig. iv.1 : Nombre de conversations par catégorie

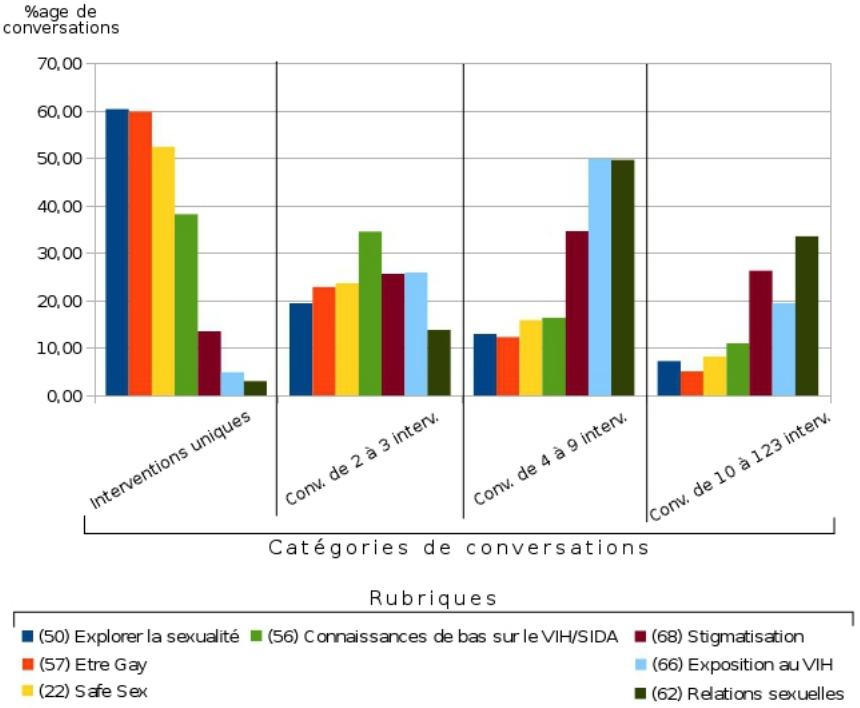


Dans l'histogramme iv.2 la répartition en *pourcentages* des conversations dans les quatre catégories, permet d'éviter le phénomène d'écrasement induit par la trop grande disparité[6] entre la rubrique (62) *Relations sexuelles etc.* et

6. Cette disparité est héritée de la répartitions des conversations dans le forum, cf. la figure iii.9, p.94.

l'ensemble des autres rubriques (plus de 1000 Conv.4-9 *vs* moins de 150 dans toutes les autres rubriques), visible dans l'histogramme IV.1.

FIG. IV.2 : Répartition des conversations par catégorie



3.1.2 *Profils de rubriques*

Par l'observation de cette répartition, il est possible de dégager trois profils de rubriques : les rubriques au profil majoritairement institutionnel, celles au profil majoritairement informel et les rubriques au profil médian.

Rubriques au profil institutionnel Les trois premières rubriques (50, 57 et 22, cf. la figure iv.2) comptent plus de la moitié de leurs conversations comme conversation à intervention unique (Conv.IU). Nous verrons dans la sous-partie 3.2 que le genre discursif des conversations à intervention unique relève du discours institutionnel rapporté. Les rubriques mentionnées héritent pour leur profil général des caractéristiques de ce genre discursif qui les compose majoritairement.

Rubrique au profil médian La rubrique (56) *Connaissances de base sur le VIH/SIDA* est une rubrique médiane, présentant un tiers de Conv.2-3, c'est la seule à en présenter une si grande proportion. Cette rubrique présente également un autre tiers de Conv.IU (cf. le paragraphe précédent).

Rubriques au profil informel Les trois dernières rubriques (62, 66 et 68, cf. la figure iv.2) présentent une majorité (un tiers à la moitié) de Conv.4-9, ainsi qu'un nombre important (un quart à un tiers) de conversations de plus de 10 interventions (Conv.10+).

3.1.3 Conclusion

Cette méthode de profilage des rubriques en fonction du nombre d'interventions des conversations qui les composent, a permis de rapprocher des rubriques entre elles, alors que la simple information de leur titre n'aurait pas permis de le faire, ce qui est particulièrement vrai dans le cas du forum vietnamien et sa cinquantaine de rubriques aux titres très variés. La représentation en pourcentage plutôt qu'en chiffres absolus nous a permis de faire émerger les faits suivants :

- trois grands profils de rubriques se dessinent distinctement ;
- la rubrique (66) *Exposition au VIH* semble avoir le même profil que la rubrique (62) *Relations sexuelles*, qui jusqu'alors concentrait toutes les attentions, du fait du foisonnement d'interventions, presque 10 fois plus important ;
- la rubrique (56) *Connaissances de base sur le VIH/SIDA* présente une caractéristique particulière qui n'est présentée par aucune autre : la grande proportion de Conv.2-3.

Dans l'objectif d'étudier les échanges discursifs informels spontanés, nous nous orienterons vers les rubriques qui présentent un nombre réduit de Conv.IU et une majorité de Conv.10+, autrement dit les rubriques auxquelles nous avons attribué un profil informel, en ce qui concerne le corpus vietnamien, il s'agira des rubriques (62), (66) et (68). Les Conv.10+ existent pourtant dans les rubriques (50), (57) et (22) et celles-ci devront faire l'objet d'attention également.

3.2 *Cas particulier des conversations à intervention unique*

Il existe au sein des forums un type de conversation à part : celles qui sont composées d'une seule intervention.

Nous avons vu au paragraphe 2.2.1 que les conversations restent théoriquement toujours ouvertes. Cependant, nous travaillons sur un corpus qui constitue une image fixe du forum, couvrant une période déterminée au moment de l'aspiration des données[7]. Nous sommes donc en mesure de parler de conversations à une seule intervention au sein de notre corpus, et de considérer ces dernières comme ne comportant aucune réponse.

7. En ce qui concerne notre corpus, cette période s'étend de 2007 à 2012, cf. le paragraphe III.2.2.2

Nous l'avons vu au paragraphe 3.1, ces conversations à intervention unique sont un critère clivant dans la détermination de profils de rubriques. Elles sont également un critère discriminant dans la catégorisation de discours au sein des forums : à la lecture, nous constatons que les interventions qui n'ont appelé aucune réponse sont systématiquement des textes provenant de sources extérieures, reproduits au moyen du procédé de *copier-coller*, publiés sur le forum au moyen de l'ouverture d'une nouvelle conversation, donc sans attache directe à aucun échange discursif existant, uniquement à une rubrique. Ces interventions n'appellent pas de réponses et n'en obtiennent pas. Il s'agit pour un internaute de publier une information pour que les autres la lisent, d'alimenter la base de connaissances commune en construction collaborative permanente qui s'enrichit à chaque intervention des membres. Ce sont donc des publications qui ont peu d'aspects en commun avec les discussions spontanées entre internautes à d'autres endroits du forum, elles ne relèvent pas du discours spontané, mais bien du discours institutionnel rapporté. Pour constituer des sous-corpus centrés sur le discours informel spontané, nous pouvons donc choisir d'écarter ces conversations à intervention unique.

Tab. iv.1 : Mise à l'écart des interventions uniques

Identifiant de rubrique	(50)	(57)	(22)	(56)	(68)	(66)	(62)
Nombre total de conversations	540	398	334	55	133	267	2032
Interventions uniques	**326**	**238**	**175**	**21**	**18**	**13**	**61**
Conversations >1 intervention	214	160	159	34	115	254	1971
Nombre total d'interventions		1088	1326	236	1108	1915	19812
Interventions non uniques		850	1151	215	1090	1902	19751

3.3 *Typologie des discours basée sur la longueur des interventions*

Nous l'avons exposé lors de la présentation des choix de structuration du corpus[8], nous avons choisi de travailler à deux niveau de segmentation du corpus en unités de textes : le niveau de la conversation et le niveau de l'intervention. Nous le constatons à présent : chaque niveau apporte des informations complémentaires. Au paragraphe 3.1, la longueur des *conversations* (en termes de nombre d'interventions) nous a permis de dessiner un profil de discours au niveau de la rubrique. Au paragraphe 3.2, l'étude du cas particulier des *conversations à intervention unique* nous a fait progresser dans l'affinage de la catégorisation de discours en fonction de la longueur des conversations. A présent, voyons en quoi la longueur des *interventions* (en termes de nombre de mots) peut nous aider à caractériser les discours.

Nous avons vu que certaines rubriques présentaient un profil informel, avec une forte majorité de conversations riches de multiples interventions. Ces conversations sont constituées d'interventions, qui sont la transcription des tours de paroles d'une conversation à plusieurs, intermédiée par ordinateurs distants, conversation qui compte autant de participants potentiels qu'existent de membres du forum. Pourtant nous constatons qu'au sein d'une conversation, la longueur des interventions varie, de quelques mots à quelques centaines de mots. De ce fait, la caractérisation des discours variera d'une intervention à l'autre, fortement influencée par sa longueur. Une conversation de ce type présente donc un profil général informel et un contenu composite, avec une diversité de textes et de discours, y compris le discours institutionnel, lequel intervient par le biais du procédé de *copier-coller*. Les internautes étayent leurs discours en reproduisant des textes provenant de sources extérieures. Nous qualifions ces interventions de discours institutionnel rappor-

8. paragraphe 2.3.2 du chapitre II, p.103

té (DIR). Le DIR peut être mobilisé au cours d'une conversation informelle spontanée, pour amener des éléments au débat. Par exemple, la conversation [VN « Comment faites-vous pour gérer vos examens de santé réguliers avec votre boulot ? »] compte 28 interventions, dont la 24ème et la 25ème sont des articles longs (comptant respectivement plus de 1500 et plus de 3000 mots) : le premier analyse les apports de la loi sur la lutte contre le VIH/SIDA, entrée en vigueur le 1er janvier 2007, par rapport à la loi de 1995. Le second reproduit le texte de loi en lui-même. Les deux interventions sont publiées à la suite par le même internaute. Ils viennent répondre à l'interrogation qui parcourt la conversation au sujet du caractère obligatoire ou non du dépistage. Ces textes, nus, sans commentaires, apportent une réponse figée : le dépistage se fait sur la base du volontariat sauf dans certains cas restreints (enquête de justice, ou certains emplois). Pourtant cette intervention ne met pas un terme à la conversation. La question posée en titre est plus vaste que la réponse qu'apportent ces textes législatifs. C'est pourquoi ils ne clôturent pas la conversation et la question de départ reste ouverte. Les interventions suivantes reprennent la conversation par de nouveaux témoignages personnels d'autres internautes. Cet exemple montre qu'au sein d'une seule conversation, les interventions ne sont pas toutes composées du même type de discours.

Pour résumer, nous faisons la distinction entre les différents types de discours suivants :

- la transcription du langage parlé sur le ton du dialogue oral, que nous qualifions de DIS,
- les témoignages personnels plus écrits, et
- la reproduction de textes institutionnels (DIR).

Les témoignages sont un discours intermédiaire : il s'agit d'interventions plus longues que la moyenne constatée en DIS, mais la longueur de l'intervention

est la seule caractéristique qu'elles partagent avec le DIR. Ces interventions étant plus écrites, nous ne pouvons plus parler de transcription du langage parlé, mais nous considérons qu'il s'agit encore de DIS.

Le critère de la longueur répartit les interventions de la manière suivante :

- Une intervention courte, de moins de 5 lignes a une probabilité quasi totale d'être du DIS.
- Les interventions de 5 à 15 lignes nécessitent davantage d'indices pour distinguer DIR et DIS.
- En ce qui concerne les interventions longues, de plus de 15 lignes, elles sont en grande majorité du DIS, mais peuvent également être un témoignage détaillé.

La longueur de l'intervention est l'indice le plus accessible, mais reste insuffisant pour qualifier finement les discours au sein des conversations.

3.4 *Typologie des discours à l'aide des segments répétés*

Nous avons décrit comment des critères de forme (nombre d'interventions par conversation, nombre de mots par intervention) pouvaient nous aider à distinguer et qualifier différents types de discours présents dans les forums. Nous allons maintenant nous appuyer sur le calcul lexicométrique des segments répétés pour qualifier ces discours.

Il s'agit ici d'analyser les segments de texte qui sont produits par plusieurs intervenants et à maintes reprises dans le corpus, pour étudier la stabilisation d'expressions dans notre corpus. Plus précisément, nous entendons par segment répété, toute suite de plus de 1 mot, répétée au moins 10 fois dans le corpus.

Lexico[9] nous permet de calculer la liste des segments répétés. Elle peut être

9. Pour une brève description, voir page 58.

exploitée sous deux angles d'analyse : la longueur du segment d'une part, sa fréquence d'autre part.[10]

- Segments répétés longs et discours institutionnel.

Le discours institutionnel est caractérisé par une forte stabilité lexico-syntaxique. Dans le but d'étudier le discours institutionnalisant qui existe au sein des échanges informels, nous devons donc nous appuyer sur des éléments indiquant une tendance à la stabilisation, par exemple des expressions figées. La probabilité pour qu'une suite de 3 mots se répète dans le corpus est statistiquement plus forte qu'une suite de 8 mots. Donc une suite de 8 mots répétée 20 fois est plus remarquable pour une recherche d'expressions figées qu'une suite de 3 mots répétée 200 fois ; et plus le segment qui se répète (pour rappel : au moins 10 fois dans le corpus) est long, plus il constitue une expression à caractère figé. La lecture de notre corpus nous amène à confirmer que les segments répétés les plus longs se trouvent dans les fragments de textes institutionnels. Ainsi, les segments répétés les plus longs héritent du trait institutionnel, trait qu'ils partagent avec le texte dans lequel ils apparaissent. Le focus pointé sur la longueur des segments sera donc en premier lieu utile dans le repérage de discours institutionnel dans les forums.

- Segments répétés fréquents et discours informel.

A l'autre extrémité du spectre, les segments répétés les plus fréquents nous renseignent sur l'usage majoritaire du vocabulaire sur le forum, par conséquent si nous déplaçons notre angle d'analyse en direction de la fréquence des segments répétés et non plus leur longueur, l'objet de l'attention sera à l'inverse la caractérisation du discours informel.

10. Des extraits de la liste des segments répétés pour chaque rubrique peuvent êtres consultés en annexe F. Les segments les plus fréquents y sont présentés par longueur de segment : d'abord les segments de longueur 2 les plus fréquents, puis les segments de longueur 3 les plus fréquents, etc. Les segments les plus longs se trouvent donc en fin de liste, et ceci pour chaque rubrique.

Le profil de la rubrique joue également un rôle dans l'analyse : selon qu'il apparaît dans une rubrique à profil institutionnel ou informel, un segment répété ne fournira pas les mêmes informations. La présence d'une expression figée longue - donc à caractère institutionnel - dans une rubrique à profil institutionnel est simplement caractéristique de la conversation, voire de la rubrique. Alors qu'apparaissant dans une rubrique à profil informel, elle permet de faire émerger les interventions institutionnels : en contexte informel, les plus longs segments se trouvent là où se niche le discours institutionnel rapporté (DIR), autrement dit les textes reproduits à partir de sources externes au forum, et qui constituent les exceptions du sous-corpus, essentiellement constitué de (DIS). Ainsi, la caractérisation des discours basée sur la longueur des interventions (cf. le paragraphe 3.3) se voit appuyée par l'étude des expressions figées.

3.4.1 *Étude des expressions figées en contexte informel*

En guise d'introduction méthodologique formulons que nous avons ici pour objectif de caractériser des textes institutionnels intervenant en contexte de discours informel. Notre attention est donc plus particulièrement portée sur les rubriques auxquelles nous avons attribué un profil informel : les rubriques (66) *Phơi nhiễm HIV* (« Exposition au VIH »), (68) *Kỳ thị và phân biệt đối xử* (« Stigmatisation et discrimination ») et (62) *Quan hệ tình dục, bao cao su, chất bôi* (« Relations sexuelles, préservatif, lubrifiant »), dans lesquelles nous chercherons à repérer la présence de discours institutionnalisant. A partir de la liste des segments répétés calculés par Lexico, triés par longueur, nous nous concentrons sur les plus longs en priorité, afin de sélectionner humainement les segments les plus pertinents.

Le segment répété le plus long dans la rubrique (68) est *biện_pháp can_thiệp*

giảm tác_hại trong dự_phòng lây_nhiễm (« mesures de réduction des dégâts dans la prévention de l'infection »). Le sous-corpus compte 11 occurrences de ce segment. Mais sa répartition est significative : 9 occurrences se trouvent dans la même intervention, les 2 dernières dans une seconde intervention. Il se trouve que ces deux interventions sont celles évoquées en exemple dans le paragraphe 3.3. Il s'agit là d'un cas typique où l'observation des segments répétés vient parfaitement corroborer l'observation de la longueur des interventions : les expressions figées longues font émerger les mêmes textes que le total de mots par intervention pour pointer le discours institutionnel au sein d'un corpus informel.

En réalité cette observation peut être rapportée à un segment de texte plus court : sur les 25 occurrences de *can_thiệp giảm tác_hại* (« intervenir pour réduire les dégâts ») 21 se trouvent dans le premier texte évoqué, 3 dans le deuxième, et la dernière occurrence nous permet de repérer un troisième texte institutionnel qui se trouve dans une autre conversation. Il s'agit également d'un texte long (880 mots). La présence simultanée de ces deux paramètres (texte long et présence d'expressions figées à caractère institutionnel) sont donc des critères puissants pour le repérage du discours institutionnel.

Le tableau suivant présente d'autres expressions figées longues trouvées en contexte informel et caractéristiques du discours institutionnel :

Expression figée	nb occ.	nb text.
Điều_trị nghiện (các) chất dạng thuốc_phiện bằng thuốc thay_thế « Le traitement de la dépendance aux substances opiacées par des médicaments de substitution »	13	1
phân_biệt đối_xử với người nhiễm HIV/AIDS « discrimination envers les personnes séropositives »	4	1
phân_biệt đối_xử với người nhiễm HIV « discrimination envers les personnes séropositives »	13	4

Les expressions figées qui donnent un trait institutionnel peuvent malgré tout figurer dans les textes informels, mais en nombre réduit. Par exemple, « người có HIV » (*séropositi.f/ve(s)*) dans la rubrique (68) a été trouvé 53 fois : 14 dans des textes de discours spontané (DIS) et 39 dans des textes de discours rapporté à caractère institutionnel (DIR). Lorsque cette expression apparaît en discours spontané, le texte n'en compte qu'une occurrence, alors que dans les textes institutionnels il peut être présent de une à seize fois dans la même intervention. Cette présence sporadique en discours spontané n'est pas suffisante pour retirer le trait institutionnel de cette expression. A l'inverse, l'expression « người có H » (*séropo(s)*) qui apparait également 53 fois dans la rubrique (68), est présent dans 4 textes rapportés, soit 12 occurrences, et 41 occurrences en discours spontané. L'expression « người có H » (*séropo(s)*) transmet donc un trait informel au texte dans lequel il apparaît, à l'inverse de l'expression « người có HIV » (*séropositi.f/ve(s)*) qui transmet son trait institutionnel. Nous devons donc considérer que certaines expressions figées partagent - c'est-à-dire reçoivent et transmettent - avec le texte dans lequel elles apparaissent le trait institutionnel, d'autres partagent le trait informel.

Tab. iv.2 : Exemples de distribution de segments répétés

Segment répété	nb. occ.	Distribution	
		informel	instit.
những người nhiễm HIV « les personnes atteintes du VIH »	20	15%	85%
người bị nhiễm HIV « personne atteinte du VIH » (connot. péjor.)	16	18,75%	81,25%
những người có HIV « les séropositifs »	10	20%	80%
những người có H « les séropos »	13	100%	0%

Nous retrouvons dans le tableau iv.2 la grande disparité de distribution

entre les segments *những người có H* et *những người có HIV*.

Pour résumer, dans notre corpus :

- les expressions figées les plus longues ont un trait institutionnel fort. Leur présence permet de repérer des textes de DIR.
- certaines expressions figées se trouvent très majoritairement en DIR à quelques exceptions près. Elles conservent leur trait institutionnel.
- d'autres expressions figées se trouvent majoritairement en DIS. Elles héritent du trait informel, qui permet de repérer des textes en DIS dans le reste du corpus.

Une nouvelle fois, nous voyons l'importance des aller-retours entre la lecture du corpus, et l'interprétation des résultats lexicométriques. Ici, la première typologie des expressions figées est constituée grâce au calcul des segments répétés puis affinée par la lecture de textes dans lesquels ceux-ci apparaissent. Cette typologie peut ensuite être généralisée à l'ensemble des textes du corpus afin de mieux les caractériser.

Pour les étapes de repérages des expression figées et leur analyse, se reporter à l'annexe J.

3.4.2 *Le discours des expert·e·s*

Certains cas de segments répétés sont à distinguer du cas général : une fois exclues les quelques erreurs techniques ou les duplications que nos procédures n'ont pas réussi à éviter, certains segments répétés longs nous donnent d'autres informations sur les discours. Par exemple, certain·e·s intervenant·e·s expert·e·s, qui répondent aux questions des internautes et auxquels un grand nombre d'internautes s'adressent directement, répètent régulièrement les mêmes conseils, en utilisant la même technique de *copier-coller* que lorsqu'il s'agit de

reproduire un texte d'une source extérieure. Dans le cas présent, la source n'est pas extérieure au forum mais antérieure à la conversation : l'intervenant·e expert·e se répète en se citant lui-même (ou un·e autre expert·e), auprès d'un nouvel internaute. En ce qui concerne le type de discours, nous considérons qu'il s'agit bien de discours institutionnel puisqu'il est tenu par un représentant reconnu du discours officiel (du fait que le forum émane d'une volonté ministérielle). Quant à savoir s'il est rapporté ou non : il n'est pas importé sur le forum à partir d'une source externe, mais il est repris par *copier-coller* à de maintes reprises et par différents intervenants. De plus, bien que les expert·e·s écrivent eux aussi de manière spontanée, le fait qu'ils répètent mot pour mot des fragments de texte entiers, rend difficile la qualification du discours comme spontané.

Outre la caractérisation du discours des expert·e·s, l'étude détaillée de ces interventions nous en apprend plus sur le discours officiel tenu sur le forum en matière de prévention, ainsi que sur son évolution au cours de l'empan temporel du corpus.

Penchons-nous sur des exemples de conseils répétés par des expert·e·s.

1. « *Loại dịch Lượng HIV Máu Rất nhiều Dịch tiết âm_đạo Nhiều Tinh dịch Nhiều Sữa mẹ Trung bình Nước ối Trung bình Nước bọt Hầu như_không có Mồ_hôi Hầu_như không có Nước_mắt Hầu_như không có Nước_tiểu Hầu_như không có* »

   « Quantité de VIH par type de fluide. Sang : élevée - Cyprine : élevée - Sperme : élevée - Lait maternel : moyenne - Liquide amniotique : moyenne - Salive : insignifiante - Sueur : insignifiante - Larmes : insignifiante »

Le fragment n°1 est répété 13 fois dans le corpus : 11 fois par l'expert nu-

méro un du forum[11], celui à qui tout le monde s'adresse (même si c'est un·e autre expert·e qui répondra), deux autres expert·e·s l'emploient 1 fois chacun. Il s'agit d'une liste de fluides corporels associés au niveau de risque qu'ils soient porteurs du VIH. Cette liste générique est utilisée pour répondre aux cas particuliers des internautes.

2. « *Để để không phải lo_lắng về việc lây_nhiễm HIV cũng_như các bệnh truyền_nhiễm lây qua đường tình_dục, bạn nên tránh xa các dịch_vụ " ăn bánh trả tiền " của người hành_nghề mại_dâm* »

   « Pour ne plus avoir à t'inquiéter à propos de la contamination au VIH et aux MST, le mieux est de ne pas avoir recours aux services sexuels des travailleu.r/se.s du sexe »

Le fragment n°2 répété 12 fois dans le corpus, 10 fois par l'expert numéro un du forum, 2 fois par un·e autre expert·e qui reprend le texte de l'expert numéro un. Ce fragment fait partie d'un texte plus long, repris sous plusieurs variantes. Ce qui est commun est l'expression *ăn bánh trả tiền* précédée de « service » et du verbe « éviter, se tenir éloigné ». L'expression *ăn bánh trả tiền* signifie littéralement « manger un gâteau et payer » et désigne les services sexuels tarifés.

11. Ce membre comptabilise plus de 3000 interventions dans le corpus, il n'y a que 3 intervenant·e·s qui dépassent le millier d'interventions dans le corpus.

3. « *Người hành_nghề mại_dâm do quan_hệ tình_dục với nhiều dạng đối_tượng phức_tạp khác_nhau, có lịch_sử không rõ_ràng về các bệnh truyền_nhiễm lây qua đường tình_dục, trong đó có HIV / AIDS. Họ rất dễ mắc các bệnh nguy_hiểm này và truyền lại cho những người lành bệnh khác có quan_hệ tình_dục với họ nhưng không đảm_bảo các điều_kiện an_toàn trong quan_hệ tình_dục.* »

   « Les travailleu.r/se.s du sexe, du fait qu'ils/elles ont des rapports sexuels avec de nombreux individus différents, ont des parcours peu clairs à propos des MST dont le VIH/SIDA. Ils présentent de forts risques de contracter des maladies dangereuses et de les transmettre à des individus sains qui ont des rapports sexuels avec eux/elles sans oser parler des moyens de protection en matière de sexe. »

4. « *Sự lo_lắng của bạn xuất_phát từ việc bạn đã tiếp_xúc quan_hệ tình_dục với người hành_nghề mại_dâm.* »

   « Ton inquiétude vient du fait que tu as eu un rapport sexuel avec une prostituée. »

Le paragraphe qui précède la plupart du temps ce fragment est le fragment n°3. Ce texte indique à l'internaute que les travailleu.r/se.s du sexe présentent un risque élevé de transmette le VIH. C'est donc le discours en quelques sorte officiel tenu sur le forum par l'expert numéro un (et d'autres qui le citent), et ce jusqu'à mi 2010. Après cette date, ce paragraphe est remplacé par une phrase plus laconique (fragment n°4) : « Ton inquiétude vient du fait que tu as eu un rapport sexuel avec une prostituée. » Ces textes (le paragraphe ou la phrase qui le remplace à partir de mi 2010, suivis du fragment commun, *ăn bánh trả tiền*) n'apparaissent jamais seuls. Ils sont toujours accompagnés d'une réponse personnalisée, adaptée au cas particulier auquel ils répondent. (*Tu as*

pris un risque réel / *Si tu as utilisé le préservatif tel que tu le décris, ta situation n'est pas à risque* / *Un préservatif déchiré constitue un risque* / *Telle pratique présente peu de risques* / *Je ne peux pas te dire si tu es contaminé, mais tu devrais aller faire un dépistage* / etc.) Cependant le discours se maintient sur la même ligne, avec le conseil qui reste inchangé (*Pour ne plus avoir à t'inquiéter à propos de la contamination au VIH et aux MST, le mieux est de ne pas avoir recours aux services sexuels des travailleu.r/se.s du sexe.*). Pourtant les pratiques à risque ne semblent pas diminuer. De nombreuses variantes de ce conseil existent. Chaque expert·e a sa formule, ses particularités typographiques mais le message reste le même.

Expert·e	Fréq.	EF	Traduction
n° 2751	93	Tránh tìm GDM/gmd	*éviter de chercher une prostituée*
	44	Tránh tìm GDM/gmd (nữa) nha (bạn/e)	*ne vas plus chercher une prostituée, hein !*
	48	Tránh tìm GMD/gmd (nữa) để k (còn) [mang/có/phải] (thêm) (những) [lo_lắng/nỗi lo]	*évite désormais de chercher une prostituée pour ne plus avoir à t'inquiéter*
numéro un	32	tránh (xa) các dịch_vụ ĂN BÁNH TRẢ TIỀN	*se tenir à l'écart des services sexuels tarifés*
	31	(Bạnh/bạn) (từ nay) (nên) Tránh/tránh (xa) các dịch_vụ ĂN BÁNH TRẢ TIỀN	*Ne vas plus acheter des services sexuels*
	13	tránh (xa) các dịch_vụ ĂN BÁNH TRẢ TIỀN là chính bạn tư bảo_vệ cho mình.	*te tenir à l'écart đes services sexuels tarifés est la meilleure manière de te protéger*
	1	Bạn lấy đây làm bài_học xương_máu cho mình đừng tìm các dịch_vụ ĂN BÁNH TRẢ TIỀN.	*Prends ça comme une leçon pour ne plus aller chercher des services sexuels tarifés*
(plusieurs)	22	tránh (xa) các dịch_vụ ăn bánh trả tiền	*se tenir à l'écart des services sexuels tarifés*
63 expert·e·s	102	tránh xa GMD	*éviter les prostituées*
10 expert·e·s	14	tránh xa gmd	*éviter les prostituées*
5 expert·e·s	12	tránh xa gái mại_dâm	*éviter les prostituées*

5. « *để được hướng_dẫn dùng thuốc (chống) (phơi nhiễm (HIV))* »

« pour te faire prescrire un traitement (post-exposition) »

Le fragment n°5 n'apparaît qu'à partir de 2011. Le corpus compte 15 occurrences, dont 11 de l'expert numéro un. Les 4 autres sont chacune d'un·e expert·e différent. Ils interviennent toujours en réponse, jamais en ouverture de conversation. Le discours tenu à propos des traitements post-exposition appelle à la modération. Les effets secondaires lourds sont régulièrement rappelés, et le segment le plus répété à ce propos est :

« *không cần [dùng/uống] thuốc (chống) phơi nhiễm (HIV/chống hiv)* »

« pas besoin de prendre un traitement post-exposition »

Les segments répétés, voire des fragments de texte relativement longs, nous ont permis de repérer et d'analyser en détail le discours institutionnalisant d'une certaine catégorie d'intervenants que nous avons appelés expert·e·s en raison de la place de leur parole au sein du forum. Il ne s'est pas agi ici d'améliorer la connaissance sur les comportements à risque face au VIH, mais plutôt celle sur les comportements langagiers des intervenant·e·s expert·e·s : les processus de normalisation des discours, par la répétition de leurs propres paroles ou celles de leurs congénères : les intervenants partageant le même type de comportement sur le forum. Le repérage de ces fragments de texte grâce à leur répétition et leur longueur nous a permis de dessiner la manière dont le discours institutionnel est relayé sur le forum, à propos des sujets principaux autour desquels tournent les conversations du forum : la prise de risque, la prostitution, ou la prise de traitement post-exposition.

4 CONCLUSION

Le travail de description du forum permet d'approfondir de plus en plus finement sa connaissance : son fonctionnement, sa structuration, puis son contenu. Afin d'accéder à la connaissance de son contenu, il a été nécessaire en premier lieu de décrire précisément sa structure hiérarchisée en rubriques, puis en conversations et enfin en interventions. Cette tâche nous a amené à distinguer pour ces trois niveaux de structuration, des profils répartis sur un continuum allant du discours institutionnel au discours informel. Nous avons relevé qu'au sein d'une rubrique au profil général informel coexistaient des conversations au profil informel (les conversation foisonnante, échanges discursifs spontanés que nous retrouverons dans le chapitre suivant) et d'autres au profil institutionnel (les conversations à intervention unique par exemple). Puis dans la même ligne, nous avons relevé qu'au sein d'une conversation au profil informel coexistaient des interventions de type informel (DIS) et d'autres de type institutionnel (DIR). En conclusion, il est difficile de composer un corpus uniquement de discours informel à partir d'un forum. Au delà des critères formels que nous avons exploités, nous avons ensuite utilisé le calcul des segments répétés pour repérer et analyser les discours institutionnalisants au sein des conversations informelles. Nous allons dans le chapitre suivant nous concentrer sur l'étude du discours informel spontané des forums.

CHAPITRE V

Analyse sémantique des discours informels

1 Introduction

Après avoir, au chapitre précédent, dressé une typologie de la variété des discours des forums, et élaboré parmi ceux-ci des critères de qualification des discours les plus *institutionnels* (conversations à intervention unique, DIR, discours des intervenant·e·s expert·e·s), nous allons dans ce chapitre nous intéresser aux discours les plus *spontanés* (conversations foisonnantes, espaces de consultation personnalisée (ECP) et enfin interventions-témoignages). Une fois décrits ces genres discursifs, que nous regroupons sous la dénomination de DIS, nous en présenterons différentes observations lexicales. Puis, dans la seconde partie de ce chapitre, nous appliquerons une méthodologie contrastive (partitionnement des corpus, et études des spécificités notamment) en confrontant les corpus des forums à ceux des sites d'information institutionnels, également en comparant le corpus vietnamien au corpus français.

2 Étude des genres discursifs informels spontanés des forums

2.1 *Étude des conversations foisonnantes*

Pour l'étude des conversations foisonnantes, nous nous appuierons aussi bien sur le corpus français que sur le corpus vietnamien. En effet, la sélection de ce type de conversation se faisant sur un critère numéral (voir le paragraphe 2.1.4.2), le procédé est reproductible quelle que soit la langue.

2.1.1 *Considérations terminologiques*

Parler de *longues conversations* serait trompeur car un grand nombre d'interventions peut avoir été produit sur une période très courte, et à l'inverse, une conversation peut s'étendre dans le temps avec une grande latence entre deux interventions. Nous étudions ici les conversations qui comptent le plus d'interventions, il serait donc plus approprié de parler de *conversations volumineuses* que *longues*. Cependant ce volume est à considérer en fonction du nombre d'interventions et non en fonction du nombre de mots, car lesdites interventions peuvent également compter un nombre très réduit de mots. C'est pourquoi nous parlerons de conversations foisonnantes.

2.1.2 *Introduction à une étude sémantique : un foisonnement provoqué par l'angoisse*

L'étude des conversations foisonnantes nous renseigne sur les thèmes attractifs secondaires, en plus des thèmes attractifs présents dans l'ensemble du corpus. En effet, des thèmes comme la prise de risque ou le traitement post-exposition se trouvent également dans les conversations moins foisonnantes. En revanche, les conversations foisonnantes concentrent d'autres thèmes at-

tractifs comme le dépistage ou l'angoisse. L'angoisse est particulièrement présente dans les titres de ces conversations foisonnantes, et ce thème se poursuit tout au long de celles-ci. C'est même cette angoisse qui explique le nombre élevé d'interventions. Les réponses rassurantes ne suffisent pas à calmer les peurs, il faut de nombreuses interventions pour que celles-ci se voient apaisées, et souvent seulement momentanément, pour reprendre de plus belle au moindre élément déclencheur. Nous développerons l'étude sémantique des conversations foisonnantes parmi les discours informels spontanés dans la partie 3 de ce chapitre.

2.1.3 *Deux types de conversations foisonnantes : les agoras et espaces de consultation personnalisés*

Parmi les conversations foisonnantes, certaines voient intervenir un grand nombre d'internautes et l'espace discursif qu'elles dessinent figurent véritablement une agora, d'autres sont centrées sur un internaute en particulier, nous les appellerons espaces de consultation personnalisée.

2.1.3.1 *Exemple de conversation foisonnante de type agora*

Prenons la conversation foisonnante la plus ancienne de notre corpus vietnamien. *QHTD bằng miệng có bị lây nhiễm HIV không ?* ([VN « Attrape-t-on le VIH avec un rapport oral ? »]) Elle est ouverte le 21 janvier 2007 par un internaute que nous appellerons *inquiet2007*, tout juste inscrit sur le forum (5 jours auparavant), et qui n'interviendra plus dans les 60 autres interventions que compte la conversation, de janvier 2007 à septembre 2012. Il ne s'agit donc pas d'une conversation de type consultation à sens unique (ECP). Les 35 internautes qui y interviennent sont venus à cette conversation par le titre, parce que le sujet les touche également et qu'il s'interrogent à ce propos. Nous

voyons que le critère du nombre d'interventions, autrement dit son aspect foisonnant, ne suffit pas à catégoriser une conversation comme un espace de consultation personnalisée réservé à l'internaute qui l'a créée.

2.1.3.2 *Définition des espaces de consultation personnalisée (ECP)*

Les conversations que nous qualifions d'espace de consultation personnalisée (ECP), bien qu'ouvertes à l'ensemble des membres, ménagent un espace du forum où un membre en particulier intervient le plus souvent et concentre la majorité des adresses des autres intervenant·e·s. Cet esquisse d'espace personnalisé résulte en partie de la volonté et des instructions des modérateurs, dans un souci de maintien de la lisibilité au sein du forum : le mot d'ordre est pour chaque membre de poursuivre intervention par intervention l'écriture progressive de son cas particulier, dans l'espace cohérent que constitue la conversation qu'iel a initié. Dans ce type de conversation, que nous appelons donc espace de consultation personnalisée (ECP), c'est son initiat·eur/rice qui acquiert le focus.

La catégorisation d'une conversation en ECP ne l'empêche pas d'intéresser de nombreux internautes, par exemple la seconde conversation initiée par *inquiet2007*, de plus de 60 interventions et qui sera classée en ECP (cf le paragraphe 2.1.4.3), a été consultée presque 10 000 fois (plus de 20 000 fois pour la première, qui est de type agora).

2.1.4 *Indices de catégorisation des conversations foisonnantes*

Pour distinguer les deux catégories de conversations foisonnantes, nous disposons de deux paramètres liés au nombre d'intervenant·e·s : l'indice de focalisation d'une part et la proportion de l'initiat·eur/rice d'autre part.

2.1.4.1 *Définition d'un indice de focalisation*

Le paramètre le plus accessible est celui que nous appelons l'indice de focalisation d'une conversation. Il s'agit de la moyenne d'interventions par intervenant·e·s[1]. Plus cette moyenne est élevée, plus la conversation tend vers un profil d'ECP, plus elle est faible, plus la conversation tend vers un profil d'agora.

2.1.4.2 *Des seuils différents entre le corpus français et le corpus vietnamien*

Le corpus français comptabilise 80 conversations de plus de 100 interventions (Conv.100+) alors que le corpus vietnamien n'en compte que 2. Il compte 35 conversations de plus de 50 interventions (Conv.50+). Nous en concluons que le seuil est à déterminer en fonction du corpus, dans notre cas vietnamien, nous considèrerons les conversations foisonnantes dès le seuil de 50.

Comparons les indices de focalisation de conversations en vietnamien et en français.

Tab. v.1 : Indice de focalisation de conversations foisonnantes

	indice de focalisation	nombre d'interventions dans la conversation
fr	5,47	104
vn	5,40	54
fr	13,5	216
vn	13,67	123

Nous voyons dans les exemples présentés dans le tableau v.1 qu'à un indice de focalisation équivalent, une conversation française mesure (en termes

1. Ces données sont calculées selon la procédure présentée en annexe G.

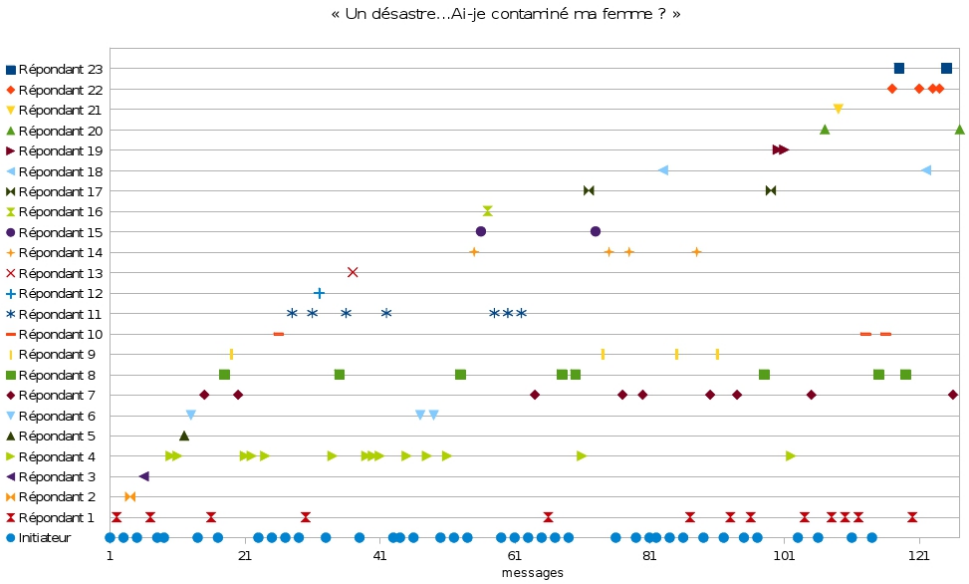
de nombre total d'interventions) le double d'une conversation vietnamienne. Comme pour les conversations les plus foisonnantes (Conv.100+ et Conv.50+), il semble que nous devions considérer sur un pied d'égalité les conversations françaises mesurant le double des conversations vietnamiennes.

Tab. v.2 : Indices de focalisation des conversations foisonnantes

	nombre de conversations foisonnantes	indice de focalisation
fr	80 Conv.100+	de 5,29 à 35,71
vn	35 Conv.50+	de 1,46 à 13,67

Le tableau v.2 illustre deux points : (i) les conversations sont plus foisonnantes en français qu'en vietnamien, et (ii) les conversations foisonnantes françaises tendent plus à devenir des ECP que les vietnamiennes : l'indice de focalisation le plus faible en français est 5,29. Il s'agit de la conversation [FR « Un désastre...Ai-je contaminé ma femme ? »], composée de 127 interventions, publiées par 24 intervenant·e·s (voir la figure v.1).

