Traductions franco-coréennes

[franco-coréen]

Cho Joon-Hyung
chojh4net@gmail.com

Résumé : L'approche quantitative nous permet d'explorer la ventilation des mots en correspondance de traduction à partir d'une segmentation des séquences textuelles dans le corpus. Avec cette méthode, nous pouvons directement comparer des mots contenus dans le corpus parallèle en langues sans parenté, bien que celles-ci n'aient aucune structure syntaxique en commun. Dans le présent article, nous présenterons comment cette méthode est applicable aux corpus parallèle en langues hétérogènes à travers l'analyse textométrique d'un couple de mots traductionnel français/coréen dans un corpus parallèle coréen-français.

Mots clés : corpus bilingues, coréen, traductologie, textométrie

Abstract : A quantitative approach enables us to explore the distribution of words in translational correspondence obtained from the segmentation of the textual sequences in a corpus. With this method, we can directly compare the words from the parallel corpus in languages without cognates, although they do not have any syntactic structure in common. In this article, we will present how this method is applicable to parallel corpora in heterogeneous languages through the textometric analysis of a couple of French/Korean translational words in a parallel French-Korean corpus.

Keywords : bilingal corpora, korean, traductology, textometrics

1 Contexte de la recherche

Les corpus parallèles bilingues, sont des corpus composés de deux textes en langues différentes dont l'un constitue la traduction de l'autre. Chacun des textes est découpé en un système d'unités de traductions qui peuvent être mises en correspondance deux à deux. Ce type de corpus est actuellement utilisé dans diverses études comparatives : stylistique comparée, lexicographie bilingue, traductologie, traitement automatique des langues, désormais TAL (cf. Véronis, 2000).

La méthode textométrique nous permet, à partir de la segmentation des séquences textuelles, d'explorer, dans chacun des volets du corpus, la ventilation des formes graphiques ainsi que les réseaux de cooccurrences autour d'une forme-pôle. Cette méthode permet, dans certains cas, d'entreprendre des analyse directes basée sur la forme graphique des unités lexicales qui entrent en rapport de traduction, écartant dans un premier temps, l'obstacle que constitue les caractéristiques syntaxiques différentes de chaque langue. Cependant, les comparaisons fructueuses entreprises à partir de textes écrits dans des langues proches deviennent plus compliquées à mettre en œuvre lorsque les bitextes associent des langues qui ne présente aucune parenté.

Dans cette étude, nous commencerons par présenter les principales caractéristiques morphosyntaxiques du coréen que nous comparerons très brièvement à celle du français (§ 2). Nous analyserons ensuite les différences quantitatives induites par ces caractéristiques pour les dépouillements de bitextes franco-coréens(§ 3). Nous envisagerons enfin l'approche textométrique des équivalences traductionnelles dans le cadre de l'étude d'un corpus parallèle coréen-français (§ 4).

2 Le coréen et son système d'écriture

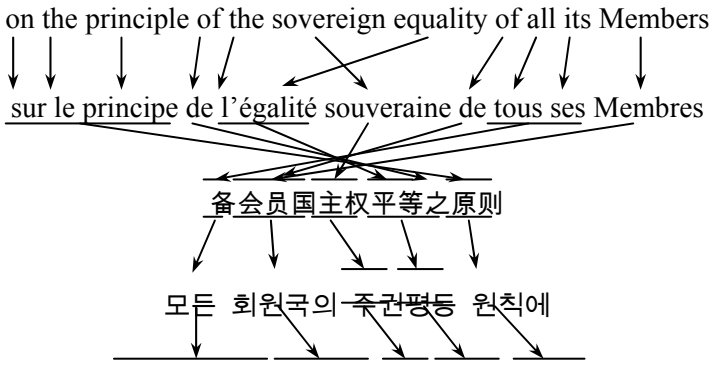
Le coréen est langue parlée en Corée par environ 72 millions de personnes. L'alphabet coréen, appelé *Hangul*, se compose fondamentalement de 24 lettres de base (14 consonnes et 10 voyelles). Mais on utilise en fait 40 lettres, si on inclut les consonnes et les voyelles doubles.