Fig. v.1 : Répartition des interventions dans la conversation [FR « Un désastre...Ai-je contaminé ma femme ? »]



Cette conversation est bel et bien un ECP et non pas une agora. Un indice de focalisation de 5,29 est donc assez élevé pour rester un indice d'ECP et trop élevé pour constituer un indice d'agora. Tout se passe comme si le genre de l'agora n'existait pas en français, puisque les indices de focalisation ne descendent pas en dessous de 5.

2.1.4.3 *Limites de l'indice de focalisation*

Un examen du côté du corpus vietnamien nous permettra d'étudier les faibles indices de focalisation à la recherche des agoras. Nous y relevons des conversations foisonnantes avec un faible indice de focalisation qui se révèlent être également des ECP. C'est le cas de la seconde conversation initiée par notre membre *inquiet2007*, 10 jours après l'ouverture de la première, pour une autre demande. *Mình stress nặng quá, giúp mình với .....* ([VN « Je stresse trop aidez-moi... »]) Cette conversation totalise plus de 60 interventions, dont 29 publiées par son initiateur, soit près de la moitié des interventions. La proportion de l'initiat·eur/rice est donc élevée et il s'agit bien d'un ECP, mais le nombre d'intervenant·e·s dans la conversation est de 21, ce qui lui donne un indice de focalisation faible. Nous ne pouvons donc pas utiliser le critère d'un indice de focalisation élevé pour distinguer de manière fiable les ECP. Le nombre d'intervenant·e·s répondant à l'initiat·eur/rice d'un ECP peut être élevé ou non, tant que l'initiat·eur/rice conserve la primauté des interventions, la conversation restera un ECP.

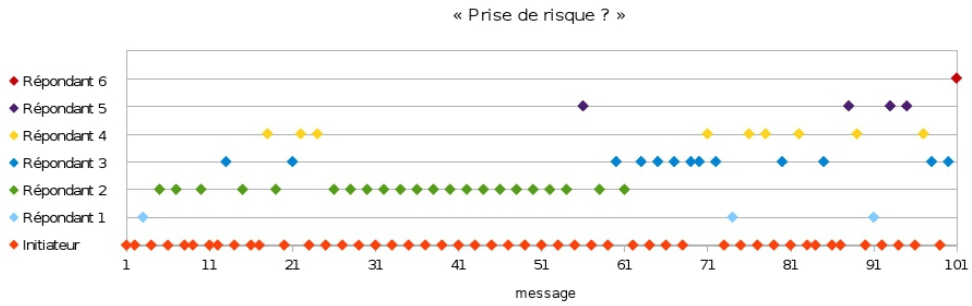
2.1.4.4 *Fiabilité de la proportion de l'initiat·eur/rice*

Le critère de l'indice de focalisation a montré ses limites pour *différentier* les agoras des ECP, alors que le critère de la proportion des interventions de

l'initiat·eur/rice semble s'avérer un bon indice de *détection* des ECP. Prenons quelques exemples.

La conversation intitulée [FR « Prise de risque ? »] (voir la figure v.2) compte 101 interventions, dont 49 par l'initiat·eur/rice de la conversation, soit presque la moitié, réparties sur l'ensemble de la chronologie.

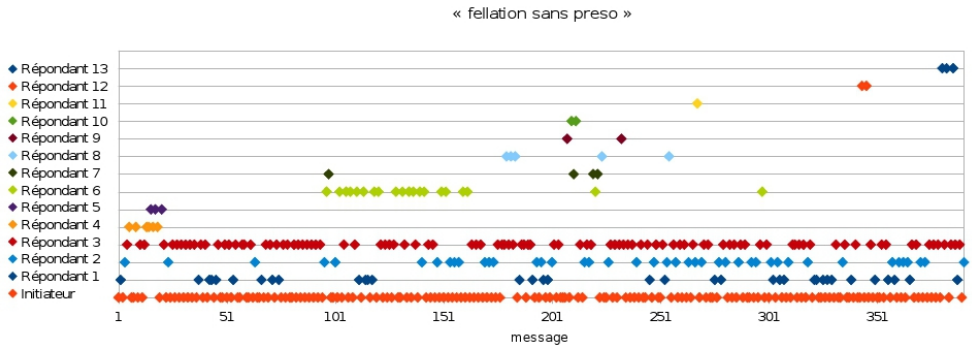
Fig. v.2 : Répartition des interventions dans la conversation [FR « Prise de risque ? »]



On considère donc qu'il s'agit d'un dialogue entre l'initiat·eur/rice d'une part, et l'ensemble des 6 autres membres qui sont intervenus dans cette conversation d'autre part. De plus, l'ensemble de la conversation s'est tenu sur 4 jours, ce qui est particulièrement court. La densité de fréquence des interventions est un indice fort du niveau d'angoisse exprimé par un membre.

Les mêmes observations peuvent être faite en sélectionnant des conversations plus conséquentes en termes de nombre d'interventions, par exemple [FR « fellation sans preso »] présente un indice de focalisation élevé (27,93). Nous retrouvons le principe du dialogue entre l'initiat·eur/rice d'une part et l'ensemble des autres intervenant·e·s d'autre part : 164 sur 391 interventions proviennent de l'initiat·eur/rice (soit 42%), et surtout : les interventions de l'initiat·eur/rice sont distribuées tout au long de la conversation.

FIG. V.3 : Répartition des interventions dans la conversation [FR « fellation sans preso »]



Autre cas d'intervenant·e prolixe et d'ECP : *garçontriste* a ouvert les deux conversations les plus foisonnantes du corpus vietnamien. De plus, cette fois, il s'agit bien de conversations avec un nombre très restreint d'intervenant·e·s (donc un fort indice de focalisation). Mais surtout, la proportion de l'initiat·eur/rice occupe entre 19,8 et 33,3% des interventions. Ces conversations sont donc restées principalement dévolues au cas de leur initiat·eur/rice, i.e. des ECP.

- *cho e hỏi gấp vấn đề này* ([VN « j'ai une question sur ce problème »]) : 96 interventions dont 32 de de l'initiat·eur/rice (*garçontriste*), seulement 11 intervenant·e·s ;

- *cho e hoi ? cai nay' cai* ([VN « j'ai une question sur ce truc »]) 106 interventions dont 21 de l'initiat·eur/rice (*garçontriste*), seulement 10 intervenant·e·s.

Ces deux conversations outre le fait qu'elles sont foisonnantes, s'étendent sur une période courte (une centaine d'interventions en un mois pour les conversations ouvertes par *garçontriste*. De plus, leur initiat·eur/rice n'est intervenu·e qu'au sein des conversations qu'iel a ouvertes, jamais ailleurs dans le forum.

Tab. v.3 : Indice de focalisation et proportion de l'initiat·eur/rice de conversations foisonnantes

titre de la conversation	nb d'intervenant·e·s	nb total d'interventions	interv. de l'initiat·eur/rice	indice de focalisation	proportion de l'initiat·eur/rice
[VN *Attrape-t-on le VIH avec un rapport oral ?*]	35	61	1	1,7	1,6%
[VN *j'ai une question sur ce truc*]	10	106	21	10,6	19,8%
[FR « Un désastre...Ai-je contaminé ma femme ? »]	24	127	39	5,3	30,7%
[VN *j'ai une question sur ce problème*]	11	96	32	8,7	33,3%
[FR « fellation sans preso »]	14	391	164	27,9	41,9%
[FR « Test VIH semble positif : panique. »]	33	550	243	17	44,2%
[VN *Je stresse trop aidez-moi...*]	21	60	29	3	48,3%
[FR « Prise de risque ? »]	7	101	49	14	48,5%

Pour résumer :

- Toutes les conversations foisonnantes ne sont pas des espaces réservés en priorité à leur initiat·eur/rice, certaines sont de véritables agoras.
- L'indice de focalisation ne suffit pas à repérer les ECP.

En revanche :

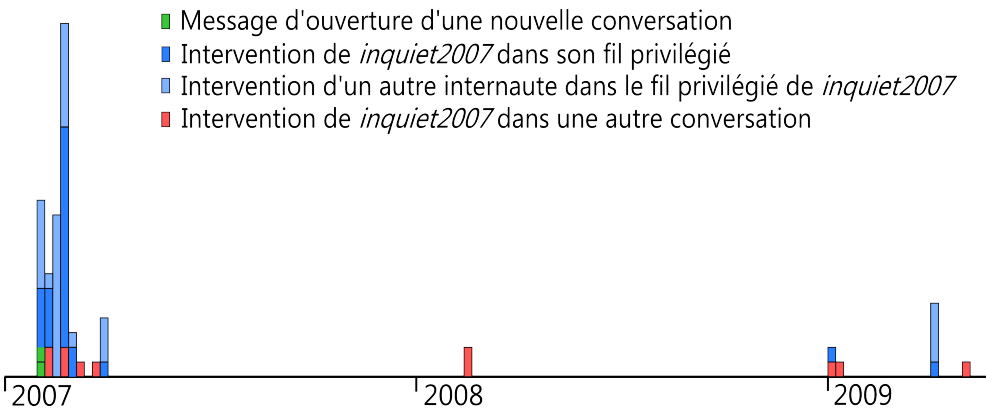
- La proportion d'interventions de l'initiat·eur/rice dans l'ensemble de la conversation est un critère fort pour ce repérage.

2.1.5 *Etude des conversations foisonnantes de type espace de consultation personnalisée (ECP)*

L'intérêt d'étudier une conversation de type ECP telle que *Mình stress nặng quá, giúp mình với .....* ([VN « Je stresse trop aidez-moi... »]) est d'appréhender un internaute sur une durée de deux ans. En effet, les interventions d'*inquiet2007* dans cette conversation s'étendent depuis son ouverture (en janvier 2007) à avril 2009. La conversation n'est plus alimentée depuis. C'est cette conversation en forme d'espace consacré à son initiat·eur/rice qui permet de porter notre attention sur cet internaute. Mais nous devons élargir le point de vue à l'ensemble du corpus, ce qui nous apprend qu'*inquiet2007* y intervient à 42 reprises, dont 1 pour lancer la première conversation évoquée ci-dessus, 29 dans la seconde. Restent 12 interventions, qui apparaissent dans 10 autres conversations. Ces 12 interventions sont très courtes, il s'agit de conseils ou d'avis émis par cet internaute en quelques mots, intervenant dans d'autres conversations en cours. *inquiet2007* a donc ouvert deux conversations, dont l'une est devenue son ECP, celui qui permet le mieux de suivre son évolution au cours de sa présence sur le forum, période qui a duré environ deux ans. L'autre conversation qu'il a ouverte s'est vue réapproprier par d'autres internautes, *inquiet2007* l'ayant totalement abandonnée. D'autre part sa présence sur le forum s'est manifestée par quelques interventions sporadiques dans des conversations menées par d'autres internautes, sous la forme de conseils expéditifs ou d'avis non argumentés. En 2009 *inquiet2007* a reçu un message des administrateurs lui indiquant qu'à la vue des avis qu'il émettait auprès d'autres membres de ce forum, il semblait ne pas avoir assez d'expérience pour donner des conseils aux autres, d'où un avertissement indiquant que ses prochains conseils seraient effacés. Il persistera à deux reprises à émettre son avis, mais il interviendra surtout, en forme de clôture de son fil privilégié,

pour indiquer que son dépistage définitif s'est avéré négatif, il n'a donc plus de souci à se faire à propos des événements qui l'avaient poussé à ouvrir cet ECP.

Fig. v.4 : Interventions de *inquiet2007*



Grâce à la représentation graphique de sa présence sur le forum (cf. la figure v.4), nous sommes en mesure de visualiser les différentes étapes de son parcours sur le forum : un foisonnement d'activité en début de période, peu de temps après la prise de risque, accompagné de nombreuses réponses d'autres internautes, puis la longue attente imposée par la période d'incubation, avec la vie réelle qui reprend son cours hors du forum, jusqu'aux résultats, qui occasionnent un retour sur le forum, pour informer les autres internautes et clôturer l'épisode.

Ce profil a également été observé dans le corpus français, notamment l'initiat·eur/rice de la conversation [FR « fellation sans preso »] (figure v.3).

Pour étudier les espaces où se déploient les discours informels spontanés, nous avons mis en évidence des récurrences dans les pratiques conversationnelles en nous penchant sur ce que nous avons appelé les ECP. Ce faisant,

nous considérions l'entité structurelle que constitue le niveau de la conversation. Nous allons dans la sous-partie suivante déplacer notre attention vers le niveau structurel inférieur et minimal, celui de l'intervention.

2.2 *Etude du genre textuel du témoignage*

Un type particulier d'intervention représente une opportunité de nous pencher sur le discours spontané : nous les qualifions d'interventions-témoignages. Il s'agit de récits personnels relatant un épisode particulier[2]. Ce sont des interventions qui ouvrent une nouvelle conversation, mais qui ne sont pas uniques (à l'inverse des Conv.IU) : elles ont entraîné interventions réactives. Mais surtout, elles suivent un scénario commun et partagent entre elles un certain nombre de thématiques. Voici les éléments partagés par ces interventions :

- Présence de marques temporelles
- Évocation de causes (alcool, prostitution, pratique sexuelle, préservatif)
- Évocation de conséquences (risque, contamination, symptômes, maladie)
- Champ lexical du regret, de la peur, de l'angoisse
- Adresse respectueuse aux autres internautes (appel à l'aide, remerciements)

Ces interventions, sont très normées car elles suivent toutes le même schéma. Les internautes se conforment aux rituels du forum en témoignant d'un mimétisme social en vue de s'intégrer à la communauté. Pour qu'une telle homogénéité se fasse jour, cela signifie que les internautes, avant de livrer leur propre témoignage, ont en premier lieu lu les témoignages d'autres internautes et adoptent les mêmes codes discursifs. Paradoxalement, c'est cette

2. En cela, les interventions-témoignages sont à rapprocher des *ego-documents* de Eensoo et Valette (2012, 2014). Cependant ceux-ci sont considérés comme des unités indépendantes et ne possèdent donc pas toutes les propriétés des interventions-témoignages, notamment leur statut dans la conversation.

même norme discursive qui invite également à la spontanéité, que nous recherchons. Nous pourrions considérer qu'intervenir sur un forum s'apparente à de la prise de parole en public, ce qui nécessite une certaine prise de distance avec l'objet ou la situation décrite. Pourtant, dans le cas de ces témoignages, les internautes décrivent des situations particulièrement intimes, ce qui place automatiquement les autres participants en position d'interlocuteurs proches, les barrières habituelles sont évitées, l'intervenant·e s'adresse aux autres participants comme à des amis intimes alors qu'iel ne sait pas encore qui va prendre part à la conversation qu'iel ouvre par son récit personnel. Les figures v.5 et v.6, présentent des exemples de témoignages personnels, où les segments de texte sont colorisés en fonction des différents champs lexicaux communs au genre textuel du témoignage qui ont été définis.

FIG. V.5 : Différentes étapes d'un témoignage personnel

Temporalité
Alcool
Pratique sexuelle
Préservatif
Prostitution
Risque, contamination, symptômes, maladie
Regret, peur, angoisse
Adresse respectueuse aux autres internautes

Cách đây 2 hôm em có quan_hệ với 1 GMD, đây là lần đầu_tiên QHTD của em. Tụi em có hôn môi nhưng ko chạm lưỡi, GMD có BJ cho em (ko có BCS) sau đó QHTD (có BCS). em thấy GMD rất cẩn_thận khi mang BCS cho e và khi xong cô ấy cũng rất cẩn_thận ngồi hẳn dậy rồi mới tháo bao cho em và em ko thấy tinh_dịch chảy ra từ BCS ạ (em cũng ko Oral sex cho cô ấy ạ). Em có đọc những bài trên diễn_đàn này và thấy trường_hợp của em là no risk ạ nhưng em vẫn cứ thấy canh_cánh trong lòng vì cô GMD đó rất ốm ko biết có bị HIV ko. 1 tháng nữa em sẽ về lại bên kia để tiếp_tục học_hành nhưng em lo quá hiz!! Các anh_chị_em ơi em có nguy_cơ nào bị nhiễm ko ạ? Em rất lo_lắng và cực_kỳ ân_hận, em thề từ nay tới già sẽ ko đụng vào GMD nữa! Xin các anh_chị coi giùm trường hợp của em ạ! Cám_ơn anh_chị_em Thân

Il y a 2 jours j'ai eu un rapport avec 1 prostituée, c'était mon premier rapport. Nous nous sommes embrassés mais sans la langue, la prostituée m'a sucé (sans préservatif) puis rapport (avec préservatif). j'ai trouvé la prostituée très précautionneuse quand elle m'a mis le préservatif et quand on a terminé elle l'a aussi été en se relevant puis en me retirant le préservatif et je n'ai pas vu de sperme couler du préservatif (je ne l'ai pas sucée). J'ai lu de nombreux posts sur ce forum et ai compris que ma situation n'est pas à risque mais je me sens quand-même hanté car la prostituée était très maigre peut-être avait-elle le VIH ? Dans 1 mois je retournerai poursuivre mes études mais je m'inquiète trop argh !! Les amis s'il vous plaît ai-je un risque d'être contaminé ? Je suis très inquiet et regrette énormément, plus jamais de ma vie je ne cèderai à la prostitution ! S'il vous plaît aidez-moi ! Merci à tous Amicalement

Fig. v.6 : Différentes étapes d'un témoignage personnel

Tình_hình là 12h đêm_hôm 28 / 12 / 2010 sau vài trận nhậu tơi_bời. mấy ông làm ở Cty rủ em đi thoải_mái thế_là mấy ỏng dẫn em tới đường nguyễn chí thanh kiếm GMD. Lúc quan_hệ thì em có dùng bao ở khách_sạn và GMD đeo cho em. Em chỉ cho vo trong AD khỏang chừng 1 đến 2 phút là rút ra vì thấy ớn_ớn. Em không biết bao có bi rách hay không. nhưng em biết là bao không bị tuột, được 4 ngày thì người em có cảm_giác ngứa khắp người, còn mấy triệu_chứng kia thì không có. Không biết em co khả_năng nhiễm HIV không vậy mấy bác. Em lo_lắng quá. Bị stress nặng e mới có 21 tuổi thôi hix hix đời_sống còn dài, em thề không_bao_giờ quan_hệ với GMD nữa, Tởn đến già luôn.

Voici ce qu'il s'est passé le soir du 28 décembre 2010 à minuit, après avoir bu jusqu'à être complètement bourré. Des collègues m'ont incité à me mettre à l'aise, c'est-à-dire qu'ils m'ont emmenés rue nguyễn chí thanh chercher une prostituée. Pendant le rapport j'ai utilisé un préservatif de l'hôtel et la prostituée me l'a mis. J'ai juste pénétré son sexe pendant environ 1 à 2 minutes avant de me retirer car je ne me sentais pas bien. Je ne sais pas si le préservatif s'est déchiré ou pas. mais je sais qu'il n'avait pas glissé, après 4 jours mon corps m'a démangé de partout, les autres symptômes je ne les ai pas eus. Pensez-vous que j'aie une probabilité d'être contaminé au VIH ?. Je suis très inquiet. Je stresse énormément je n'ai que 21 ans argh la vie est encore longue, je jure que plus jamais je n'aurai de rapport avec une prostituée, Je vais le regretter toute ma vie.

A partir de ces premières interventions, nous pouvons établir un scénario commun, composé d'une série de champs lexicaux.

0a) Phrase d'introduction, formule d'adresse aux internautes (j'ai pris un risque, j'en appelle à vos conseils)

1) Situation temporelle

2) Alcool + amis, collègues, situation de perte de contrôle

3) Lieu de prostitution

4) Pratique sexuelle

5) Risques de contamination

6) Symptômes

7) Angoisse/regrets.

0b) Phrase de conclusion, adresse aux internautes (aidez-moi s'il vous plaît)

Tous les témoignages ne contiennent pas l'intégralité des étapes du scénario, mais la majorité d'entre elles.

A partir de cette étude préliminaire, nous pouvons constituer une liste de formes lexicales pour chaque thématique. Les listes non exhaustives qui ont été constituées figurent en annexe K.

Ces listes lexicales permettent de repérer d'autres interventions similaires. Le premier critère de repérage des interventions-témoignages est leur statut d'intervention initiative. Des sous-corpus sont donc constitués avec uniquement des interventions initiatives. D'autre part, nous nous concentrons sur les rubriques auxquelles nous avons attribué un profil informel (voir la sous-partie 3.1 du chapitre IV, page 125). Plus particulièrement, les résultats que nous présentons dans cette sous-partie ont été obtenus pour la rubrique [(62) Relations sexuelles, Préservatif, Lubrifiant], celle qui attire la majorité des échanges sur le forum, mais des similarités ont été observées dans les autres

rubriques auxquelles nous avons attribué le profil informel, à l'inverse des observations sur les rubriques au profil institutionnel.

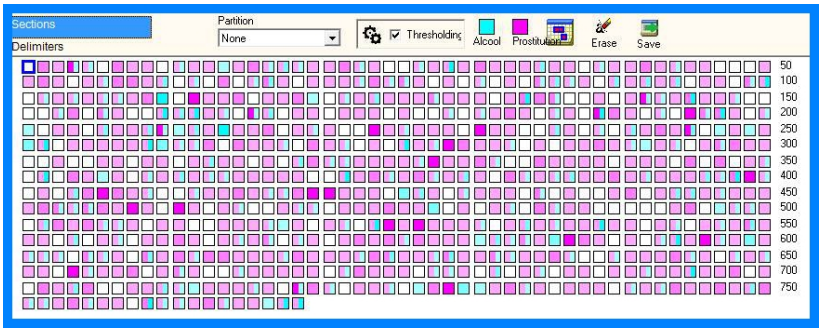
Nous utilisons la carte des sections de Lexico[3] pour explorer ces sous-corpus. En optant pour un sectionnement en phrases, nous pouvons également entrer dans le détail et la succession des étapes évoquées. Les pratiques sexuelles et le préservatif sont cités en début d'intervention, dans la phase de description de la situation, à propos de laquelle l'intervenant·e recherche des réponses. Les risques sont évoqués en deuxième partie d'intervention, les regrets et la peur sont exprimés en fin d'intervention, lors de l'appel à d'autres internautes, comme pour demander pardon pour une faute commise, avec l'espoir irrationnel que le regret exprimé puisse réduire les risques réels.

Le personnage du/de la TS est présent de manière régulièrement répartie dans les interventions-témoignages : que ce soit en cooccurrence avec la description des pratiques ou dans l'évocation de la peur et des regrets. La liste de formes relative au champ lexical de la prostitution inclut des expressions a priori peu évidentes, telles que *quán cà fe gọi đâu* (« café shampoing »). Pourtant, cette observation corrobore bien celles de Nguyễn (1997) recensant les lieux de prostitution, et que nous avons exposées au chapitre 1 (p. 16). Nous vérifions dans le corpus, et donc dans les comportements décrits par les internautes, que ces lieux de services proposent également des services sexuels.

Nous constatons, dans les interventions-témoignages du corpus vietnamien, une forte corrélation de la thématique de l'alcool avec celle de la prostitution. L'alcool est cité en élément déclencheur, dé-responsabilisant. La carte des sections par conversations de Lexico (v.7) indique que la majorité des conversations contiennent ces deux thématiques.

3. Pour une brève description, voir page 58.

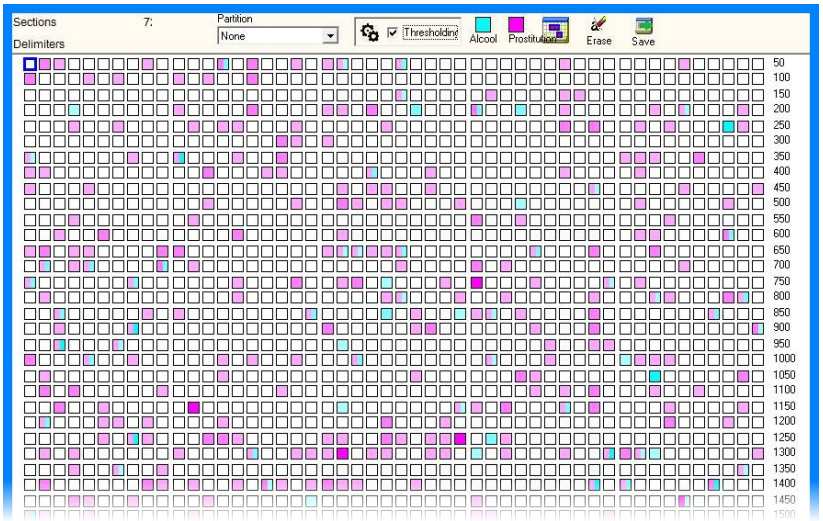
Fig. v.7 : Alcool et prostitution dans les conversations de la rubrique 62. Chaque conversation est représentée par un carré, celles qui contiennent des formes ou expressions que nous avons listées comme représentatives de la thématique de l'alcool sont colorées en bleu et de manière similaire, celles qui abordent la thématique de la prostitution sont colorée en rose.



Il est ainsi aisé de visualiser à quel point les deux thématiques parcourent l'ensemble du corpus (ici, la rubrique 62). Peu de conversations y échappent.

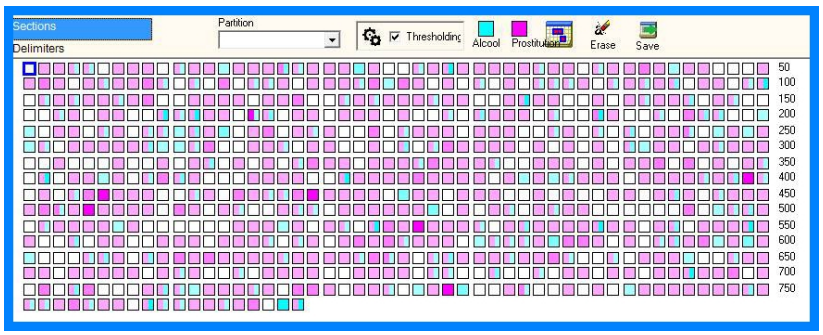
Si nous entrons dans le détail d'une segmentation du corpus en intervention, nous perdons l'impression de répartition uniforme. La carte des sections par intervention (v.8) donne un rendu plus dilué de la présence de ces thématiques.

Fig. v.8 : Alcool et prostitution dans les interventions de la rubrique 62. Chaque intervention est représentée par un carré



En revanche nous retrouvons la présence massive des deux thématiques dans le sous-corpus des interventions initiatives (v.9). Cette observation confirme que ces thématiques sont principalement abordées en ouverture de conversation, au sein des interventions initiatives, celles-ci présentant de plus une forte probabilité d'être des interventions-témoignages (champ lexicaux du témoignage, position en ouverture de conversation).

Fig. v.9 : Alcool et prostitution dans les interventions initiales de la rubr. 62. Chaque carré représente l'intervention initiative d'une nouvelle conversation.



L'examen des interventions-témoignages nous amène ainsi à constater que le VIH est fortement associé à la prostitution dans les discours informels spontanés en vietnamien. Ce n'est pas le cas dans le corpus informel français. En vietnamien, le VIH semble majoritairement perçu comme une conséquence néfaste d'un acte sexuel répréhensible. En effet, si le VIH n'est plus officiellement considéré comme un fléau social, la prostitution reste illégale. Les institutions s'efforcent de lutter contre la stigmatisation - encore forte - des personnes séropositives mais continuent à condamner les services sexuels tarifés.

Nous constatons, dans un corpus vietnamien des années 2007-2012, une perception du VIH comme une sorte de châtiment suite à des actes sexuels répréhensibles. Nous sommes tentée de rapprocher cette perception de celle des années 1990 en France, où le VIH restait fortement associé à l'homosexua-

lité, encore largement condamnée socialement.

La prostitution étant prohibée, les internautes vietnamiens semblent devoir justifier le fait d'y avoir recours. C'est dans ce but qu'intervient selon nous la thématique de la perte de contrôle, dont l'alcool est la cause majoritairement représentée dans notre corpus.

2.3 *Observations lexicales sur le discours informel spontané*

2.3.1 *Transcription de l'oralité*

Les discours des forums sont produits directement au format numérique, pour la quasi totalité par le bais d'une interface de saisie de texte au clavier, donc sous forme écrite et sans présence corporelle donc sans possibilité d'utilisation des autres moyens d'expression que le texte. Pourtant, la possibilité pour les autres internautes d'accéder de manière quasi instantanée à cette expression, et d'y répondre avec la même instantanéité, donne un aspect polylogique direct à ces échanges, et leur confère des traits associés à l'oralité. D'où l'emploi de l'expression "forum de *discussion*".

L'absence du corps et notamment du visage dans ces discussions est palliée pour une part par l'utilisation d'émoticônes.

Par ailleurs, la situation s'apparentant à l'oralité, les internautes ont souvent recours au registre du langage parlé, qu'ils transcrivent dans l'interface du forum. Il en résulte un grand nombre d'interjections, de longues séries de ponctuations fortes, d'adresses aux autres, et dans le cas du vietnamien de lexique réservé à l'oral, des variantes phonétiques, et une très forte proportion d'abréviations. A titre d'illustration, quelques exemples sont présentés ci-dessous.

- interjections :

  - *huhu* (« bouhouh »),
  - *hix / hic* (« argh »),
  - *hihi* (« héhé »),
  - *phùùù* (« pffiou »)
- séries de ponctuations fortes : !!!!!!!, ????
- adresses aux autres :
  - *tất cả các bạn* (« vous tous »),
  - *Thân !* (« Amicalement ! »)
- lexique oral : *mình, tớ* au lieu de *em, tôi*, etc. pour se désigner soi
- variantes phonétiques :
  - *wa* pour *quá / qua*,
  - *wen* pour *quên* (« oublier »),
  - *hok* pour *không* (négation),
  - *phải hông ?* pour *phải không ?* (« n'est-ce pas ? »),
  - *bít* pour *biết* (« savoir »),
  - *pác* pour *bác* (« vous »),
  - *j* pour *gì* (« quoi »)
- abréviations :
  - *e* pour *em* (« je »),
  - *đc* pour *được* (marq. positif),
  - *ko, k, kô* et *kg* pour *không* (négation),
  - *les* pour « lesbienne »
- dont acronymies :
  - *H* pour *HIV* (« VIH »),
  - *XN* et *xn* pour *xét nghiệm* (« examen », « dépister »)
  - *VGB* pour *Viêm gan B* (« hépatite B », abrégé en français par VHB pour Virus de l'Hépatite B)

Exemple d'utilisation des abréviations :

« *lúc đó*	*tôi*	*ko*	*biết*	*tc*	*đó*	*là ty* »
		(*ko* pour *không*)		(*tc* pour *tình cảm*)		(*ty* pour *tình yêu*)
à ce moment-là	je	ne savais pas que		cette affection		était de l'amour

Ce lexique, ainsi que les émoticônes sont absolument spécifiques au discours informel spontané (DIS) et représentent de bons critères pour le distinguer du discours institutionnel rapporté (DIR). A l'inverse, d'autres éléments sont caractéristiques du DIR, comme par exemple *GMT+* du fait que le DIR est majoritairement constitué d'articles copié-collés contenant une date

et une heure exprimée en temps universel. Ce critère est typiquement hérité du contexte technique dans lequel le discours est produit.

Nous pouvons également observer dans le corpus des variantes typographiques dûes à l'influence de la norme de saisie du vietnamien qui est utilisée[4].

- Si la norme utilisée est VNI, par exemple la touche 6 est utilisée pour l'accent circonflexe et la touche 9 pour la barre du đ :
  - *d9c* pour *đc* (*được*, marq. positif),
  - *kh6ng* pour *không* (négation),
  - *y6en tâm* pour *yên tâm* (« se tranquilliser »),
  - *L6au* pour *Lâu* (« longtemps »)
- Si la norme utilisée est Telex, le double appui sur une touche est utilisé pour ajouter un accent circonflexe à une voyelle, et les lettres absentes de l'alphabet vietnamien mais présentes sur les claviers standards querty ou azerty (f, j, w, également r, s, x, présents dans l'alphabet mais jamais en finale) servent à ajouter les diacritiques :
  - *maf* pour *mà* (« mais »),
  - *banj* pour *bạn* (« ami »),
  - *nguwoif* pour *người* (« personne »),
  - *năgj* pour *nặng* (« fort »)
  - *xét nghiemj* pour *xét nghiệm* (« dépistage »),
  - *trieejuj chứng* pour *triệu chứng* (« symptôme »),
  - *cuj thể* pour *cụ thể* (« réalité »),
  - *hành vj* pour *hành vị* (« acte »),
  - *dcj* pour *được* (marq. positif),

Ainsi, une phrase telle que *sao anh laij mois vaayj bay gio em lo lam* indique qu'aucun outil de saisie du vietnamien n'a été utilisé : les 2 premiers mots ne nécessitaient pas de diacritique, lors de la saisie des 3 mots suivants l'intervenant·e ne s'est pas rendu compte que l'option saisie du vietnamien

4. A ce sujet, voir la section 5.2.3 dans l'état de l'art, page 64

n'était pas activée, et pour les 5 derniers mots, l'intervenant·e s'en est rendu compte et a écrit sans utiliser aucun diacritique. Il en est de même avec la phrase *thif hoj nois ngi ngof con gi nua ma anh* : les 5 premiers mots sont écrits en s'attendant qu'un outil de saisie convertisse les lettres en diacritiques et les 5 derniers sans attendre de conversion et donc en se passant de diacritiques. Ces variations sont difficilement saisissables par les outils de textométrie et ces occurrences échappent aux calculs. Heureusement, il s'agit d'un phénomène marginal : la lecture du vietnamien étant difficile en l'absence des diacritiques, les internautes évitent au maximum de les omettre dans leurs publications, et s'ils y sont contraints, ils s'excusent auprès de leurs lecteurs :

#member512

Sorry may laptop cua minh mua o nuoc ngoai , nen ko co su dung dau duoc , mong cac bac thong cam
« Désolé.e mon ordinateur portable a été acheté à l'étranger, donc je ne peux utiliser les accents, j'espère que vous me pardonnerez »

mong cac Bac thong cam , may laptop cua em ko danh duoc dau tieng Viet
« j'espère que vous me pardonnerez, mon ordinateur portable ne peut pas écrire les accents »

#member836

em ko viet ' ko dc cac ' anh tho ˆ ng cam ?
« je peux pas écrire pardonnez-moi »

#member943

ko co ' vietkey sorry cac ' pac nha
« j'ai pas vietkeys désolé.e ! »

#member915

E xl tại k viêt Vietkey k pải vì e k biêt viêt mà máy nhà e hư vietkey mât e xl
« Pardon de pas écrire en Vietkeys, c'est pas parce que je sais pas mais ça marche plus sur mon ordi dsl »

#member1705

em khong biet sao , em khong viet tieng viet duoc , mong cac bac khong cam
« je sais pas pourquoi, je peux pas écrire le vietnamien, j'espère que vous me pardonnerez »

#member2078

Em dang onl bang dt nen ko viet tieng viet duoc . moi nguoi thong cam
« Je suis sur mon téléphone donc je peux pas écrire le vietnamien. j'espère que vous me pardonnerez »

2.3.2 *Euphémisations*

Dans la langue vietnamienne en général, les expressions fréquemment utilisées dans un domaine sont l'objet d'un processus d'abréviation systématiquement adopté. On aura notamment recours aux sigles pour remplacer des expressions figées, comme par exemple :

- *BCS* pour *bao cao su* (« préservatif »)
- *NCH* pour *người có HIV* (« personne atteinte du VIH », « séropositif »)

Et même :

- *H* pour *HIV* (« VIH »).

Mais le recours plus singulier à des procédés langagiers imagés et détournés, très caractéristique du corpus informel vietnamien, vise, selon nous, à effectuer une mise à distance purement lexicale (métaphorisation, euphémisation), au sein même de la zone anthropique proximale[5] – distance qui permet de contourner le caractère tabou d'un sujet de discussion, alors que ce caractère relève précisément de sa dimension identitaire ou proximale. En d'autres termes, la distance produite par ces procédés permet de s'affranchir des barrières séparant les sujets tabous de la verbalisation.

5. Voir le paragraphe 3.1.1.

2.3.2.1 *Techniques de distanciation*

Parmi les techniques de mise à distance, nous recensons les suivantes :

- les abréviations permettent de ne pas écrire en toutes lettres des mots ou expressions présentant un caractère tabou :
  - *NCH* pour *người có HIV* (« personne atteinte du VIH »)
  - *BCS* pour *bao cao su* (« préservatif »)
  - *QHTD* pour *quan hệ tình dục* (« rapport sexuel »)
  - *XT* pour *xuất tinh* (« éjaculation »)
  - *TD* pour *thủ dâm* (« masturbation »)
  - *HJ* pour *hand job* (« masturbation »)
  - *BJ* pour *blow job* (« fellation »)
  - *OS* pour *oral sex* (« fellation », « sexe oral »)

  A force d'usage, ces sigles présentent tous une forte tendance à l'acronymie et perdent ainsi leur capitalisation. Ainsi, les occurrences de xt, qhtd, bj, os, etc. sont nombreuses et fonctionnent comme des unités lexicales.
- les emprunts à l'anglais permettent d'établir une distance avec l'objet du discours et d'éviter d'employer l'équivalent vietnamien, trop direct :
  - *sex* pour *tình dục* (« sexe »)
  - *boy bi* pour *lưỡng tính* (« bisexuel »)
  - *blow job* et *bj* pour « fellation »
  - *oral sex* et *os* pour « fellation, sexe oral »
- d'autres procédés linguistiques permettant le contournement, l'allusion :
  - *tình *ục* pour *tình dục* (« sexe »)
  - *ngủ* (« dormir ») pour « coucher »
  - *chuyện ấy* (« cette histoire ») pour « rapport sexuel »
  - *cái đó* (« ce truc ») pour « pénis »
  - *chỗ ấy* (« cet endroit ») pour « entrejambe », « sexe »
- enfin, les métaphores sont très nombreuses :

expression	**traduction littérale**	**signification**
lên tới đỉnh	« monter au sommet »	« avoir un orgasme », « jouir »
lên mây	« monter aux nuages »	« atteindre le septième ciel », « jouir »
nếm trái cấm	« goûter le fruit défendu »	« avoir son premier rapport sexuel »
thổi ken	« souffler dans la trompette »	« faire une fellation »
sung	« arme à feu »	« pénis, sexe masculin »
nổ sung	« tirer un coup de feu »	« éjaculer »
áo mưa	« cape de pluie »	« préservatif »
chân trần	« pieds nus »	« sans préservatif »
cậu nhỏ	« petit gars »	« sexe masculin »
cô bé	« petite fille »	« sexe féminin »
chim	« oiseau »	« sexe masculin »

2.3.2.2 *Exemples de thèmes euphémisés*

Parmi les thématiques les plus sujettes aux procédés linguistiques de distanciation, nous retenons les suivantes.

- **l'anatomie**
  - *DV* pour *dương_vật* (« pénis »)
  - *AD* pour *âm_đạo* (« vulve » / « vagin »)
  - *chỗ ấy* (« cet endroit ») pour « entrejambe »
  - *chỗ quan trọng nhất ấy* (« cet endroit le plus important ») pour « entrejambe »
  - *chỗ / vùng nhạy_cảm* (« zone sensible ») pour « parties intimes »
  - *cái đó* (« cette chose-là ») pour « sexe masculin »
  - *cái của quý* (« la précieuse possession ») pour « pénis »
  - *súng* (« arme à feu ») pour « pénis »
  - *chim* (« oiseau ») pour « pénis »
  - *"cậu nhỏ"* (« petit gars ») pour « pénis »
  - *"cô bé"* (« petite fille ») pour « sexe féminin » (mais peut aussi désigner une personne, souvent partenaire sexuelle)
  - *hột* (« graine », « pépin ») pour « clitoris »
- **la sexualité**
  - *có lỡ "trót dại"* (« commettre la "faute" ») pour soit « perdre sa virginité », soit « tromper », soit « avoir un rapport non protégé », soit tout simplement « consommer », « coucher ».
  - *(có) chuyện ấy / đó* (« (avoir) cette histoire ») pour « coucher »
  - *trong chuyện đó* (« pendant l'histoire ») pour « pendant l'amour »
  - *sex* (« sexe »)
  - *qhtd* pour *quan_hệ tình_dục* (« rapport »)

- *làm tình* (« faire l'amour »)
- *ngủ* (« dormir ») pour « coucher »
- *ăn / nếm / xơi trái cấm* (« goûter le fruit défendu », « croquer la pomme »)
- *"cửa sau"* (« "porte arrière" ») sodomie
- *tinh binh* (« soldats d'élite ») pour « sperme »
- *xuất binh* (« envoyer les soldats ») pour « éjaculer »
- *bắn* (« tirer » (un coup de feu, une flèche)) pour « éjaculer »
- *"thăng hoa"* (« sublimation ») pour « orgasme »
- *"đạt tới đỉnh"* (« atteindre le sommet ») pour « jouir »
- *lên tới đỉnh* (« monter jusqu'au sommet ») pour « jouir »
- *lên mây* (« monter aux nuages ») pour « jouir »
- *boy bi* (« bisexuel »)
- *tgt3* (*thế giới thứ ba*) « troisième monde » pour « milieu homosexuel »

· **l'infidélité, la prostitution, la pornographie**

- *ăn bánh trả tiền* (« manger un gâteau et payer », « manger à emporter ») pour « aller voir une prostituée »
- *ăn phở* (« manger une soupe phở ») pour « avoir une relation extra-conjugale »
- *ăn cơm* (« manger du riz ») pour « être fidèle »
- *ăn chay* (« manger végétarien ») pour « n'avoir ni rapport sexuel, ni pratique masturbatoire », « être abstinent »
- *trang web đen* (« pages web noires ») pour « sites pornos »

· **la séropositivité au VIH**

- *NCH* (« séropo »)
- *người có H* (« séropo »)
- *người h* (« séropo »)

2.3.2.3 *Institutionnalisation des euphémismes*

L'euphémisation n'est pas réservée strictement au discours informel. Le discours institutionnel en a également un usage, mais moins systématique. Certaines formes ou expressions (par exemple *cậu nhỏ*, *xuất binh*, etc.) apparaissent à parts égales dans le corpus institutionnel et dans le corpus informel.

Une exploration diachronique du corpus visant à déterminer s'il existe un processus d'institutionnalisation de certaines variations lexicales, nées des

phénomènes de mise à distance, révèle que ce processus est difficilement observable dans le corpus. Par exemple, nous ne pouvons pas parler d'institutionnalisation de l'expression *cậu nhỏ* car la forme apparaît chronologiquement dans le corpus institutionnel avant celui du forum. A l'inverse, l'expression *chuyện ấy* apparaît d'abord dans le corpus informel, entre guillemets, puis est largement utilisée en institutionnel, tout en continuant à l'être en informel mais minoritairement. Pourtant cela ne suffit pas pour conclure à une institutionnalisation de la forme, du fait du trop petit nombre d'occurrences. L'expression *ăn bánh trả tiền* est utilisée une fois sur le forum en 2007 par un expert, puis 3 fois en institutionnel en 2008, enfin, elle est utilisée sur le forum 93 fois entre juillet 2009 et novembre 2012, par des experts. Nous avons montré[6] que le discours des experts relevait du discours institutionnel ou institutionnalisant, malgré le fait qu'il soit produit dans le forum. Cette expression n'est donc pas utilisée en discours informel spontané. *cái đó* pourrait constituer un exemple d'institutionnalisation : il apparaît 51 fois dans le corpus informel entre janvier 2007 et avril 2008, puis 2 fois entre guillemets en institutionnel en mai 2008, puis dans les 3 mois qui suivent 23 fois, sans guillemets en informel, avec guillemets en institutionnel. Enfin, dans la suite du corpus il est quasiment systématiquement utilisé sans guillemets, que ce soit en institutionnel ou en informel. *chỗ ấy* suit le même schéma dans le corpus.

L'exploration diachronique des euphémismes permet d'étudier l'évolution de leur emploi au sein des différents discours, mais n'est pas suffisante pour conclure à des phénomènes d'institutionnalisation.

6. Voir le paragraphe 3.4.2 du chapitre IV.

2.3.3 *Étude des titres*

Le titre de chaque conversation est donné par son initiat·eur/rice : l'internaute qui ouvre une nouvelle conversation rédige son titre, qui sera valable pour toute la durée de la conversation. Celui-ci nous donne donc des informations sur le ou les thème(s) de la conversation, mais également sur l'état d'esprit du membre du forum qui l'a ouverte. Le tableau v.4 présente les formes ou groupes de formes (rassemblées par proximité sémantique) les plus fréquent·e·s dans les titres des conversations du corpus français et celles du corpus vietnamien. Ces formes (et groupes de formes) sont triées par ordre décroissant de leur proportion par rapport au nombre total d'occurrences dans le corpus où elles apparaissent (celui des titres en français ou celui des titres en vietnamien).

L'étude comparée du lexique des titres des conversations permet de pointer les spécificités du français et du vietnamien. Un corpus pour chaque langue est étudié séparément, puis les pourcentages sont mis en regard. Cette étude est faite avec TXM, qui inclut la ponctuation dans la liste des formes. En premier lieu, nous notons qu'en français la ponctuation forte occupe les proportions les plus élevées alors qu'en vietnamien, des groupes de formes surpassent ces ponctuations. A propos des ponctuations fortes nous notons également que le corpus français présente le double de points d'interrogation comparé aux points d'exclamation, alors qu'en vietnamien, l'exclamation dans les titres prend le pas sur l'interrogation (du moins sa manifestation explicite par le caractère de ponctuation qui lui est dédié). Dans le forum français, les internautes présentent leur question dès le titre alors qu'en vietnamien, l'expression de l'angoisse prend le pas sur l'exposition des interrogations qui la provoquent.

Tab. v.4 : Le lexique des titres des conversations

Forum français	% des occ.		Forum vietnamien
		4,55	« rapport » (*QHTD / quan_hệ / Quan_hệ / sex / QH / qh / qhtd*)
?	3,96		
		3,83	« préservatif » (*BCS / bcs / bao_cao_su / bao*)
		2,65	!
		2,53	« prostituée » (*GMD / gmd / mại_dâm*)
		2,23	?
!	2,1		
...	1,89		
risque	1,22		
test	1,06		
		0,99	« fellation » (*miệng / Oral / OS / oral*)
« fellation » (+ *Fellation / FELLATION*)	0,84		
		0,56	...
« préservatif » (+ *preservatif / capote*)	0,53		
« rapport » (+ *Rapport / rapports / Relation / relations*)	0,34	0,34	« risque » (*nguy_cơ*)
TPE	0,2	0,2	« TPE » (*PEP*)
		0,12	« test » (*xét_nghiệm*)
prostituée	0,03		

La proportion des thèmes les plus représentés au sein des titres diffère également. Sur le modèle des titres des rubriques (donnés par les créateurs du forum), ceux des conversations (donnés par les internautes) sont plus génériques en français qu'en vietnamien. Le thème le plus fréquent des titres français est la prise de risque, qui nous place dès le titre à un niveau d'abstraction, dénotant une volonté de catégorisation, puis en deuxième lieu vient le thème

du dépistage, autrement dit l'acte médical qui permettra de résoudre l'interrogation à propos de la contamination. En vietnamien, les titres nous placent à un niveau concret par l'évocation des circonstances de la prise de risque : le préservatif, le rapport sexuel, le personnage de la prostituée. L'étude des titres semble donc confirmer les observations faites sur les interventions-témoignages (voir la sous-partie 2.2).Le thème de la prostitution est relégué loin des priorités dans le corpus français car ce n'est plus un thème fortement lié aux problématiques du VIH. Les rapports bucco-génitaux en revanche concentrent plus d'interrogations : ils sont plus évoqués dans les titres français que ne l'est le préservatif. Le traitement post-exposition (TPE) n'est pas encore un thème fort, en français comme en vietnamien, ce qui est peut-être dû au fait que notre corpus est borné à 2012.

3 ANALYSE DES DISCOURS INFORMELS PAR CONTRASTE

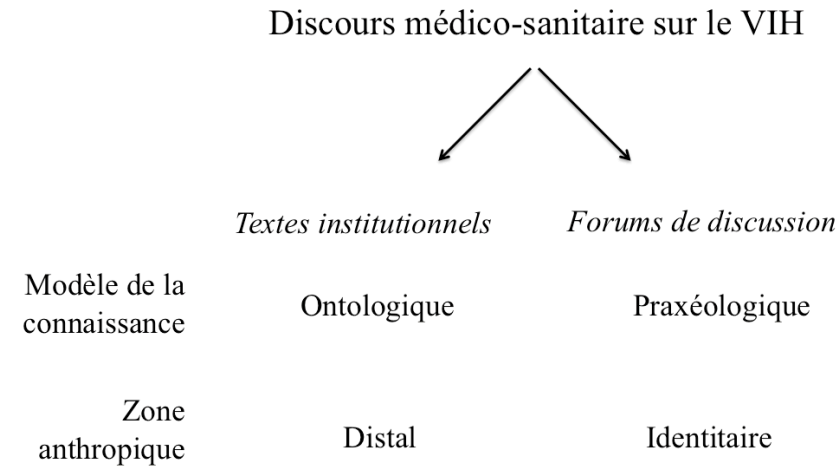
Nous posons comme postulat de départ l'aspect lacunaire des connaissances institutionnalisées, et sur cette base, travaillons dans le sens d'un principe de complétion. Nous estimons que les connaissances stabilisées - desquelles il résulte un discours institutionnalisé - doivent sans cesse être enrichies et actualisées grâce à l'apport de la praxis - de laquelle il résulte des discours en perpétuelle évolution. Il s'agit donc de confronter ces discours afin de mettre en évidence les lacunes du premier, grâce à la mise en contraste avec les seconds. Nous souhaitons analyser ce que le contraste permet de faire émerger pour utiliser la complétion comme méthodologie de construction des connaissances.

Dans la lignée de Slodzian et Valette (2009), nous faisons l'hypothèse que les discours institutionnels médicaux sont stables quant à leurs prescriptions car fondés sur les pratiques normalisées et consensuelles d'une médecine scientifique internationale. Les discours de prévention sont, de fait, homogènes et peu productifs car ils sont adossés à des connaissances médicales dont les progrès sont lents. Par ailleurs, la réappropriation des discours institutionnels sur les forums de discussion, implique hétérogénéité et productivité. Ainsi, les discours informels complètent les informations dispensées par les discours institutionnels en nous renseignant sur la variété des pratiques, peu couvertes par les discours institutionnels. Notre objectif est de montrer en quoi les discours informels apportent une vision du domaine différente et complémentaire des discours institutionnels, en reflétant plus précisément les préoccupations des internautes.

3.1 *Distanciation et intimité : des univers de référence distincts*

Les spécificités macroscopiques mettent en évidence une opposition majeure entre (i) ce qui relève du conceptuel, du monde des références, en un mot, de l'ontologie propre au discours médical, dans les corpus institutionnels, (ii) un univers de pratiques, d'expériences et de témoignages, qui relèveraient d'une praxis de la maladie ou du risque pathologique, dans les corpus informels (cf. la figure v.10).

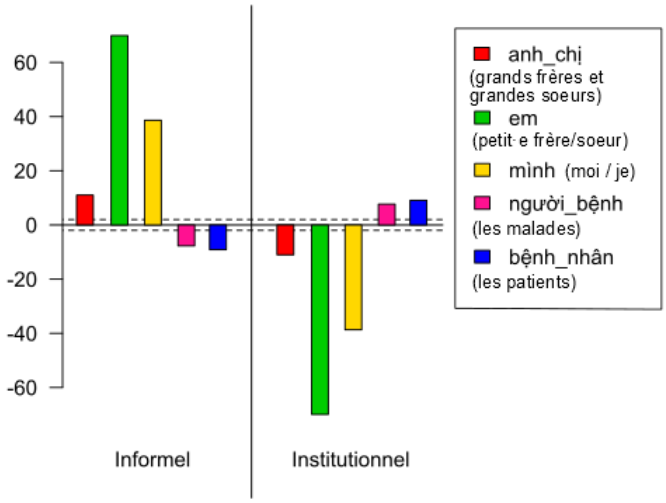
Fig. v.10 : Articulations conceptuelle et anthropique des discours de prévention médico-sanitaire



Cette opposition fondamentale entre discours institutionnels et discours informels issus du web social a déjà été mise en évidence par Slodzian et Valette (2009) à propos du discours de prévention du tabagisme et semble ici se confirmer : le discours institutionnel est caractérisé par des opérations de catégorisation, d'abstraction et de généralisation, tandis que le discours informel se caractérise, à l'inverse, par la description de pratiques concrètes et variées sur un mode narratif et particularisant. Par exemple, dans le corpus vietnamien, les discours institutionnels désignent de grandes catégories de

population, comme *người bệnh* (« les malades ») ou *bệnh nhân* (« les patients »), tandis que les discours informels privilégient l'expression de la première et de la deuxième personne : *em* (« petit·e frère/sœur », qui, dans l'usage vietnamien, signifie « moi » ou « je », avec une connotation déférente), *mình* (« moi », « je » dans le langage courant), *anh chị* (« grand·e·s frères/sœurs », i.e. « vous »)

FIG. V.11 : Les acteurs dans le corpus vietnamien



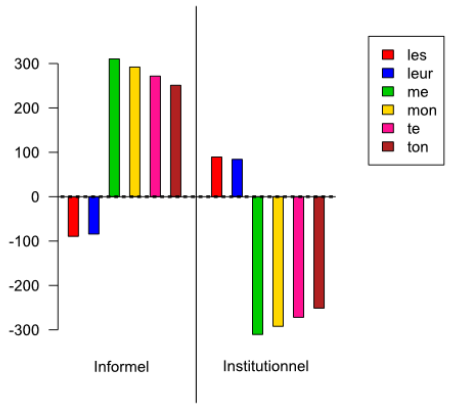
3.1.1 *Zones anthropiques distales, proximales, identitaires*

On observe par ailleurs que les acteurs des discours institutionnels évoluent dans une *zone anthropique distale*[7], c'est-à-dire qu'ils construisent une dis-

7. Nous empruntons à Rastier (2001) le concept de zone anthropique.

tance entre l'énonciateur et l'objet, il est question d'aide aux malades, de populations à risques, de prévention contre de nouvelles contaminations, mais ni de l'*ethos* du malade ni des conséquences psychologiques de l'exposition, de la peur de l'exposition, ou de la contamination. À l'inverse, les acteurs des discours informels évoluent dans une zone anthropique identitaire (*j'ai pris un risque*, *je flippe*) et proximale (*aidez-moi*, *vous me conseillez quoi ?*). L'analyse des variables morpho-syntaxiques, dans le corpus français, corrobore cette opposition anthropique : la deuxième personne du pluriel (vouvoiement) et la troisième personne du pluriel sont caractéristiques des discours institutionnels, tandis que les première et deuxième personnes du singulier sont propres aux discours informels (cf. la figure v.12).

FIG. V.12 : Les pronoms personnels et les adjectifs possessifs dans le corpus français



Unité	F	f_Informel	Score	f_Institutionnel	Score
me	2918	2546	310,4454	372	-310,4454
mon	2266	2040	292,1154	226	-292,1154
te	1234	1209	271,7665	25	-271,7665
ton	1158	1132	251,0863	26	-251,0863
m'	1917	1698	223,985	219	-223,985
ma	1533	1383	200,4529	150	-200,4529
ta	630	620	143,5966	10	-143,5966
mes	906	832	132,3577	74	-132,3577
tes	460	454	107,2486	6	-107,2486
vous	2938	1851	18,5206	1087	-18,5206
sa	1374	833	5,0004	541	-5,0004
ses	865	474	-0,3093	391	0,3093
son	1642	900	-0,3162	742	0,3162
vos	410	214	-0,8321	196	0,8321
notre	532	262	-2,2914	270	2,2914
nos	404	189	-3,1901	215	3,1901
s'	2797	1370	-9,7037	1427	9,7037
votre	705	270	-18,187	435	18,187
nous	2406	1099	-19,1245	1307	19,1245
leurs	497	122	-42,4111	375	42,4111
leur	1537	465	-84,1619	1072	84,1619
les	17059	8063	-89,3847	8996	89,3847

L'opposition marquée entre un univers distal, abstrait et désincarné et un univers intime, concret et lié à des pratiques, permet d'identifier des thèmes sémantiques contrastifs. On prendra ci-dessous l'exemple de deux d'entre eux : les effets secondaires et la sérologie. Les cooccurrents des « effets secondaires » (*tác dụng phụ*), dans le corpus institutionnel, où le concept est

peu abordé, sont *chóng mặt* (« vertiges ») ou *thuốc mới* (« nouveau(x) médicament(s) »). Au contraire les effets secondaires sont très présents dans le corpus informel et montrent l'importance des perceptions intimes (identitaires). Les cooccurrents expriment là encore des procès : *mọc* (« apparaître »), *gây* (« provoquer »), *thấy* (« ressentir », « se sentir ») et des éléments à valeur thymique *ghê gớm* (« horrible »).

La sérologie, dans le corpus institutionnel, est exprimée de façon très distanciée, montrant là encore la fonction objectivante des genres privilégiés dans les discours institutionnels ; par exemple : *tải lượng vi rút* (« charge virale »), *lấy* (« contaminer »), etc. Dans le corpus informel l'ancrage énonciatif est fondamentalement différent et indique que la sérologie est quelque chose de vécu, d'ancré dans la réalité : nous avons (i) des marqueurs de spatialité : *cơ sõ* (« centre »), *bệnh viện* (« hôpital »), *Tràng An* (un grand hôpital privé au Vietnam), *đi* (« aller »), *tại, chỗ* (locatifs), etc. ; et (ii) des marqueurs de temporalité : *6 tuần* (« 6 semaines »), *3 tháng* (« 3 mois »), *sau* (« après »), *giai đoạn* (« période »), etc.

3.1.2 *Ontologie et praxéologie*

L'examen des spécificités macroscopiques du corpus français confirme les observations faites sur le corpus vietnamien, non seulement au niveau des formes graphiques, mais aussi au niveau des catégories syntaxiques. On constate ainsi une sous-représentation notable des verbes dans le corpus institutionnel (67 000 occurrences contre 113 000 dans le corpus informel). Cette observation corrobore vraisemblablement l'opposition entre l'ontos, dont la catégorie grammaticale privilégiée est le substantif, et la praxis, souvent caractérisés par les verbes exprimant un procès. Dans le corpus français informel, on relève de nombreux verbes d'action spécifiques au domaine (passer,

protéger, immuniser, choper, etc.).

Dans le corpus vietnamien informel, on ne peut guère interpréter l'usage qui est fait des verbes en raison de la morphologie isolante du vietnamien. En revanche, les verbes d'action propres au domaine que nous relevons en abondance sont, eux, très spécifiques au corpus informel (*rác* (« déchirer »), *xét nghiệm* (« faire un examen (sérologie) »), *sản sinh* (« produire ») (des anticorps), etc.). Nous notons également une forte présence des verbes à valeur thymique : *yên tâm* (« se tranquilliser »), *lo, lo lắng* (« s'inquiéter »), *mong* (« espérer »), etc.

Dans le corpus institutionnel français, non seulement les verbes sont moins représentés, mais ils ne font pas partie des mêmes classes sémantiques, expression d'un discours volontariste ou interventionniste : permettre, proposer, faciliter, etc. ; réaliser, agir, travailler, constituer, créer, etc. ; engager, consacrer, mobiliser, financer, etc. ; lutter, intervenir, etc. ; améliorer, renforcer, évoluer, augmenter, assurer, etc. Ces verbes sont propres, selon nous, aux genres discursifs « interventionnistes ». En revanche les verbes spécifiques au domaine sont rares dans le corpus (entraîner (des conséquences, des effets secondaires), traiter (une maladie)). Tout se passe comme si nous nous situions au plus haut niveau d'abstraction possible. Les mêmes observations peuvent être réalisées à propos du corpus vietnamien institutionnel : *điều trị* (« traiter »), *chăm sóc* (« surveiller »), *tránh* (« éviter »), *hỗ trợ* (« aider »), *phòng* (« lutter »), *tiếp cận* (« avoir accès »), etc.

3.2 *Tabous, déviance et écart à la norme dans le corpus vietnamien*

L'analyse thématique du corpus vietnamien met en évidence un hiatus marqué au niveau des événements à risque abordés dans les forums de discussions et dans les textes institutionnels, alors que dans le corpus français,

ces différences sont nettement moins marquées. Nous constatons dans le contexte vietnamien, une réelle relation de complétion entre les différents genres discursifs : dans les textes institutionnels, les risques de transmission sont la sexualité en général (*tình dục*), le partage de seringues (*dùng chung bơm kim tiêm*), la transmission de la mère à l'enfant (*mẹ truyền qua con*), ou la transfusion sanguine (*bị truyền máu*), tandis que dans les forums de discussion, les risques de transmissions évoqués sont les rapports sexuels (*quan hệ, QHTD* — à opposer à la sexualité en générale, cf. supra), les prostituées (*cô, GDM, cô gái*), et les accidents de protection, principalement le préservatif (*bao cao su, BCS*) qui sera déchiré, périmé, absent, etc.

Par exemple, en contexte institutionnel, les cooccurrents de *bao cao su* (« préservatif ») le décrivent suivant la norme implicite relative à son utilisation : fonction « protection » (*an toan*), accessibilité « distributeur automatique » (*máy bán bao cao su tự đọng*), usage (*sử dụng bao cao su đúng cách*, littéralement « utiliser correctement le préservatif »). En contexte informel, sont exprimées des situations s'écartant de cette norme : surreprésentation des verbes signifiant se déchirer, utiliser, retirer, être en contact, périmé, etc.

3.3 *La perception du temps*

3.3.1 *Opposition des temps verbaux*

L'étude des parties du discours dans le corpus du forum français apporte des éléments supplémentaires à la description du genre qui a été décrit pour le forum vietnamien au paragraphe 2.2. La surreprésentation de l'imparfait confirme que les genres narratifs y sont privilégiés : l'imparfait est le temps verbal du récit, du témoignage, du retour d'expérience, de la description des actions successives du locuteur, qui les raconte ensuite.

« *J'étais un peu ivre* »

« *toi ou ton mari était porteur avant d'être en couple* »

« *La question me turlupinait* »

« *On m'annonçait que j'étais séropositive* »

S'ajoutent à ces critères verbaux divers marqueurs substantivaux de temps (*hier soir*, *ce matin*, etc., et dans le corpus informel vietnamien, *rồi* (marque du passé), *ngày sau* (« le lendemain »), etc.).

À l'inverse, le participe présent est caractéristique du corpus institutionnel :

« *limitant ainsi le risque de transmission* »

« *la sexualité des femmes vivant avec le virus* »

« *les hommes ayant des relations sexuelles avec des hommes* »

Le patron syntaxique [NOM + VER :ppre] (substantif suivi d'un verbe au participe présent) collecte 936 occurrences dans le corpus institutionnel français, aucune dans le corpus informel.

Tout se passe comme si la temporalité n'était pas exprimée dans les discours institutionnels.

3.3.2 *Atemporalité institutionnelle*

A la vérité, dans les corpus français comme vietnamien, les discours institutionnels substituent à la temporalité deux états atemporels auxquels correspondent deux publics cibles distincts : les personnes non contaminées d'une part et les personnes séropositives d'autre part. Les deux états correspondent donc à la période avant contamination, sur laquelle porte le discours de prévention, et à la période après contamination, auquel correspond un discours

médical, établissant une description de la pathologie. A titre d'exemple, en vietnamien,

- l'état préventif (discours de prévention) actualise des syntagmes tels que *Ai sẽ là người nhiễm HIV ?* (« qui peut attraper le VIH ? »), *Con đường lây nhiễm* (« les voies de contamination »), *Sử dụng bao cao su đúng cách* (« utilisation correcte du préservatif »), *hành vi nguy cơ* (« pratique à risque ») ;

- l'état pathologique (discours médical relatif à la pathologie) actualise des syntagmes tels que *Đã nhiễm HIV* (« contaminé par le VIH »), *ăn uống tốt* (« manger sainement »), *tập thể dục* (« faire du sport »), *điệu trị ARV* (« traitements antirétroviraux »), *Người bệnh là nạn nhân* (« les malades sont des victimes »).

Cette observation cruciale est rendue possible par l'examen différentiel du corpus informel : la temporalité n'y est pas perçue de la même manière. C'est ce que nous verrons dans le paragraphe suivant.

3.3.3 *La fenêtre sérologique : le temps ressenti et mesuré*

La réaction du corps soumis à un virus est de produire des anticorps pour lutter contre celui-ci. Dans le cas du virus d'immunodéficience humaine, le corps va produire des anticorps mais ceux-ci ne seront pas capables d'éradiquer l'infection. L'incertitude ne réside pas à ce niveau mais demeure dans le fait de savoir si le virus a été transmis ou non lors d'un événement représentant un risque, car il est difficile à détecter. Ainsi, les tests de dépistage ne cherchent pas à détecter le virus lui-même mais la présence des anticorps produits en cas d'entrée de celui-ci dans le corps. Pour ajouter à la difficulté, ces anticorps ne sont pas détectables immédiatement après l'infection.

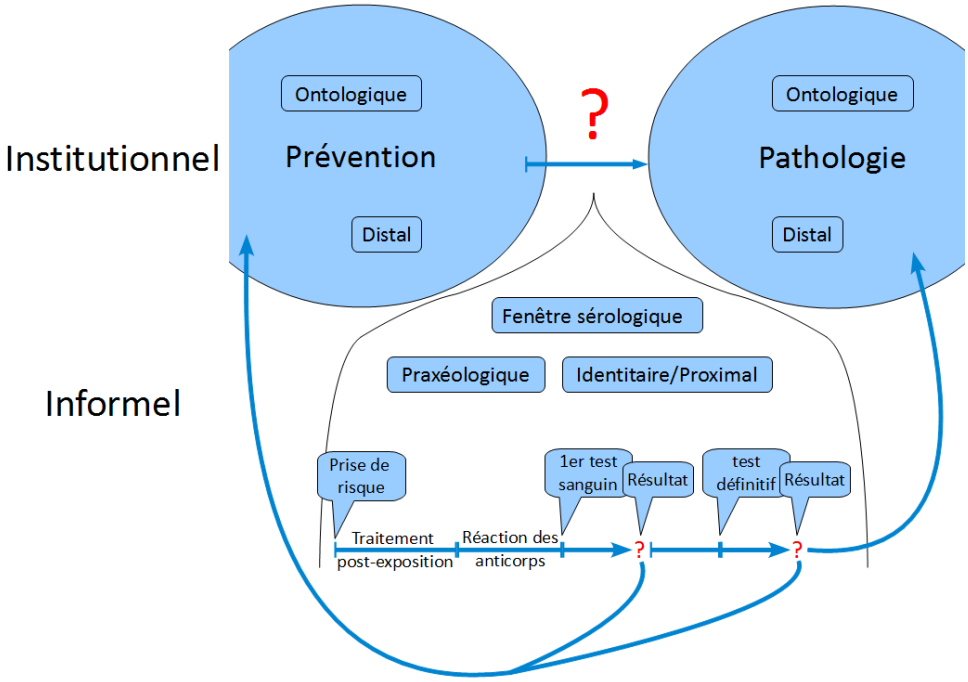
La fenêtre sérologique désigne la période pendant laquelle les anticorps produits pour lutter contre une infection au VIH sont encore indétectables.

Cette période dure en moyenne 3 semaines. Le dépistage doit donc être réalisé au minimum après ce délais. Si aucune trace d'anticorps n'est détectée par le dépistage, la période est étendue à 6 semaines pour assurer avec certitude que le virus n'as pas été transmis. A la fin de cette période, une nouvelle sérologie fixera la réponse. Depuis 1998[8], il est possible de prendre un traitement post-exposition, au mieux avant la 4e heure, au plus tard dans les 48h après une exposition, et ce pendant 4 semaines. Dans ce cas, le dépistage ne peut être fiable que 12 semaines après la fin du traitement, soit 4 mois après l'exposition.

Entre les deux états atemporels du discours institutionnel, entre l'instant précis de la prise de risque, et celui aussi précis du résultat définitif de la sérologie, cette période superbement ignorée des discours institutionnels, cristallise toutes les inquiétudes et interrogations des personnes concernées. C'est cette période que couvrent et investissent dans une large mesure les forums de discussion.

8. En France et pour tous, grâce aux efforts d'Act-Up Paris, voir http://www.actupparis.org/article3030.html.

Fig. v.13 : Les temporalités dans le discours sur le VIH



A cette fenêtre sérologique correspond dans le forum un foisonnement d'échanges saturés de marqueurs dysphoriques. Si on analyse plus en détail cet entre-deux, on y décèle un enchaînement d'intervalles temporels très précis. Par exemple, l'instant de la prise de risque met un terme à l'état préventif, mais n'entraîne pas pour autant un passage à l'état pathologique. Le temps se distend fortement à ce moment-là et les marqueurs de temporalité abondent. Ils ressortissent à deux temps :

- le temps réel (ou chronologique) : il se décompose objectivement en phases médicales (d'une part le traitement post-exposition à prendre dans les 48 heures après l'exposition ; puis les semaines d'incubation où la contamination n'est pas encore détectable, avant la possibilité de procéder à un examen sanguin qui indiquera l'état sérologique et qui déterminera ainsi le passage

vers l'état pathologique ou le retour à l'état préventif.

- le temps vécu, où la mesure prend une grande importance. L'infection n'étant pas détectable pendant les premières semaines, les sujets sont confrontés à une période d'attente incompressible mais d'une durée variable d'une situation à l'autre, les plongeant dans un état d'impuissance et d'incertitude. Cet état est marqué par les signes de la dysphorie d'une part, mais aussi une abondance de commentaires-exutoires. Celle-ci sera observable dans le corpus par la forte présence d'un vocabulaire de la mesure (*semaine*, *jours*, *après*, *đủ ba thắng* (« trois mois entiers »), *bao giờ* (« à quel moment ? »), etc. Voir les figures v.14 et v.15), et des verbes porteurs de procès et d'itération (*refaire*, *commencer*, *retourner*, etc.).

Fig. v.14 : La temporalité dans le corpus vietnamien

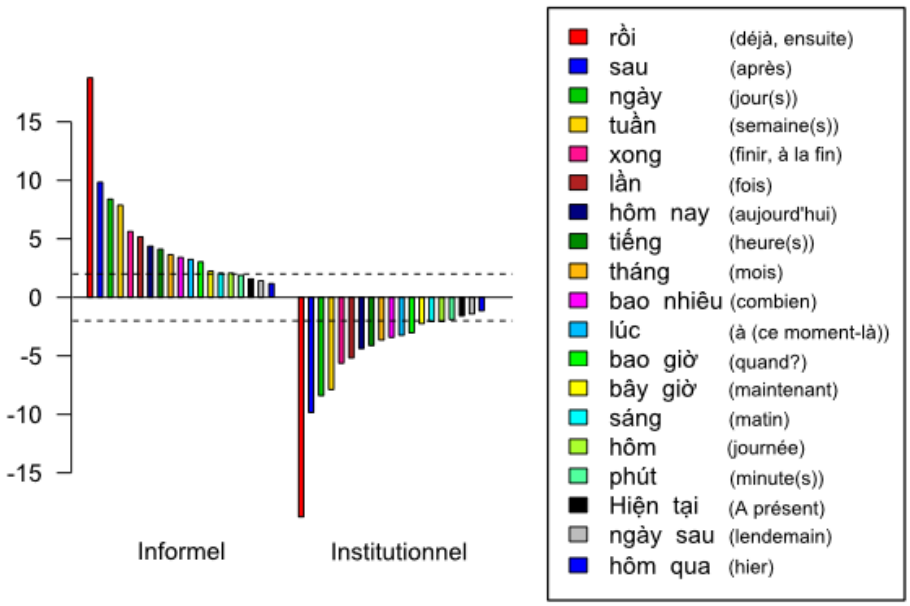
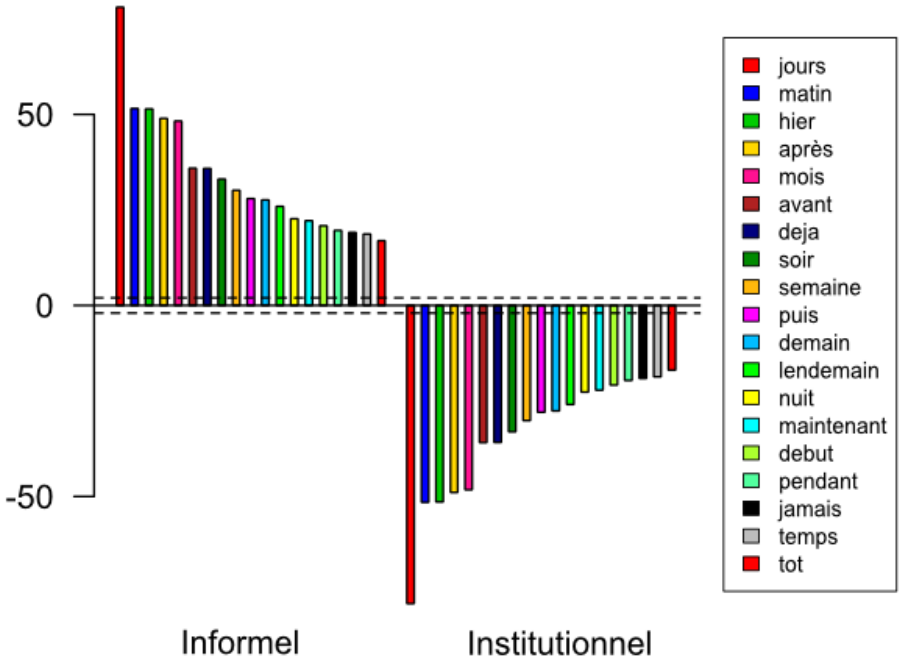


Fig. v.15 : La temporalité dans le corpus français



4 CONCLUSION

Nous avons consacré ce dernier chapitre à la description des discours informels spontanés des forums de discussion. En conservant notre positionnement dans une perspective de complétion des discours institutionnels, nous nous sommes attachée à dresser un panorama des genres discursifs qui se développent spécifiquement dans les environnements des forums. En premier lieu, nous avons porté notre attention sur le critère du foisonnement des interventions, en considérant qu'il constituait un signe particulier des discours informels spontanés (DIS). Après avoir défini les conversations foisonnantes,

nous en avons distingué deux types : l'agora et l'espace de consultation personnalisée. L'agora attire une multitude d'intervenant·e·s différent·e·s, sans que le primat ne soit accordé à un·e intervenant·e en particulier. Dans ce cas, l'initiat·eur/rice de la conversation n'a pas de statut qui le distingue des autres intervenant·e·s. Ce type de conversation foisonnante est marginal dans le corpus. Le type de conversation foisonnante majoritaire sur nos forums a été désigné par l'expression d'espace de consultation personnalisée (ECP), du fait que ces conversations sont le lieu privilégié pour un·e internaute (leur initiat·eur/rice), d'échanger à propos de son cas personnel. Les autres membres, eux, interviennent majoritairement en réponse aux questions posées par celui qui a ouvert l'ECP. Nous avons montré : (i) qu'il était possible de s'appuyer sur la proportion des interventions de l'initiat·eur/rice pour repérer les ECP parmi les conversations foisonnantes et (ii) les avantages à analyser les ECP pour étudier sur le long terme des cheminements d'internautes confrontés au VIH, notamment, en réaction à des épisodes de prise de risque. Nous avons ensuite consacré notre analyse aux interventions-témoignages dans le forum vietnamien, en esquissant des caractéristiques communes, permettant d'extraire du corpus de nouvelles occurrences du même genre discursif. Ces caractéristiques comprennent : (i) le statut d'intervention initiative et (ii) la présence, suivant un scénario commun, de champs lexicaux que nous avons définis. Le sous-corpus des interventions-témoignages a ensuite été examiné à l'aide de cartes de sections pour révéler des faits d'ordre sociologique.

Après avoir, dans la première partie du chapitre, décrit les espaces du forum où se se déploient les discours informels spontanés, nous avons mené des analyses sémantiques sur ces discours, proposant des observations sur les spécificités des pratiques discursives.

Enfin, nous avons appliqué une méthodologie contrastive pour faire émer-

ger les opportunités de complétion des discours institutionnels. Cette méthodologie a permis la mise en relief des endroits où les discours informels spontanés (DIS) prenaient le relais sur ceux-ci : les DIS opposent à la généralisation, l'abstraction, la distanciation caractéristiques des discours institutionnels la description de pratiques concrètes, identitaires et proximales. Pour illustrer ce contraste, nous nous sommes intéressée à la perception du temps. Aux deux états atemporels dessinés par les discours institutionnels, les DIS opposent un foisonnement intense d'échanges caractérisés par la mesure et la scansion d'un temps vécu, foisonnement que nous avons désigné par l'expression fenêtre sérologique, puisqu'elle correspond à la période d'incertitude entre le moment d'une prise de risque et celui du résultat du dépistage.

Conclusion générale et perspectives

L'objectif de cette thèse était de proposer des méthodes d'analyse des discours spontanés natifs du web, qui viennent concurrencer les détenteurs traditionnels de l'autorité en matière de construction des connaissances : les discours institutionnels sont fortement contraints par leur vocation à produire de la norme et de la stabilité. Les discours spontanés du web social réinventent la production de sens commun : leur massivité entraîne de nouvelles formes de régularités, que l'analyse outillée permet de faire émerger. L'analyse de ces discours devait révéler les lacunes des discours institutionnels. Pour appliquer ces méthodes, nous nous sommes intéressée aux forums de discussion sur le VIH.

Acquis de la thèse

Cette thèse a permis l'acquisition d'expertise sur plusieurs plans. Pour la constitution de corpus, nous avons exploité la régularité structurale des forums afin de rendre reproductibles les procédures ; Ce travail se veut également une contribution technique à la manipulation des corpus de données numériques et leurs outils d'exploration ; Concernant le traitement automa-

tique de la langue vietnamienne, nous avons montré qu'il est possible de mener une variété d'analyses sur des données textuelles malgré la faible dotation en outils de la langue traitée. Enfin, l'exploration des données des forums a permis la construction de connaissances, absentes des discours institutionnels, à propos des pratiques et des comportements des jeunes internautes vietnamiens confrontés à l'épidémie de VIH, ce qui était la proposition originelle.

Modélisation des discours natifs du web

Partant du postulat que les discours institutionnels sont stables et normés, nous avons cherché à observer en quelle mesure ils sont concurrencés par les discours natifs du web social. Nous avons montré que, si ces discours sont détachés des contraintes propres aux discours institutionnels, ils s'élaborent dans des environnements spécifiques qui influent sur leur production.

C'est pourquoi nous nous sommes attachée dans le quatrième chapitre à exposer les caractéristiques structurelles de l'environnement étudié, i.e. le forum, en décrivant ses niveaux de segmentation (rubriques, conversations, interventions) comme autant d'objets discursifs. L'organisation structurelle des forums a une fonction de lisibilité, pour faciliter l'accès aux connaissances, dans le foisonnement des interactions qui s'y déploient. Néanmoins, les contraintes discursives sont limitées et il en résulte une diversité de genres discursifs.

Nous avons exploité une variété d'indicateurs pour établir une typologie des genres discursifs des forums. Nous avons utilisé des paramètres formels, tels que

- le nombre d'interventions par conversation pour distinguer – puis décrire – conversations à intervention unique et conversations foisonnantes,

- ou la proportion d'interventions de l'initiat·eur/rice pour repérer les espaces de consultation personnalisée (ECP).

Nous avons également utilisé des indicateurs lexicaux, tels que les segments répétés longs, ou expressions figées, voire les fragments de texte répétés, pour isoler un discours des expert·e·s et observer les phénomènes d'institutionnalisation.

Ces différentes observations ont permis d'organiser les genres discursifs des forums sur un continuum du plus normé (interventions uniques, discours des expert·e·s) au plus spontané (ECP, interventions-témoignages).

A partir de cette typologie, nous avons, dans le dernier chapitre, concentré notre attention sur les discours spontanés natifs du web, tels que les conversations foisonnantes, et parmi celles-ci les ECP ; ainsi que les interventions-témoignages. Pour étudier ces discours nous avons eu recours à des logiciels d'ADT afin de produire des observations quantitatives à analyser. Nous avons ainsi pu étudier de manière diachronique l'institutionnalisation de formes lexicales ; Par l'étude des titres de conversations, nous avons proposé une manière de percevoir les thèmes principalement abordés ; Nous avons également élaboré des listes thématiques et décrit les étapes caractéristiques qui ont permis d'automatiser le repérage des interventions-témoignages parmi les interventions initiatives.

Modélisation des discours sur le VIH

En prenant appui sur notre terrain applicatif des forums sur le VIH, nous avons montré comment s'articulent et interagissent les trois variations du triangle discursif.

Le canon, que constitue la recherche médico-sanitaire mondialisée, insti-

tue des connaissances à propos des traitements et des modes de transmission de l'épidémie de VIH, connaissances vouées à être intégrées par le reste de la société, sans pour autant que soit théorisée leur appropriation.

La vulgate, que constituent les instances publiques et les acteurs de la prévention d'une société en particulier, s'empare du discours canonique global pour le diffuser dans la société à laquelle elle s'adresse, en adjoignant à ces connaissances scientifiques des valeurs morales, spécifiques audit contexte socioculturel, pour produire des discours à caractère officiel, stable et normé, voués à être publicisés, diffusés, répétés, voire matraqués. Dans les forums, la vulgate s'incarne d'abord dans le DIR, autrement dit des productions discursives à caractère officiel (textes juridiques, articles scientifiques) externes au forum et reproduites à l'identique par le procédé du *copier-coller*, procédé rendu possible par le caractère numérique des discours, et qui influe non seulement sur leur diffusion mais également sur leur production. D'autre part, la vulgate s'incarne dans la parole des intervenant·e·s expert·e·s, qui s'autocitent en reproduisant des interventions précédentes qu'eux-même ou d'autres expert·e·s ont publiées précédemment pour répondre à une situation similaire. Là encore le procédé du *copier-coller* est utilisé. Plus largement que ce procédé éditorial, DIR et discours des expert·e·s possèdent le caractère répétitif que nous attribuons à la vulgate, en raison de leur inscription dans une norme.

La doxa, constituée dans notre étude par les échanges informels produits nativement sur les forums, vient combler par un foisonnement de productions discursives informelles et spontanées (DIS) les espaces discursifs laissés vacants par la vulgate. Ce phénomène est constaté notamment lors de la période qualifiée de fenêtre sérologique, qui concentre toutes les incertitudes des internautes, alors qu'elle n'est pas investie par les discours institutionnels : ceux-ci lui opposent deux états atemporels, celui de la prévention d'une part

et celui du traitement d'autre part. La mesure du temps est donc également un champ spécifique à la doxa, autrement dit, la doxa couvre la dimension proximale des discours sur le VIH. Nous avons montré dans le contexte vietnamien par quels procédés les acteurs de la doxa s'affranchissent des contraintes discursives imposées par les valeurs morales de la vulgate (tabous) pour verbaliser les sujets de l'intime. Ces procédés sont d'ordre stylistique (euphémisations) mais prennent également la forme de déresponsabilisation, notamment par le biais du champ lexical de la perte de contrôle. Par ces procédés, les intervenant·e·s sont en mesure de s'exprimer sur tout sujet, y compris ceux que les tabous de la vulgate occultent. En étudiant ces discours nous avons ainsi montré à quel point le VIH était associé à la prostitution dans les discours de la doxa.