2.1 Caractéristiques linguistiques

Le coréen se distingue du chinois et du japonais, qui appartiennent à la même région culturelle et géographique par quelques caractéristiques typographiques et grammaticales.
On trouve ci-dessous, à titre d'exemple, cinq traductions, commentées grammaticalement, d'un même article de la *Charte des Nations Unies* (chapitre I, article 2.1): anglais, français, chinois, coréen et japonais.[1]

- The Organization *(sujet)* [[is based] *(verbe)* [on the principle of the sovereign equality of all its Members] *(complement)*] *(prédicat)*. (anglais)
- L'Organisation *(sujet)* [[est fondée] *(verbe)* [sur le principe de l'égalité souveraine de tous ses Membres] *(complément)*] *(prédicat)*. (français)
- 本组织 *(sujet)* [[系基于] *(verbe)* [备会员国主权平等之原则] *(complément)*] *(prédicat)*. (chinois)
- 기구는 *(sujet)* [[모든 회원국의 주권평등 원칙에] *(complément)* [기초한다] *(verbe)*] *(prédicat)*. (coréen)
- この機構は *(sujet)*, [[そのすべての加盟国の主権平等の原則に] *(complément)* [基礎をおいている] *(verbe)*] *(prédicat)*. (japonais)

Bien qu'il s'agisse de langues différentes, l'anglais et le français partagent, en plus de l'alphabet latin, des structures syntaxiques proches. En revanche, les trois dernières langues orientales possèdent des caractéristiques qui diffèrent fortement des premières et qui ne sont pas les mêmes à l'intérieur du second groupe. En premier lieu, les trois langues orientales utilisent depuis longtemps certains caractères chinois à des fins de communication. Mais ceux-ci se diffèrent dans chaque cas par la prononciation et la forme.
En coréen moderne, les caractères chinois (*hanja* caractères phonétiques, idéogrammes indispensables à l'écriture du chinois et du japonais) ont pour rôle principal d'aider à lever de nombreuses ambiguïtés sémantiques qui résultent de la transcription en *Hangul* des mots chinois.



[1] Les versions anglais/français/chinois de la Charte de l'ONU se trouvent sur le site officiel de l'ONU (http://www.un.org). Les versions coréenne et japonaise, peuvent être consultées respectivement sur les sites du Ministère des affaires étrangères et du commerce en Corée (http://www.mofat.go.kr) et sur celui du Centre d'information des Nations Unies au Japon (http://www.unic.or.jp/know/kensyo.htm).

そのすべての加盟国の主権平等の原則に

Par ailleurs, le chinois possède fondamentalement une structure phrastique qui n'est pas sans rapport avec les deux premières langues occidentales (sujet-verbe-complément), alors que le coréen et le japonais recourent à une structure phrastique inverse (sujet-complément-verbe). Par contre, le chinois suit, pour la position des attributs, un ordre identique à celui des deux langues orientales.

Le coréen fait partie, avec le japonais et le turc, des langues agglutinantes caractérisées par la combinaison des radicaux avec des particules auxiliaires qui déterminent les propriétés grammaticales des radicaux. Comme nous le verrons plus loin, ces particularités entraînent des conséquences importantes au plan quantitatif. Le grand nombre des formes différentes dans les textes coréens dépouillés en formes graphiques résulte avant tout de cette agglutination des particules auxiliaires aux radicaux qui complique singulièrement l'analyse morphologique.

2.2 Les caractéristiques typographique

Le coréen moderne utilise généralement les signes de ponctuation occidentaux pour marquer les limites de la phrase et celles de la proposition. Il utilise de surcroît quelques ponctuations coréennes comme 「 」, 『 』 pour noter les titres d'œuvres. On note aussi quelques différences entre la ponctuation du coréen et celle du français : par exemple, le coréen utilise pour les citations des guillemets anglais (“ ”) au lieu des guillemets français (« »).
Comme en français et en anglais, les mots coréens sont séparés par des espaces. Les corpus de textes coréens se prêtent donc sans grande difficulté à la segmentation automatique en mot par la sélection d'un ensemble de *délimiteurs* (signes de ponctuation et espace).

La structure syllabique originale du coréen est caractérisé par la combinaison de 2 à 3 lettres par syllabe, disposées en *carré virtuel*, on recense effectivement 11 172 combinaisons de ce type qui peuvent être identifiées à des caractères. La version actuelle de *Lexico3* n'accepte pas encore la table *Unicode*. Elle rencontre, de ce fait, des problèmes pour afficher simultanément le coréen et le français.