Perspectives

L'ensemble du travail a été mené avec une volonté de reproductibilité des procédures. Ainsi, la modélisation des genres discursifs des forums peut être exploitée pour analyser d'autres forums de discussion. Sur un autre plan, l'automatisation des traitements de corpus leur donne un caractère intrinsèquement reproductible et des possibilités d'améliorations itératives.

En travaillant, dans une optique de comparabilité des données, à partir de deux langues, deux contextes socioculturels distincts, nous avons mis en évidence des fonctionnements communs. Se dessine par conséquent une perspective de généralisation des capacités d'analyse à d'autres contextes linguistiques et socioculturels.

La tâche de collecte des données peut être exécutée de manière automatique, sur d'autres rubriques que celles sélectionnées pour les corpus de la thèse, ou sur un empan temporel différent. Elle peut également être adaptée

à d'autres forums, en intégrant leurs spécificités interfacielles.

La tâche d'adaptation automatique des corpus à différents outils d'ADT peut également faire l'objet d'un élargissement. En effet, si *Lexico 5* et *TXM* ont satisfait les analyses exposées dans le manuscrit, les programmes que nous avons développés présentent la possibilité d'adapter les corpus à d'autres outils, notamment *Weka* ou *Calico*, et quelques développements simples permettraient l'élargissement à des outils supplémentaires comme *Iramuteq* ou d'autres encore.

La tâche d'anonymisation automatique peut également être améliorée pour intégrer plus de pratiques discursives concernant les variations d'écriture des pseudonymes.

La prise en compte des dernières avancées dans le traitement automatique de la langue vietnamienne pourrait notamment intégrer un étiqueteur syntaxique, afin d'enrichir les possibilités d'analyse des discours vietnamiens.

Enfin, nos travaux se situent dans un champ de recherche plus large, celui de l'automatisation des traitements du langage naturel, c'est pourquoi nos efforts de modélisation étaient menés ici dans la perspective d'une automatisation. Les récurrences observées pourraient faire l'objet d'intégration dans des procédures de TAL robuste et servir de base à la conception d'un système de complétion automatique des discours institutionnels par les discours natifs du web social.

Annexes

ANNEXE A

Étapes de constitution du corpus

L'ensemble des commandes à exécuter et leurs options est présenté page suivante sous la forme d'un tableau, les commentaires à propos du tableau le suivent.

Langages et scripts informatiques

L'essentiel des programmes a été développé dans le langage de programmation PERL, donc ils sont exécutables à partir de tous les systèmes d'exploitation (Windows, Mac, Unix). Leur exécution doit être appelée à la ligne de commande dans une invite de commandes. Nous avons notamment eu recours au module `Web::Scraper` pour développer le programme de collecte des données depuis le web. Le programme de transformation des corpus en vue de leur import dans les différents logiciels d'analyse est également développé en PERL. Le segmenteur du vietnamien a été développé par ses concepteurs en Java, c'est donc dans ce langage qu'il a été adapté pour notre usage. MySQL a été utilisé pour le recours aux bases de données. Enfin, des scripts simples interprétables par le *shell* (l'interpréteur de commandes) ont été ajoutés pour faciliter l'exécution des programmes assortis de leurs options, gérer leur enchaînement (lancement consécutif ou en parallèle), et l'arborescence des répertoires. Ayant essentiellement travaillé en environnement Windows, ceux-ci sont au format Batch (le langage shell de Windows), mais il est tout-à-fait possible d'exécuter les programmes principaux sans cette couche, à partir de n'importe quel environnement.

Tab. A.1 : Chaîne de traitement des corpus

FORUM	SITE
Parcours en ligne et **sélection des rubriques** à traiter	
vn : 22 ; 50 ; 56 ; 62 ; 66 ; 68 fr : 1 ; 8 ; 14 ; 17 ; 24	vn : (id des rubriques donnés arbitrairement) 01 ; 02 ; 03 ; 04 ; 05 fr : pas de rubrique
	Récupération manuelle du code source pour chaque rubrique (opération longue car la page est créée dynamiquement) source02ThongTin.html source03ThuocDieuTri.html source04ChuyenBaNguoi.html source05ChuyenHaiNguoi.html
Positionnement dans le répertoire Corpus	
Création d'un répertoire par rubrique (nom rép. = id rubrique)	
`>mkdir 22 50 56 62 66 68`	`>mkdir 01 02 03 04 05`
Récupération d'une url par conversation dans chaque rubrique (plusieurs processus peuvent être lancés en parallèle)	
getTopics.pl (nécessite une connexion à Internet)	generateArticlesUrls.pl (processus hors ligne)

`>perl getTopics.pl 22 get`
`>perl getTopics.pl 50 get`
`>perl getTopics.pl 56 get`
`>perl getTopics.pl 62 get`
`>perl getTopics.pl 66 get`
`>perl getTopics.pl 68 get`
	`>perl generateArticlesUrls.pl source02ThongTin.html 02`
	`>perl generateArticlesUrls.pl source03ThuocDieuTri.html 03`
	`>perl generateArticlesUrls.pl source04ChuyenBaNguoi.html 04`
	`>perl generateArticlesUrls.pl source05ChuyenHaiNguoi.html 05`

Un fichier urls.txt a été généré dans chaque répertoire	

FORUM	SITE

Récupération des données (nécessite une connexion à Internet)
(ne pas lancer plusieurs processus en parallèle car le fichier authors.csv est commun)

getPosts.pl	getArticles.pl
`>perl getPosts.pl get section 22`	
`>perl getPosts.pl get section 50`	`>perl getArticles.pl ws 02`
`>perl getPosts.pl get section 56`	`>perl getArticles.pl ws 03`
`>perl getPosts.pl get section 62`	`>perl getArticles.pl ws 04`
`>perl getPosts.pl get section 66`	`>perl getArticles.pl ws 05`
`>perl getPosts.pl get section 68`	

Les données brutes ont été récupérées pour chaque rubrique
et la liste des intervenants est stockée dans authors.csv

Segmentation du vietnamien dans chaque fichier
(plusieurs processus peuvent être lancés en parallèle)
(les sous-répertoires de rubriques seront créés automatiquement dans VnTok)

```
>java -jar vnTokenizer.jar -i 22 -o VnTok/22
>java -jar vnTokenizer.jar -i 50 -o VnTok/50
>java -jar vnTokenizer.jar -i 56 -o VnTok/56
>java -jar vnTokenizer.jar -i 62 -o VnTok/62
>java -jar vnTokenizer.jar -i 66 -o VnTok/66
>java -jar vnTokenizer.jar -i 68 -o VnTok/68
>java -jar vnTokenizer.jar -i 01 -o VnTok/01
>java -jar vnTokenizer.jar -i 02 -o VnTok/02
>java -jar vnTokenizer.jar -i 03 -o VnTok/03
>java -jar vnTokenizer.jar -i 04 -o VnTok/04
>java -jar vnTokenizer.jar -i 05 -o VnTok/05
```

FORUM	SITE
Lancement de generate.pl pour chaque rubrique et pour chaque logiciel (ne pas lancer plusieurs processus en parallèle car le fichier metadata.csv est commun)	

Parcours en ligne et sélection des rubriques à traiter

(Pour une description détaillée de cette étape non automatisée, se référer aux paragraphes 2.1.2 et 2.2.1) En ce qui concerne nos forums, chaque rubrique en ligne est identifiée par un chiffre, qui sert à la construction de son URL, et lui servira également d'identifiant dans notre corpus. Il faut donc en premier lieu retenir les identifiants des rubriques à aspirer, et créer un répertoire par rubrique, nommé avec ce même identifiant. Pour les corpus institutionnels, les identifiants sont donnés arbitrairement.

Collecte des données

Pour chaque rubrique, la première étape est d'établir la liste des conversations qui y ont été ouvertes, grâce au programme `getTopics.pl`. Cette liste contiendra l'URL de chaque conversation. Un empan temporel étant défini au préalable (dans la fonction `inRange` du fichier `MySubs.pl`), ne sont retenues que les conversations qui ont été ouvertes dans cet intervalle de temps. (Se référer au paragraphe 2.2.2) De même pour les articles institutionnels, le programme `getArticles.pl` génère la liste des URL des pages à récupérer à partir du fichier HTML constitué pour chaque rubrique. Après cette étape il existe donc un répertoire par rubrique, portant comme nom l'identifiant de la rubrique, et dans chaque répertoire un fichier contenant la liste des URL des conversations qui ont été ouvertes au sein de la rubrique. Une fois cette liste établie, il est possible de mettre à jour cette liste ultérieurement, avec l'option `update` au lieu de `get`.

Dès lors, la collecte à proprement parler peut démarrer, avec le programme `getPosts.pl` (et `getArticles.pl` pour l'institutionnel). Chaque conversation de la rubrique traitée est récupérée dans un fichier XML enregistré dans le répertoire de la rubrique. Ceci est fait pour toutes les rubriques à traiter. A cette étape, une liste des intervenant·e·s est également constituée (pseudonyme + identifiant en ligne).

Le fichier XML d'une conversation contient les métadonnées suivantes : l'identifiant de la conversation et son titre, et pour chaque intervention : son identifiant, son aut.eur/rice (pseudonyme + identifiant), la date de sa publication, le contenu du texte. Dans le cas du corpus vietnamien, des informations supplémentaires ont été conservées dans la version brute du corpus, mais supprimées au moment de la segmentation. Ces informations sont les citations et les remerciements. Ces derniers sont l'équivalent des *j'aime* de Facebook, ou des *pouces bleus* de Youtube par exemple. Si l'intervention a été *remerciée*, le nombre de remerciements est enregistré, ainsi que les pseudonymes des membres qui ont remercié.

Segmentation du vietnamien

Dans le cas du corpus vietnamien, c'est à cette étape qu'intervient la tâche de segmentation, avec l'aide du logiciel `vnTokenizer.jar` (voir le paragraphe 3.1). Celui-ci étant open-source, son code a pu être personnalisé pour l'adapter aux spécificités du corpus et l'enrichir de nouvelles fonctionnalités correspondant à ces spécificités : gestion de la structure XML et de l'arborescence par rubriques, normalisation des émoticônes, suppression des citations (ceci pour éviter les textes en doublons, mais elles sont conservées comme information annexe dans la version brute du corpus).

A la fin de cette étape, les corpus vietnamiens et français se présentent sous des formes équivalentes.

Mise en forme des corpus pour leur import dans les logiciels d'exploration

Une fois le corpus aspiré et stocké localement sous forme brute (XML), il faut le transformer de manière à ce qu'il soit interprétable par le logiciel avec lequel on compte l'explorer, chaque logiciel impose un format d'import spécifique.

Un script Batch `generate.bat` permet de lancer le programme principal (`generate.pl`) pour plusieurs rubriques. Autrement, le programme principal `generate.pl` peut être exécuté directement, une fois pour chaque rubrique. Dans ce cas il faut créer dans le répertoire CorpusVn/ ou CorpusFr/ un répertoire du nom du logiciel pour lequel on veut transformer le corpus, et pour TXM un sous-répertoire topics/ ou posts/ selon que l'on veut générer un corpus segmenté en conversations ou en interventions ; pour Weka un sous-répertoire pour chaque rubrique. Ces tâches de gestion de l'arborescence sont prises en charge par le script Batch. Que ce soit avec le script Batch ou sans, il faut préciser les options choisies :

- pour quel logiciel le corpus doit être mis en forme
  (Lexico, TXM, Weka, Calico)
- quel est le corpus à mettre en forme
  (français, vietnamien, institutionnel, informel)
- quelles sont les rubriques à traiter
  (l'ensemble du corpus ou un sous-ensemble)
- les données doivent-elles être anonymisées.

De plus, pour TXM il faut déterminer le niveau de segmentation du corpus (conversation ou intervention, voir le paragraphe 3.2.2) ; et pour Lexico il faut préciser si la rubrique traitée constituera à elle seule un corpus ou si elle doit être ajoutée à un corpus constitué de plusieurs rubriques (nommé `CorpusLexicoTotal.txt`).

Pour un récapitulatif des formats de corpus pour chaque logiciel d'exploration, se reporter au tableau A.2. Soient

- nr le nombre de rubriques du corpus
- nc le nombre de conversations dans le corpus
- ni le nombre d'interventions dans le corpus
- nj le nombre d'interventions dans la conversation

Tab. A.2 : Récapitulatif des formats de corpus

	Lexico	TXM
Format	1 fichier .txt balisé -> nr balises ¤ -> nc balises # -> ni balises µ	1 fichier metadata.csv -> entête + nc ou ni lignes nc ou ni fichiers .txt
Exemple	¤ <section=XX> <title=...> # <topic=xxx> <title=...> <...*topic metadata*...> µ <postID=xxx> <...*post metadata*...> [texte du message 1] µ <postID=xxx> <...post metadata...> [texte du message 2]	id,sectionid,topicid,topictitle,rank, authorname,authorid,date,month,corpus "xxx","XX","xxx","...","n","...","xxx","20x xxxxxx","20xx_xx...","in...el" "xxx","XX", "xxx", "...","n","...","xxx","2 0xxxxxxx","20xx_xx...", "in...el" --- [texte du message 1] --- [texte du message 2]

	Weka	Calico
Format	nr répertoires -> nc fichiers .arff -> entête + nj lignes	nr fichiers .xml -> ni balises <message>
Exemple	@relation XX @attribut postID numeric @attribute author string @attribute date numeric @attribute text string @data xxx,...,20xxxxxxx,"[texte du message 1]" xxx,...,20xxxxxxx,"[texte du message 2]"	<?xml version="1.0"encoding="utf-8" ?> <forum> <name>...XX...</name> <message id="xxx"> <header> <datetime>20xx-x...</datetime> <author id="xxx">...</author> <subject id="xxx">...</subject> </header> <body> <content type="text">[texte du message 1]</content> </body> </message> <message id="xxx"> <header>...</header> <body>...</body> </message> ... </forum>

Suppression des doublons dans le corpus français

Le forum français donne la possibilité de classer une même conversation dans plusieurs rubriques. La conversation se trouvera donc dupliquée dans chacune des rubriques auxquelles elle a été associée. Pour éliminer ces doublons de manière automatisée nous avons eu recours à la puissance des bases de données en écrivant un script MySQL[1] (voir l'annexe 4). Il faut pour l'utiliser avoir généré la version pour TXM du corpus : en effet le fichier metadata.csv est nécessaire à l'exécution du script SQL, ainsi qu'un répertoire contenant l'ensemble des conversations toutes rubriques confondues.

1. Script qui a déjà pu être utilisé par d'autres doctorants pour la constitution de leur corpus.

ANNEXE B

Utilisation de Web ::Scraper

Ce module PERL sert à récupérer sélectivement des éléments du code-source d'une page web (au format html). Pour indiquer les éléments que l'on souhaite récupérer il faut repérer les nœuds et les attributs correspondants. Plusieurs conventions d'écriture sont au choix.

La plus basique est celle du CSS :

- le `.` sert à désigner l'attribut `class` d'un noeud
- le `#` sert à désigner l'attribut `id` d'un noeud
- les noeuds enfants suivent en étant séparés par un espace.

Par exemple `'div.postdiv div div table'` désigne le noeud `table` descendant du noeud `div class="postdiv"`

Pour des sélections plus complexes il faut passer à Xpath :

- il faut débuter l'expression par `//`
- les attributs sont nommés en toutes lettres ainsi : `[@attribut=" "]`
- les noeuds suivants sont séparés par `/`

Par exemple `'//div[@class="postdiv"]/div/div/table'` désigne le noeud `table` descendant du noeud `div class="postdiv"`

- Il est également possible d'utiliser des sélecteurs plus précis : `contains` et `starts-with`

Par exemple `'//table[contains(@class, "postContainer")]'`
ou `'//div[starts-with(@id, "dvThanks")]/a/u'`

Pour encadrer les expressions les simples quotes (`'`) peuvent être utilisés. Si les doubles quotes (`"`) sont utilisés il faut despécialiser `@` et `'` par `\`.
Par exemple `'//table[contains(@class, "postContainer")]'`
équivaut à `"//table[contains(\@class, \'postContainer\')]"`

ANNEXE C

Recours aux bases de données

L'interface utilisée est MySQL Workbench.

Créer une nouvelle connexion. Dans la fenêtre de gauche, cliquer sur Startup / Shutdown.

Le script `DeleteDuplicateTopics.sql` dans `ProcessingFlow\CorpusFr` crée une nouvelle database (éventuellement en changer le nom si on exécute le script plusieurs fois).

Avant de l'exécuter, il faut avoir une version pour TXM totale du corpus à traiter, toutes rubriques confondues : dans un même répertoire toutes les conversations quelle que soit leur rubrique. Il faut adapter le chemin de ce répertoire dans le script SQL. Il faut également créer une version du fichier metadata.csv auquel on retire la première ligne l'en-tête, et que l'on nomme metadataSansEntete.csv.

Le script `DeleteDuplicateTopics.sql` peut ensuite être exécuté. Des conversations existant sous plusieurs rubriques, ne sont conservées que celles dont la rubrique a le plus petit identifiant (Si une conversation est classée dans les rubriques 1 et 8, le script ne conserve que l'entrée de la rubrique 1, si une une conversation est classée dans les rubriques 8 et 14, le script ne conserve que l'entrée de la rubrique 8,etc.).

Une fois la table de résultat obtenue, dans la fenêtre qui s'affiche avec la table, cliquer sur Export (recordset to an external file), enregistrer en .csv

Ne pas oublier de remplacer dans la ligne d'en-tête `topicID` par `id` avant d'importer le corpus dans TXM.

ANNEXE D

Extrait de metadata.csv - Corpus segmenté par conversation

Ce fichier de métadonnées est créé lorsque le script generate.pl est exécuté pour générer une version du corpus importable dans le logiciel TXM. TXM propose plusieurs formats d'import. L'option TXT+CSV a été choisie afin de pouvoir exploiter les métadonnées rassemblées dans le fichier CSV. A titre d'illustration, voici représenté un extrait d'un fichier CSV (sous forme de tableau pour une meilleure lisibilité). Le script generate.pl propose deux types de segmentation du corpus : en conversation ou en intervention. L'extrait présenté correspond à la première.

Légende :

- id : identifiant de la conversation
- Titre : titre de la conversation, donné par son initiateur
- A1 : identifiant de l'initiateur de la conversation
- D1 : date d'ouverture de la conversation
- TM : nombre total d'interventions dans la conversation
- TA : nombre total d'intervenants dans la conversation

id	Titre	A1	D1	TM	TA	...
29503	Thắc mắc cần trợ giúp	3765	20070124213736	2	2	
29792	Có lây nhiễm HIV không khi đã lỡ...	1206	20070201124846	1	1	
29909	Quan hệ không xài BCS và xuất tinh ngoài	3812	20070202233710	7	6	
30225	HIV qua 1 lan phoi nhiem	3200	20070207060611	3	1	
31287	nguy cơ lây nhiễm H của tôi có cao không ?	3965	20070226220639	5	5	
33764	Em muốn hỏi về bệnh lậu và bệnh giang mai.	4193	20070402091334	3	3	
35253	cho mình hỏi về lông sinh dục	4423	20070503113445	3	3	
35254	Bệnh lậu - phải làm sao	4424	20070503131450	8	7	
36093	cho em hỏi gấp cac pác ơi !!!!	4519	20070519140758	3	3	
36640	Y' KIEN CHUNG,MONG CAC BAN CO KINH NGHIEM GOP' Y'	4563	20070525122340	6	6	
40555	Tặng BCS	4815	20070627235210	10	9	
40826	Trước khi hành kinh có phải huyết trắng ra nhiều ?	4838	20070630111240	1	1	
43884	Bao cao su co thuc su an toan ko ?	3255	20070724231636	8	7	
44927	ORAL XEX va tay	5040	20070804112353	1	1	
44928	ORAL XEX va tay	5040	20070804112446	2	2	
46162	Bệnh hiểm	83	20070817195621	4	4	
48720	Bao lâu mới đi xét nghiệm được các bệnh lây qua đường tình dục như giang mai, lậu....	5390	20070909130326	3	3	
49939	help me, ban gai em chua co kinh sau khi „„	5601	20070921235807	41	13	
52175	bi lau co the bi H ko	5898	20071024093714	4	3	
53535	Quảng cáo bao cao su !	0	20071031204947	5	4	
54216	Herpes sinh dục là gì ?	2126	20071106193006	2	2	
54823	Giới trẻ và tình dục an toàn	2947	20071109153908	1	1	
55090	Mấy anh Pro ơi ... nhờ tư vấn giúp em lần đầu QHTD !	5565	20071111232857	38	14	
60257	dùng thuốc tránh thai postinor 2 liệu có an toàn ? (trả lời gấp hộ em)	814	20071204113418	13	5	
60342	Quan hệ tình dục đường miệng - Những điều cần lưu ý	0	20071204190602	1	1	
60359	Tìm hiểu về Bệnh zona	0	20071204202236	1	1	
62159	Bệnh lậu mãn tính và hiện tượng tiết dịch niệu đạo	0	20071217195512	1	1	
62646	cho e hoi ? cai nay' cai	6233	20071220104238	106	10	
63548	cho e hỏi gấp vấn đề này	6233	20071226122520	96	11	
63591	BCS OK có tránh thai tốt không ?	6527	20071226182726	8	5	
64242	loi' cam mon chan thanh'	6602	20080102132336	3	2	
64756	Cần giúp đỡ về HIV/AIDS	6632	20080106103425	7	6	
70535	Cho em hoi ve chat luong bao cao su OK ?	6824	20080219113058	11	8	
...						

ANNEXE E

Extrait de metadata.csv - Corpus segmenté par intervention

Ce fichier de métadonnées est créé lorsque le script generate.pl est exécuté pour générer une version du corpus importable dans le logiciel TXM. TXM propose plusieurs formats d'import. L'option TXT+CSV a été choisie afin de pouvoir exploiter les métadonnées rassemblées dans le fichier CSV. A titre d'illustration, voici représenté un extrait d'un fichier CSV (sous forme de tableau pour une meilleure lisibilité). Le script generate.pl propose deux types de segmentation du corpus : en conversation ou en intervention. L'extrait présenté correspond à la seconde.

Légende :

- id : identifiant de l'intervention
- tid : identifiant de la conversation
- Titre : titre de la conversation
- R : rang de l'intervention dans la conversation
- A : auteur de l'intervention
- D : date de publication de l'intervention

id	tid	Titre	R	A	D	...
29503	29503	Thắc mắc cần trợ giúp	1	3765	20070124213736	
29909	29909	Quan hệ không xài BCS và xuất tinh ngoài	1	3812	20070202233710	
30082	29909	Quan hệ không xài BCS và xuất tinh ngoài	2	3806	20070205190026	
30100	29909	Quan hệ không xài BCS và xuất tinh ngoài	3	2824	20070205230614	
30122	29909	Quan hệ không xài BCS và xuất tinh ngoài	4	83	20070206124246	
30125	29909	Quan hệ không xài BCS và xuất tinh ngoài	5	3812	20070206130610	
30198	29909	Quan hệ không xài BCS và xuất tinh ngoài	6	3082	20070206235150	
30225	30225	HIV qua 1 lan phoi nhiem	1	3200	20070207060611	
30996	30225	HIV qua 1 lan phoi nhiem	2	3200	20070224062506	
30997	29503	Thắc mắc cần trợ giúp	2	3200	20070224062658	
31287	31287	nguy cơ lây nhiễm H của tôi có cao không ?	1	3965	20070226220639	
31290	31287	nguy cơ lây nhiễm H của tôi có cao không ?	2	3557	20070226222423	
31315	31287	nguy cơ lây nhiễm H của tôi có cao không ?	3	3200	20070227054951	
31316	30225	HIV qua 1 lan phoi nhiem	3	3200	20070227055200	
31317	29909	Quan hệ không xài BCS và xuất tinh ngoài	7	3200	20070227055338	
33469	31287	nguy cơ lây nhiễm H của tôi có cao không ?	4	628	20070329101753	
33764	33764	Em muốn hỏi về bệnh lậu và bệnh giang mai.	1	4193	20070402091334	
33843	33764	Em muốn hỏi về bệnh lậu và bệnh giang mai.	2	4075	20070403162004	
35253	35253	cho mình hỏi về lông sinh dục	1	4423	20070503113445	
35254	35254	Bệnh lậu - phải làm sao	1	4424	20070503131450	
35265	35254	Bệnh lậu - phải làm sao	2	737	20070503182059	
35300	35254	Bệnh lậu - phải làm sao	3	4424	20070504155813	
36093	36093	cho em hỏi gấp cac pác ơi !!!!	1	4519	20070519140758	
36640	36640	Y' KIEN CHUNG,MONG CAC BAN CO KINH NGHIEM GOP' Y'	1	4563	20070525122340	
39167	36640	Y' KIEN CHUNG,MONG CAC BAN CO KINH NGHIEM GOP' Y'	2	4001	20070616101947	
39444	36640	Y' KIEN CHUNG,MONG CAC BAN CO KINH NGHIEM GOP' Y'	3	3664	20070618072720	
40555	40555	Tặng BCS	1	4815	20070627235210	
40572	36640	Y' KIEN CHUNG,MONG CAC BAN CO KINH NGHIEM GOP' Y'	4	4143	20070628012703	
40758	36640	Y' KIEN CHUNG,MONG CAC BAN CO KINH NGHIEM GOP' Y'	5	883	20070629130508	
40927	36640	Y' KIEN CHUNG,MONG CAC BAN CO KINH NGHIEM GOP' Y'	6	4691	20070701084755	
41417	36093	cho em hỏi gấp cac pác ơi !!!!	2	4862	20070704095735	
41428	35253	cho mình hỏi về lông sinh dục	2	4862	20070704101944	
41436	36093	cho em hỏi gấp cac pác ơi !!!!	3	3603	20070704104122	
41437	35253	cho mình hỏi về lông sinh dục	3	3603	20070704104329	
...						

ANNEXE F

Segments répétés les plus fréquents par rubrique

Tab. F.1 : Rubrique (62) Relations sexuelles, préservatif, lubrifiant

F	L	Segment	Traduction
4681	2	có nguy_cơ	avoir un risque
2859	2	em có	j'ai / ai-je
2665	2	không có	il n'y a pas
2042	2	cho em	permettez-moi de
1950	2	của bạn	ton / à toi
1807	2	mọi người	vous tous
1671	2	cô ấy	elle / cette femme
1665	2	của em	mon / à moi
1604	2	đi xét_nghiệm	faire un dépistage
1524	2	nhiễm HIV	contaminé au VIH
1328	2	có bị	[affirmation] + [élément négatif]
1306	2	cô ta	elle / cette femme
1291	2	bị nhiễm	contaminé.e
1278	2	mình có	j'ai / ai-je
1267	2	của mình	mon / à moi
1267	2	các bạn	les amis
1266	2	các anh	les grands frères
1236	2	quan_hệ với	rapport avec
1219	2	3 tháng	3 mois
1156	2	bạn không	tu n'as pas
1072	2	không biết	ne sais pas
1057	2	sau đó	ensuite / après ça
987	2	ko có	
967	2	với GMD	
963	2	cho mình	
954	2	có dùng	
948	2	6 tuần	
887	2	thì bạn	
867	2	e có	

Tab. F.1 : Rubrique (62) Relations sexuelles, préservatif, lubrifiant (suite)

F	L	Segment	Traduction
864	2	sau khi	
848	2	bạn có	
823	2	đi xn	
797	2	hay không	
782	2	em không	
761	2	dùng BCS	
759	2	không cần	
756	2	thì em	
743	2	em hỏi	
733	2	giúp em	
731	2	có quan_hệ	
731	2	bạn đã	
722	2	em đã	
715	2	bạn nên	
713	2	khi quan_hệ	
687	2	bị rách	
681	2	rất nhiều	
667	2	em cũng	
667	2	các anh_chị	
642	2	không phải	
637	2	không sao	
623	2	không bị	
621	2	quan_hệ tình_dục	
612	2	các bệnh	
605	2	trường_hợp của	
601	2	cũng không	
599	2	lúc đó	
592	2	có đi	
585	2	thì không	
577	2	cho bạn	
576	2	1 lần	
565	2	đó có	
563	2	đó em	

F	L	Segment	Traduction
1369	3	không có nguy_cơ	ne pas avoir de risque
566	3	cho em hỏi	permet.s/tez-moi de demander
529	3	bị nhiễm HIV	contaminé au VIH
475	3	có quan_hệ với	avoir un rapport avec
444	3	bạn không có	tu n'as pas
418	3	quan_hệ với GMD	rapport avec une prostituée
322	3	có dùng BCS	utilisé un préservatif / protégé
308	3	ko có nguy_cơ	sans risque / pas de risque
307	3	k có nguy_cơ	sans risque / pas de risque
304	3	Bạn không có	Tu n'as pas
283	3	có bị nhiễm	contaminé
283	3	giúp em với	m'aider à propos de
282	3	có nguy_cơ lây_nhiễm	avoir un risque d'attrapper
277	3	có sử_dụng BCS	utilisé un préservatif / protégé
277	3	sau 3 tháng	après 3 mois
260	3	em có nguy_cơ	j'ai un risque de
257	3	có nguy_cơ nhiễm	avoir un risque d'être contaminé
253	3	Không có nguy_cơ	Pas de risque
248	3	em có bị	ai-je + [élément négatif]
247	3	em có quan_hệ	j'ai eu un rapport
242	3	trong trường_hợp này	dans cette situation
239	3	tư_vấn giúp em	me conseiller / conseil pour m'aider
238	3	của bạn là	pour toi c'est
232	3	em có đi	j'ai été / je suis allé
231	3	của em có	à moi il y a
216	3	qua đường tình_dục	par voie sexuelle
207	3	trường_hợp của em	mon cas
204	3	Trường_hợp của bạn	Ton cas
202	3	thuốc phơi nhiễm	traitement post-exposition
196	3	của cô ấy	
196	3	nên đi xét_nghiệm	
188	3	là có nguy_cơ	
186	3	cho mình hỏi	
184	3	là không có	
184	3	trường_hợp của bạn	
179	3	của bạn không	
178	3	mình có bị	
177	3	cô ấy có	
176	3	sau đó em	
175	3	có phải là	
175	3	sau 6 tuần	
173	3	bạn có nguy_cơ	
172	3	anh cho em	
171	3	tư_vấn cho em	
170	3	cô ta có	
166	3	bạn không cần	
166	3	không cần dùng	
166	3	những gì bạn	

Rubrique (62) Relations sexuelles, préservatif, lubrifiant (suite)

F	L	Segment	Traduction
376	4	bạn không có nguy_cơ	tu n'as pas de risque
290	4	Bạn không có nguy_cơ	Tu n'as pas de risque
188	4	có quan_hệ với GMD	avoir un rapport avec une prostituée
183	4	em có quan_hệ với	j'ai eu un rapport avec
133	4	có bị nhiễm HIV	contaminé au VIH
119	4	Bạn k có nguy_cơ	T'as pas de risque
112	4	lây qua đường tình_dục	sexuellement transmissible
108	4	anh cho em hỏi	permet.s/tez-moi de demander
100	4	là không có nguy_cơ	ne comporte pas de risque
100	4	thuốc chống phơi nhiễm	traitement post-exposition
94	4	có nguy_cơ nhiễm HIV	avoir un risque d'être contaminé au VIH
94	4	dùng thuốc phơi nhiễm	prendre un traitement post-exposition
90	4	những gì bạn kể	ce que tu as dis
89	4	cho em lời khuyên	donnez-moi un conseil
88	4	của bạn không có	à toi n'a pas
82	4	nói lên điều gì	ne rien dire
81	4	thì không có nguy_cơ	alors il n'y a pas de risque
79	4	quan_hệ với gái mại_dâm	rapport avec une prostituée
78	4	mình có quan_hệ với	j'ai eu un rapport avec
77	4	cho em hỏi là	permettez-moi de demander la chose suivante
77	4	tránh xa các dịch_vụ	se tenir à l'écart des services
76	4	từ ngày có nguy_cơ	depuis le jour de la prise de risque
75	4	có quan_hệ với 1	avoir un rapport avec 1
74	4	trong đó có HIV	
73	4	ăn bánh trả tiền	
72	4	mình có bị nhiễm	
71	4	bị nhiễm HIV không	
71	4	bệnh lây qua đường	
69	4	không có nguy_cơ nhiễm	
68	4	bạn k có nguy_cơ	
67	4	bạn ko có nguy_cơ	
67	4	tư_vấn giúp em với	
66	4	có nguy_cơ nhiễm hiv	
66	4	không có nguy_cơ lây_nhiễm	
66	4	QHTD có sử_dụng BCS	
63	4	anh_chị cho em hỏi	
62	4	có sử_dụng BCS là	
62	4	các bệnh lây qua	
62	4	smileybiggrin [x 4]	
61	4	quan_hệ tình_dục không an_toàn	
59	4	có nguy_cơ lây_nhiễm HIV	
58	4	không cần dùng Pep	
58	4	mọi người tư_vấn giúp	
58	4	sử_dụng BCS là an_toàn	
58	4	smileyhuh [x 4]	
58	4	smileylaugh [x 4]	

Rubrique (62) Relations sexuelles, préservatif, lubrifiant (suite)

F	L	Segment	Traduction
75	5	em có quan_hệ với GMD	j'ai eu un rapport avec une prostituée
63	5	bệnh lây qua đường tình_dục	maladie(s) sexuellement transmissible(s)
49	5	có sử_dụng BCS là an_toàn	utilisé un préservatif c'est protégé
48	5	6 tuần và 3 tháng	6 semaines et 3 mois
46	5	kể từ ngày có nguy_cơ	raconter depuis le jour de la prise de risque
41	5	có bị nhiễm HIV không	contaminé au VIH ?
41	5	không nói lên điều gì	ne rien dire
41	5	với GMD có dùng BCS	protégé avec une prostituée
40	5	có quan_hệ với gái mại_dâm	eu un rapport avec une prostituée
39	5	các anh cho em hỏi	permettez-moi de demander
38	5	6 tuần và 12 tuần	6 semaines et 12 semaines
37	5	thì bạn không có nguy_cơ	alors tu n'as pas de risque
36	5	có bị nhiễm HIV ko	contaminé au VIH ?
36	5	không có nguy_cơ nhiễm hiv	pas de risque d'être contaminé au vih
36	5	uống thuốc chống phơi nhiễm	prendre un traitement post-exposition
36	5	smileylaugh [x 5]	[émoticône qui rit] [x 5]
35	5	mình có bị nhiễm HIV	
35	5	biết mình có bị nhiễm	
35	5	bệnh tình_dục trong đó có	
35	5	smileybiggrin [x 5]	
35	5	dịch_vụ ăn bánh trả tiền	
34	5	mình có quan_hệ với GMD	
34	5	các bệnh tình_dục trong đó	
34	5	truyền_nhiễm lây qua đường tình_dục	
33	5	có bị nhiễm HIV hay	
33	5	nói lên được điều gì	
33	5	bệnh lây_truyền qua đường tình_dục	
32	5	các bệnh lây_truyền qua đường	
31	5	em có QHTD với GMD	
31	5	bị nhiễm HIV hay không	
31	5	Các anh cho em hỏi	
30	5	em có quan_hệ với 1	
29	5	có nguy_cơ bị nhiễm HIV	
29	5	của em có cao không	
29	5	dùng thuốc phơi nhiễm HIV	
29	5	trường_hợp bạn không có nguy_cơ	
28	5	có quan_hệ với GMD có	
28	5	em có nguy_cơ bị nhiễm	
28	5	bạn tự bảo_vệ cho mình	
28	5	không có nguy_cơ nhiễm HIV	
28	5	Bạn không có nguy_cơ nhiễm	
28	5	những gì bạn kể thì	

Rubrique (62) Relations sexuelles, préservatif, lubrifiant (suite)

F	L	Segment	Traduction
34	6	Trường_hợp của bạn không có nguy_cơ	Ton cas n'est pas à risque
23	6	QHTD có sử_dụng BCS là an_toàn	rapport protégé c'est sans risque
22	6	có bị nhiễm HIV hay không	contaminé ou non au VIH / contaminé au VIH ?
21	6	không có nguy_cơ nên không cần	pas de risque donc pas besoin de
20	6	không có nguy_cơ trong trường_hợp này	pas de risque dans ce cas de figure
20	6	smileybiggrin [x 6]	[émoticône qui rit fort] [x 6]
19	6	các dịch_vụ ăn bánh trả tiền	
19	6	ngày kể từ ngày có nguy_cơ	
19	6	bệnh lây qua đường tình_dục khác	
19	6	gmd nữa để k mang thêm	
19	6	Tránh tìm gmd nữa nha bạn	
18	6	Bạn không có nguy_cơ nhiễm hiv	
18	6	smileylaugh [x 6]	
18	6	Tránh tìm GMD nữa nha bạn	
17	6	không nói lên được điều gì	
17	6	của bạn là không có nguy_cơ	
17	6	QHTD với GMD có dùng BCS	
17	6	Những khó_khăn của người nhiễm HIV	
17	6	smileyphtqr [x 6]	
17	6	Ai sẽ là người nhiễm HIV	
16	6	em có quan_hệ với gái mại_dâm	
16	6	em có quan_hệ với GMD có	
16	6	để k mang thêm nỗi lo	
16	6	xa các dịch_vụ ăn bánh trả	
16	6	Tránh tìm GMD nữa để k	

F	L	Segment	Traduction
28	7	các bệnh tình_dục trong đó có HIV	les maladies sexuelles dont le VIH
18	7	các bệnh truyền_nhiễm lây qua đường tình_dục	les maladies sexuellement transmissibles
17	7	tìm gmd nữa để k mang thêm	ne plus aller chercher une prostituée pour ne plus
15	7	gmd nữa để k mang thêm lo_lắng	ne plus ... prostituée pour ne plus s'inquiéter
15	7	xa các dịch_vụ ăn bánh trả tiền	loin des services sexuels tarifés
15	7	Những gì bạn kể không có nguy_cơ	Ce que tu as décrit est sans risque
15	7	smileyphtqr [x 7]	[émoticône qui applaudit] [x 7]
15	7	chuẩn của Việt_Nam về thời_gian xét_nghiệm HIV	??? vietnamien(ne) sur le délais de dépistage du VIH
14	7	tìm GMD nữa để k mang thêm	ne plus chercher une prostituée pour ne plus
13	7	để biết rõ tình_trạng sức_khỏe của mình	pour connaître clairement son état de santé
13	7	GMD nữa để k mang thêm lo_lắng	ne plus ... prostituée pour ne plus s'inquiéter
12	7	các bệnh lây qua đường tình_dục khác	les autres maladies sexuellement transmissibles
12	7	được hướng_dẫn dùng thuốc phơi nhiễm HIV	se faire prescrire un traitement post-exposition
12	7	để được hướng_dẫn dùng thuốc phơi nhiễm	pour se faire prescrire un traitement post-exposition
11	7	và các bệnh lây qua đường tình_dục	et les maladies sexuellement transmissibles
11	7	các bệnh về tình_dục trong đó có	les maladies sexuelles dont
11	7	bệnh về tình_dục trong đó có HIV	maladies sexuelles dont le VIH
11	7	smileybiggrin [x 7]	[émoticône qui rit fort] [x 7]
11	7	CLB Spam của những member đang chờ	Club des membres qui attendent

Tab. F.2 : Rubrique (68) Stigmatisation et discrimination

F	L	Segment
364	2	nhiễm HIV
187	2	người nhiễm
170	2	các bạn
168	2	những người
149	2	bị nhiễm
144	2	người có
130	2	có nguy_cơ
130	2	không có
127	2	mọi người
115	2	của mình
107	2	của bạn
104	2	của người
100	2	có H
89	2	có HIV
88	2	lây_nhiễm HIV
87	2	em có
78	2	về HIV
75	2	với người
73	2	rất nhiều
70	2	cho em
69	2	HIV không
67	2	sự kỳ_thị
65	2	không phải
64	2	có bị
64	2	là một
64	2	cho mình
64	2	đó là
63	2	không biết
63	2	mình có
62	2	của em
62	2	đi xét_nghiệm
61	2	người bị
61	2	cho người
61	2	móng tay
61	2	phân_biệt đối_xử
56	2	HIV là
54	2	mình bị
54	2	bị HIV
53	2	người khác
53	2	cho bạn
51	2	cũng có

Rubrique (68) Stigmatisation et discrimination (suite)

F	L	Segment
153	3	người nhiễm HIV
90	3	bị nhiễm HIV
53	3	người có HIV
51	3	người có H
34	3	những người có
30	3	với người nhiễm
29	3	không có nguy_cơ
28	3	của người nhiễm
27	3	người bị nhiễm
26	3	cho em hỏi
26	3	phân_biệt đối_xử với
26	3	chất dạng thuốc_phiện
25	3	can_thiệp giảm tác_hại
24	3	không phải là
24	3	những người nhiễm
23	3	có cái nhìn
22	3	đối_xử với người
21	3	thuốc kháng HIV
21	3	dự_phòng lây_nhiễm HIV
20	3	với người có
20	3	với những người
18	3	và phân_biệt đối_xử
18	3	bị lây_nhiễm HIV
17	3	có bị nhiễm
17	3	em có bị
17	3	kỳ_thị và phân_biệt
17	3	kỳ_thị với người
17	3	trường_hợp của bạn
16	3	là những người
16	3	cho người nhiễm
15	3	nhiễm HIV không
15	3	về căn_bệnh này
15	3	sống chung với
15	3	smileyeusaboohoo [x 3]
14	3	có dính máu
14	3	và các bạn
14	3	cho người khác
13	3	bị nhiễm H
13	3	một trong những
13	3	anh cho em
13	3	theo quy_định của
13	3	quy_định tại khoản
13	3	kiến_thức về HIV
13	3	Mong các bạn