2.3 Encodage des textes coréens pour Lexico3

Le *couteau suisse* de *Lexico3* permet d'afficher les caractères coréens lorsqu'ils sont encodés avec la table de caractères *win-949*, basée sur l'*ASCII*, qui correspond au codage « Coréen Wansung ». Mais, dans le cas du traitement informatique d'un corpus multilingue constitué par des couples *langues occidentales /langues orientales*, les outils informatiques ont du mal à afficher simultanément les caractères correspondant aux deux systèmes d'écriture.

3 Le corpus

Pour illustrer notre propos, nous avons sélectionné un corpus de textes juridiques constitué par une série de conventions, protocoles, chartes, etc., publiés à propos du thème des droits de l'homme, par le Haut-Commissariat des Nations Unies aux droits de l'homme, le Conseil de l'Europe, la Commission Interaméricaine des Droits de l'Homme et le Bureau International du Travail.[2]

---

[2] On peut consulter les textes originaux du corpus *Droit* sur les sites suivants :
Haut-Commissariat des Nations Unies aux droits de l'homme (http://www.ohchr.org/french);
Conseil de l'Europe (http://conventions.coe.int/Treaty/FR/v3DefaultFRE.asp);

Le corpus ***Droits*** se compose de deux volets : le premier est constitué par le texte orignal en français, le second par sa traduction en coréen. Les traductions coréennes ont été officiellement publiées par la représentation de l'UNESCO en Corée et par la Commission nationale des Droits de l'Homme de Corée.[3] Signalons que les traductions coréennes n'ont pas été réalisées directement à partir des textes français mais à partir de leurs équivalents anglais. Cependant, dans la mesure où l'anglais et le français sont les deux langues officielles de ces organisations qui effectuent pour leur compte des traductions de qualité, nous avons considéré, pour cette expérience, que le bitexte franco-coréen pouvait être considéré comme un corpus parallèle de bonne qualité.

Le corpus ***Droits*** a déjà été aligné au niveau des phrases. Il ne contient aucune balise véhiculant des informations linguistiques à l'exception de quelques caractères spéciaux portant sur la structure des textes et sur leur alignement en phrases: le paragraphe (§), la phrase (#), l'indice des phrases alignées ($) et les lettres en majuscules contenues dans l'original (*)[4]. Une relecture attentive du corpus nous a permis de corriger certaines erreurs de traduction. Certains fragments absents dans l'un des volets ont été supprimés dans le volet correspondant pour constituer un corpus d'expérimentation acceptable.

français	coréen
§$# *article 1 §$# *tous les êtres humains naissent libres et égaux en dignité et en droits. $# *ils sont doués de raison et de conscience et doivent agir les uns envers les autres dans un esprit de fraternité. §$# *article 2 §$# *chacun peut se prévaloir de tous les droits et de toutes les libertés proclamés dans la présente *déclaration, sans distinction aucune, notamment de race, de couleur, de sexe, de langue, de religion, d'opinion politique ou de toute autre opinion, d'origine nationale ou sociale, de fortune, de naissance ou de toute autre situation.	§$# 제 1 조 §$# 모든 사람은 태어날 때부터 자유롭고, 존엄성과 권리에 있어서 평등하다. $# 사람은 이성과 양심을 부여받았으며 서로에게 형제의 정신으로 대하여야 한다. §$# 제 2 조 §$# 모든 사람은 인종, 피부색, 성, 언어, 종교, 정치적 또는 그 밖의 견해, 민족적 또는 사회적 출신, 재산, 출생, 기타의 지위 등에 따른 어떠한 종류의 구별도 없이, 이 선언에 제시된 모든 권리와 자유를 누릴 자격이 있다.

Tableau 1 :
Extrait du corpus ***Droits***

Le corpus ***Droits*** se compose de quarante parties qui correspondent chacune à une convention ou à un protocole. Les deux volets du corpus comptent respectivement 7 867 phrases françaises et 7 947 phrases coréennes. Le nombre de couples des phrases alignées est de 7 721, en raison des divers types de correspondances entre les phrases alignées. Pour

---

Commission interaméricaine des Droits de l'Homme (http://www.cidh.org/docdebase.htm); et
Bureau international du Travail (http://www.logos-net.net/ilo/150_base/fra/instr/afri_2.htm).

[3] Il est possible d'obtenir les textes traduits en coréen que nous avons utilisés sur les sites suivants:
UNESCO en Corée (http://www.unesco.or.kr/hrtreaty), Commission nationale des Droits de l'Homme de Corée (http://humanrights.go.kr/eng/index.jsp).

[4] Les caractères identiques contenus dans les textes originaux ont été remplacés par d'autres signes de ponctuation.

cette étude lexicométrique, les deux textes ont été segmentés en occurrences de formes graphiques afin d'obtenir une première comparaison des caractéristiques lexicales des deux langues, sur la base de ce type de segmentation.[5]

Partie	Occurrences	Formes	Hapax	Fréq. Max	Forme Max
français	214 313	7 821	2 548	12 576	de
coréen	114 006	21 068	11 732	1 642	또는

Tableau 2 :
Principales caractéristiques lexicométriques du corpus ***Droits***

Le Tableau 2 montre que la taille du volet français, mesurée en occurrence de formes graphiques, est près deux fois supérieure à celle du volet coréen. A l'inverse, le nombre des formes du volet coréen est 3 fois plus élevé que celui qui a été calculé pour le volet français. Le volet coréen compte beaucoup plus d'hapax[6] que le volet français, conséquence des particularités morphologiques propres à la langue coréenne que nous avons mentionnées plus haut. Dans le volet coréen, plus de la moitié des formes, soit 55,7 % des formes graphiques, apparaissent en tant qu'hapax, ce qui contraste avec le taux de 32,6 % calculé pour le volet français.

3.1 Accroissement du vocabulaire

L'étude de l'apparition de nouvelles formes graphiques au fil du corpus confirme les différences quantitatives entrevues plus haut entre le coréen et le français. La courbe d'accroissement de vocabulaire calculée simultanément pour les deux volets du corpus (Figure 1) montre que la croissance du vocabulaire français s'épuise plus rapidement que celle du vocabulaire coréen[7]. De plus, l'accroissement du vocabulaire français devient de plus en plus faible au fur et à mesure que l'on avance dans le texte, alors que la courbe qui correspond au texte coréen maintient une pente relativement stable. Plus que le texte français, le texte coréen voit sans cesse apparaître de nouvelles formes graphiques.

[5] Les présents travaux, y compris la segmentation du corpus, ont été effectués à l'aide du logiciel *Lexico 3*, développé par le *CLA2T* (Centre de Lexicométrie et d'Analyse Automatique des Textes), Université de la Sorbonne Nouvelle - Paris 3. (http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/).

[6] Les hapax sont les formes dont la fréquence est égale à un dans le corpus..

[7] Signalons que ce corpus particulier montre un accroissement du vocabulaire relativement constant pour un texte français. Cela est sans doute,à mettre sur le compte d'une certaine hétérogénéité des documents rassemblés dans le corpus à partir de sources diverses, bien que concernant le thème des *droits de l'homme*.

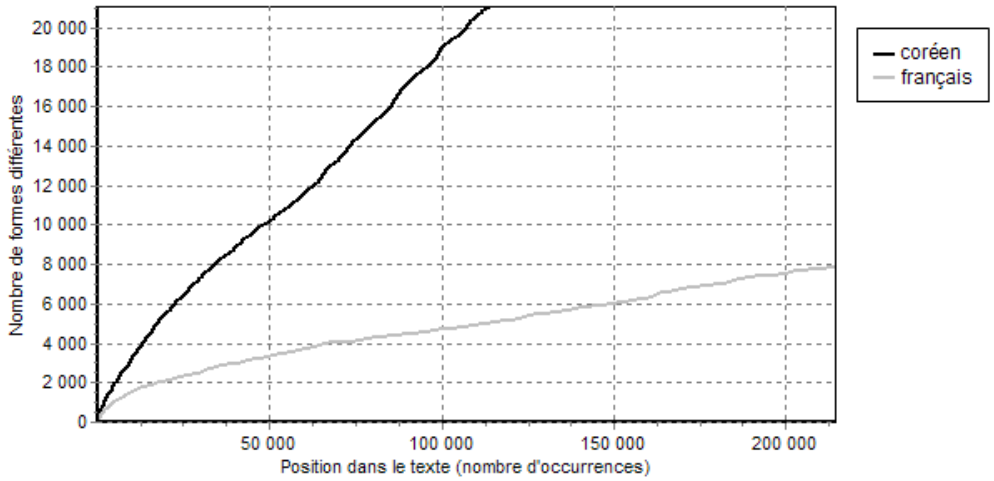


Figure 1:

Accroissement de vocabulaire dans les deux volets du corpus ***Droits***

3.2 Diagramme de Pareto

Le diagramme de Pareto, figure 2, permet de visualiser la gammes des fréquence du vocabulaire pour chacune des deux langues rassemblées dans le corpus ***Droits*** [8].