Rubrique (68) Stigmatisation et discrimination (suite et fin)

F	L	Segment
28	4	với người nhiễm HIV
23	4	của người nhiễm HIV
21	4	phân_biệt đối_xử với người
20	4	những người nhiễm HIV
18	4	các chất dạng thuốc_phiện
17	4	bởi quản_trị viên
16	4	người bị nhiễm HIV
16	4	kỳ_thị và phân_biệt đối_xử
14	4	smileyeusaboohoo [x 4]
13	4	những người có H
13	4	biện_pháp can_thiệp giảm tác_hại
12	4	anh cho em hỏi
11	4	có bị nhiễm HIV
11	4	mình bị nhiễm HIV
11	4	cho người nhiễm HIV
10	4	của người có HIV
10	4	bị phơi nhiễm với
10	4	những người có HIV
10	4	gây ra hội_chứng suy_giảm
10	4	suy_giảm miễn_dịch mắc phải

Rubrique (68) Stigmatisation et discrimination (suite et fin)

F	L	Segment
17	5	nghiện các chất dạng thuốc_phiện
15	5	Sửa bởi quản_trị viên
13	5	smileyeusaboohoo [x 5]
10	5	các biện_pháp can_thiệp giảm tác_hại
22	6	can_thiệp giảm tác_hại trong dự_phòng lây_nhiễm
13	6	phân_biệt đối_xử với người nhiễm HIV
13	6	chất dạng thuốc_phiện bằng thuốc thay_thế
12	6	smileyeusaboohoo [x 6]
17	7	can_thiệp giảm tác_hại trong dự_phòng lây_nhiễm HIV
11	7	biện_pháp can_thiệp giảm tác_hại trong dự_phòng lây_nhiễm
11	7	smileyeusaboohoo [x 7]
10	8	smileyeusaboohoo [x 8]

Tab. F.3 : Rubrique (66) Exposition au VIH

F	L	Segment	Traduction
452	2	uống thuốc	prendre un médicament
437	2	phơi nhiễm	exposition à une contamination
266	2	có nguy_cơ	avoir un risque
238	2	cho em	permet.tez/s-moi
231	2	em có	j'ai
210	2	mọi người	vous tous
205	2	nhiễm HIV	contaminé.e au VIH
200	2	sau khi	après avoir
170	2	dùng thuốc	prendre un médicament
151	2	1 tháng	1 mois
150	2	em hỏi	je demande
148	2	của em	à moi /ma/mon
146	2	dùng PEP	prendre un TPE
141	2	không có	ne pas avoir
140	2	thuốc phơi	médicament exposition
138	2	của bạn	à toi/ton/ta
125	2	đi xét_nghiệm	faire une dépistage
115	2	các bạn	les amis
114	2	anh Tuấn	grand frère Tuấn
113	2	em đã	j'ai
113	2	không biết	ne pas savoir
113	2	3 tháng	3 mois
112	2	em uống	je prends/ j'ai pris (un médicament)
112	2	tác_dụng phụ	effet(s) secondaire(s)
102	2	các anh	les grands frères
99	2	bị nhiễm	contaminé.e (passif négatif)
99	2	sau đó	ensuite
96	2	mình có	j'ai
96	2	của mình	à moi /ma/mon
96	2	quan_hệ với	rapport avec
95	2	cho bạn	pour toi
93	2	điều_trị phơi	traitement expooser
87	2	dùng pep	prendre un tpe
86	2	mua thuốc	acheter médicament
84	2	trong thời_gian	dans l'intervalle de temps
84	2	loại thuốc	type de médicament
83	2	không cần	pas besoin
82	2	đã uống	bu (boire au passé)
81	2	uống PEP	prendre un TPE
81	2	e có	j'ai

Rubrique (66) Exposition au VIH (suite)

F	L	Segment	Traduction
135	3	thuốc phơi nhiễm	traitement post-exposition
111	3	cho em hỏi	permet.tez/s-moi de demander
91	3	điều_trị phơi nhiễm	traitement post-exposition
81	3	phơi nhiễm HIV	exposition au VIH
74	3	chống phơi nhiễm	contre l'exposition
63	3	thuốc chống phơi	médicament contre exposer
46	3	không cần dùng	pas besoin de prendre
45	3	uống thuốc phơi	prendre un traitement exposer
42	3	đúng và đủ	à heures fixes et sans en oublier
41	3	không có nguy_cơ	il n'y a pas de risque
39	3	dùng thuốc phơi	prendre un traitement exposer
39	3	bị phơi nhiễm	contaminé.e
36	3	và mọi người	et vous tous
35	3	có quan_hệ với	avoir un rapport avec
35	3	tác_dụng phụ của	effet(s) secondaire(s) de
33	3	quan_hệ với GMD	rapport avec une prostituée
32	3	bị nhiễm HIV	contaminé.e au VIH
29	3	của em có	ma/mon/ à moi comporte
29	3	khi có nguy_cơ	quand il y a une prise de risque
29	3	phụ của thuốc	secondaire(s) du médicament
27	3	em uống thuốc	j'ai pris (un médicament)
27	3	sau khi uống	après avoir pris (un médicament)
26	3	có nguy_cơ lây_nhiễm	avoir un risque de se faire contaminer
26	3	uống đúng giờ	prendre (un médicament) à l'heure
26	3	anh cho em	permet.tez/s-moi
26	3	giúp em với	aide.z-moi à propos de
25	3	đã uống thuốc	avoir pris un médicament
25	3	tháng thứ 3	le 3ème mois
24	3	không phải là	il est faux que
24	3	sau 3 tháng	après 3 mois
22	3	có hại cho	avoir des effets néfastes sur
22	3	trong thời_gian điều_trị	pendant la durée du traitement
22	3	tác_dụng của thuốc	effet(s) du médicament
21	3	có tác_dụng phụ	avoir des effets secondaires
21	3	em đã uống	j'ai pris (un médicament)
21	3	uống thuốc chống	prendre un médicament contre
21	3	e uống thuốc	j'ai pris un médicament
21	3	trường_hợp của em	mon cas

Rubrique (66) Exposition au VIH (suite et fin)

F	L	Segment	Traduction
59	4	thuốc chống phơi nhiễm	traitement post-exposition
43	4	uống thuốc phơi nhiễm	prendre un traitement post-exposition
38	4	dùng thuốc phơi nhiễm	prendre un traitement post-exposition
24	4	thuốc phơi nhiễm HIV	traitement post-exposition au VIH
23	4	tác_dụng phụ của thuốc	effet(s) secondaire(s) du médicament
20	4	uống thuốc chống phơi	prendre un traitement contre exposer
18	4	anh cho em hỏi	permet.tez/s-moi de demander
17	4	điều_trị phơi nhiễm HIV	traitement post-exposition
16	4	Tuấn và mọi người	Tuấn et vous tous
15	4	cho em hỏi thêm	permet.tez/s-moi de poser une autre question
14	4	em có quan_hệ với	j'ai eu un rapport avec
14	4	uống đúng và đủ	prendre à heures fixes et sans en oublier
14	4	cho em hỏi là	permet.tez/s-moi de demander la chose suivante
13	4	phác_đồ điều_trị phơi nhiễm	traitement post-exposition
19	5	uống thuốc chống phơi nhiễm	prendre un traitement post-exposition
10	5	dùng thuốc chống phơi nhiễm	prendre un traitement post-exposition
10	5	cần dùng thuốc phơi nhiễm	besoin de prendre un traitement post-exposition

Tab. F.4 : Rubrique (56) Connaissances de base sur le VIH/SIDA

F	L	Segment
286	2	nhiễm HIV
91	2	bị nhiễm
74	2	người nhiễm
71	2	những người
69	2	không có
58	2	có HIV
58	2	có nguy_cơ
53	2	người có
51	2	quan_hệ tình_dục
50	2	người khác
47	2	lây_nhiễm HIV
46	2	của mình
44	2	qua đường
43	2	của người
41	2	với người
40	2	sau khi
36	2	cho người
34	2	HIV và
34	2	hay không
33	2	người bị
33	2	là một
33	2	xét_nghiệm HIV
33	2	mọi người
32	2	các bệnh
31	2	Chưa rõ
30	2	của họ
29	2	về HIV
28	2	nguy_cơ lây_nhiễm
28	2	of HIV
28	2	phơi nhiễm
27	2	có khả_năng
27	2	Sửa bởi
26	2	HIV không
26	2	của bạn
26	2	một người
26	2	bạn tình
26	2	3 tháng
26	2	virus HIV
26	2	tiếp_xúc với

Rubrique (56) Connaissances de base sur le VIH/SIDA (suite et fin)

F	L	Segment
67	3	bị nhiễm HIV
57	3	người nhiễm HIV
29	3	người có HIV
28	3	qua đường tình_dục
19	3	Người nhiễm HIV
18	3	cho người khác
17	3	có nguy_cơ lây_nhiễm
17	3	người bị nhiễm
17	3	nguy_cơ nhiễm HIV
16	3	bị phơi nhiễm
16	3	Bạn đã bao_giờ
15	3	nhiễm HIV sẽ
15	3	lây qua đường
15	3	QHTD không an_toàn
14	3	nguy_cơ lây_nhiễm HIV
14	3	quan_hệ tình_dục với
13	3	trên 1 tháng
13	3	tình_dục với người
13	3	sang giai_đoạn AIDS
12	3	không phải là
12	3	khi quan_hệ tình_dục
12	3	chuyển sang giai_đoạn
11	3	of HIV transmission
10	3	viêm gan B
10	3	bên ngoài cơ_thể
16	4	từ mẹ sang con
13	4	người bị nhiễm HIV
12	4	lây_truyền qua đường tình_dục
11	4	quan_hệ tình_dục với người
11	4	kéo_dài trên 1 tháng
11	4	chuyển sang giai_đoạn AIDS

Tab. F.5 : Rubrique (22) Safe sex

F	L	Segment	Traduction
384	2	tránh thai	contraception
351	2	quan_hệ tình_dục	rapport sexuel
296	2	qua đường	par voie
260	2	không có	il n'y a pas
225	2	là một	est un.e
217	2	đường tình_dục	voie sexuelle
209	2	của mình	[possessif 1ère pers. sing.]
202	2	sau khi	après avoir
189	2	nữ nhân	femme
165	2	nam nhân	homme
164	2	các bệnh	les maladies
160	2	của bạn	[possessif 2ème pers. sing.]
148	2	những người	les/des personnes
145	2	trước khi	avant de
143	2	bạn tình	partenaire
135	2	không phải	il ne faut pas / c'est faux
128	2	làm cho	rendre
128	2	lây_truyền qua	transmis.e(s) par
123	2	chuyện ấy	rapport (litt.. : cette histoire)
121	2	khi quan_hệ	pendant le rapport
116	2	các bạn	les amis
116	2	mà không	mais pas
114	2	đó là	ça c'est
113	2	cả hai	tous les deux
112	2	đã có	avait
110	2	có một	avoir un
110	2	có những	avoir des
109	2	thuốc tránh	médicament préventif
106	2	cũng là	est/sont aussi
104	2	rất nhiều	très nombreux
104	2	nhiễm HIV	contaminé au VIH
103	2	có nguy_cơ	avoir un risque
103	2	cho rằng	dire que
101	2	không biết	ne pas savoir
100	2	có tác_dụng	
100	2	là những	
100	2	mang thai	
97	2	không được	
96	2	mắc bệnh	
94	2	có quan_hệ	
93	2	smileybiggrin [x 2]	
90	2	trong khi	
90	2	một lần	
89	2	các loại	
88	2	cũng không	
88	2	đây là	
87	2	cũng có	
87	2	đi khám	

Rubrique (22) Safe sex (suite)

F	L	Segment	Traduction
209	3	qua đường tình_dục	par voie sexuelle
115	3	lây_truyền qua đường	transmis.e(s) par voie
108	3	thuốc tránh thai	médicament contraceptif / pilule
81	3	" yêu "	"(s')aimer" (faire l'amour)
61	3	biện_pháp tránh thai	moyen(s) de contraception
57	3	smileybiggrin [x 3]	émoticône qui rit fort [x 3]
55	3	lây qua đường	transmis.e(s) par voie
54	3	không phải là	il est faux que
50	3	có quan_hệ tình_dục	avoir un rapport sexuel
41	3	là một trong	est un.e parmi
41	3	khi quan_hệ tình_dục	pendant le rapport
41	3	bao quy đầu	prépuce
40	3	các bệnh lây_truyền	les maladies transmissibles
35	3	bệnh lây qua	maladies transmissibles par
34	3	quan_hệ tình_dục với	rapport sexuel avec
33	3	của nữ nhân	de(s / la) femme(s)
33	3	một trong những	un.e parmi
30	3	tình_dục không an_toàn	sexuel non protégé
30	3	quan_hệ tình_dục không	rapport sexuel non
30	3	da quy đầu	prépuce
30	3	smileylaugh [x 3]	émoticône qui rit [x 3]
29	3	cho cả hai	pour tous les deux
29	3	smileydrink [x 3]	émoticône qui boit de l'alcool [x 3]
28	3	các loại thuốc	les sortes de médicaments
28	3	" áo_mưa "	"capote" (litt.. : vêtement de pluie)
27	3	Thuốc tránh thai	Médicament(s) contraceptif(s)
26	3	qua đường hậu_môn	par voie anale
25	3	của nam nhân	des hommes
25	3	bị nhiễm HIV	contaminé au VIH
24	3	quan_hệ tình_dục bằng	rapport sexuel par
24	3	tránh thai khẩn_cấp	contraceptif d'urgence
24	3	qua đường miệng	par voie orale
23	3	có rất nhiều	il y a beaucoup
23	3	không có gì	pas de quoi
23	3	uống tránh thai	prendre contraceptif
22	3	chứ không phải	et non pas
22	3	lây_lan qua đường	transmissible(s) par voie
21	3	được coi là	qu'on appelle
21	3	cũng là một	est aussi un.e
21	3	ngày rụng trứng	jour d'ovulation

Rubrique (22) Safe sex (suite)

F	L	Segment	Traduction
99	4	lây_truyền qua đường tình_dục	transmissible(s) par voie sexuelle
80	4	bệnh lây_truyền qua đường	maladie(s) transmissible(s) par voie
52	4	" chuyện ấy "	"cette histoire" (le rapport)
45	4	lây qua đường tình_dục	transmissible(s) par voie sexuelle
27	4	là một trong những	est un.e parmi
27	4	smileybiggrin [x 4]	émoticône qui rit fort [x 4]
25	4	" cậu nhỏ "	"petit gars" (sexe masculin)
22	4	quan_hệ tình_dục không an_toàn	rapport sexuel non protégé
21	4	smileydrink [x 4]	émoticône qui boit de l'alcool [x 4]
20	4	lây_lan qua đường tình_dục	transmissible(s) par voie sexuelle
18	4	các bệnh lây qua	les maladies transmissibles par
17	4	quan_hệ tình_dục bằng miệng	rapport bucco-génital
15	4	các biện_pháp tránh thai	les moyens contraceptifs
15	4	một biện_pháp tránh thai	un moyen contraceptif
15	4	dùng thuốc tránh thai	prendre un médicament contraceptif
14	4	thuốc uống tránh thai	prendre un médicament contraceptif
14	4	lây_nhiễm qua đường tình_dục	transmissible(s) par voie sexuelle
14	4	nhiễm_trùng lây_truyền qua đường	??? transmissible(s) par voie
13	4	bệnh qua đường tình_dục	maladie(s) par voie sexuelle
13	4	quan_hệ tình_dục với người	rapport sexuel avec une personne
13	4	thuốc tránh thai khẩn_cấp	pilule du lendemain
13	4	qua đường tình_dục và	par voie sexuelle et
13	4	uống thuốc tránh thai	prendre un médicament contraceptif
12	4	lây_truyền qua đường TD	transmissible(s) par voie sexuelle
12	4	smileylaugh [x 4]	émoticône qui rit [x 4]
12	4	cắt bao quy đầu	circoncire
11	4	có tác_dụng tránh thai	avoir un effet abortif
11	4	bạn tình của mình	mon/ma partenaire
11	4	tình_dục qua đường hậu_môn	sexe anal
11	4	cả nam và nữ	homme comme femme
11	4	càng sớm càng tốt	le plus tôt c'est le mieux
10	4	là biện_pháp tránh thai	est une / sont des mesure(s) contraceptive(s)
10	4	" cô bé "	"petite fille" (sexe féminin)
10	4	sử_dụng thuốc tránh thai	utiliser un contraceptif
10	4	dễ dính bầu nhất	le(s) plus susceptible(s) de mettre en cloque

F	L	Segment	Traduction
79	5	bệnh lây_truyền qua đường tình_dục	MST
39	5	các bệnh lây_truyền qua đường	les maladies transmissibles par voie
33	5	bệnh lây qua đường tình_dục	MST
20	5	smileydrink [x 5]	émoticône qui boit [x 5]
17	5	smileybiggrin [x 5]	émoticône qui rit fort [x 5]
15	5	làm " chuyện ấy "	faire "[possessif] + affaire"
11	5	nhiễm_trùng lây_truyền qua đường TD	transmissible(s) par voie sexuelle
38	6	các bệnh lây_truyền qua đường tình_dục	les MST
19	6	smileydrink [x 6]	émoticône qui boit [x 6]
17	6	các bệnh lây qua đường tình_dục	les MST
15	6	những bệnh lây_truyền qua đường tình_dục	des MST
10	6	smileybiggrin [x 6]	émoticône qui rit fort [x 6]
18	7	smileydrink [x 7]	émoticône qui boit [x 7]
17	8	smileydrink [x 8]	émoticône qui boit [x 8]
16	9	smileydrink [x 9]	émoticône qui boit [x 9]
15	10	smileydrink [x 10]	émoticône qui boit [x 10]
14	11	smileydrink [x 11]	émoticône qui boit [x 11]

Rubrique (50) Explorer la sexualité

F	L	Segment
488	2	bạn sẽ
397	2	là một
381	2	của mình
309	2	không có
297	2	những người
293	2	của bạn
234	2	không phải
222	2	rất nhiều
222	2	quan_hệ tình_dục
215	2	có những
197	2	đó là
196	2	người đàn_ông
175	2	tránh thai
171	2	người khác
164	2	phải là
159	2	một người
158	2	trước khi
154	2	sau khi
153	2	sẽ có
152	2	của người
152	2	mọi người
146	2	các bạn
141	2	người phụ_nữ
140	2	có một
140	2	với những
139	2	có nhiều
139	2	làm cho
137	2	của các
137	2	Bạn sẽ
136	2	là những
136	2	mà không
132	2	cũng là
129	2	nhiều người
129	2	cũng có
127	2	Cuối tuần
124	2	cô ấy
118	2	trong những
116	2	bạn tình
116	2	ảnh_hưởng đến
115	2	nhiều hơn
114	2	dẫn đến
113	2	gây ra
112	2	không biết
111	2	không được
109	2	cho rằng
108	2	sẽ rất
106	2	không nên
104	2	hay không
(...)		

Rubrique (50) Explorer la sexualité (suite)

F	L	Segment
103	3	không phải là
77	3	qua đường tình_dục
63	3	smileybiggrin [x 3]
60	3	thuốc tránh thai
59	3	bạn sẽ có
54	3	lây_truyền qua đường
51	3	một trong những
51	3	bệnh lây_truyền qua
51	3	biện_pháp tránh thai
45	3	bao quy đầu
44	3	là một trong
42	3	được coi là
42	3	smileylaugh [x 3]
41	3	bạn sẽ rất
40	3	có quan_hệ tình_dục
40	3	có rất nhiều
36	3	" yêu "
36	3	từ đầu tuần
35	3	nam và nữ
33	3	Đây cũng là
31	3	ngay từ đầu
30	3	với những người
30	3	chuyện ấy "
30	3	chứ không phải
30	3	Cuối tuần này
29	3	của những người
29	3	một người đàn_ông
29	3	sự phát_triển của
29	3	smileydrink [x 3]
28	3	trong tuần này
28	3	sẽ giúp bạn
28	3	với một người
28	3	đạt cực khoái
27	3	của bạn sẽ
27	3	của người khác
27	3	của người phụ_nữ
27	3	sẽ có những
26	3	là những người
26	3	các cặp vợ_chồng
26	3	các mối quan_hệ
26	3	1 – 19
26	3	quan_hệ tình_dục với
26	3	5 – 21
26	3	4 – 21
26	3	6 – 22
26	3	7 – 22
26	3	8 – 22
26	3	Đây là một
26	3	12 – 20
(...)		

Rubrique (50) Explorer la sexualité (suite)

F	L	Segment
51	4	lây_truyền qua đường tình_dục
50	4	bệnh lây_truyền qua đường
37	4	là một trong những
36	4	smileybiggrin [x 4]
29	4	“ chuyện ấy ”
26	4	smileylaugh [x 4]
23	4	ngay từ đầu tuần
22	4	“ tự sướng ”
22	4	smileydrink [x 4]
20	4	thì bị phạt tù
20	4	bị phạt tù từ
17	4	smileycool [x 4]
15	4	các biện_pháp tránh thai
15	4	smileyeusapray [x 4]
14	4	cũng không phải là
14	4	dùng thuốc tránh thai
14	4	chứ không phải là
14	4	viêm gan siêu_vi B
14	4	chú_ý nhiều hơn đến
13	4	không phải là một
13	4	smileyphtqr [x 4]
12	4	trong các trường_hợp sau
12	4	giữa nam và nữ
12	4	lây qua đường tình_dục
11	4	cả nam và nữ
11	4	bệnh lây qua đường
11	4	mang thai ngoài ý_muốn
11	4	càng sớm càng tốt
11	4	hẹp bao quy đầu
10	4	bạn sẽ có những
10	4	các mối quan_hệ xã_hội
10	4	có_thể đạt cực khoái
10	4	nào đi tới biển
10	4	nang lông bã nhờn

Rubrique (50) Explorer la sexualité (suite et fin)

F	L	Segment
47	5	bệnh lây_truyền qua đường tình_dục
27	5	các bệnh lây_truyền qua đường
20	5	smileydrink [x 5]
19	5	thì bị phạt tù từ
16	5	smileybiggrin [x 5]
16	5	smileylaugh [x 5]
14	5	smileycool [x 5]
14	5	smileyeusapray [x 5]
10	5	smileyphtqr [x 5]
24	6	các bệnh lây_truyền qua đường tình_dục
19	6	smileydrink [x 6]
13	6	Phạm_tội thuộc một trong các trường_hợp
13	6	smileyeusapray [x 6]
11	6	smileylaugh [x 6]
11	6	smileycool [x 6]
18	7	smileydrink [x 7]
12	7	smileyeusapray [x 7]
17	8	smileydrink [x 8]
11	8	Phạm_tội thuộc một trong các trường_hợp sau đây
11	8	smileyeusapray [x 8]
16	9	smileydrink [x 9]
10	9	smileyeusapray [x 9]
15	10	smileydrink [x 10]
10	10	Gây tổn_hại cho sức_khoẻ của nạn_nhân mà tỷ_lệ thương_tật từ
14	11	smileydrink [x 11]

Tab. F.6 : Rubrique (57) Etre Gay

F	L	Segment	Traduction
1056	2	những người	
779	2	người đồng_tính	
686	2	của mình	
672	2	là một	
483	2	một người	
404	2	đồng giới	
397	2	không phải	
391	2	không có	
374	2	mọi người	
329	2	là người	
285	2	đó là	
283	2	quan_hệ tình_dục	
279	2	với những	
275	2	người khác	
267	2	đồng_tính nam	
260	2	phải là	
241	2	là những	
239	2	có một	
235	2	rất nhiều	
231	2	có những	
228	2	người đàn_ông	
228	2	của người	
226	2	của họ	
219	2	không biết	
202	2	chỉ là	
199	2	cũng là	
196	2	cho rằng	
190	2	những gì	
187	2	người có	
187	2	nhiều người	
186	2	với nhau	
186	2	đã có	
184	2	mình là	
183	2	của tôi	
181	2	của những	
179	2	cũng không	
178	2	các bạn	
171	2	nhiễm HIV	
169	2	cũng có	
167	2	của một	

Rubrique (57) Etre Gay (suite)

F	L	Segment	Traduction
215	3	những người đồng_tính	
208	3	không phải là	
131	3	người đồng_tính nam	
120	3	tình_dục đồng giới	
108	3	với những người	
99	3	là người đồng_tính	
90	3	của những người	
82	3	là một người	
78	3	qua đường tình_dục	
71	3	một người đàn_ông	
64	3	cho những người	
61	3	là những người	
60	3	người cùng giới	
58	3	những gì mà	
58	3	những người có	
57	3	của người đồng_tính	
55	3	những người đàn_ông	
55	3	chứ không phải	
50	3	người dị tính	
50	3	phải là một	
49	3	mình là người	
47	3	có quan_hệ tình_dục	
47	3	quan_hệ tình_dục với	
46	3	bị nhiễm HIV	
46	3	dị tính luyến_ái	
46	3	luyến_ái đồng giới	
45	3	những người bạn	
45	3	cũng là một	
45	3	lây_truyền qua đường	
44	3	tất_cả mọi người	
43	3	người đồng giới	
42	3	chỉ là một	
41	3	những người khác	
41	3	một trong những	
41	3	xu_hướng tình_dục của	
40	3	là một trong	
40	3	của một người	
39	3	với người khác	
39	3	tình_dục của mình	
38	3	một cái gì	
38	3	ĐTLA và LTLA	

Rubrique (57) Etre Gay (suite)

F	L	Segment	Traduction
47	4	quan_hệ tình_dục đồng giới	
45	4	không phải là một	
43	4	lây_truyền qua đường tình_dục	
40	4	những người đồng_tính nam	
33	4	của những người đồng_tính	
32	4	một cái gì đó	
31	4	bệnh lây_truyền qua đường	
29	4	mình là người đồng_tính	
28	4	chứ không phải là	
27	4	là một trong những	
26	4	người ĐTLA và LTLA	
26	4	xu_hướng tình_dục của mình	
24	4	với người cùng giới	
20	4	những gì mà mình	
20	4	với những người đồng_tính	
20	4	cho những người đồng_tính	
18	4	là một người đàn_ông	
17	4	những người dị tính	
17	4	quan_hệ tình_dục đường miệng	
16	4	người dị tính luyến_ái	
16	4	đã chỉ ra rằng	
16	4	tình_dục đồng giới nam	
16	4	quan_hệ tình_dục qua hậu_môn	
16	4	định_kiến và phân_biệt đối_xử	
16	4	smileyeusapray [x 4]	
15	4	những người đồng giới	
15	4	bởi quản_trị viên	
15	4	dành cho người đồng_tính	
14	4	quan_hệ tình_dục không an_toàn	
14	4	nhóm nam quan_hệ tình_dục	
14	4	tháng … năm …	
14	4	lây qua đường tình_dục	
13	4	có quan_hệ tình_dục với	
13	4	với người khác giới	
13	4	tình_dục qua đường hậu_môn	
13	4	tình_dục với người cùng	
13	4	quan_hệ tình_dục với người	
12	4	về xu_hướng tình_dục của	
12	4	đồng_tính không phải là	
12	4	Người_yêu bạn thật lòng	

Rubrique (57) Etre Gay (suite et fin)

F	L	Segment	Traduction
29	5	bệnh lây_truyền qua đường tình_dục	
26	5	các bệnh lây_truyền qua đường	
15	5	smileyeusapray [x 5]	
13	5	có quan_hệ tình_dục đồng giới	
13	5	… tháng … năm …	
12	5	bệnh lây qua đường tình_dục	
11	5	tình_dục với người cùng giới	
11	5	nam quan_hệ tình_dục đồng giới	
11	5	Người_yêu bạn thật lòng sẽ	
10	5	về xu_hướng tình_dục của mình	
24	6	các bệnh lây_truyền qua đường tình_dục	
14	6	smileyeusapray [x 6]	
10	6	nhóm nam quan_hệ tình_dục đồng giới	
13	7	Nếu bạn thật_sự yêu một người thì	
13	7	smileyeusapray [x 7]	
10	7	ngày … tháng … năm …	
12	8	smileyeusapray [x 8]	
11	9	smileyeusapray [x 9]	
10	10	smileyeusapray [x 10]	

Rubriques au profil informel (62, 66, 68)

F	L	Segment	Traduction
5077	2	có nguy_cơ	comporter/présenter un risque
3177	2	em có	j'ai / ai-je
2936	2	không có	ne pas avoir / il n'y a pas
2350	2	cho em	permet.s/tez-moi de
2195	2	của bạn	ton / ta / tes
2144	2	mọi người	vous tous
2093	2	nhiễm HIV	contaminé.e au VIH
1875	2	của em	mon / ma / mes
1791	2	đi xét_nghiệm	faire un dépistage
1751	2	cô ấy	elle / cette femme
1552	2	các bạn	les amis
1539	2	bị nhiễm	contaminé.e
1478	2	của mình	mon / ma / mes
1470	2	có bị	[affirmation] + [élément négatif]
1437	2	mình có	j'ai / ai-je
1413	2	các anh	les grands frères
1369	2	cô ta	elle / cette femme
1359	2	3 tháng	3 mois
1343	2	quan_hệ với	rapport avec
1283	2	bạn không	tu + [négation]
1248	2	không biết	ne pas savoir
1202	2	sau đó	ensuite / après ça
1108	2	sau khi	après avoir
1106	2	ko có	ne pas avoir / il n'y a pas
1100	2	cho mình	permet.s/tez-moi de
1041	2	với GMD	avec une prostituée
1015	2	có dùng	avoir utilisé
1013	2	uống thuốc	prendre un médicament
1000	2	6 tuần	6 semaines
987	2	thì bạn	alors tu
959	2	bạn có	tu as
956	2	e có	j'ai / ai-je
929	2	em hỏi	je demande
915	2	phơi nhiễm	exposition
909	2	hay không	ou pas
881	2	đi xn	faire un dépistage
867	2	không cần	pas besoin
861	2	em không	je + [négation]
858	2	em đã	je + [passé]
848	2	thì em	alors je
(...)			

Rubriques au profil informel (62, 66, 68) (suite)

F	L	Segment	Traduction
1439	3	không có nguy_cơ	ne pas (re)présenter de risque
703	3	cho em hỏi	permet.s/tez-moi de demander
651	3	bị nhiễm HIV	contaminé.e au VIH
511	3	có quan_hệ với	avoir un rapport avec
462	3	bạn không có	tu n'as pas
456	3	quan_hệ với GMD	rapport avec une prostituée
340	3	thuốc phơi nhiễm	traitement post-exposition
336	3	có dùng BCS	avoir utilisé un préservatif
331	3	người nhiễm HIV	personne séropositive
325	3	ko có nguy_cơ	ne pas (re)présenter de risque
318	3	có nguy_cơ lây_nhiễm	présenter un risque d'être contaminé.e
318	3	giúp em với	aide(z)-moi à propos de
316	3	k có nguy_cơ	ne pas (re)présenter de risque
314	3	Bạn không có	Tu n'as pas
307	3	có bị nhiễm	être contaminé.e
306	3	sau 3 tháng	après 3 mois
289	3	có sử_dụng BCS	avoir utilisé un préservatif
282	3	có nguy_cơ nhiễm	présenter un risque d'être contaminé.e
282	3	em có bị	je suis / suis-je +[élément négatif]
282	3	em có nguy_cơ	j'ai / ai-je un risque
272	3	trong trường_hợp này	dans ce cas
271	3	của em có	mon / ma / mes [...] présente(nt)
265	3	em có quan_hệ	j'ai eu un rapport
261	3	của bạn là	ton / ta / tes [...] est / sont
260	3	Không có nguy_cơ	Il n'y a pas de risque
259	3	tư_vấn giúp em	aide(z)-moi / conseille(z)-moi
253	3	em có đi	je suis allé.e
238	3	trường_hợp của em	mon cas
236	3	qua đường tình_dục	par voie sexuelle
215	3	Trường_hợp của bạn	Ton cas
213	3	không cần dùng	pas besoin d'utiliser
213	3	nên đi xét_nghiệm	devoir faire un dépistage
211	3	anh cho em	permet.s/tez-moi de
211	3	trường_hợp của bạn	ton cas
207	3	cho mình hỏi	permet.s/tez-moi de demander
205	3	không phải là	il est faux que
204	3	là không có	cela ne représente pas
200	3	của cô ấy	son / sa / ses (de cette femme)
198	3	có phải là	est-ce vrai que
194	3	sau đó em	ensuite je
(...)			

Rubriques au profil informel (62, 66, 68) (suite)

F	L	Segment	Traduction
388	4	bạn không có nguy_cơ	tu n'as pas de risque
299	4	Bạn không có nguy_cơ	Tu n'as pas de risque
198	4	có quan_hệ với GMD	avoir un rapport avec une prostituée
198	4	em có quan_hệ với	j'ai eu un rapport avec
161	4	thuốc chống phơi nhiễm	traitement post-exposition
145	4	có bị nhiễm HIV	être contaminé.e au VIH
138	4	anh cho em hỏi	permet.s/tez-moi de demander
133	4	dùng thuốc phơi nhiễm	prendre un traitement post-exposition
123	4	lây qua đường tình_dục	transmis par voie sexuelle
122	4	Bạn k có nguy_cơ	Tu n'as pas de risque
110	4	có nguy_cơ nhiễm HIV	présenter un risque d'être contaminé.e au VIH
104	4	là không có nguy_cơ	cela ne représente pas un risque
98	4	những gì bạn kể	ce que tu dis / as dit
94	4	cho em lời khuyên	donne(z)-moi un conseil
92	4	của bạn không có	ton / ta /tes .. n'a/ont pas
92	4	cho em hỏi là	permet.s/tez-moi de demander si
88	4	thì không có nguy_cơ	alors il n'y a pas de risque
87	4	từ ngày có nguy_cơ	depuis le jour de la prise de risque
86	4	quan_hệ với gái mại_dâm	rapport avec une prostituée
83	4	nói lên điều gì	ne rien dire
82	4	có quan_hệ với 1	avoir un rapport avec 1
82	4	mình có quan_hệ với	j'ai eu un rapport avec
80	4	bị nhiễm HIV không	contaminé.e au VIH ?
80	4	trong đó có HIV	dont le VIH
80	4	uống thuốc phơi nhiễm	prendre un traitement post-exposition
79	4	tránh xa các dịch_vụ	éviter les services
77	4	ăn bánh trả tiền	sexe tarifé
(...)			

Rubriques au profil informel (62, 66, 68) (suite)

F	L	Segment	Traduction
81	5	của bạn không có nguy_cơ	ton / ta /tes [...] ne présent(ent) pas de risque
80	5	em có quan_hệ với GMD	j'ai eu un rapport avec une prostituée
68	5	bệnh lây qua đường tình_dục	maladie(s) sexuellement transmissible(s)
56	5	uống thuốc chống phơi nhiễm	prendre un traitement post-exposition
56	5	6 tuần và 3 tháng	6 semaines et 3 mois
54	5	kể từ ngày có nguy_cơ	compter à partir de la prise de risque
53	5	các bệnh lây qua đường	les maladie(s) transmissible(s) par voie
50	5	các anh cho em hỏi	permettez-moi de demander
49	5	có sử_dụng BCS là an_toàn	as utilisé un préservatif alors c'est sans risque
48	5	Trường_hợp bạn không có nguy_cơ	Ton cas ne présente pas de risque
47	5	tình_dục trong đó có HIV	sexuelle(s) dont le VIH
46	5	có bị nhiễm HIV không	suis-je contaminé au VIH
44	5	có quan_hệ với gái mại_dâm	avoir un rapport avec une prostituée
43	5	nên tránh xa các dịch_vụ	devoir éviter les services
42	5	có nguy_cơ trong trường_hợp này	ce cas présente un risque
42	5	không nói lên điều gì	ne rien dire
42	5	với GMD có dùng BCS	protégé avec une prostituée
42	5	Các anh cho em hỏi	Permettez-moi de demander
40	5	của em có cao không	mon / ma / mes [...] est/sont-il/elle(s) élevé.e(s)
40	5	smileylaugh [x 5]	[émoticône qui rit] [x 5]
(...)			

Rubriques au profil informel (62, 66, 68) (suite)

F	L	Segment	Traduction
48	6	các bệnh lây qua đường tình_dục	les maladies sexuellement transmissibles
34	6	các bệnh tình_dục trong đó có	les maladies sexuelles dont
34	6	Trường_hợp của bạn không có nguy_cơ	Ton cas ne présente pas de risque
31	6	các bệnh lây_truyền qua đường tình_dục	les maladies sexuellement transmissibles
31	6	bệnh tình_dục trong đó có HIV	maladies sexuelles dont le VIH
29	6	nữa để k mang thêm lo_lắng	ne plus [...] pour ne plus s'inquiéter
29	6	chính bạn tự bảo_vệ cho mình	toi seul.e assures ta protection
26	6	bệnh truyền_nhiễm lây qua đường tình_dục	maladie(s) sexuellement transmissible(s)
25	6	bạn nên tránh xa các dịch_vụ	tu dois éviter les services
25	6	Những khó_khăn của người nhiễm HIV	Les difficultés des séropositifs
25	6	Ai sẽ là người nhiễm HIV	Qui sera séropositif
24	6	gì bạn kể không có nguy_cơ	ce que tu dis ne présente pas de risque
23	6	có bị nhiễm HIV hay không	contaminé au VIH ?
23	6	QHTD có sử_dụng BCS là an_toàn	rapport protégé est sans risque
22	6	ngày kể từ ngày có nguy_cơ	jour(s) à compter du jour de la prise de risque
22	6	can_thiệp giảm tác_hại trong dự_phòng lây_nhiễm	réduire les dégâts dans la prévention de l'infection
21	6	không có nguy_cơ nên không cần	ne pas présenter de risque donc ne pas nécessiter
21	6	không có nguy_cơ trong trường_hợp này	il n'y a pas de risque dans ce cas
(...)			

F	L	Segment	Traduction
29	7	các bệnh tình_dục trong đó có HIV	les maladies sexuelles dont le VIH
22	7	các bệnh truyền_nhiễm lây qua đường tình_dục	les maladies sexuellement transmissibles
21	7	là chính bạn tự bảo_vệ cho mình	c'est toi seul.e qui assures ta protection
17	7	tìm gmd nữa để k mang thêm	ne plus aller voir une prostituée pour ne plus ressentir
17	7	can_thiệp giảm tác_hại trong dự_phòng lây_nhiễm HIV	réduire les dégâts dans la prévention de l'infection au VIH
16	7	chuẩn của Việt_Nam về thời_gian xét_nghiệm HIV	norme vietnamienne sur la période de dépistage du VIH
15	7	xa các dịch_vụ ăn bánh trả tiền	loin des services sexuels tarifés
15	7	gmd nữa để k mang thêm lo_lắng	ne plus [...] prostituée pour ne plus être inquiet
15	7	Những gì bạn kể không có nguy_cơ	Ce que tu racontes ne représente pas un risque
15	7	smileyphtqr [x 7]	[émoticône qui applaudit] [x 7]
14	7	tìm GMD nữa để k mang thêm	ne pas retourner chercher une prostituée pour ne plus ressentir de
13	7	để biết rõ tình_trạng sức_khỏe của mình	pour connaître clairement son propre état de santé
13	7	GMD nữa để k mang thêm lo_lắng	ne plus [...] prostituée pour ne plus s'inquiéter
13	7	các bệnh truyền_nhiễm lây qua đường tình_dục	les maladies sexuellement transmissibles
12	7	và các bệnh lây qua đường tình_dục	et les maladies sexuellement transmissibles
12	7	các bệnh lây qua đường tình_dục khác	les autres maladies sexuellement transmissibles
12	7	được hướng_dẫn dùng thuốc phơi nhiễm HIV	se faire prescrire un traitement post-exposition au VIH
12	7	để được hướng_dẫn dùng thuốc phơi nhiễm	pour se faire prescrire un traitement post-exposition
(...)			

Rubriques au profil informel (62, 66, 68) (suite)

F	L	Segment	Traduction
17	8	ĂN BÁNH TRẢ TIỀN là chính bạn tự	SEXE TARIFÉ c'est toi seul.e
16	8	Tránh tìm gmd nữa để k mang thêm	Évite de retourner chercher une prostituée pour ne plus ressentir de
15	8	các dịch_vụ ĂN BÁNH TRẢ TIỀN là	les services SEXUELS TARIFÉS c'est
15	8	Tài_liệu chuẩn của Việt_Nam về thời_gian xét_nghiệm HIV	Document normatif vietnamien sur le délai de dépistage du VIH
14	8	tránh xa các dịch_vụ ăn bánh trả tiền	éviter les services sexuels tarifés
14	8	smileyphtqr [x 8]	[émoticône qui applaudit] [x 8]
14	8	Cố lên 6 thang là okie lắm rồi	Courage pendant 6 mois et tout sera ok !
13	8	Tránh tìm GMD nữa để k mang thêm	Évite de retourner chercher une prostituée pour ne plus ressentir de
12	8	cũng_như các bệnh truyền_nhiễm lây qua đường tình_dục	tout comme les maladies sexuellement transmissibles
11	8	lây_nhiễm các bệnh tình_dục trong đó có HIV	attraper les maladies sexuelles dont le VIH
10	8	các bệnh về tình_dục trong đó có HIV	les maladies sexuelles dont le VIH
10	8	để được hướng_dẫn dùng thuốc phơi nhiễm HIV	pour se faire prescrire un traitement post-exposition au VIH
10	8	smileyeusaboohoo [x 8]	[émoticône qui pleure] [x 8]
(...)			