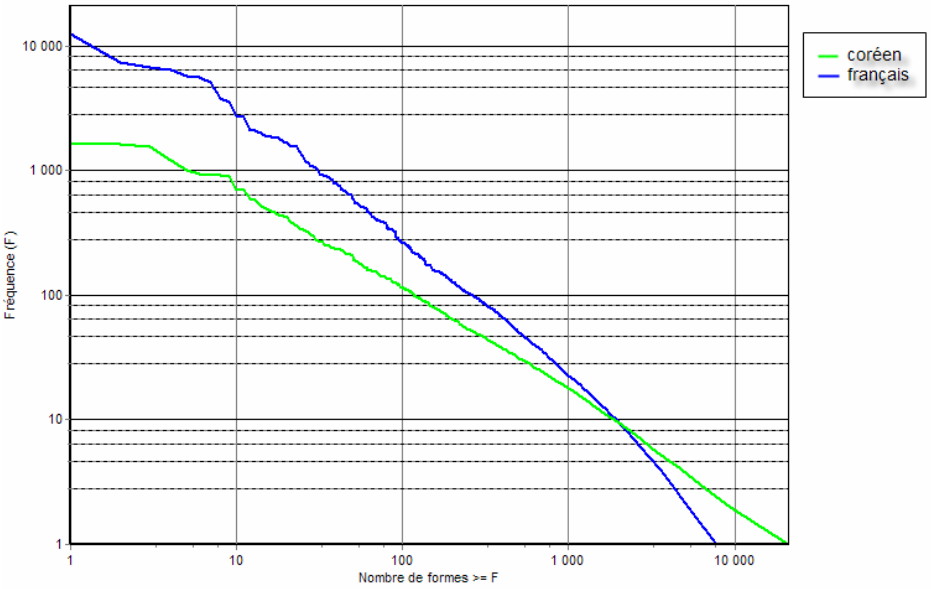


Figure 2 :

Diagramme de Pareto pour les deux volets du corpus ***Droits***

[8] « Le diagramme de Pareto fournit une représentation très synthétique des renseignements contenus dans la gamme des fréquences. […] Sur l'axe vertical, gradué selon une échelle logarithmique, on porte la fréquence de répétition *F*, qui varie donc de 1 à *Fmax*, la fréquence maximale du corpus. Sur l'axe horizontal, gradué selon la même échelle logarithmique, on porte, pour chacune des valeurs de la fréquence *F* comprises entre 1 et *Fmax*, le nombre *N(F)* des formes répétées au moins *F* fois dans le corpus. La courbe obtenue est donc une courbe cumulée. » (Lebart et Salem, 1994 : 48)

Les différences que l'on peut constater aux deux extrémités du Diagramme confirment que le français utilise plus de formes de haute fréquence et moins d'hapax que le coréen. Ainsi, le taux de formes ayant plus de 10 occurrences atteint 24,6 % pour le français, tandis qu'il est environ de 8,8 % pour le coréen. Près de 91,2 % des formes graphiques du coréen ont une fréquence inférieure à 9 occurrences.

Les résultats statistiques présentés ci-dessus conduiraient à penser que le coréen utilise un grand nombre de mots monosémiques. Comme nous l'avons déjà signalé, il s'agit sans doute d'un artefact lié à la segmentation en formes graphiques que nous avons opérée sur la base de la distinction entre caractères délimiteurs et caractères non-délimiteurs. Nous reporterons à une autre étude l'analyse de l'incidence des propriétés agglutinantes que avons mentionné plus haut sur les calculs de fréquence.

Cet obstacle lié a la segmentation en formes graphiques peut cependant être contourné, pour les analyses qui suivent, par un repérage systématique, utilisant notamment le langage des expressions régulières qui offre une possibilité de repérer les différentes compositions réalisées à partir d'un même radical.

4 Analyse des équivalences traductionnelles français/coréen

Pour l'analyse textométrique, les textes sont d'abord segmentés en occurrences de formes graphiques qui sont ensuite regroupées par type. Les corpus textuels ainsi découpés permettent d'observer directement des formes ou des séquences textuelles sans référence particulières aux structures syntaxiques particulières des langues considérées.

Les résultats obtenus à l'aide du calcul statistique à partir de textes qui entrent en correspondance de traduction, constituent des données parallèles particulièrement précieuses pour les études contrastives[9]. Les travaux lexicométriques de M. Zimina (Zimina 2000) portant sur des corpus parallèles français-anglais constitués de documents concernant la *Convention de sauvegarde des Droits de l'Homme et des libertés fondamentales*, ont illustré les possibilités de cette méthode pour contribuer à l'alignement des unités correspondantes dans les deux volets du corpus. En comparant les fréquences globales et locales des termes français et de leurs traductions anglaises, ils ont mis en évidence des similarités distributionnelles entre les répartitions des termes des deux volets. D'autre part, l'analyse multidimensionnelle des formes qui entrent en rapport de cooccurrence avec un terme-pôle a permis de mettre en lumière des similarités distributionnelles qui concernent les réseaux de cooccurrences.

Le français et le coréen sont deux langues qui n'ont aucune parenté structurelle et qui, de plus, utilisent des caractères différents. Ces différences interdisent de s'appuyer sur la ressemblance des formes graphiques pour comparer la ventilation de termes qui entrent en rapport de traduction dans les deux langues. L'approche lexicométrique est-elle susceptible d'apporter un éclairage intéressant pour l'étude des corpus parallèles coréen-français ?

Dans ce qui suit, nous montrerons l'utilité de la méthode textométrique, sur l'exemple de l'analyse d'un ensemble de formes qui entrent en rapport de traduction dans le corpus français/coréen ***Droits***.

[9] Des analyses lexicométriques de ce type ont été réalisées à propos de corpus parallèles, parmi lesquelles : Martinez et Zimina (2002), Salem (2004) et Zimina (2000, 2002, 2004a, 2004b), etc.

4.1 Etude de l'équivalence traductionnelle homme/'사람 saram'

La forme *homme* est, en français, une forme polysémique capable de désigner plusieurs concepts du générique au particulier. Dans des contextes ordinaires, cette forme est fréquemment traduite en coréen par les quatre formes : '사람 saram', '인간 ingan' ; '인류 illyu' (en fr., *humanité*) ; '남자 namja' (en fr., l'antonyme de *femme*).

A l'inverse de ce qui se passe pour les confrontations entre langues proches comme le français et l'anglais pour lesquelles les comparaisons peuvent s'appuyer sur des ressemblances typographiques (*homme/human*, *administration/administration, etc.*), les confrontations entre textes français et coréens ne peuvent s'appuyer sur des ressemblance de ce type. Pour recenser l'ensemble des équivalences traductionnelles d'un terme particulier appartenant à un des volet du corpus, il est nécessaire d'examiner, autant que possible, l'ensemble du vocabulaire de l'autre volet. On peut optimiser ce genre de recherche en s'appuyant sur la fréquence et la répartition des formes attestées dans chacun des volets du corpus.

Le nom commun français connaît deux variations grammaticales, le singulier et le pluriel. Dans le volet français du corpus ***Droits***, la forme singulière *homme* compte 1 046 occurrences et son pluriel *hommes* 41 occurrences. En coréen, le nom commun est susceptible de prendre un assez grand nombre de variations au plan de la forme graphique.

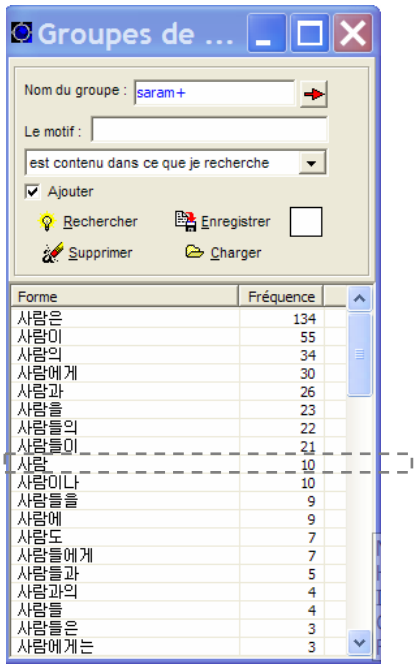


Figure 3 :
Groupe de formes *saram* dans le volet coréen du corpus ***Droits***

Le mot coréen *saram* connaît deux principales variations grammaticales *saram* (singulier) et *saram**deul*** (pluriel). Dans le volet coréen, nous nous trouvons du fait de la structure agglutinante de la langue coréenne, de nombreuses occurrences qui concernent également la forme *saram* : '사람은 saram*eun*' (F=134), '사람이 saram*i*' (F=55), '사람의 saram*ui*' (F=34), '사람에게 saram*ege*' (F=30), '사람과 saram*gwa*' (F=26), '사람을 saram*eul*' (F=23), '사람들의 saram*deurui*' (F=22), etc. Dans notre corpus (cf. figure 3), ces formes trouvent,