Rubriques au profil informel (62, 66, 68) (suite et fin)

F	L	Segment	Traduction
13	9	Tránh tìm gmd nữa để k mang thêm lo_lắng	Évite de retourner chercher une prostituée pour ne plus t'inquiéter
13	9	smileyphtqr [x 9]	[émoticône qui applaudit] [x 9]
12	9	Tránh tìm GMD nữa để k mang thêm lo_lắng	Évite de retourner chercher une prostituée pour ne plus t'inquiéter
12	10	smileyphtqr [x 10]	[émoticône qui applaudit] [x 10]
14	11	nếu kq bên phường chưa chắc cú thì đi qua bệnh	si ton centre local n'est pas sûr, vas dans un
13	11	không phải lo_lắng về việc lây_nhiễm HIV cũng_như các bệnh truyền_nhiễm	devoir ne pas s'inquiéter à propos d'une contamination au VIH ou aux maladies transmissibles
13	11	Loại dịch Lượng HIV Máu Rất nhiều Dịch tiết âm_đạo Nhiều	Quantité de VIH par type de fluide. Sang : élevée - Cyprine : élevée
17	11	ĂN BÁNH TRẢ TIỀN là chính bạn tự bảo_vệ cho mình	SEXE TARIFÉ est la meilleure manière de te protéger
15	11	các dịch_vụ ĂN BÁNH TRẢ TIỀN là chính bạn tự bảo_vệ	les services SEXUELS TARIFÉS est la meilleure manière de te protéger
11	11	smileyphtqr [x 11]	[émoticône qui applaudit] [x 11]
10	11	tránh các dịch_vụ ĂN BÁNH TRẢ TIỀN là chính bạn	éviter les services SEXUELS TARIFÉS est la meilleure manière de te

Rubriques au profil institutionnel (22, 50, 57)

F	L	Segment	Traduction
1501	2	những người	les personnes
1294	2	là một	est un.e
1276	2	của mình	à soi / mon /ma / mes
960	2	không có	ne pas avoir / il n'y a pas
856	2	quan_hệ tình_dục	rapport(s) sexuel(s)
798	2	người đồng_tính	personne homosexuelle
766	2	không phải	c'est faux
717	2	một người	une personne
601	2	mọi người	chaque personne
596	2	đó là	ceci est
593	2	bạn sẽ	tu/vous [+futur]
588	2	của bạn	ton / ta / tes
563	2	tránh thai	contraception / contraceptif
561	2	rất nhiều	de très nombreu.ses/x
556	2	có những	(y) avoir des
509	2	phải là	vrai que
502	2	người đàn_ông	les hommes
499	2	người khác	les autres
499	2	qua đường	par voie
496	2	là người	est une personne
489	2	có một	avoir un.e
477	2	là những	sont des
477	2	với những	avec des
466	2	của người	d'une / des personne(s)
466	2	sau khi	après avoir
440	2	các bạn	les amis
437	2	cũng là	être aussi
432	2	không biết	ne pas savoir
428	2	đồng giới	homosexuel.le
413	2	bạn tình	partenaire sexuel.le
408	2	cho rằng	déclarer
401	2	đã có	avoir eu
395	2	nhiều người	de nombreuses personnes
393	2	mà không	mais [négatif]
385	2	cũng có	avoir aussi
377	2	đường tình_dục	voie sexuelle
375	2	chỉ là	être seulement
369	2	cũng không	non plus
369	2	trước khi	avant de
367	2	làm cho	rendre / conduire / faire que
(...)			

Rubriques au profil institutionnel (22, 50, 57) (suite)

F	L	Segment	Traduction
365	3	không phải là	il est faux que
364	3	qua đường tình_dục	par voie sexuelle
221	3	những người đồng_tính	les personnes homosexuelles
214	3	lây_truyền qua đường	transmissible(s) par voie sexuelle
168	3	thuốc tránh thai	médicament contraceptif / pilule
162	3	bệnh lây_truyền qua	maladie(s) transmise(s) par voie
151	3	với những người	avec / envers les personnes
137	3	có quan_hệ tình_dục	avoir un /des rapport(s) sexuel(s)
134	3	người đồng_tính nam	gay / homme homosexuel
131	3	tình_dục đồng giới	sexe homosexuel
125	3	là một trong	être un parmi
125	3	của những người	des personnes
125	3	một trong những	un parmi
121	3	" yêu "	(litt. : "s'aimer") avoir un rapport
120	3	smileybiggrin [x 3]	[émoticône qui rit fort] [x 3]
113	3	một người đàn_ông	un homme
113	3	biện_pháp tránh thai	moyen(s) contraceptif
107	3	quan_hệ tình_dục với	rapport(s) sexuel(s) avec
107	3	chứ không phải	et non pas
105	3	là một người	est une personne
102	3	là người đồng_tính	être homosexuel.le
100	3	có rất nhiều	avoir de nombreu.ses/x
99	3	là những người	sont des personnes
94	3	các bệnh lây_truyền	les maladies transmises
93	3	những người có	des / les personnes (qui) ont
91	3	cũng là một	est aussi un.e
89	3	được coi là	être appelé / dénommé
86	3	khi quan_hệ tình_dục	pendant un / les rapport(s) sexuel(s)
86	3	bao quy đầu	prépuce
85	3	" chuyện ấy	(litt. "cette histoire) rapport sexuel
84	3	chuyện ấy "	(litt. cette histoire") rapport sexuel
84	3	lây qua đường	transmis par voie
83	3	những người đàn_ông	les / certains hommes
81	3	cho những người	pour les / certaines personnes
78	3	không có gì	il n'y a pas de quoi / ne rien avoir
77	3	trong đó có	dont
76	3	chỉ là một	être seulement un
74	3	phải là một	vrai que c'est un
74	3	bị nhiễm HIV	contaminé.e au VIH
74	3	nam và nữ	hommes et femmes
(...)			

Rubriques au profil institutionnel (22, 50, 57) (suite)

F	L	Segment	Traduction
193	4	lây_truyền qua đường tình_dục	transmis par voie sexuelle
161	4	bệnh lây_truyền qua đường	maladie(s) transmise(s) par voie
91	4	là một trong những	être un.e parmi
82	4	“ chuyện ấy ”	(litt. ”cette histoire”) rapport sexuel
71	4	lây qua đường tình_dục	transmis par voie sexuelle
63	4	smileybiggrin [x 4]	[émoticône qui rit fort] [x 4]
61	4	không phải là một	n’est pas un.e
56	4	bệnh lây qua đường	maladie(s) transmise(s) par voie
50	4	“ cậu nhỏ ”	(litt. : ”petit gars”) sexe masculin
50	4	quan_hệ tình_dục đồng giới	rapport(s) homosexuel(s)
50	4	chứ không phải là	et non pas
43	4	smileydrink [x 4]	[émoticône qui boit] [x 4]
40	4	những người đồng_tính nam	les / des hommes homosexuels
39	4	quan_hệ tình_dục không an_toàn	rapport(s) sexuel(s) non protégé(s)
38	4	smileylaugh [x 4]	[émoticône qui rit] [x 4]
35	4	của những người đồng_tính	des personnes homosexuelles
33	4	một cái gì đó	quelque chose
33	4	các bệnh lây qua	les maladies transmises par voie
31	4	smileyeusapray [x 4]	[émoticône qui prie] [x 4]
30	4	các biện_pháp tránh thai	les moyens contraceptifs
30	4	quan_hệ tình_dục với người	rapport(s) sexuel(s) avec une / des personne(s)
29	4	có quan_hệ tình_dục với	avoir un / des rapport(s) sexuel(s) avec
29	4	mình là người đồng_tính	je suis homosexuel.le
29	4	dùng thuốc tránh thai	prendre un médicament contraceptif / la pilule
28	4	là một người đàn_ông	est un homme
28	4	cũng không phải là	ce n’est pas non plus
28	4	cả nam và nữ	(les) hommes aussi bien que (les) femmes
28	4	qua đường tình_dục và	par voie sexuelle et
27	4	với người cùng giới	avec une personne du même sexe
27	4	“ tự sướng ”	(litt. : ”s’auto-contenter”) se masturber / masturba
26	4	người ĐTLA và LTLA	personne(s) homosexuel.le(s) et bisexuel.le(s)
26	4	đã chỉ ra rằng	avoir montré que
26	4	xu_hướng tình_dục của mình	sa propre / son orientation sexuelle
25	4	tình_dục qua đường hậu_môn	sexuel.le par voie anale
25	4	lây_nhiễm qua đường tình_dục	transmis.e(s) par voie sexuelle
(...)			

F	L	Segment	Traduction
155	5	bệnh lây_truyền qua đường tình_dục	maladie(s) sexuellement transmissible(s)
92	5	các bệnh lây_truyền qua đường	les maladies transmises par voie
54	5	bệnh lây qua đường tình_dục	maladie(s) sexuellement transmissible(s)
40	5	smileydrink [x 5]	[émoticône qui boit] [x 5]
33	5	smileybiggrin [x 5]	[émoticône qui rit fort] [x 5]
32	5	các bệnh lây qua đường	les maladies transmises par voie
29	5	smileyeusapray [x 5]	[émoticône qui prie] [x 5]
23	5	thì bị phạt tù từ	encour(en)t une peine d'emprisonnement de
22	5	smileylaugh [x 5]	[émoticône qui rit] [x 5]
18	5	làm " chuyện ấy "	(litt. : faire "cette histoire") avoir un rapport
16	5	cũng là một trong những	est aussi un.e parmi
15	5	có quan_hệ tình_dục với người	avoir un / des rapport(s) sexuel(s) avec une personne
14	5	quan_hệ tình_dục qua đường hậu_môn	rapport(s) sexuel(s) par voie anale
14	5	smileycool [x 5]	[émoticône tranquille] [x 5]
13	5	có quan_hệ tình_dục đồng giới	avoir un /des rapport(s) homosexuel(s)
13	5	… tháng … năm …	telle date
13	5	bệnh lây_lan qua đường tình_dục	maladie(s) sexuellement transmissible(s)
13	5	bệnh lây_nhiễm qua đường tình_dục	maladie(s) sexuellement transmissible(s)
12	5	tình_dục với người cùng giới	sexe avec une personne du même sexe
12	5	qua đường tình_dục và HIV	par voie sexuelle et le VIH
12	5	nạn_nhân mà tỷ_lệ thương_tật từ	la/les victime(s) avec un taux d'infirmité de
11	5	là một trong những nguyên_nhân	est une des causes
11	5	nam quan_hệ tình_dục đồng giới	homme rapport(s) homosexuel(s)

Rubriques au profil institutionnel (22, 50, 57) (suite)

F	L	Segment	Traduction
11	5	nhiễm_trùng lây_truyền qua đường TD	infection(s) sexuellement transmissible(s)
11	5	smileyphtqr [x 5]	[émoticône qui applaudit] [x 5]
11	5	Người_yêu bạn thật lòng sẽ	Celui/Celle qui t'aime de tout son coeur + [futur]
10	5	trong " chuyện ấy "	(litt. pendant "cette histoire") pendant le rapport
10	5	về xu_hướng tình_dục của mình	à propos de sa propre orientation sexuelle
10	5	trên bảo dưới không nghe	(litt. : le haut commande, le bas n'écoute pas) impuissance
10	5	Ngày_mai vẫn còn là bí_ẩn	Demain cela restera mystérieux
86	6	các bệnh lây_truyền qua đường tình_dục	les maladies sexuellement transmissibles
38	6	smileydrink [x 6]	[émoticône qui boit] [x 6]
30	6	các bệnh lây qua đường tình_dục	les maladies sexuellement transmissibles
27	6	smileyeusapray [x 6]	[émoticône qui prie] [x 6]
19	6	những bệnh lây_truyền qua đường tình_dục	les maladies sexuellement transmissibles
19	6	bệnh lây_truyền qua đường tình_dục và	maladie(s) sexuellement transmissible(s) et
17	6	smileybiggrin [x 6]	[émoticône qui rit fort] [x 6]
15	6	Phạm_tội thuộc một trong các trường_hợp	Cas de crime parmi
14	6	smileylaugh [x 6]	[émoticône qui rit] [x 6]
12	6	bệnh lây_truyền qua đường tình_dục khác	autre(s) maladie(s) sexuellement transmissible(s)
11	6	mắc bệnh lây_truyền qua đường tình_dục	attraper une maladie sexuellement transmissible
11	6	smileycool [x 6]	[émoticône tranquille] [x 6]
10	6	nhóm nam quan_hệ tình_dục đồng giới	groupe des hommes ayant des rapports homosexuels
10	6	ngày … tháng … năm …	telle date

F	L	Segment	Traduction
36	7	smileydrink [x 7]	[émoticône qui boit] [x 7]
25	7	smileyeusapray [x 7]	[émoticône qui prie] [x 7]
15	7	và các bệnh lây_truyền qua đường tình_dục	et les maladies sexuellement transmissibles
13	7	Nếu bạn thật_sự yêu một người thì	Si tu/vous aime.s/z vraiment une personne alors
11	7	bệnh lây_truyền qua đường tình_dục và HIV	maladie(s) sexuellement transmissible(s) et le VIH
11	7	mắc các bệnh lây_truyền qua đường tình_dục	attraper les maladies sexuellement transmissibles
10	7	các bệnh lây_truyền qua đường tình_dục và	les maladies sexuellement transmissibles et
34	8	smileydrink [x 8]	[émoticône qui boit] [x 8]
23	8	smileyeusapray [x 8]	[émoticône qui prie] [x 8]
13	8	Phạm_tội thuộc một trong các trường_hợp sau đây	Cas de crime parmi les suivants
32	9	smileydrink [x 9]	[émoticône qui boit] [x 9]
21	9	smileyeusapray [x 9]	[émoticône qui prie] [x 9]
10	9	Vậy tại_sao chúng_ta không sống hết_mình cho ngày hôm_nay	Alors pourquoi ne profitons-nous pas du moment présent
30	10	smileydrink [x 10]	[émoticône qui boit] [x 10]
19	10	smileyeusapray [x 10]	[émoticône qui prie] [x 10]
10	10	Gây tổn_hại cho sức_khoẻ của nạn_nhân mà tỷ_lệ thương_tật từ	Provoquer des préjudices sur la santé de la victime à un taux d'infirmité de
28	11	smileydrink [x 11]	[émoticône qui boit] [x 11]
17	11	smileyeusapray [x 11]	[émoticône qui prie] [x 11]

ANNEXE G

Calcul du niveau de personnalisation des conversations foisonnantes

Sont énumérées ci-dessous les étapes de calcul du niveau de personnalisation des conversations foisonnantes.

1. Accéder à la liste des métadonnées par conversation, triée par nombre d'interventions (.../ProcessingFlow/CorpusXx/TXM/topics/metadata.csv à ouvrir dans un tableur, trier par `NBmsg`)
2. Etablir la liste des plus longues conversations (sélectionner les conversations de plus de 50 interventions dans le corpus vietnamien, de plus de 100 interventions dans le corpus français, les copier dans une nouvelle table)
3. ajouter à cette nouvelle table une colonne affichant le résultat du calcul suivant : `NBmsg` / `NBauthors`

ANNEXE H

Constitution des diagrammes de répartition des interventions

Sont énumérées ci-dessous les étapes de constitution des diagrammes de répartition des interventions d'une conversation par intervenant.

1. Copier depuis metadata.csv (posts) dans un tableur temporaire les colonnes `rang`,`author`,`authorid`,`date` de la conversation souhaitée.
2. Déplacer la colonne `rang` après la colonne `date`. Copier les deux colonnes, collage spécial dans la table finale : transposer en ligne (sous forme d'en-tête du tableau)
3. Soit n le nombre d'intervenants dans la conversation (indiqué sur la ligne de metadatas.csv (topics) correspondant à la conversation).
4. Copier et transposer en ligne dans une autre table temporaire la colonne `author`, la dupliquer n fois, copier le tout dans notepad++
5. Remplacer toutes les `\t` par des ,
6. Copier le premier auteur, sélectionner la première ligne, remplacer par 0 toutes les occurrences du premier auteur dans la ligne (cliquer sur l'option "Dans la sélection"), puis remplacer toutes les autres occurrences par rien (décocher l'option "Dans la sélection"), recopier l'auteur en début de ligne suivi d'une virgule (NE PAS OUBLIER LA VIRGULE).
7. Copier le deuxième auteur, sélectionner la deuxième ligne, remplacer par 1 toutes les occurrences du deuxième auteur dans la ligne (cliquer sur l'option "Dans la sélection"), puis remplacer toutes les autres occurrences par rien (décocher l'option "Dans la sélection").
8. Recommencer n fois l'étape 7. Supprimer toutes les lignes vides une fois terminé.
9. Insérer une colonne à gauche du tableau.
10. Recopier les données transformées dans la table finale, sous l'en-tête date-rang.
11. Insérer deux colonnes à droite de la première colonne. Dans la première nouvelle colonne (la 2ème du tableau), recopier les auhtorid à la main (n opérations).

12. Dans la 2ème nouvelle colonne (la 3ème du tableau), en première ligne après l'en-tête (soit 3ème ligne 3ème colonne du tableau) écrire Initiateur, en deuxième ligne Répondant 1, dupliquer cette deuxième cellule jusqu'à Répondant n-1.
13. Sélectionner toutes les lignes sauf les 2 lignes d'en-tête, Trier les données par la 2ème colonne, en ordre décroissant. Attention : les dizaines sont classées après 1 etc.
14. Sélectionner les données de la 2ème ligne-3ème colonne à la dernière cellule, (Configuration OpenOffice Calc) menu Insertion/Diagramme : diagramme dispersion, points seuls. Séries de données en lignes, première colonne comme étiquette. Légende à gauche, Titre et sous-titre (traduction du titre), titre de l'axe X : messages. Ne toucher à rien de plus et ajuster le diagramme visuellement une fois celui-ci généré (Axe Y : pas d'étiquette, intervalle de 1 – Axe X : Valeur max, intervalle,etc. - Légende : entrée en face des lignes - ...).

ANNEXE I

Calcul du nombre d'interventions par membre

Pour étudier les intervenants les plus prolifiques du corpus, il faut compter le nombre d'interventions de chacun et trier ces nombres par ordre décroissant. Le nombre d'interventions par intervenant est accessible dans le fichier metadata.csv du corpus segmenté par intervention (voir l'annexe E, où les colonnes exploitées ici sont résumées par la colonne *A* pour auteur. En réalité les colonnes du véritable fichier sont nommées *authorid* et *authorname*).

Dans Open Calc :

- ouvrir metadata.csv (par interventions) du corpus total
- sélectionner toute la colonne *authorname* (reproduire la procédure avec la colonne *authorid* si l'on veut conserver les identifiants associés)
- cliquer sur Données/Filtres/Filtre standard
  - *Nom de champ* : authorname (ou authorid)
  - *Condition* : =
  - *Valeur* : .*
  - *Plus d'options*, cocher :
    - *Caractère générique*
    - *La plage contient des étiquettes de colonne*
    - *Sans doublons*
    - *Copier le résultat vers...* une colonne vide du tableur.

  On obtient une colonne avec une ligne par intervenant. (reproduire la procédure avec la colonne *authorid* si l'on veut conserver les identifiants associés)
- Dans une nouvelle colonne écrire la formule

$$= NB.SI(\$X\$2 : \$X\$m; Yn)$$

Cette formule indique que pour la cellule Yn (Y étant la colonne obtenue par le calcul précédent, donc Yn un des intervenants du corpus) on

veut compter le nombre de contenus identiques dans la zone du tableau allant de la cellule X2 à la cellule Xm (ici, la colonne initiale des intervenants). Le calcul est plus rapide et plus fiable sur des chiffres que sur des caractères alphanumériques, donc privilégier la colonne authorid si on l'a calculée.

- Il faut ensuite reproduire la formule pour chaque ligne (en double-cliquant sur le petit carré en bas à droite de la cellule sélectionnée).

$$= NB.SI(\$X\$2 : \$X\$m; Yn + 1)$$

$$= NB.SI(\$X\$2 : \$X\$m; Yn + 2)$$

etc.

- Pour finir, trier les colonnes obtenues par nombre décroissant d'interventions, pour accéder aux profils les plus prolifiques en haut de liste. Si la colonne authorname a été utilisée plutôt que la colonne authorid, vérifier en fin de liste les éventuelles erreurs de calcul (dues à des caractères alphanumériques problématiques, et indiquées comme ayant o occurrences).

ANNEXE J

Qualification des discours à l'aide des expressions figées (EF)

Cette annexe présente l'utilisation des expressions figées, leur longueur et leur fréquence, dans le processus de qualification des discours des forums, en institutionnels ou informels. Les étapes sont d'abord listées puis la procédure est illustrée avec un exemple.

1. (Lexico) Calcul des segments répétés ;
2. Sélection humaine des EF pertinentes ;
3. (Lexico) établissement de la liste des variantes lexicales de chaque EF ;
4. pour chaque EF :
   - constitution d'une expression régulière incluant toutes ses variantes ;
   - (TXM) concordance à partir de l'expression régulière constituée ;
   - (TXM) étude en plein texte, en vue de l'affectation du trait institutionnel ou informel ;
   - qualification en DIR ou DIS de textes contenant l'EF par application de son trait institutionnel/informel.

Prenons un exemple avec l'ensemble des variantes lexicales désignant les *maladies sexuellement transmissibles* (MST ou STD en anglais) :

Un premier parcours de l'ensemble des segments répétés permet d'établir la liste suivante :

các / những	bệnh	lây_truyền	qua đường tình_dục
		lay	qua đường tình_dục
		lây	qua dg tình_dục
		lây_nhiễm	qua đường tình_dục
		lây_lan	qua đường tình_dục
		LTQDTD	
		LTQĐTD	
		LTTD	
		STDs	

L'expression régulière constituée afin de rechercher dans TXM cette expression figée et ses variantes est la suivante :

```
[word="các|những"][word="bệnh"](([word="lây_truyền|lây|lây_nhiễm|
 lây_lan"][word="qua"][word="đường|dg"][word="tình_dục|TD"])|
 ([word="LTQDTD|LTQĐTD|LTTD|STDs"]))
```

Tous ces segments commencent par *các bệnh* ou *những bệnh* (« les maladies »). Nous pouvons donc avoir de nouveau recours à la liste des segments répétés de Lexico afin d'accéder à un aperçu direct des segments les plus fréquents. Pour cela, nous faisons un premier tri alphabétique de la liste totale des segments répétés (quelle que soit leur longueur), afin de sélectionner uniquement ceux commençant par *các / những bệnh*. Puis nous trions les segments sélectionnés par fréquence. Dans notre exemple :

Fréq.	Segment répété
109	các bệnh tình_dục
89	các bệnh lây_truyền qua đường tình_dục
62	các bệnh lây qua đường tình_dục
59	các bệnh khác
30	các bệnh truyền_nhiễm lây qua đường tình_dục
30	các bệnh STDs
29	các bệnh STD
26	các bệnh về tình_dục
15	những bệnh lây_truyền qua đường tình_dục
13	các bệnh nguy_hiểm
11	các bệnh qua đường tình_dục

Cette approche nous permet d'en savoir plus sur les pratiques discursives les plus adoptées dans le forum concernant la désignation des MST.

- La formule *các bệnh tình_dục* (« les maladies sexuelles ») n'avait pas été trouvée par le premier repérage manuel, alors que c'est la formule la plus largement utilisée.
- C'est également le cas de *các bệnh về tình_dục* (« les maladies en lien avec la sexualité »).
- D'autres segments très fréquemment utilisés ont été repérés mais ils sont trop génériques pour être retenus comme variantes désignant les MST (*các bệnh khác* (« les autres maladies ») ou *các bệnh nguy_hiểm* (« les maladies dangereuses ») notamment).
- Nous remarquons d'autre part que deux termes que nous avions considérés comme interchangeables sur l'axe paradigmatique apparaissent l'un à la suite de l'autre : *truyền_nhiễm* (« transmettre/transmissible ») et *lây* (« attraper/attrapable »). Le corpus compte 30 occurrences de *các bệnh truyền_nhiễm lây qua đường tình_dục* (« les maladies qui se transmettent et s'attrapent par voie sexuelle »), il nous faudra explorer en contexte le type de discours ou le profil des intervenants ayant recours à cette formule.

Pour intégrer ces précisions, il nous faut modifier l'expression régulière de recherche dans TXM.

```
[word="các|những"][word="bệnh"](((((word="truyền_nhiễm"]|[])
([word="lây_truyền|lây|lây_nhiễm|lây_lan"]|[])(([word="qua"]
[word="đường|dg"])|[]))|[word="về"])[word="tình_dục|TD"])|
[word="LTQDTD|LTQĐTD|LTTD|STDs|STD"])
```

A présent que l'expression régulière est établie, nous pouvons étudier les contextes en utilisant l'outil de concordance de TXM.

Nous avons fait l'hypothèse[1] que les segments répétés les plus longs étaient plus caractéristiques du discours institutionnel rapporté (DIR) et que les segments répétés les plus fréquents étaient plus caractéristiques du discours infomel spontané (DIS). Voyons si ce fait se vérifie dans le cas des MST :

Fréq.	Segment répété	Lg
109	các bệnh tình_dục	3
11	các bệnh qua đường tình_dục	5
89	các bệnh lây_truyền qua đường tình_dục	6
13	các bệnh nguy_hiểm	3

Le segment court *các bệnh tình_dục* et ses variantes de même longueur *các bệnh TD* et *những bệnh tình_dục* totalisent 139 occurrences, dont 115 dans la rubrique la plus informelle : (62) Relations sexuelles, préservatif, lubrifiant. Ce segment, d'autant plus qu'il est relativement court, devrait donc être caractéristique du discours informel. Pourtant, les intervenants qui utilisent ces formules sont de profil expert[2] Cette expression figée doit donc rester du côté du discours institutionnalisant des conseils d'experts.

Un segment plus long et moins fréquent, *các bệnh qua đường tình_dục* devrait logiquement se situer du côté du discours institutionnel. Il est principalement utilisé par des experts (6 sur 10 intervenants), mais pas exclusivement.

1. cf. p.134
2. cf le paragraphe 3.4.2, p.138.

ANNEXE K

Listes de formes par champs lexicaux du genre témoignage

Les listes sont constituées humainement à partir du lexique des textes repérés comme appartenant au genre du témoignage, puis complétées par un parcours du dictionnaire et de la liste des segments répétés, calculés à l'aide de Lexico. La fonction *Groupe de forme* est ensuite utilisée pour constituer une liste par champ lexical, enregistrée au format .lst

Adresse aux internautes			
Xin	xin	Xin chào	Chào
ạ	ơi		
anh chị em	Các anh chị em ơi	các ACE	
anh / chị	Các anh / chị	các anh các chị	các anh chị
mấy anh chị	mấy anh	mấy bác	các bác
mọi người	Mọi người	Các bạn	các bạn
tất cả các bạn	tất cả các thành viên		tất cả mọi người
trong diễn đàn	tham gia diễn đàn		
có một chuyện	chuyện	Tình hình là	tình hình là
thưa với	là như này	trường hợp	thắc mắc
cần	Xin cho	hỏi	muốn hỏi
ý kiến	tư vấn	tư vấn cho	lời khuyên
vui lòng tư vấn	giải đá	hãy cho	hãy
giúp	giúp ơn	giúp ơn với	
dùm	coi giùm	Liệu	Làm ơn
mong	Mong	mong nhận được	Mong hồi âm sớm từ
XIn cảm ơn	Cám ơn	Cảm ơn rất nhiều	Thân

Mesure du temps					
Khi	khi	Lúc	lúc	Sau	sau
lúc đó	sau đó	trước	trước khi	trong khi	trong lúc
từ	đến	Cách đây	Cho đến jo	Hiện giờ	hiện tại
Thời gian	thời gian	lâu	ngắn	gần	gần đây
rồi	chưa	đã	lần	lúc xong việc	khi xong
Khoảng	khoảng	12h	21h30	1,5	30
chiều	tối	đêm	hôm	ngày	ngày sau
phút	giờ	tuần	tháng		

Préservatif
BCS
bao
bao cao su
bcs
áo mưa
durex
baocaosu

Alcool - Perte de contrôle			
rượu	bia	nhậu	đi nhậu
say	say xỉn	quá chén	bí tỉ
trận nhậu	vài trận nhậu tơi bời		
uống rượu say	đi uống vài chai bia		
không	làm chủ	(được)	bản thân
ko		(đc)	mình
k		(dc)	hành vi
kô		(dược)	hành động
koo			cảm xúc
không thể			
rủ			
với thằng bạn		thằng bạn	
mấy ông làm ở cty		mấy ỏng	
có việc đi xa			

Prostitution				
gái mại dâm				cà phê ôm
GMD	gái gọi	đụng vào GMD		cafe ôm
gmd	làng chơi	kiếm GMD		cf ôm
Gmd	bán dâm	quan hệ với GMD		Karaoke ôm
GmD	tiếp viên	đi dịch vụ		karaoke ôm
GMd	Tiếp viên	đi chơi GMD		nhà nghỉ
cô GMD	cô ấy	đi chơi ″ gái ″		bia ôm
				có gmd
				cà fe gội đầu
				gội đầu maxsa
				nhà nghỉ quan hệ
			(quán)	nhà nghỉ qhtd
				matxa
				matsxa
				massage
				massa
				masa
				masage
				masager
				mát sa
				mát xa
				mx
				tẩm quất

Pratiques sexuelles				
quan hệ	oral sex	HJ	dương vật	miệng
quan hệ tình dục	OS	handjob	âm đạo	kích thích
QHTD	ory	hj	DV	xuất tinh
qh	os	hand job	AD	tinh dịch
QH	Oral sex		dv	nước bọt
qhtd	thổi kèn		ad	hôn
Quan hệ	ngậm		Dương vật	môi
Quan hệ tình dục	mút		Dv	hôn môi
	blow( )job		Âm đạo	lưỡi
	BJ			mồm
	bj			tháo
				lần đầu tiên
				làm chuyện ấy

Angoisse
canh cánh trong lòng
lo
lo lắng
lo quá
lo lắm
rất lo
rất lo lắng
rất lo láng
đang rất lo lắng
lo lắng quá
lo lắng lắm
rất là lo sợ
sợ liệu
sợ
sợ lắm
ko thể ngủ
mất ăn mất ngủ
ảnh hưởng rất nhiều đến công việc
hiz !!
hix hix
hic hic
hu hu
cực kỳ ân hận
thề
từ nay tới già sẽ ko … nữa !
không bao giờ … nữa
Tởn đến già luôn
xin thề cạch tới già
thề là cạch đến già !
Bị stress nặng
bj stress quá
bị stress quá
chết
ko thể nào em đuổi suy nghĩ đấy ra khỏi đầu được
cầu giời phù hộ cho là
cầu giời phù hộ
k dám

Symptômes	
triệu chứng	đi ngoài
cảm giác	bị loét
đau nhức	bi phồng lên
đau buốt	bị bỏng
buốt	vỡ ra
đau	đi ngoài liên tục
bị đau	khớp
bị xước	đốt sống
rát	thắt lưng
ngứa	đau nhức
chảy máu	nổi hạch
nổi	vã mồ hôi đêm
bị sốt	không ngủ được
vã mồ hôi đêm	ăn uống
tăng cân	sức khỏe
ăn uống	vết xước
ăn ngủ	đi tiểu
ốm	cơ thể
gầy	khắp người
khỏi	họng
bình thường	hạch
bt	dương vật
ổn định	chân
phát hiện	tay
đau bụng	bụng
thấy	gốc
có mủ	đầu
bị rối loạn	ở gốc
tiêu hóa	ở đầu
đau bụng	ở tai phải

ANNEXE L

Lexique vietnamien

Tab. L.1 : Lexique vietnamien de linguistique et TAL

Linguistique et TAL	
học máy thống kê	apprentissage automatique/machine
thuật toán	algorithme
ranh giới từ	frontière de mots
phân tích	analyse
cú pháp	syntaxe
phân tích cú pháp	analyse syntaxique
công cụ	outil
gán nhãn	étiqueter/-tage
từ loại	catégorie grammaticale (POS)
công cụ gán nhãn từ loại	outil d'étiquetage morpho-syntaxique
bộ gán nhãn từ loại	étiquetage morpho-syntaxique, POS tagging
đồng tham chiếu	co-référence
nhập nhằng	ambigu(-ïté)
cấu trúc	structure
cụm từ	phrase

Tab. L.2 : Lexique vietnamien du web

Internet	
máy điện toán	ordinateur
nhu liệu	logiciel
cập nhật	mise à jour
máy chủ	serveur
lên mạng	aller sur Internet (litt. monter sur le réseau)
cư dân mạng	internautes (litt. : communauté du/en réseau)
dạo quanh một vòng các trang mạng	faire un tour sur la toile
trang mạng	page web, site
trang báo	blog
diễn đàn / DD	forum
mạng xã hội	réseau social
các mạng lưới xã hội	les réseaux sociaux
tiểu blog	micro-blogging (twitter)
đăng ký	s'inscrire
đăng nhập	se connecter
đăng xuất	se déconnecter
mật khẩu	mot de passe
tiểu sử	profil (information de membre)
đăng	publier
biểu tượng	smiley (icône, symbole)
nhắn tin	sms / message
bình luận	commenter/-taire
trích dẫn	citer
lick chuột vào	cliquer sur
bàn bạc	délibérer, discuter
người điều hành	administrateur
ngoài đời	dans la vie réelle

Tab. L.3 : Lexique extrait du corpus vietnamien

terme/expression	traduction
Famille	
bậc phụ huynh	parents
phu quân / phu nhân	mari / femme
gia đình	famille
ba mẹ	parents
chồng / vợ	mari / femme

Tab. L.3 : Lexique extrait du corpus vietnamien (suite)

terme/expression	traduction
Relations amoureuses / couple	
vợ chồng	couple (litt. femme mari)
hẹn hò với	sortir avec
cuộc hẹn hò	rendez-vous, aventure, relation
mối tính	flirt
Hẹn Tốc Độ	Speed Dating
sinh tình cảm với	tomber amoureux de, avoir des sentiments pour
ngã vào	tomber, craquer pour / tomber amoureux de
nhay cảm	sensible (physiquement ou émotionellement)
lãng mạn	romantique
mối tình kiểu như hàn quốc	(litt. amour de type coréen) amour-haine ?
một bờ vai để tựa vào mỗi khi mỏi mệt	une épaule sur laquelle s'appuyer lorsqu'on est fatigué
bị chồng phản bội	trompée par son mari
bạo hành gia đình	violence domestique
thuốc an thần	anti-dépresseurs
tan vỡ	casser, rompre, arrêter une relation
đơn phương	divorce, séparation
li dị	divorcer (de)
chửi	insulter, dénigrer
non kém	immature
về đòi vợ tha thứ	demander pardon à sa femme
người bạn đời của mình	son/sa compagnon/-gne
cuộc sống gia đình của họ trở nên nồng ấm và tràn đầy hạnh phúc hơn	leur vie de couple est devenue plus chaleureuse et remplie de bonheur
Reproduction	
khi thai sổ	pendant la grossesse
rụng trứng	ovulation (litt. chute de l'oeuf)
triệt sản	stérilis(er/ation)
vô sinh	stérile
dậy thì	puberté
tuổi mới lớn	adolescence
thụ tinh trong ống nghiệm	FIV
mang thai hộ	GPA

Tab. L.3 : Lexique extrait du corpus vietnamien (suite)

terme/expression	traduction
	Corps
cơ thể	corps
thân hình	formes, corps
mông	fesses, cul
bìu	scrotum, bourse
dương vật (DV)	pénis
âm vật	clitoris, parfois sexe féminin
âm đạo	vagin
âm hộ	vulve
môi / mép	lèvre(s) / lèvres comme un tout qui fait bordure de qqch
(cổ) tử cung	(col de) l'utérus
bộ phận sinh dục	parties génitales
dịch tiết âm đạo / dịch âm_đạo / dịch âm hộ	secrétions vaginales
tinh dịch	sperme
tinh trùng	spermatozoïde(s)
tinh hoàn	testicule(s)
mào tinh hoàn	épididyme
nước ối	liquide amniotique
nước bọt	salive
mồ hôi	transpiration
nước mắt	larmes
nước tiểu	urine
máu (chảy máu)	sang (saigner)
mủ (mưng mủ)	pus (suppurer)
núm vú	seins
niệu đạo	urètre
bao quy đầu	prépuce
tinh hoàn	testicule
mào tinh hoàn	épididyme
diểm G	point G
ngực	poitrine
phẫu thuật tạo lại hình dáng	chirurgie esthétique
dịch vụ thẩm mĩ	soins de beauté
nâng ngực	augmentation mammaire
làm hẹp âm đạo	rétrécissement du vagin
đường tiết niệu	voies urinaires
cơ vòng	sphincter

Tab. L.3 : Lexique extrait du corpus vietnamien (suite)

terme/expression	traduction
Pratiques sexuelles	
cuộc sống chăn gối	vie sexuelle
chuyện ấy	sexe, rapports sexuels (litt. cette histoire)
chuyện phòng the	sexualité (litt. histoire(s) de chambre, ce qu'il se passe dans la chambre)
chuyện giường chiếu	idem (litt. : histoire(s) de lit le soir)
việc phòng sự	idem (litt. histoire(s) de chambre)
nếm trái cấm	goûter le fruit défendu
đéo	baiser / aller se faire foutre
ham muốn tình dục	libido
tương sinh tương khắc	lois d'engendrement et destruction
kích dục	aphrodisiaque
vuốt ve	caresser
tư thế sinh hoạt	position(s) sexuelle(s)
tư thế quan hệ	idem
màn dạo đầu	préliminaires
lén lút	en cachette
ém	dissimulé, sous silence
phẩm đồi trụy	pornographie (littéralement : œuvre sale)
quan hệ tình dục	relations sexuelles
khẩu dâm	rapport (litt. échange) sexuel
các hành vi tình dục	pratiques sexuelles
bạn tình	partenaire sexuel
đối tác	partenaire
thủ dâm	masturbation
đạt được	atteindre
khoái cảm	plaisir
khoái cảm / cục khoái	orgasme
đạt (được) khoái cảm / đạt tới đỉnh	atteindre l'orgasme
lên đỉnh	atteindre l'orgasme (litt. monter au sommet)
ái ân	contentement, plaisir
tiếng thét nho nhỏ	petits cris
rên xiết	gémir
thích thú	jouir
bằng tay / qua đường miệng / hậu môn	manuel.le / oral.e / anal.e
có rung	vibrant
rung	vibrer/-ation

Tab. L.3 : Lexique extrait du corpus vietnamien (suite)

terme/expression	traduction
Pratiques sexuelles (suite)	
nam giới vào tình trạng bị động	(litt. : l'homme passe en situation passive) l'homme se laisse faire (masturbation, fellation ?)
cương cứng	érection
vùng kích dục (trên cơ thể)	zone(s) érogène(s)
vuùng nhay cảm trên cơ thể	zone(s) sensible(s)
không thể "làm ăn" gì được	ne pas arriver à (litt. faire du business)
mất hứng	perdre l'excitation
xuất tinh (sớm)	éjaculation (précoce)
hoang mang	panique
xâm nhập	pénétrer/-tration
cọ xát	frottement, friction
MST	
bệnh lây qua đường tình dục	maladie(s) sexuellement transmissible(s)
bệnh nhiễm trùng	infection
nhiễm trùng	s'infecter/-tion
(vi) khuẩn	bactérie
xâm nhập	intrusion
lây truyền	transmettre/-ission
giải độc	désintoxiquer/cation
ủ bệnh	incubation
kháng_thể	anticorps
cấp tính	aigu
biểu hiện	manifestation (donc ici symptôme)
triệu chứng	symptôme
hội chứng	symptôme
xử lý	traitement
âm tính	négatif
dương_tính	positif
dương_tính với HIV	(séro)positif au VIH
hệ miễn dịch	système immunitaire
nhiễm trùng cấp	infection aigüe
sơ nhiễm	primo-infection
hạch to	lymphadénopathie
bị ngứa	ressentir des démangeaisons
viêm nang lông	folliculite
lậu cầu / bệnh lậu	blennorragie (gonorrhée, chaude-pisse)

Tab. L.3 : Lexique extrait du corpus vietnamien (suite)

terme/expression	traduction
MST (suite)	
khiến	attenter à, faire du mal
đau rát	douloureux
tổn thương	blessure
dễ bị tổn thương	fragile
viêm	inflammation
tiêu chảy	diarrhée
ung thư	cancer
sốt thương hàn	fièvre typhoïde
HPV	papillomavirus
VGB (Viêm gan B)	hépatite B, appelée aussi VHB
kháng thể	anticorps
Préservatif	
bao cao su	préservatif (litt. sac en latex)
"bao"	capote
loại bao cao su không có núm tròn	modèle de préservatif sans réservoir
vòng cuốn	anneau
được bôi trơn	lubrifié
thành phần chất bôi	lubrifiant
chất bôi trơn tan trong nước	lubrifiant hydrosoluble
dùng "áo mưa"	se protéger (utiliser un préservatif) (litt. utiliser un vêtement de pluie)
QHTD có dùng BCS (quan hệ tình dục có dùng bao cao su)	rapport(s) sexuel(s) protégé(s)
tháo bcs	retirer le préservatif
tránh thai	contraceptif
dự phòng	prévenir (préventif/prévention)
Prostitution	
ân ái	sexe
chuyên ân ái	professionnel du sexe
mại dâm	prostitution
bán dâm	prostitution
MSW (male sex worker)	prostitué
NBD (người bán dâm) di động	prostituée mobile (dans la rue, les parcs)
GMD (gái mại dâm)	prostituée
hooker	prostituée
wor	prostituée

Tab. L.3 : Lexique extrait du corpus vietnamien (suite)

terme/expression	traduction
Prostitution (suite)	
làng chơi	prostituée
tiếp viện	hôtesse
gái gọi	escort (litt. fille qu'on appelle)
mát xa / matxa / massa	massage
tẩm quất	massage
trang web đen	sites porno
LGBT	
quan hệ đồng giới	relation/rapport homosexuel.le
(miêu tả) xu hướng tình dục	(définir son) orientation sexuelle
cổng cao đẳng	le milieu homosexuel
tgt3 (thế giới thứ ba)	milieu homosexuel / homosexualité (litt. troisième monde)
đông tính nam	homosexuel, gay
đông tính nữ	homosexuelle, lesbienne
dị tính	hétérosexuel.le
lưỡng tính	bisexuel.le
boy bi	bisexuel
MSM (men who have sex with men)	homosexuel, gay
bóng	homosexuel, gay
bóng kín	homo dans le placard, non "outé"
bóng lộ	extraverti, folle, ou trans MTF
ẻo lả	effeminé
bê đê	pédé
thằng đồng dâm	l'homo (péj.)
Transexualité	
nhận dạng	identité
giới tính	sexe (masculin/féminin)
giống	genre
Nam chuyển giới thành nữ (đã qua phẫu thuật chuyển đổi giới tính)	trans MTF (passée par l'opération de réassignation sexuelle)
Nữ chuyển giới thành nam	trans FTM
Nam giới ăn mặc và có cử chỉ nữ giới (chưa qua phẫu thuật chuyển đổi giới tính)	Travesti, trans MTF (n'étant pas passée par l'opération de réassignation sexuelle)
bộ phận sinh dục	organes génitaux
nhiễm sắc thể	chromosome

Tab. L.3 : Lexique extrait du corpus vietnamien (suite)

terme/expression	traduction
LGBT (suite)	
Acceptation / Discrimination	
chấp nhận	accepter
thấu hiểu	comprendre
cởi mở	ouvert (d'esprit)
vị tha	altruiste
chấp nhận	accepter
công nhận hôn nhân đồng giới	reconnaissance du mariage entre personnes du même sexe
kì thị, sự kỳ thị	stigmatisation / discrimination
phân biệt đối xử	discriminer
phân biệt giới tính	sexiste
xì căng đan	scandale
kẻ khả nghi	le suspect
cô đơn	esseulé
em lại đang bị nghi là gay	je suis soupçonné d'être gay
chưa có can đảm nói sự thật ấy	pas encore eu le courage de le dire

ANNEXE M

Liste des émoticônes

Dans le corpus vietnamien, les émoticônes sont remplacées par une combinaison du préfixe `smiley` et du nom de leur image sur le forum.