pour la plupart, une fréquence supérieure à celle de la forme *saram* laquelle ne compte que 10 occurrences.
Dans le cadre du dépouillement en formes graphiques à partir de la sélection de caractères délimiteurs, la variation graphique associée à un nom commun français provient de la marque éventuelle du pluriel par rapport au singulier. Dans le cas d'un texte coréen cette variation est augmentée par la combinaison possible avec différents mots fonctionnels ou particules auxiliaires[10]. C'est la raison pour laquelle le dépouillement des textes coréens génère, comme nous l'avons déjà signalé au § 2, beaucoup plus de formes et d'hapax[11] que celui des textes équivalents français.
Faute de posséder une procédure de segmentation adaptée à la morphologie de la langue coréenne, il est nécessaire, pour repérer des traductions possibles du terme *homme,* d'examiner, au delà de la chaîne de caractères isolée *saram*, les occurrences de toutes les formes contenant la séquence de caractères *saram*.
Pour venir à bout de cette tâche, le concept de *Type généralisé* (***TGen)*** va se révéler d'une grande utilité[12]. Le ***TGen*** *homme+* (désormais *homme_fr*) nous permet de rassembler les variations de la forme *homme* attestées dans le volet français du corpus (*hommes* et *hommes*). De la même façon, on constitue le ***TGen*** *saram+* en rassemblant toutes les occurrences contenant *saram*. Nous pouvons faire de même pour chacune des formes traductionnelles coréennes mentionnées ci-dessus et rassembler l'ensemble de ces occurrences du corpus coréen dans un ***TGen*** *homme_co* que nous allons comparer au ***TGen*** français *homme_fr*.

TGen	Fréquence
saram+	428
ingan+	135
illyu+	18
namja+	0
Total	581

Tableau 3 :
Fréquence des mots traductionnels coréens correspondants au type *homme_fr* dans le volet coréen du corpus ***Droits***

La comparaison des fréquences de chaque sous-groupe de formes du ***TGen*** *homme_co* révèle que, dans le corpus ***Droits***, les types *saram+* et *ingan+* sont nettement plus fréquents pour traduire le terme français *homme* (Tableau 3), Au contraire, la fréquence du ***TGen*** *namja+* est nulle dans la présente enquête. Ce résultat peut laisser penser que la forme *homme* n'est jamais utilisée comme antonyme de *femme* dans le corpus ***Droits.***
La question qui reste posée est celle de comprendre les raisons qui peuvent être à l'origine de l'écart fréquentiel entre les deux ***TGen*** *homme_fr* (F=1 087) et *homme_co* (F=581). Dans

[10] Dans nos exemples, '-은eun' (nominatif), '-이i' (nominatif), '-의eui' (génitif), '-에게ege' (datif) '-을eul' (accusatif) ; '-과gwa' (conjonction) appartiennent aux particules auxiliaires. Elles ne définissent que la position du nom dans une phrase et n'entraînent aucun changement au plan sémantique. Ce phénomène est un des traits particuliers des langues agglutinantes telles que le coréen et le japonais.

[11] Dans l'état actuel, bien que la forme coréenne ait une seule occurrence, il serait difficile d'affirmer que cette forme est un hapax. Par exemple, les formes coréennes '사람들도saramdeuldo' et '사람으로sarameuro' ont une seule occurrence dans le corpus ***Droits***. En pratique, nous recensé 12 hapax contenant '사람' dans le volet coréen du corpus ***Droits***.

[12] Le TGen (Type généralisé) est un ensemble d'occurrences sélectionnées parmi les occurrences du texte. (cf. Lamalle et Salem, 2002).

ce qui suit, nous allons chercher ces raisons à partir de l'exploration des fréquences locales de ces deux ***TGen*** dans les parties du corpus.