Dans le corpus français, elles sont remplacées par `EMOT_titre_ICONE`, où titre est celui de l'image sur le forum.

Tab. M.1 : Émoticônes du forum vietnamien

étiquette	image	titre	description
flapper	msp_flapper.gif	Flapper	tire la langue
	tongue.gif		tire la langue
=d>	eusa_clap.gif	Applause	applaudit
	ph34r.gif		applaudit
[thumbup]	msp_thumbup.gif	ThumpUp	lève le pouce et lunettes de soleil
[lol]	msp_lol.gif	LOL	se renverse de rire, se roule de rire
[love]	msp_love.gif	Love	coeurs dans les yeux
:d/	eusa_dance.gif	Dancing	danse indexes levés
[-o<	eusa_pray.gif	Pray	prie
[-x	eusa_naughty.gif	Shame on you	paas bien, non du doigt
#-o	eusa_doh.gif	d'oh !	se tape le front (?)
:-"	eusa_whistle.gif	Whistle	siffle
:-#	eusa_silenced.gif	Silenced	bouche cousue
:-\$	eusa_shhh.gif	Shhh	chut
:^o	eusa_liar.gif	Liar	menteur
:-k	eusa_think.gif	Think	perplexe
:-s	eusa_eh.gif	Eh ?	surpris
[-(	eusa_snooty.gif	Not talking	refus
[angry]	msp_angry.gif	Angry	en colère
[biggrin]	msp_biggrin.gif	BigGrin	grand sourire
[blink]	msp_blink.gif	Blink	hallucine

Tab. M.1 : Émoticônes du forum vietnamien (suite)

étiquette	image	titre	description
[blush]	msp_blushing.gif	Blushing	rougit
[bored]	msp_bored.gif	Bored	saoulé, fatigué
[confused]	msp_confused.gif	Confused	incompréhension
[cool]	msp_cool.gif	Cool	lunettes de soleil
	cool.gif		lunettes de soleil
[crying]	msp_crying.gif	Crying	yeux pleins de larmes
[cursing]	msp_cursing.gif	Cursing	pete un câble
[drool]	msp_drool.gif	Drool	bave
[glare]	msp_wink.gif	Glare	clin d'oeil ?
[glare]	msp_wink.gif	Wink	clin d'oeil ?
[huh]	msp_huh.gif	Huh	surpris
[laugh]	msp_laugh.gif	Laugh	rit
[mad]	msp_mad.gif	Mad	en colère
[mellow]	msp_mellow.gif	Mellow	neutre
	mellow.gif		ange coquin
[omg]	msp_ohmy.gif	OhMyGod	crie
	ohmy.gif		bouche bée
[razz]	msp_razz.gif	Razz	langue qui pend
[rolleyes]	msp_rolleyes.gif	RollEyes	lève les yeux au ciel
[sad]	msp_sad.gif	Sad	triste
[scared]	msp_scared.gif	Scared	effrayé, dégouté
[sleep]	msp_sleep.gif	Sleep	ferme les yeux
[smile]	msp_smile.gif	Smile	sourit
	happy.gif		sourit
	smile.gif		sourit
[sneaky]	msp_sneaky.gif	Sneaky	fâché, regard de travers
[thumbdn]	msp_thumbdn.gif	ThumbDown	baisse le pouce
[tongue]	msp_tongue.gif	Tongue	tire la langue
[unsure]	msp_unsure.gif	Unsure	hésite, incertain
[woot]	msp_woot.gif	Woot	agréablement surpris
[wub]	msp_wub.gif	Wub	amoureux
](*,)	eusa_wall.gif	Brick wall	tête contre le mur
=;	eusa_hand.gif	Speak to the hand	bloque de la main
=p	eusa_drool.gif	Drool	bave
8-[	eusa_shifty.gif	Anxious	anxieux
o :)	eusa_angel.gif	Angel	auréole, angélique
	wbyim10.gif		bisou ?
	wbyim16.gif		diable
	dry.gif		croise les bras, ignore

Tab. M.2 : Émoticônes du forum français

code	image
EMOT_1_ICONE	/images/smilies/Forum65.gif
EMOT_54_ICONE	/images/smilies/54.gif
EMOT_applaud_ICONE	/images/smilies/Applaudissements02.gif
EMOT_applaud01_ICONE	/images/smilies/Applaudissements04.gif
EMOT_applaud02_ICONE	/images/smilies/MDR59.gif
EMOT_arc_ICONE	/images/smilies/Amour18.gif
EMOT_aspirateur_ICONE	/images/smilies/aspirateur.gif
EMOT_atonna10_ICONE	/images/smilies/Grrrr25.gif
EMOT_barman_ICONE	/images/smilies/barman.gif
EMOT_brancard1_ICONE	/images/smilies/brancard1.gif
EMOT_bravo_raison_ICONE	/images/smilies/bravo_raison.gif
EMOT_bravo_ICONE	/images/smilies/Super21.gif
EMOT_bricolo_ICONE	/images/smilies/bricoleur02.gif
EMOT_cachesouslachaise_ICONE	/images/smilies/cachesouslachaise.gif
EMOT_caddy_courses07_ICONE	/images/smilies/caddy_courses07.gif
EMOT_cadeau07_ICONE	/images/smilies/cadeau07.gif
EMOT_camp_ICONE	/images/smilies/Inclassable66.gif
EMOT_carre_etonne_ICONE	/images/smilies/carre_etonne.gif
EMOT_champagne_ICONE	/images/smilies/champa07e.gif
EMOT_chapeau_ICONE	/images/smilies/Chapeau1.gif
EMOT_chapeau01_ICONE	/images/smilies/Chapeau2.gif
EMOT_cheval_baton_ICONE	/images/smilies/cheval_baton.gif
EMOT_chinois_ICONE	/images/smilies/Chinois.gif
EMOT_civiere_snoopy_ICONE	/images/smilies/civiere_snoopy.gif
EMOT_clin_d10_ICONE	/images/smilies/icon_e_surprised.gif
EMOT_coeur10_ICONE	/images/smilies/icon_cool.gif
EMOT_colere10_ICONE	/images/smilies/Inclassable27.gif
EMOT_conten10_ICONE	/images/smilies/Panneau32.gif
EMOT_coucou_ICONE	/images/smilies/Coucou07.gif
EMOT_course_ICONE	/images/smilies/FaireSesCourses.gif
EMOT_danse_ICONE	/images/smilies/Danseurvert.gif
EMOT_deguisee_ICONE	/images/smilies/deguisee.gif
EMOT_dodo3_ICONE	/images/smilies/Dodo30.gif
EMOT_dodo_ICONE	/images/smilies/Dodo01.gif
EMOT_dodo01_ICONE	/images/smilies/Dodo08.gif
EMOT_dodo02_ICONE	/images/smilies/Dodo13.gif
EMOT_dodo03_ICONE	/images/smilies/Dodo42.gif
EMOT_facteur07b_ICONE	/images/smilies/facteur07b.gif
EMOT_faim_ICONE	/images/smilies/Faim00.gif

Tab. M.2 : Émoticônes du forum français (suite)

code	image
EMOT_faim01_ICONE	/images/smilies/Faim02.gif
EMOT_faim02_ICONE	/images/smilies/Faim32.gif
EMOT_faim03_ICONE	/images/smilies/Faim35.gif
EMOT_fatigu10_ICONE	/images/smilies/Grrrr06.gif
EMOT_fete_ICONE	/images/smilies/Fete02.gif
EMOT_feuilles07_ICONE	/images/smilies/feuilles07.gif
EMOT_fleur10_ICONE	/images/smilies/Hein85.gif
EMOT_fleurs_ICONE	/images/smilies/image002.gif
EMOT_fou10_ICONE	/images/smilies/Hopital23.gif
EMOT_fouetter_ICONE	/images/smilies/fouetter.gif
EMOT_fumeo1_ICONE	/images/smilies/Fumeur07.gif
EMOT_fumeo2_ICONE	/images/smilies/Fumeur08.gif
EMOT_fumeo3_ICONE	/images/smilies/Fumeur13.gif
EMOT_fumeo4_ICONE	/images/smilies/Fumeur05.gif
EMOT_gay_flag_ICONE	/images/smilies/Gay_flagflagGF.gif
EMOT_grr_black_ICONE	/images/smilies/scorn.gif
EMOT_grrrr04_ICONE	/images/smilies/Grrrr27.gif
EMOT_grrrr05_ICONE	/images/smilies/Grrrr28.gif
EMOT_grrrr06_ICONE	/images/smilies/Mecontent18.gif
EMOT_grrrr06_ICONE	/images/smilies/Mecontent18.gif
EMOT_grrrr07_ICONE	/images/smilies/Mecontent20.gif
EMOT_gyrophare2_ICONE	/images/smilies/gyrophare2.gif
EMOT_heino1_ICONE	/images/smilies/Hein15.gif
EMOT_heino2_ICONE	/images/smilies/Hein25.gif
EMOT_heino3_ICONE	/images/smilies/Hein40.gif
EMOT_heino34_ICONE	/images/smilies/Hein42.gif
EMOT_heino5_ICONE	/images/smilies/Hein38.gif
EMOT_hein38_ICONE	/images/smilies/Hein38.gif
EMOT_hello_ICONE	/images/smilies/Sourire65.gif
EMOT_hosto_ICONE	/images/smilies/hopital15.gif
EMOT_hum_violet_ICONE	/images/smilies/hum_violet.gif
EMOT_infirmier_ICONE	/images/smilies/Infirmier.gif
EMOT_jessie-james_ICONE	/images/smilies/jessie-james.gif
EMOT_kangourou_ICONE	/images/smilies/kangourou.gif
EMOT_kisso2_ICONE	/images/smilies/338.gif
EMOT_kitty_dance_ICONE	/images/smilies/sAni_kittydance.gif
EMOT_langue_ICONE	/images/smilies/Langue02.gif
EMOT_langue10_ICONE	/images/smilies/image001.gif
EMOT_lapinpaq_ICONE	/images/smilies/lapinpaq.gif
EMOT_licorne_mini_ICONE	/images/smilies/licorne_mini.gif

Tab. M.2 : Émoticônes du forum français (suite)

code	image
EMOT_love_ICONE	/images/smilies/Amour03.gif
EMOT_love01_ICONE	/images/smilies/Amour21.gif
EMOT_love03_ICONE	/images/smilies/image006.gif
EMOT_malade_ICONE	/images/smilies/114.gif
EMOT_mdr_fille_ICONE	/images/smilies/mdr_fille.gif
EMOT_mdr_ICONE	/images/smilies/mdr182.gif
EMOT_mdr03_ICONE	/images/smilies/MDR99.gif
EMOT_mdr05_ICONE	/images/smilies/MDR55.gif
EMOT_merci_ICONE	/images/smilies/Super19.gif
EMOT_monstre_ICONE	/images/smilies/monstre.gif
EMOT_music03_ICONE	/images/smilies/Musique63.gif
EMOT_non01_ICONE	/images/smilies/NonNon01.gif
EMOT_nul_ICONE	/images/smilies/SuperNul05.gif
EMOT_ok_5sur5_ICONE	/images/smilies/ok_5sur5.gif
EMOT_ok_dixsurdix_ICONE	/images/smilies/ok_dixsurdix.gif
EMOT_ok_ICONE	/images/smilies/Accord01.gif
EMOT_onsetel_ICONE	/images/smilies/onsetel.gif
EMOT_optic_ICONE	/images/smilies/Opticien.gif
EMOT_panda_coeur_ICONE	/images/smilies/panda_coeur.gif
EMOT_panda_lol_ICONE	/images/smilies/panda_lol.gif
EMOT_peur_ICONE	/images/smilies/Peureux05.gif
EMOT_peur01_ICONE	/images/smilies/Peureux08.gif
EMOT_peur02_ICONE	/images/smilies/Peureux02.gif
EMOT_peur03_ICONE	/images/smilies/Peureux04.gif
EMOT_peur10_ICONE	/images/smilies/image016.gif
EMOT_peur210_ICONE	/images/smilies/image013.gif
EMOT_pierre_feuilleciseaux_ICONE	/images/smilies/pierre_feuilleciseaux.gif
EMOT_pirogue_ICONE	/images/smilies/pirogue.gif
EMOT_pleure10_ICONE	/images/smilies/image011.gif
EMOT_plonger_ICONE	/images/smilies/plonger.gif
EMOT_plongeur_ICONE	/images/smilies/plongeur.gif
EMOT_pompes_ICONE	/images/smilies/Pompes01.gif
EMOT_poubelle_ICONE	/images/smilies/poubelle.gif
EMOT_pourqwa_ICONE	/images/smilies/pourqwa.gif
EMOT_poussiere_ICONE	/images/smilies/poussiere.gif
EMOT_priere_ciel_ICONE	/images/smilies/priere_ciel.gif
EMOT_prof_ICONE	/images/smilies/prof.gif
EMOT_psy_ICONE	/images/smilies/psychanalyste.gif
EMOT_renard_ICONE	/images/smilies/renard.gif
EMOT_saut_cheshire_ICONE	/images/smilies/saut_cheshire.gif

Tab. M.2 : Émoticônes du forum français (suite)

code	image
EMOT_saut_lapin_ICONE	/images/smilies/saut_lapin.gif
EMOT_sauto8_ICONE	/images/smilies/saut08.gif
EMOT_serpent_mordlakeu_ICONE	/images/smilies/serpent_mordlakeu.gif
EMOT_sonic_ICONE	/images/smilies/sonic.gif
EMOT_sos_ICONE	/images/smilies/Coucou20.gif
EMOT_soso1_ICONE	/images/smilies/Inclassable16.gif
EMOT_sourire_ICONE	/images/smilies/Sourire45.gif
EMOT_souristourne._ICONE	/images/smilies/souristourne.gif
EMOT_sparadrap_nez_ICONE	/images/smilies/sparadrap_nez.gif
EMOT_succube_ICONE	/images/smilies/succube.gif
EMOT_sunshine_ICONE	/images/smilies/sunshine.gif
EMOT_supporter_corne_ICONE	/images/smilies/supporter_corne.gif
EMOT_surprise_fille_ICONE	/images/smilies/surprise_fille.gif
EMOT_tapedansboite_ICONE	/images/smilies/tapedansboite.gif
EMOT_tchino1_ICONE	/images/smilies/Alcooliques04.gif
EMOT_temps_mort_ICONE	/images/smilies/Temps_mort.gif
EMOT_temps_ICONE	/images/smilies/Temps01.gif
EMOT_tempso1_ICONE	/images/smilies/Temps02.gif
EMOT_tempso2_ICONE	/images/smilies/Temps04.gif
EMOT_tempso3_ICONE	/images/smilies/Temps18.gif
EMOT_terre_ICONE	/images/smilies/Inclassable69.gif
EMOT_titanic_ICONE	/images/smilies/Inclassable53.gif
EMOT_tortue_ICONE	/images/smilies/tortue.gif
EMOT_toubib_ICONE	/images/smilies/DocteurVieux.gif
EMOT_toubibo1_ICONE	/images/smilies/Docteur.gif
EMOT_tristeo1_ICONE	/images/smilies/Triste11.gif
EMOT_tristeo3_ICONE	/images/smilies/Triste20.gif
EMOT_urticaire_ICONE	/images/smilies/urticaire.gif
EMOT_vieux_ICONE	/images/smilies/Vieux.gif
EMOT_yes_ICONE	/images/smilies/Super11.gif
EMOT_yeso1_ICONE	/images/smilies/Super13.gif
EMOT_yeux_vrille_ICONE	/images/smilies/yeux_vrille.gif
EMOT_z_sms_interdit_ICONE	/images/smilies/z_sms_interdit.gif
EMOT_zen_ICONE	/images/smilies/MDR53.gif
EMOT_zen10_ICONE	/images/smilies/Forum07.gif
EMOT_zz_ICONE	/images/smilies/Bonhomme_gym.gif
EMOT_Choqué_ICONE	/images/smilies/icon_eek.gif
EMOT_Clindoeil_ICONE	/images/smilies/icon_e_wink.gif
EMOT_Confus_ICONE	/images/smilies/icon_e_confused.gif
EMOT_Cow_ICONE	/images/smilies/icon_e_geek.gif

Tab. M.2 : Émoticônes du forum français (suite)

code	image
EMOT_Diable_ICONE	/images/smilies/icon_evil.gif
EMOT_Embarrassé_ICONE	/images/smilies/icon_redface.gif
EMOT_M.Vert_ICONE	/images/smilies/icon_mrgreen.gif
EMOT_Neutre_ICONE	/images/smilies/icon_neutral.gif
EMOT_pause_ICONE	/images/smilies/tasse_cafe.gif
EMOT_Sourire_ICONE	/images/smilies/icon_e_smile.gif
EMOT_Tirelalangue_ICONE	/images/smilies/icon_razz.gif
EMOT_Trèscontent_ICONE	/images/smilies/icon_e_biggrin.gif
EMOT_Triste_ICONE	/images/smilies/icon_e_sad.gif

Glossaire

ADT analyse statistique des données textuelles. Exploration quantitative et statistique de productions langagières sous forme de données numériques. 56–58, 73, 107, 108, 110, 113, 197, 200, 296–298, 301

agora conversation foisonnante dont l'indice de focalisation est faible : plus il est faible plus le nombre d'intervenant·e·s dans la conversation est élevé. 149–151, 153, 156, 193, 296, *Par opposition à :* ECP

ALT alanine transaminase : enzyme présente dans le foie. Lorsque celui-ci est blessé, cette enzyme est relâchée dans le sang. Un test sanguin permet donc d'établir un diagnostic sur l'état du foie, en fonction du taux d'ALT. (source : Medlineplus.gov). 301

antirétroviral traitement médical agissant en perturbant la réplication du rétrovirus sur de nouvelles cellules. 17, 18, 74, 299, 301

avatar image choisie par chaque membre du forum pour représenter son identité de manière visuelle. Élément constitutif de son profil. 81, 102, 103, 297

canon terme emprunté à Sarfati (2008) désignant le discours de l'institution, le discours instituant le sens commun, et que nous situons dans un rapport de triangle discursif avec les deux autres variations du discours définies par Sarfati, que sont la vulgate et la doxa. 40–43, 197, 294, 299

CD4 cluster de différenciation 4. Le CD4 est une glycoprotéine exprimée à la surface des lymphocytes T4. C'est le récepteur utilisé par le VIH pour infecter ses cellules cibles (les lymphocytes T4, les monocytes et les macrophages).(source : Wikipedia). 297, 301

conversation niveau de structuration intermédaire des forums, dont la création est laissée au libre choix des internautes. Une *conversation* est ouverte au sein d'une rubrique par un membre du forum. Elle est composée au minimum d'un titre et d'une intervention initiative. Mais le nombre d'interventions qui la composent est théoriquement infini car elle-même est théoriquement toujours ouverte (sauf si un administrateur du forum décide de la fermer). 50, 52–54, 56, 80, 81, 83, 86, 95, 96, 102–106, 109, 111, 113, 115, 118, 120, 122–133, 135, 136, 139, 144, 145, 148–160, 164–166, 176, 177, 193, 196, 197, 207–210, 217, 293–298, 301, *Synonyme :* fil de discussion

conversation à intervention unique (Conv.IU) cas particulier de conversation, ne comportant qu'une seule intervention, aucune intervention réactive (le corpus étant temporellement borné). 54, 124, 126, 128–131, 145, 147, 159, 196, 197, 296, 301

conversation foisonnante conversation comptabilisant plus de n interventions, n pouvant varier d'un forum à l'autre. 145, 147–153, 156, 157, 192, 193, 196, 197, 261, 293, 295, 296, 309

cooccurrence présence statistiquement significative de deux formes dans la même fenêtre contextuelle. 56, 59

discours informel discours produit en dehors du cadre institutionnel, sur les réseaux sociaux numériques. Il s'agit d'énoncés natifs du web. Etant donné qu'ils sont produits dans le cadre de polylogues asynchrones médiés par ordinateurs, ils partagent certains traits avec le langage parlé et sont plus ou moins fortement la transcription d'une oralité. 3, 4, 6, 39, 44, 120, 130, 134, 135, 145, 147, 294, 296, 298, 301

discours informel spontané transcription du langage parlé sur le ton du dialogue oral. Type de discours présent sur les forums. Opposition à discours institutionnel rapporté. 6, 117, 129, 130, 132, 133, 135, 137, 138, 145, 147, 149, 166, 192–195, 197, 198, 294, 298, 301

discours institutionnel discours produit par les acteurs des institutions. Discours stable et normé. 3, 4, 6, 32–38, 42, 74, 75, 117, 128, 130, 131, 134, 136, 139, 144, 145, 192, 194–196, 200, 294, 296, 298, 301

discours institutionnel rapporté reproduction par *copié-collé* de textes institutionnels, depuis des sources extérieures au forum ou antérieures. Type de discours présent sur les forums. Opposition à discours informel spontané. 6, 117, 118, 128, 130, 132, 133, 135, 137, 138, 145, 147, 198, 294, 301

discours natif du web social (terme emprunté à Paveau, 2012b). Discours produit directement dans les environnements du web 2.0, sous forme textuelle (voire agrémentée d'éléments audiovisuels) et assortie de métadonnées, donc plurisémiotique. 200, 295, 297

doxa terme emprunté à Sarfati (2008) désignant le discours des acteurs en dehors de l'institution, qui ont intégré le sens commun institué par le canon et transmis par la vulgate et l'adaptent à leur réalité sociale pour produire leur discours. Nous situons la *doxa* dans un rapport de triangle discursif avec les deux autres variations du discours définies par Sarfati, que sont le canon et la vulgate. 40, 42, 43, 117, 198, 199, 293, 299

émoticône signe qui imite une émotion, afin de la rendre perceptible au cours de l'énonciation d'un contenu, dans le cadre d'une communication médiée par un ordinateur (Halté, 2013). les *émoticônes* participent donc au sens des productions discursives natives du web. Celles-ci prennent la forme soit d'une suite de caractères de ponctuation et alphabétiques, soit d'une icône, le plus souvent de la hauteur d'un interligne, insérée

dans le texte. Il est souvent fait référence aux *émoticônes* par leur équivalent anglais *smileys*. 54, 102, 103, 107, 124, 167, 168, 208, 285

énoncé natif du web (voir discours natif du web social). 47, 50, 294

espace de consultation personnalisée conversation foisonnante dont l'initiat·eur/rice est le/la principal intervenant·e. Conversation menée par son initiat·eur/rice, les autres intervenant·e·s venant répondre/réagir à ses interventions. 147, 149–158, 193, 197, 293, 295, 296, 301, *Par opposition à :* agora

expert numéro un du forum vietnamien intervenant·e expert·e comptabilisant plus de 3000 interventions dans le corpus, il n'y a que 3 intervenant·e·s qui dépassent le millier d'interventions dans le corpus. 139–141, 143, 144

expert·e (voir intervenant·e expert·e). 118

fenêtre sérologique intervalle temporel caractérisé par une grande incertitude, entre le moment d'un risque d'exposition au VIH et le résultat du dépistage par prise de sang (sérologie). 188, 190, 194, 198

fil de discussion (voir conversation). 104, 123

fléau social phénomène social qualifié de néfaste pour la société et que les institutions s'efforcent de combattre. Au Viêt Nam les fléaux sociaux désignent aussi bien la prostitution que la corruption, le proxénétisme, la toxicomanie, les jeux d'argent, les vols, également l'homosexualité. Cette liste a des frontières floues et la politique est plus ou moins répressive selon les époques et les phénomènes. Par exemple le VIH a fait partie de cette liste, mais les institutions ont évolué en vue de réduire les discriminations envers les personnes séropositives et n'emploient plus la qualification de fléau social concernant le VIH. Ce n'est pas le cas de la prostitution ou de la toxicomanie. 16, 19, 27–29, 166

forum de discussion plateforme web permettant des échanges discursifs entre internautes. Les discussions y sont archivées ce qui permet une communication asynchrone et une consultation ultérieure. La somme des discussions archivées constitue donc une base de connaissance enrichie en permanence et mise à disposition de tout internaute à tout moment. Cette mise en commun d'expériences individuelles fait des *forums* le lieu d'élaboration et de co-construction de savoirs experts. Etant donné que ce sont leurs utilisateurs qui sont les créateurs des contenus, les *forums* font partie du web 2.0. 2, 4–6, 32, 39, 44, 46, 47, 49–55, 73, 74, 76, 80–83, 85, 90, 91, 98, 99, 102–104, 106, 107, 113, 117–123, 125, 126, 129, 130, 133–135, 139, 141, 144, 145, 147, 167, 192, 193, 195–200, 207, 294, 297

indice de focalisation nombre d'interventions d'une conversation, divisé par le nombre de ses participants. Autrement dit : la moyenne d'interventions par intervenant·e. Plus cette moyenne est élevée, plus le nombre d'intervenant·e·s est réduit, plus la conversation tend par conséquent vers un profil d'espace de consultation personnalisée. Plus cette moyenne est basse, plus le nombre

d'intervenant·e·s est important, plus la conversation tend par conséquent vers un profil d'agora. 150–156, 293, 309

informel (voir discours informel). 39, 99, 117, 120, 122, 127, 131, 134–138, 145

initiat·eur/rice membre qui ouvre une conversation en publiant l'intervention initiative et le titre de celle-ci. 104, 105, 150, 153–158, 176, 193, 197, 295, 296, 307, 309

institutionnel (voir discours institutionnel). 117, 120, 127, 132, 134–138, 145, 294

intervenant·e membre participant à une conversation en y publiant une/des intervention(s). 81, 103, 105, 112, 140, 150–156, 160, 164, 169, 170, 193, 199, 207, 293, 295, 296

intervenant·e expert·e membre du forum qui a publié un grand nombre d'interventions. Nous considérons un membre comme *expert·e* lorsque le corpus comptabilise un nombre d'interventions publiées par ce membre supérieur à 50. Le corpus compte autour de 80 expert·e·s sur plus de 4700 intervenant·e·s. Le terme d'expert·e n'est donc pas basé sur une véritable expertise validée scientifiquement, mais sur la prolixité d'un membre dans le cadre restreint du forum. 3 expert·e·s comptabilisent plus de 1000 interventions dans le corpus. 138–140, 142–144, 147, 197, 198, 295

intervention tour de parole dans le polylogue que constitue la conversation. Elle est le résultat d'un acte de publication de la part d'un membre du forum. Il s'agit d'un texte produit par ledit membre du forum et qui constitue l'unité minimale de segmentation dans nos corpus. L'*intervention* s'insère dans une conversation en possédant un rang dans la chronologie de cette dernière. La date attribuée à l'*intervention* correspond à l'instant précis de sa publication en ligne par son auteur. 51–54, 56, 60, 81, 92, 97, 101, 103–106, 109, 111–113, 115, 118, 122–126, 128–133, 135–137, 139, 140, 145, 148–157, 159, 163–166, 192, 193, 196–198, 207–209, 219, 263, 265, 293–297, 301, *Synonyme :* message

intervention initiative (terme emprunté à Marcoccia, 2003). Intervention qui ouvre une conversation. Elle est publiée par l'initiat·eur/rice de la conversation. 54, 104–106, 123, 124, 163, 166, 193, 197, 293, 296

intervention réactive intervention qui en suit chronologiquement une autre dans une conversation. 54, 104, 124, 159, 294, 297

intervention unique (voir conversation à intervention unique). 54, 124, 130, 197

intervention-témoignage intervention initiative de conversation non-IU et le plus souvent foisonnante, qui suit un scénario discursif établi et décrit une situation intime. Une conversation ouverte par une *intervention-témoignage* devient le plus souvent un ECP. 104, 147, 159, 163, 164, 166, 178, 193, 197, 271

lexicométrie exploration statistique de données lexicales. La discipline de l'ADT tarde à fixer sa terminologie (Mayaffre, 2004), mais, que l'approche des données se fasse par le prisme du lexique, du texte ou du

discours, les calculs tendent tous vers le même objectif d'exploration statistiques de productions langagières sous forme de données numériques. 56, *Synonymes :* textométrie & logométrie

logométrie exploration statistique de données discursives. La discipline de l'ADT tarde à fixer sa terminologie (Mayaffre, 2004), mais, que l'approche des données se fasse par le prisme du lexique, du texte ou du discours, les calculs tendent tous vers le même objectif d'exploration statistiques de productions langagières sous forme de données numériques. 56, *Synonymes :* lexicométrie & textométrie

lymphocyte T CD4+ cellules humaines possédant des récepteurs CD4, par lesquels le VIH s'introduit. L'infection par le VIH conduit à la réduction progressive du nombre de ces cellules. Ce nombre est donc utilisé pour suivre la progression du VIH et l'efficacité des traitements. (source : Wikipedia). 293, 299

médias sociaux (voir web 2.0). 38

membre du forum internaute inscrit sur le forum, identifié par un profil. 81, 85, 86, 103–105, 112, 131, 140, 176, 207, 265, 293, 296–298

message (voir intervention). 81, 209, 210

natif du web (voir discours natif du web social). 1, 110, 117, 195–197, 294

ONUSIDA programme de l'Organisation des Nations Unies créé le 1er décembre 1995 pour coordonner les efforts de lutte contre la pandémie de VIH/SIDA à l'échelle mondiale. Le sigle anglophone est UNAIDS. Ce programme produit un rapport annuel recensant les données chiffrées et estimations pour chaque région du monde et une analyse de l'état de la pandémie. 15–17

profil ensemble des informations d'un membre permettant de l'identifier : pseudonyme, avatar, statut, nombre d'interventions, etc. 293, 297

pseudonyme chaîne de caractère choisie par chaque membre du forum pour représenter son identité de manière textuelle. Autrement dit, le nom qu'il se choisit, avec lequel les autres membres s'adresseront ou feront référence à lui. Élément constitutif de son profil. 103, 112, 207

quốc ngữ système d'écriture de la langue vietnamienne sur la base d'un alphabet latin, accompagné d'un ensemble de diacritiques. Système officiel en vigueur depuis 1918 pendant l'époque coloniale, et réofficialisé par les autorités vietnamiennes à l'indépendance. 57, 63–66

réaction (voir intervention réactive). 54, 104

réponse intervention qui en suit chronologiquement une autre dans une conversation. 54, 123, 129, 130, 144, *Synonyme :* intervention réactive

réseaux sociaux numériques ensemble des applications web servant à constituer un réseau social virtuel, en permettant aux internautes d'élaborer une identité sociale en ligne et d'interagir entre eux. 2, 44, 48–52, 294, 299

rétrovirus virus qui parasite des cellules du corps infecté en mutant constamment. Exemple de rétrovirus : le VIH. 293, 299, *voir aussi :* antirétroviral

rubriquage compartimentation en rubriques. 119, 120

rubrique niveau de structuration le plus haut et le plus figé des forums étudiés. Les *rubriques* sont créées par les administrateurs du forum et non par n'importe quel membre du forum. Elles sont déterminées de manière verticale et constituent le cadre dans lequel les internautes interviennent. Les conversations sont ouvertes au sein d'une rubrique. 52, 53, 56, 80–84, 86, 89–100, 103, 104, 107, 109, 113, 118–126, 128–131, 134, 135, 145, 196, 199, 207–210, 293, 298

sagesse des foules théorie popularisée par Surowiecki (2004), qui présuppose que l'information produite par un grand groupe d'individus est plus pertinente que celle produite par le plus expert des individus du groupe. 39

segment répété suites de formes graphiques, non séparées par une ponctuation forte, qui apparaissent plus d'une fois dans un corpus de textes (Salem, 1986). 56, 58, 133–138, 144, 145, 197, 219

sens commun concept emprunté à Sarfati (2008) : ensemble des dispositifs normatifs distinctifs d'une société, et, par suite, ensemble des normes mises en œuvre dans les discours et les textes. 34, 40, 195, 293, 294, 299

séronégati·f/ve non-porteu·r/se du virus, par exemple du VIH. *voir aussi :* séropositi·f/ve

séropositi·f/ve porteu·r/se du virus, ici, pour parler des séropositi·f/ve·s au VIH. 31, 78, 85, 86, 89, 90, 166, *voir aussi :* séronégati·f/ve

SIDA Syndrome d'ImmunoDéficience Acquise, la phase terminale du VIH. 15, 19, 20, 27, 28, 30, 32, 74, 75, 78, 85, 86, 89–91, 93, 95, 114, 115, 128, 129, 132, 141, 297, 302

SIS association Sida Info Service, `http://sida-info-service.org`. 78, 85, 86

site d'information institutionnel site émergeant de la volonté des pouvoirs publics de proposer un espace de présentation et de transmission des connaissances dans le domaine concerné (i.e. le VIH), sous la forme d'articles d'actualité ou d'information. Espace d'expression des discours institutionnels. 79, 99, 100, 147, 298, 299

site-vitrine (voir site d'information institutionnel). 98

spontané (voir discours informel spontané). 130, 197

textométrie exploration statistique de données textuelles. La discipline de l'ADT tarde à fixer sa terminologie (Mayaffre, 2004), mais, que l'approche des données se fasse par le prisme du lexique, du texte ou du

discours, les calculs tendent tous vers le même objectif d'exploration statistiques de productions langagières sous forme de données numériques. 56, *Synonymes :* lexicométrie & logométrie

TPE Traitement Post-Exposition. Traitement antirétroviral à prendre dans les 48 heures après une exposition au VIH. 90, 301, 302

triangle discursif rapports de complétion entre les trois niveaux discursifs définis par Sarfati (2008) : le canon, la vulgate et la doxa. (voir la figure II.1, page 43). 42, 197, 293, 294, 299

URL adresse web. Chaîne de caractères permettant d'accéder à un·e page / site / plateforme web / image. 76, 77, 81, 99–102, 113, 207

VIH Virus d'Immunodéficience Humaine. Le VIH est un rétrovirus, c'est-à-dire qu'il parasite des cellules du corps humain (ici, principalement les lymphocytes T4) en mutant constamment, ce qui le rend très difficilement traitable. Le *VIH* est une MST, il se transmet par voie sexuelle mais aussi par voie sanguine, par l'allaitement et pendant l'accouchement. 4, 5, 14–17, 19–21, 27, 28, 30–32, 73–75, 77, 78, 85, 86, 89–91, 93–95, 99, 114, 115, 121, 128, 129, 132, 135, 137, 139–142, 144, 166, 193, 195–199, 293, 295, 297–299, 301, 302

vitrine institutionnelle (voir site d'information institutionnel). 75, 99

vulgate terme emprunté à Sarfati (2008) désignant le discours transmis par les acteurs de l'institution dans un cadre moins formel, le discours institué du sens commun, et que nous situons dans un rapport de triangle discursif avec les deux autres variations du discours définies par Sarfati, que sont le canon et la doxa. 40–43, 75, 117, 198, 199, 293, 294, 299

web social ère de l'avènement de l'interactivité dans l'évolution du World Wide Web. Internet devient un espace de socialisation dans lequel chaque internaute a la possibilité d'établir des liens avec les autres, dialoguer, échanger des informations, exprimer des opinions, autrement dit : de déployer une identité sociale au sein des réseaux sociaux numériques, et par ces actions être act.eur/rice.s de la construction des contenus internet. 1–5, 39, 73, 195, 196, 200, 294, 295, 297, 299, *voir aussi :* web 2.0

web 2.0 expression désignant l'évolution du World Wide Web, à partir de 2005, vers une simplification d'utilisation (la complexification technologique étant masquée par des interfaces). Cette évolution a rendu possible l'interactivité entre les internautes, qui a permis la naissance du web social, et entraîné une massification des échanges et des contenus produits sur Internet. Le 2.0, calqué sur la numérotation des logiciels, est essentiellement marketing ou sociologique, il ne désigne rien d'ordre technologique en particulier. NB : le *web 2.0* tend de plus en plus à être désigné par le terme global de *médias sociaux*. 48–51, 294, 295, 297

Acronymes

AD analyse du discours. 2, 34, 37, 38, 47, 49
ADT analyse statistique des données textuelles. 56–58, 73, 107, 108, 110, 113, 197, 200, *Glossaire :* ADT
ALT ALanine Transaminase. En vietnamien : men gan. *Glossaire :* ALT
ARV AntiRétroViraux. 18, *Glossaire :* antirétroviral

CD4 cluster de différenciation 4. *Glossaire :* CD4
Conv.10+ conversations de plus de 10 interventions. 126, 128, 129
Conv.100+ conversations de plus de 100 interventions. 151, 152
Conv.2-3 conversations de 2 à 3 interventions. 126, 128, 129
Conv.4-9 conversations de 4 à 9 interventions. 126–128
Conv.50+ conversations de plus de 50 interventions. 151, 152
Conv.IU conversation à une seule intervention. 124, 126, 128, 129, 301, *Glossaire :* conversation à intervention unique

DIR discours institutionnel rapporté. 117, 118, 131–133, 135, 137, 138, 145, 147, 198, *Glossaire :* discours institutionnel rapporté
DIS discours informel spontané. 117, 132, 133, 135, 137, 138, 145, 147, 192, 194, 198, *Glossaire :* discours informel spontané

ECP espace de consultation personnalisée. 147, 149–158, 193, 197, 296, *Glossaire :* espace de consultation personnalisée

HCMV Hồ Chí Minh-Ville. 13, 14
HIV human immunodeficiency virus. *En français :* VIH

IST infection sexuellement transmissible. 85, 86, 89, 91, 95, 115
IU (voir Conv.IU). 296

LGBT Lesbien·ne·(s), Gay(s), Bisexuel·le·(s) et Trans. 94, 95, 115

MST maladie sexuellement transmissible. 140–142, 299

OMS Organisation Mondiale de la Santé. 18

PEP Post Exposure Profilaxis. 121, *En français :* TPE

RESET Recherches En Sciences sociales sur InternET. 47
RSF Reporters Sans Frontières. 10
RSVN République Socialiste du Viêt Nam. 8, 9

SHS sciences humaines et sociales. 1–3, 73
SIDA Syndrome d'ImmunoDéficience Acquise. 15, 19, 20, 27, 28, 30, 32, 74, 75, 78, 85, 86, 89–91, 93, 95, 114, 115, 128, 129, 132, 141, 297, *Glossaire :* SIDA

TAL traitement automatique des langues. 5, 57, 60, 200
TIC technologies de l'information et de la communication. 24, 32, 73
TPE traitement post-exposition. 86, 90, 121, *Glossaire :* TPE
TS travailleu·se/r·s sexuel·le·s. 14, 15, 28, 164

UDI usag·er/ère·s de drogues injectables. 14, 15, 28

VIH virus d'immunodéficience humaine. 4, 5, 14–17, 19–21, 27, 28, 30–32, 73–75, 77, 78, 85, 86, 89–91, 93–95, 99, 114, 115, 121, 128, 129, 132, 135, 137, 139–142, 144, 166, 193, 195–199, 295, 297–299, *Glossaire :* VIH

Table des matières

Sommaire v

Introduction 1

Chapitre 1 Le Viêt Nam comme axe de recherche 7

1 Le contexte socio-culturel vietnamien 7

1.1 Contexte historique 7

1.2 Situation politique : l'abus des libertés démocratiques est dangereux pour la santé 8

1.3 Situation démographique : une population jeune et dense 11

1.4 Mondialisation et VIH 14
1.4.1 Drogues injectables et VIH 14
1.4.2 Prostitution et VIH 15
1.4.3 Évolution de l'épidémie du VIH dans la population vietnamienne 16
1.4.4 Accès aux traitements 17

2 Une société en tension 18

2.1 Traditions et tabous : un discours lacunaire sur la sexualité 19

2.2 Modernisation et accroissement de l'écart entre les générations 20
2.2.1 L'éducation par les pairs 20

2.3 Utilisation d'Internet et des TIC 22

3 Le discours institutionnel vietnamien en question 25

3.1 L'état tout-puissant 25
3.1.1 La politique du parti unique 25
3.1.2 Le verrouillage des médias 26

3.2 La lutte contre les fléaux sociaux 27
3.2.1 Les fléaux sociaux 27

3.2.2 La propagande 29
3.2.3 Les efforts pour diversifier les moyens de lutte 29

4 Conclusion 31

Chapitre II État de l'art 33

1 Points de vue sur les discours institutionnels 33

1.1 Le rôle de l'institution dans la construction des connaissances 33

1.2 Analyse du discours institutionnel 34
1.2.1 Une fonction de production de sens commun 34
1.2.2 Une fonction de figement des connaissances et d'autorité prescriptive 35

2 Reconfigurations de l'autorité dans la construction des connaissances 38

2.1 Autorité gnoséologique selon l'analyse du discours 38
2.1.1 Application au domaine médico-sanitaire 41
2.1.2 Prescription *vs* description des connaissances 43

2.2 Autorité gnoséologique selon la sociologie 45

3 L'analyse des discours informels 47

3.1 Spécificités des forums 49
3.1.1 Formalisation de la structuration des forums 52
3.1.2 Le TAL confronté à l'écriture sur les forums 55

4 Méthodes et outils de corpus 56

4.1 Segmentation du corpus pour la fouille contrastive 56

4.2 Analyse des données textuelles 56
4.2.1 Lexico (5) 57
4.2.1.1 Lexico et le vietnamien 57
4.2.1.2 Segments répétés et carte des sections 58
4.2.2 TXM 59
4.2.2.1 Cooccurrences, expressions régulières et tri des données 59

5 Le traitement automatique du vietnamien 60

5.1 Les acteurs du traitement automatique du vietnamien 61

5.2 Les spécificités du vietnamien et ses conséquences en TAL 62
5.2.1 L'écriture de la langue vietnamienne 62
5.2.2 L'encodage du quốc ngữ 63
5.2.3 La saisie du quốc ngữ au clavier 64
5.2.4 La segmentation 65

5.2.5 L'étiquetage morpho-syntaxique 68

Chapitre III Constitution des données 73

1 Préambule : le choix des sources 73

1.1 Choix des sources vietnamiennes 74

1.2 Choix des sources françaises 77

2 Du web au corpus 80

2.1 Étude des données en ligne 80
2.1.1 Deux genres textuels à organiser en un corpus homogène 80
2.1.1.1 Analyse technodiscursive des forums 81
2.1.1.2 Analyse des sites d'information institutionnels 82
2.1.2 Etude de l'organisation des forums en rubriques thématiques 82

2.2 Délimitation des données à collecter 91
2.2.1 Sélection des rubriques à analyser 91
2.2.2 Choix de l'empan temporel 96
2.2.3 Choix concernant le corpus institutionnel 98
2.2.3.1 Collecte des données sur le site d'information institutionnel HIV Online 99
2.2.3.2 Collecte des données sur le site d'information institutionnel Sida Info Service 100

2.3 Choix de structuration des données en corpus 101
2.3.1 Gestion des différents types d'éléments sémiotiques 101
2.3.2 Profondeur de segmentation du corpus : la conversation ou l'intervention comme unité minimale ? 103

3 Traitement des données pour l'analyse 107

3.1 Choix des outils de traitement automatique des langues 107
3.1.1 Choix des outils de segmentation du vietnamien 107
3.1.2 Choix des outils d'étiquetage morpho-syntaxique 107
3.1.3 Positionnement quant au nettoyage des données 108

3.2 Mise en forme pour les logiciels d'ADT 108
3.2.1 Mise en forme des corpus pour Lexico 109
3.2.2 Mise en forme des corpus pour TXM 110

3.3 Anonymisation 112

4 Description de la chaîne de traitement 113

5 Description quantitative du corpus 113

Chapitre IV Cartographie d'un forum de discussion 117

1 Introduction et définitions 117

2 Description structurale des forums 118

2.1 Rubriques 118
2.1.1 Coquille vide ou colonne vertébrale : le cadre institutionnel 120
2.1.2 Actualisation et enrichissement du rubriquage 120

2.2 Conversations et Interventions 122
2.2.1 Description d'une conversation 123
2.2.2 Description d'une intervention 124

3 Typologie des discours 124

3.1 Typologie des rubriques basée sur le nombre d'interventions par conversation 125
3.1.1 Catégories de conversations en fonction de leur nombre d'interventions 125
3.1.2 Profils de rubriques 127
3.1.3 Conclusion 128

3.2 Cas particulier des conversations à intervention unique 129

3.3 Typologie des discours basée sur la longueur des interventions 131

3.4 Typologie des discours à l'aide des segments répétés 133
3.4.1 Étude des expressions figées en contexte informel 135
3.4.2 Le discours des expert·e·s 138

4 Conclusion 145

Chapitre v Analyse sémantique des discours informels 147

1 Introduction 147

2 Étude des genres discursifs informels spontanés des forums 148

2.1 Étude des conversations foisonnantes 148
2.1.1 Considérations terminologiques 148
2.1.2 Introduction à une étude sémantique : un foisonnement provoqué par l'angoisse 148
2.1.3 Deux types de conversations foisonnantes : les agoras et espaces de consultation personnalisés 149
2.1.3.1 Exemple de conversation foisonnante de type agora 149
2.1.3.2 Définition des espaces de consultation personnalisée (ECP) 150
2.1.4 Indices de catégorisation des conversations foisonnantes 150
2.1.4.1 Définition d'un indice de focalisation 151
2.1.4.2 Des seuils différents entre le corpus français et le corpus vietnamien 151

2.1.4.3 Limites de l'indice de focalisation 153
2.1.4.4 Fiabilité de la proportion de l'initiat·eur/rice 153
2.1.5 Etude des conversations foisonnantes de type espace de consultation personnalisée (ECP) 157
2.2 Etude du genre textuel du témoignage 159
2.3 Observations lexicales sur le discours informel spontané 167
2.3.1 Transcription de l'oralité 167
2.3.2 Euphémisations 171
2.3.2.1 Techniques de distanciation 172
2.3.2.2 Exemples de thèmes euphémisés 173
2.3.2.3 Institutionnalisation des euphémismes 174
2.3.3 Étude des titres 176
3 Analyse des discours informels par contraste 179
3.1 Distanciation et intimité : des univers de référence distincts 180
3.1.1 Zones anthropiques distales, proximales, identitaires 181
3.1.2 Ontologie et praxéologie 184
3.2 Tabous, déviance et écart à la norme dans le corpus vietnamien 185
3.3 La perception du temps 186
3.3.1 Opposition des temps verbaux 186
3.3.2 Atemporalité institutionnelle 187
3.3.3 La fenêtre sérologique : le temps ressenti et mesuré 188
4 Conclusion 192
Conclusion générale et perspectives 195
Annexes 201
A Étapes de constitution du corpus 203
B Utilisation de Web ::Scraper 211
C Recours aux bases de données 213
D Extrait de metadata.csv - Conversations 215
E Extrait de metadata.csv - Interventions 217
F Segments répétés les plus fréquents par rubrique 219
G Calcul du niveau de personnalisation des conversations foisonnantes 261

H Constitution des diagrammes de répartition des intervention 263

I Calcul du nombre d'interventions par membre 265

J Qualification des discours à l'aide des expressions figées 267

K Listes de formes par champs lexicaux du genre témoignag 271

L Lexique vietnamien 275

M Liste des émoticônes 285

Glossaire 293

Acronymes 301

Table des matières 303

Liste des tableaux 309

Table des figures 311

Bibliographie 313

Liste des tableaux

Tableau II.1 Touches du clavier pour marquer les tons 65

Tableau III.1 Premier niveau de structuration sur le site vietnamien 85
Tableau III.2 Premier niveau de structuration sur le site français 85
Tableau III.3 Deuxième niveau de structuration sur le site français 86
Tableau III.5 Rubriques sélectionnées sur le forum vietnamien 93
Tableau III.6 Rubriques sélectionnées sur le forum français 95
Tableau III.7 Rubriques institutionnelles sélectionnées sur le site vietnamien 99
Tableau III.8 Modifications sur le site vietnamien 100
Tableau III.9 Langages et programmes informatiques utilisés 113
Tableau III.10 Description quantitative des corpus 114

Tableau IV.1 Mise à l'écart des interventions uniques 130
Tableau IV.2 Exemples de distribution de segments répétés 137

Tableau V.1 Indice de focalisation de conversations foisonnantes 151
Tableau V.2 Indices de focalisation des conversations foisonnantes 152
Tableau V.3 Indice de focalisation et proportion de l'initiat·eur/rice de conversations foisonnantes 156
Tableau V.4 Le lexique des titres des conversations 177

Tableau A.1 Chaîne de traitement des corpus 204
Tableau A.2 Récapitulatif des formats de corpus 209

Tableau F.1 Rubrique (62) Relations sexuelles, préservatif, lubrifiant 219
Tableau F.2 Rubrique (68) Stigmatisation et discrimination 226
Tableau F.3 Rubrique (66) Exposition au VIH 229
Tableau F.4 Rubrique (56) Connaissances de base sur le VIH/SIDA 232
Tableau F.5 Rubrique (22) Safe sex 234
Tableau F.6 Rubrique (57) Etre Gay 242

Tableau L.1 Lexique vietnamien de linguistique et TAL 275

Table des figures

Figure I.1 Panneau de propagande 12
Figure I.2 Pourcentage d'internautes en France et au Viêt Nam 22
Figure I.3 Abonnements à la téléphonie mobile cellulaire 23

Figure II.1 Le triangle discursif 43

Figure III.1 Page d'accueil de HIV Online 76
Figure III.2 Logo, titre et sous-titre de HIV Online 77
Figure III.3 Page d'accueil de Sida Info Service 79
Figure III.4 Forums sur *forum.hiv.com.vn* 83
Figure III.5 Rubriques sur *forum.hiv.com.vn* 84
Figure III.6 Forums et rubriques sur *forum.sida-info-service.org* 84
Figure III.7 Répartition des interventions par rubrique sur le forum français 92
Figure III.8 Répartition des interventions par rubrique sur le forum vietnamien 92
Figure III.9 Rubriques sélectionnées sur le forum vietnamien 94
Figure III.10 Rubriques sélectionnées sur le forum français 95
Figure III.11 Nombre d'interventions publiées par année 97
Figure III.12 Nombre de découvertes de séropositivité par année. 98

Figure IV.1 Nombre de conversations par catégorie 126
Figure IV.2 Répartition des conversations par catégorie 127

Figure V.1 Répartition des interventions dans la conversation [FR « Un désastre...Ai-je contaminé ma femme ? »] 152
Figure V.2 Répartition des interventions dans la conversation [FR « Prise de risque ? »] 154
Figure V.3 Répartition des interventions dans la conversation [FR « fellation sans preso »] 155
Figure V.4 Interventions de inquiet2007 158
Figure V.5 Témoignage personnel 161
Figure V.6 Témoignage personnel 162
Figure V.7 Alcool et prostitution par conversations 165

Figure v.8 Alcool et prostitution par interventions 165
Figure v.9 Alcool et prostitution par interventions initiales 166
Figure v.10 Articulations des discours 180
Figure v.11 Acteurs du corpus vietnamien 181
Figure v.12 Acteurs du corpus français 183
Figure v.13 Les temporalités dans le discours sur le VIH 190
Figure v.14 La temporalité dans le corpus vietnamien 191
Figure v.15 La temporalité dans le corpus français 192

Bibliographie

Barbara Watson Andaya. From Temporary Wife to Prostitute : Sexuality and Economic Change in Early Modern Southeast Asia. *Journal of Women's History*, 9(4), pages 11–34, 1998. doi : 10.1353/jowh.2010.0225. URL https://www.researchgate.net/publication/236812906_From_Temporary_Wife_to_Prostitute_Sexuality_and_Economic_Change_in_Early_Modern_Southeast_Asia.

Tim Berners-Lee. The Semantic Web. *Scientific American Magazine*, may 2001.

Marie-Ève Blanc. Campagne de prévention de l'épidémie de sida au Viêt Nam : représentation des risques, institutionnalisation de la prévention et enjeux socio-politiques. *In* Marie-Eve Blanc, Laurence Husson et Evelyne Micollier (dir.) : *Sociétés Asiatiques face au Sida*, Recherches Asiatiques, pages 171–192. L'Harmattan, 2001a.

Marie-Ève Blanc. Du modèle confuscéen à la quête de l'égalité entre homme et femme. *In* Claude Bontems (dir.) : *Mariage-Mariages*, pages 407–442. PUF, 2001b.

Marie-Ève Blanc. Sex Education for Vietnamese Adolescents in the Context of HIV/AIDS Epidemic : The NGOs, the School, the Family and the Civil Society. *In* Evelyne Micollier (dir.) : *Sexual Cultures in East Asia. The Social Construction of Sexuality and Sexual Risk in a Time of AIDS*, pages 241–262. Taylor & Francis, 2004.

Marie-Ève Blanc. Le pluralisme médicamenteux face à l'épidémie de VIH/sida au Viêt Nam. *Moussons*, 15, 2010. URL http://moussons.revues.org/337. Mis en ligne le 01 octobre 2012.

Marie-Eve Blanc, Laurence Husson et Evelyne Micollier (dir.). *Sociétés Asiatiques face au Sida*. Recherches Asiatiques. L'Harmattan, 2001.

Sergio Bolasco, Isabella Chiari et Luca Giuliano (dir.). *Statistical Analysis of Textual Data - 10th International Conference JADT*, Rome, 9-11 juin 2010. Edizioni Universitarie di Lettere Economia Diritto.

Claude Bontems (dir.). *Mariage-Mariages*. PUF, Paris, 2001.

Evelyne Broudoux. Construction de l'autorité informationnelle sur le web. *In* N. W. Lund R. Skare et A. Vårheim (dir.) : *A Document (Re)turn : Contributions from a Research Field in Transition*, pages 265–278. Peter Lang, Frankfurt, 2007. URL http://archivesic.ccsd.cnrs.fr/file/index/docid/120710/filename/AutorInfo.pdf.

Ilonka Brügemann et Barbara Franklin. *Love and the Risk of AIDS for Women in Vietnam : A Qualitative Study of Women in the Hanoi Area*. CARE International in Vietnam, 1995. URL https://books.google.fr/books?id=rac2HAAACAAJ. pas lu.

Xuân Hao Cao. Les linguistes vietnamiens et la phonologie de leur langue. *Revue Phonologie et linéarité,SELAF n°spécial 18*, page 18, 1985.

Dominique Cardon. Le design de la visibilité. un essai de cartographie du web 2.0. *Réseaux*, 152, pages 93–137, 2008. doi : 10.3917/res.152.0093. URL www.cairn.info/revue-reseaux1-2008-6-page-93.htm.

Dominique Cardon. Vertus démocratiques de l'internet. *La Vie des idées*, nov 2009. URL http://www.laviedesidees.fr/Vertus-democratiques-de-l-Internet.html. ISSN : 2105-3030.

Dominique Cardon. *La démocratie Internet : promesses et limites*. La République des idées. Seuil, 2010. ISBN 9782021026917. EAN : 9782021026917.

Francis Chateauraynaud. Visionnaires à rebours. des signaux faibles à la convergence de séries invisibles, dec 2007. URL http://www.gspr-ehess.com/documents/FC_Visionnaires-a-rebours-dec-2007.pdf.

Francis Chateauraynaud et Josquin Debaz. Veille sociologique et flux d'informations numériques. *In 9e Journées Francophones "Extraction et Gestion des Connaissances"*, jan 2009. URL https://hal.archives-ouvertes.fr/hal-00492950/document.

R W. Connell, R. Connell et G W. Dowsett. *Rethinking Sex : Social Theory and Sexuality Research*. Temple University Press, 1993. ISBN 9781566390736. URL https://books.google.fr/books?id=jBy5QgAACAAJ.

John W. Creswell. *Research Design : Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE Publications, 2nde edition, 2003. ISBN 9780761924425.

Nanette J. Davis (dir.). *Prostitution : An International Handbook on Trends, Problems, and Policies*. Greenwood Press, Westport, Connecticut, 1993.

Diên Dinh. Building a training corpus for word sense disambiguation in the English-to-Vietnamese Machine Translation. *In Proceedings of Workshop on Machine Translation in Asia*, pages 26–32, 2002.

URL https://www.researchgate.net/publication/228698285_Building_a_training_corpus_for_word_sense_disambiguation_in_English-to-Vietnamese_Machine_Translation.

Diên Dinh et Kiem Hoang. Pos-tagger for English-Vietnamese Bilingual Corpus. *In Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts : Data Driven Machine Translation and Beyond - Volume 3*, HLT-NAACL-PARALLEL '03, pages 88–95, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. doi : 10.3115/1118905.1118921. URL http://dx.doi.org/10.3115/1118905.1118921.

Diên Dinh, Kiem Hoang et Van Toan Nguyen. Vietnamese word segmentation. *In Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium*, Hitotsubashi Memorial Hall, National Center of Sciences, Tokyo, Japan, November 27-30 2001. URL https://www.researchgate.net/publication/220706845_Vietnamese_Word_Segmentation.

Quang Thắng Dinh, Hồng Phương Lê, Thị Minh Huyền Nguyễn, Cẩm Tú Nguyễn, Mathias Rossignol et Xuân Lương Vũ. Word segmentation of Vietnamese texts : a comparison of approaches. 6th international conference on Language Resources and Evaluation - LREC 2008, may 2008. URL https://hal.inria.fr/inria-00334760.

Cory Doctorow. Metacrap : Putting the torch to seven straw-men of the meta-utopia. *The WELL*, 2001. URL http://www.well.com/~doctorow/metacrap.htm.

Judith Donath, Karrie Karahalios et Fernanda Viégas. Visualizing conversation. *Journal of Computer-Mediated Communication*, 4(4), 1999. URL www.ascusc.org/jcmc/vol4/issue4/donath.html.

Fernande Dupuis, Robert Kapitan et François Daoust. Expérience d'entrainement de TreeTagger et d'intégration à l'interface web de SATO. *In* Sergio Bolasco, Isabella Chiari et Luca Giuliano (dir.) : *Proceedings of 10th International Conference JADT 2010*, Rome, 9-11 juin 2010. Edizioni Universitarie di Lettere Economia Diritto. URL http://www.ledonline.it/ledonline/JADT-2010/allegati/JADT-2010-1011-1020_176-Dupuis.pdf.

Egle Eensoo et Mathieu Valette. Sur l'application de méthodes textométriques à la construction de critères de classification en analyse des sentiments. *In* G. Antoniadis, H. Blanchon et G. Sérasset (dir.) : *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, volume 2, pages 367–374, Grenoble, 4-8 juin 2012.

Egle Eensoo et Mathieu Valette. Sémantique textuelle et TAL : un exemple d'application à l'analyse des Sentiments. *In* D. Ablali, S. Badir et D. Du-

CARD (dir.) : *Documents, textes, œuvres*, Rivages linguistiques. Presses Universitaires de Rouen, 2014.

Egle EENSOO et Mathieu VALETTE. Une méthodologie de sémantique de corpus appliquée à des tâches de fouille d'opinion et d'analyse des sentiments : étude sur l'impact de marqueurs dialogiques et dialectiques dans l'expression de la subjectivité. *Texto ! Textes et Cultures*, XX, 2015. URL http://www.revue-texto.net/docannexe/file/3688/eensoo_valette_xx.3.2015.pdf.

Barbara FRANKLIN. *Risk of AIDS in Vietnam : An Audience Analysis of Urban Men and Sex Workers, with Guidelines for Prevention*. CARE International in Vietnam, 1993. URL https://books.google.fr/books?id=397MYgEACAAJ.

Marie-Carmen GARCIA. *Amours clandestines. Sociologie de l'extraconjugalité durable*. collection « Sexualités ». Presses universitaires de Lyon, Lyon, 2016. URL http://presses.univ-lyon2.fr/produit.php?id_produit=1004.

Pierre HALTÉ. *Les marques modales dans les chats : étude sémiotique et pragmatique des émoticônes et des interjections dans un corpus de conversation synchrones en ligne*. Thèse de doctorat, Université de Lorraine, 2013. URL http://docnum.univ-lorraine.fr/public/DDOC_T_2013_0308_HALTE.pdf.

Serge HEIDEN, Jean-Philippe MAGUÉ et Bénédicte PINCEMIN. TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. *In* Sergio BOLASCO, Isabella CHIARI et Luca GIULIANO (dir.) : *Statistical Analysis of Textual Data - Proceedings of 10th International Conference JADT 2010*, volume 2 of *C*, pages 1021–1032, Rome, 9-11 juin 2010. Edizioni Universitarie di Lettere Economia Diritto. URL http://www.ledonline.it/ledonline/JADT-2010/allegati/JADT-2010-0341-0354_023-Pincemin.pdf. Logiciel disponible sur http ://textometrie.ens-lyon.fr/.

Bao Quoc Hô, Jean-Pierre CHEVALLET, Marie-France BRUANDET et Thi Bich Thuy DONG. An approach to Vietnamese Information Retrieval. *8th International Workshop on Academic Information Networks and Systems. Etude commune de Global Sharing of Digital Assets of Scientific et de Cultural Information*, page 6, 2002.

Bao Quoc Hô, Jean-Pierre CHEVALLET et Marie-France BRUANDET. Mise en place d'un Système de Recherche d'Informations en vietnamien. *TALN 2003 Traitement Automatique des Langues Naturelles*, page 12, 11-14 juin 2003. URL http://mrim.imag.fr/publications/2003/TALN_2003_Bao_Quoc_HO_CLIPS-IMA.pdf. Atelier "multilinguisme", Batz-sur-Mer, France.

Océane Hô DINH et Mathieu VALETTE. Textes institutionnels et textes informels issus du web social. observations sur la construction et la sanction

des connaissances dans des corpus français et vietnamien du domaine sanitaire. *In* Julien Longhi et Georges-Elia Sarfati (dir.) : *Les discours institutionnels en confrontation, Contribution à l'analyse des discours institutionnels et politiques*, Espaces Discursifs, pages 147–165. L'Harmattan, Paris, 2014a.

Océane Hô Dinh et Mathieu Valette. Analyse différentielle des discours de prévention du VIH : textes institutionnels et textes informels en français et en vietnamien. *In Actes des 12e Journées internationales d'analyse statistique des données textuelles (JADT 2014)*, pages 277–287, Paris, 3-6 juin 2014b. URL http://lexicometrica.univ-paris3.fr/jadt/jadt2014/01-ACTES/23-JADT2014.pdf.

Mark Hochhauser et James H. Rothenberger. *AIDS education*, volume III. Wm. C. Brown, Dubuque, Iowa, 1992.

David A. Huffaker et Sandra L. Calvert. Gender, Identity, and Language Use in Teenage Blogs. *Journal of Computer-Mediated Communication*, 10(2), 2005. doi : 10.1111/j.1083-6101.2005.tb00238.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2005.tb00238.x/full#b8.

Agata Jackiewicz et Marko Vidak. Étude sur les mots-dièse. *Congrès Mondial de Linguistique Française*, 8, pages 2033 – 2050, 2014. URL https://www.shs-conferences.org/articles/shsconf/pdf/2014/05/shsconf_cmlf14_01198.pdf.

Todd D. Jick. Mixing Qualitative and Quantitative Methods : Triangulation in Action. *Administrative Science Quarterly*, 24(4), pages 602–611, 1979. URL http://www.jstor.org/stable/2392366.

Paula-Frances Kelly. What is Known about Gender, the Constructs of Sexuality and Dictates of Behaviour in Vietnam as a Confucian and Socialist Society and their Impact on the Risk of HIV/AIDS Epidemic. *In* Evelyne Micollier (dir.) : *Sexual Cultures in East Asia : The Social Construction of Sexuality and Sexual Risk in a Time of AIDS*. Taylor & Francis, 2004.

Thu Hong Khuât. Study on Sexuality in Vietnam : the known and the unknown issues. *South and East Asia Regional Working Papers*, 11 :Hanoi : Population Council, 1998.

Alice Krieg-Planque et Claire Oger. Discours institutionnels. Perspectives pour les sciences de la communication. *Mots. Les langages du politique [En ligne]*, 94, pages 91–96, 2010. URL http://mots.revues.org/19870. Mis en ligne le 06 novembre 2012.

Hồng Phương Lê. Vers une grammaire électronique du vietnamien. page 65, 2005.

Hồng Phương Lê, Thi Minh Huyen Nguyen, Laurent Romary et Azim Roussanaly. A Lexicalized Tree-Adjoining Grammar for Vietnamese. page 6, 2006.

Hồng Phương Lê, Azim Roussanaly et Thị Minh Huyền Nguyễn. A Hybrid Approach to Word Segmentation of Vietnamese Texts. *Language and Automata Theory and Applications*, art. Berlin, 2008.

Thi Quy Lê. *Prostitution in Vietnam*. 1993a.

Thi Quy Lê. Some ideas about prostitution in Vietnam. *In Joining Forces to Further Shared Vision*, Washington DC, October 1993b.

Pierre Lafon. Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1(1), pages 127–165, 1980. ISSN 0243-6450. doi : 10.3406/mots.1980.1008. URL http://www.persee.fr/doc/mots_0243-6450_1980_num_1_1_1008.

Pierre Lafon. Analyse lexicométrique et recherche des cooccurrences. *Mots*, 3 (1), pages 95–148, 1981. ISSN 0243-6450. doi : 10.3406/mots.1981.1041. URL http://www.persee.fr/doc/mots_0243-6450_1981_num_3_1_1041.

Aurélien Lauf. *Propagation du buzz sur Internet – Identification, analyse, modélisation et représentation dans un contexte de veille*. Linguistics, Institut National des Langues et Civilisations Orientales (INALCO), PARIS, 2014.

Ludovic Lebart et André Salem. *Statistique textuelle*. Dunod, 1994. ISBN 9782100022397.

Julien Longhi. L'hybridation du discours institutionnel à l'épreuve du numérique : renouvellement et reconfiguration de la parole institutionnelle. *In* Julien Longhi et Georges-Elia Sarfati (dir.) : *Les discours institutionnels en confrontation. Contribution à l'analyse des discours institutionnels et politiques*, Espaces Discursifs, pages 167–188. L'Harmattan, 2014. URL https://halshs.archives-ouvertes.fr/halshs-00989072/document.

Julien Longhi et Georges-Elia Sarfati (dir.). *Les discours institutionnels en confrontation. Contribution à l'analyse des discours institutionnels et politiques*. L'Harmattan, 2014.

Michel Marcoccia. Parler politique dans un forum de discussion. *Langage et société*, 2(104), pages 9–55, 2003. URL https://www.cairn.info/revue-langage-et-societe-2003-2-page-9.htm.

Michel Marcoccia. L'analyse conversationnelle des forums de discussion : questionnements méthodologiques. *Les Carnets du Cediscor*, 8, pages 23–37, 2004. URL http://cediscor.revues.org/220.

Damon Mayaffre. *Paroles de président. Jacques Chirac (1995-2003) et le discours présidentiel sous la Ve République*. Honoré Champion, Paris, 2004.

Damon Mayaffre. L'analyse de données textuelles aujourd'hui : du corpus comme une urne au corpus comme un plan. *Lexicometrica*, pages 1–12, 2007. URL https://halshs.archives-ouvertes.fr/hal-00551468/document.

Evelyne Micollier. Social Significance Of Commercial Sex Work : Implicitly Shaping A Sexual Culture ? *In* Evelyne Micollier (dir.) : *Sexual Cultures in East Asia. The social construction of sexuality and sexual risk in a Time of AIDS*, Asian Studies, pages 3–22. Taylor & Francis, 2004a. URL https://halshs.archives-ouvertes.fr/halshs-01070604/document.

Evelyne Micollier (dir.). *Sexual Cultures in East Asia : The Social Construction of Sexuality and Sexual Risk in a Time of AIDS*. Taylor & Francis, 2004b. ISBN 9781134393503. URL https://books.google.fr/books?id=LUJT6q8Zy0wC.

Émilie Née (dir.). *Méthodes et outils informatiques pour l'analyse du discours*. Didact Méthodes. Presses Universitaires de Rennes, 2017. URL http://www.pur-editions.fr/detail.php?idOuv=4428. Domaine : Sciences humaines et sociales - Information-Communication. Auteurs : Emilie Née and Christine Barats ans Serge Fleury and Jean-Marc Leblanc and Frédérique Sitri and Marie Veniard.

Vinh Long Ngo. Vietnam. *In* Nanette J. Davis (dir.) : *Prostitution : An International Handbook on Trends, Problems, and Policies*, page 428. Greenwood Press, Westport, Connecticut, 1993.

Thanh Bon Nguyen, Thi Minh Huyen Nguyen, Laurent Romary et Xuan Luong Vu. Lexical descriptions for vietnamese language processing. page 8, 2005.

Thanh V. Nguyen, Hoang K. Tran, Thanh T.T. Nguyen et Hung Nguyen. Word segmentation for vietnamese text categorization : An online corpus approach. *RIVF 06 Conference Internationale Associant Chercheurs Vietnamiens et Francophones en Informatique*, page 6, 2006.

Thi Minh Huyen Nguyen. Vers la génération d'un lexique bilingue (français-vietnamien). 2000.

Thi Minh Huyen Nguyen, Laurent Romary et Xuan Luong Vu. Une étude de cas pour l'étiquetage morpho-syntaxique de textes vietnamiens. page 10, 2003.

Văn Chỉnh Nguyễn. *Social Change in Rural Vietnam : Children's Work and Seasonal Migration*, volume 13 of *Political and Social Change Working Paper Series*, chapter In search of work : socio-economic change and seasonal migration in northern Vietnam, pages 39–65. Department of Political and Social Change, Division of Politics and International Relations, Research School of Pacific and Asian Studies, Australian National University, Canberra, ACT, 1997.

Claire Oger et Caroline Ollivier-Yaniv. Analyse du discours institutionnel et sociologie compréhensive : vers une anthropologie des discours institutionnels. *Mots. Les langages du politique [En ligne]*, 71, pages 125–145, 2003. URL http://mots.revues.org/8423. Mis en ligne le 05 mai 2008.

Claire Oger et Caroline Ollivier-Yaniv. Conjurer le désordre discursif. Les procédés de « lissage » dans la fabrication du discours institutionnel. *Mots. Les langages du politique [En ligne]*, 81, pages 63–77, 2006. URL http://mots.revues.org/675. Mis en ligne le 01 juillet 2008.

Claire Oger, Caroline Ollivier-Yaniv et Marie-Anne Paveau. Discours militaire sur les médias. *Langage & société*, 4(94), pages 5–7, dec 2000. doi : 10.3917/ls.094.0005. URL www.cairn.info/revue-langage-et-societe-2000-4-page-5.htm.

Phillipe Papin et Laurent Passicousset. *Vivre avec les Vietnamiens*. Guide. Archipel, 2010. ISBN 9782809804515. URL https://books.google.fr/books?id=PriHrVo72fMC.

Marie-Anne Paveau. Un dictionnaire d'analyse du discours numérique (dadn). Technologies discursives, [Carnet de recherche], dec 2012a. URL http://technodiscours.hypotheses.org/?p=245.

Marie-Anne Paveau. Genre de discours et technologie discursive. tweet, twittécriture et twittérature. *Pratiques*, 157-158, pages 7–30, 2012b. URL https://hal.archives-ouvertes.fr/hal-00824817/document.

Marie-Anne Paveau. Du contexte à l'environnement : une approche écologique du discours. Journée Doscila, apr 2013a. URL http://penseedudiscours.hypotheses.org/11322.

Marie-Anne Paveau. Technodiscursivités natives sur twitter. une écologie du discours numérique. *Epistémé (Revue internationale de sciences humaines et sociales appliquées, Séoul)*, 9, pages 139–176, 2013b. URL https://hal-univ-paris13.archives-ouvertes.fr/hal-00859064/document.

Bénédicte Pincemin. Sémantique interprétative et textométrie. *Texto !*, XVII (3) :Christophe Cusimano, 2012. URL http://www.revue-texto.net/docannexe/file/3049/pincemin_texto11.pdf. 21 pages.

François Rastier. *Arts et sciences du texte*. PUF, Paris, 2001.

André Salem. La typologie des segments répétés dans un corpus, fondée sur l'analyse d'un tableau croisant mots et textes. *Les Cahiers de l'Analyse des Données*, 9(4), pages 489–500, 1984.

André Salem. Segments répétés et analyse statistique des données textuelles. *Histoire & Mesure*, I(2), pages 5–28, 1986. URL https://www.researchgate.net/publication/30431328_Segments_repetes_et_analyse_statistique_des_donnees_textuelles.

André Salem. *Pratique des segments répétés, Essai de statistique textuelle*. Klincksiek, Paris, 1987.

Georges-Elia Sarfati. Pragmatique linguistique et normativité : Remarques sur les modalités discursives du sens commun. *Langages*, 2(170), pages 92–108, 2008. doi : 10.3917/lang.170.0092. URL www.cairn.info/revue-langages-2008-2-page-92.htm.

Georges-Elia Sarfati. L'emprise du sens. Note sur les conditions théoriques et les enjeux de l'analyse des discours institutionnels. *In Les discours institutionnels en confrontation. Contribution à l'analyse des discours institutionnels et politiques*, pages 13–46. L'Harmattan, 2014.

Emilie Saunier, Olivier Vanhée et Géraldine Bois. Les réaménagements actuels de l'autorité prescriptive dans le secteur littéraire : le cas des blogs de lecteurs. *In Hyperchoix et prescription culturelle*, Saint-Cloud, France, nov 2014. Peter Lang. URL halshs-01089764.

Ferdinand de Saussure. *Écrits de linguistique générale*. Gallimard, Paris, 2002.

Helmut Schmid. Improvements in part-of-speech tagging with an application to german. *In Proceedings of the ACL SIGDAT-Workshop*, Dublin, Ireland, 1995. URL http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf.

Alexandre Serres. L'évaluation de l'information à l'heure du web 2.0 : entre changement et continuité. Journée d'étude Médiadix, dec 2010. URL http://urfist.enc.sorbonne.fr/sites/default/files/JournéeMediadix-Urfist_ASerres_Evaluation_information_web2_0.pdf.

Clay Shirky. Ontology is overrated. march 2005. URL http://shirky.com/writings/ontology_overrated.html. O'Reilly ETech conference.

Mohamed Sidir, Nadine Lucas et Emmanuel Giguet. De l'analyse des discours à l'analyse structurale des réseaux sociaux : une étude diachronique d'un forum éducatif. *Revue des Sciences et Technologies de l'Information et de la*

Communication pour l'Education et la Formation (STICEF), 13, pages 20, 2006. URL https://halshs.archives-ouvertes.fr/hal-00696387/document.

Monique Slodzian et Mathieu Valette. Connaissances prescrites et connaissances décrites ? l'apport de la sémantique des textes. *In* Khaldoun Zreik (dir.) : *Patrimoine 3.0, Actes du 12e Colloque International sur le Document Electronique (CIDE.12)*, pages 129–141. Europia Productions, Paris, 2009.

James Surowiecki. *The Wisdom of Crowds : Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Doubleday ; Anchor, 2004. traduction française publiée en 2008 : La Sagesse des Foules.

Abbas Tashakkori et Charles Teddlie. *Handbook of Mixed Methods in Social & Behavioral Research*. SAGE Publications, 2003.

Étienne Brunet (dir.). *Méthodes quantitatives et informatiques dans l'étude des textes*, Nice, 5-8 juin 1985 1986. Slatkine, Genève. en hommage à Charles Muller.

Đặng Kiều Minh Trần, Nga Hằng et Hà Loan Vương. *Giáo dục giới tính và hạnh phúc lứa đôi*. Nhà Xuất Bản Tổng hợp Hậu Giang, Ho Chi Minh Ville, 1990. Education sexuelle et bonheur conjugal.

Duc Tran Tuan, Nicolas Garcelon, Denis Delamare et Pierre Le Breux. Lexicographie et recherche d'informations medicales par croisement de langues : une approche socioterminologique d'un lexique trilingue. page 5, 2004.

Ian Walters. Dutiful daughters and temporary wives : Economic dependency on commercial sex in vietnam. *In* Evelyne Micollier (dir.) : *Sexual Cultures in East Asia : The Social Construction of Sexuality and Sexual Risk in a Time of AIDS*. Taylor & Francis, 2004.

Patrick Wilson. *Second-hand knowledge : An inquiry into cognitive authority*, volume 44 of *Contributions in librarianship and information science*. Greenwood Press, Westport, Connecticut, USA, 1983.

Ivan Wolffers. La recherche sur les maladies sexuellement transmissibles en asie face aux défis lancés par le VIH et le SIDA. *In* Marie-Eve Blanc, Laurence Husson et Evelyne Micollier (dir.) : *Sociétés Asiatiques face au SIDA*, pages 147–168. L'Harmattan, 2001.

Ivan Wolffers, Paula Kelly et Anke van der Kwaak. Sex work in times of aids : Caught between the visible and the invisible. *In* Evelyne Micollier (dir.) : *Sexual Cultures in East Asia : The Social Construction of Sexuality and Sexual Risk in a Time of AIDS*. Taylor & Francis, 2004.

Océane Hồ Đình

Caractérisation différentielle de forums de discussion sur le VIH en vietnamien et en français

Résumé

Les discours normés que produisent les institutions sont concurrencés par les discours informels ou faiblement formalisés issus du web social. La démocratisation de la prise de parole redistribue l'autorité en matière de connaissance et modifie les processus de construction des savoirs. Ces discours spontanés sont accessibles par tous et dans des volumes exponentiels, ce qui offre aux sciences humaines et sociales de nouvelles possibilités d'exploration. Pourtant elles manquent encore de méthodologies pour appréhender ces données complexes et encore peu décrites. L'objectif de la thèse est de montrer dans quelle mesure les discours du web social peuvent compléter les discours institutionnels. Nous y développons une méthodologie de collecte et d'analyse adaptée aux spécificités des discours natifs du numérique (massivité, anonymat, volatilité, caractéristiques structurelles, etc.). Nous portons notre attention sur les forums de discussion comme environnements d'élaboration de ces discours et appliquons la méthodologie développée à une problématique sociale définie : celle de l'épidémie du VIH/SIDA au Viêt Nam. Ce terrain applicatif recouvre plusieurs enjeux de société : sanitaire et social, évolutions des mœurs, concurrence des discours. L'étude est complétée par l'analyse d'un corpus comparable de langue française, relevant des mêmes thématique, genre et discours que le corpus vietnamien, de manière à mettre en évidence les spécificités de contextes socioculturels distincts.

Mots-clés : humanités numériques, web social, forums de discussion, discours institutionnel, corpus comparables, analyse contrastive, analyse de données textuelles, linguistique de corpus, sémantique, textométrie, traitement automatique du vietnamien, Viêt Nam, VIH/SIDA, santé sexuelle

Abstract

The standard discourse produced by official organisations is confronted with the unofficial or informal discourse of the social web. Empowering people to express themselves results in a new balance of authority, when it comes to knowledge and changes the way people learn. Social web discourse is available to each and everyone and its size is growing fast, which opens up new fields for both humanities and social sciences to investigate. The latter, however, are not equipped to engage with such complex and little-analysed data. The aim of this dissertation is to investigate how far social web discourse can help supplement official discourse. In it we set out a method to collect and analyse data that is in line with the characteristics of a digital environment, namely data size, anonymity, transience, structure. We focus on forums, where such discourse is built, and test our method on a specific social issue, ie the HIV/AIDS epidemic in Vietnam. This field of investigation encompasses several related questions that have to do with health, society, the evolution of morals, the mismatch between different kinds of discourse. Our study is also grounded in the analysis of a comparable French corpus dealing with the same topic, whose genre and discourse characteristics are equivalent to those of the Vietnamese one: this two-pronged research highlights the specific features of different socio-cultural environments.

Keywords: Digital Humanities, Social Web, Forums, Institutional Discourse, Comparable Corpora, Contrastive Analysis, Textual Data Analysis, Corpus Linguistics, Semantics, Textometrics, Vietnamese Natural Language Processing, Vietnam, HIV/AIDS, Sexual Health