4.2. Comparaison des fréquences locales dans les parties du corpus

Le type *homme_fr* compte 1 087 occurrences dans le volet français du corpus. Comme nous l'avons vu, la fréquence du *TGen* correspondant dans le volet coréen, *homme_co*, est beaucoup moins élevée (581 occurrences). Pour expliquer cet écart important, il est nécessaire d'explorer les fréquences locales du couple *homme_fr*/*homme_co* dans les parties du corpus. L'exploration de la variation des fréquences locales nous permettra de comprendre les raisons de cette disparité globale.

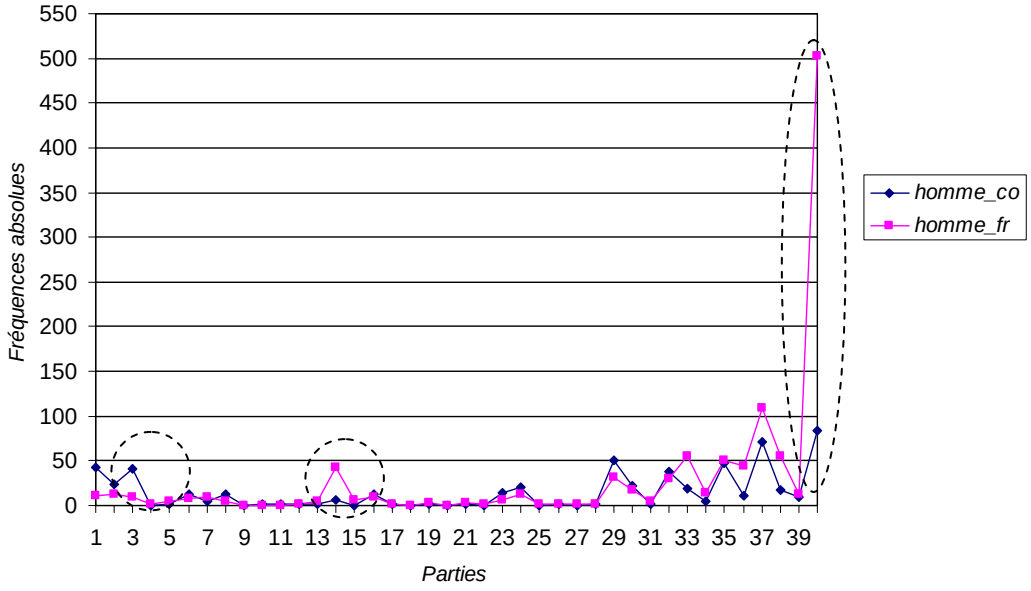


Figure 4 :
Fréquences locales des deux types *homme_fr* et *homme_co*
dans les quarante parties du corpus ***Droits***

Comme nous l'avons signalé plus haut, le corpus ***Droits*** est constitué de quarante parties. On voit, sur la Figure 4, que les deux courbes présentent un profil distributionnel similaire à quelques exceptions près. Le TGen *homme_fr* ne dépasse la cinquantaine d'occurrences que dans quelques parties. Dans les parties 37 et 40, *homme_fr* compte respectivement 109 occ. et 502 occ. Les parties 04, 09-12, 17-18, 20, 22 contiennent au maximum occurrence.
Dans le volet coréen, la fréquence locale du TGen *homme_co* dans chaque partie reste inférieurs à 50 occurrences, à l'exception des parties 37 et 40, dans lesquelles leur fréquence atteint respectivement 71 et 83 occurrences.
Les parties 4-5, 9-10, 12, 15, 17-22, 25, 27 comptent chacune une occurrence au plus. On a répertorié au tableau 4 des parties du corpus pour lesquelles la différence fréquentielle entre les deux volets est particulièrement importante.

Parties	01	03	14	33	36	37	38	40
homme_fr	11	10	42	55	44	109	56	502
homme_co	42	41	7	19	11	71	18	83

Tableau 4 :

Extrait des fréquences locales de *homme_fr* et *homme_co*
dans les parties du corpus *Droits*

Une cartographie textuelle permet de visualiser, au niveau de chaque section, la présence ou l'absence des occurrences de chacun des *TGens*. La carte des sections (Figure 5) montre des écarts dans la répartition des *TGens homme_fr* et *homme_co* entre les deux volets du corpus ***Droits***. Dans chacun des volets de la carte des sections, un carré représente une séquence (en général une phrase) alignée avec une sélection appartenant à l'autre volet du corpus[13]. Dans le volet français, la ventilation du *TGen homme_fr* est représentée par des carrés noirs ; celle du *TGen homme_co* est représentée par des carrés vert foncé dans le volet coréen.

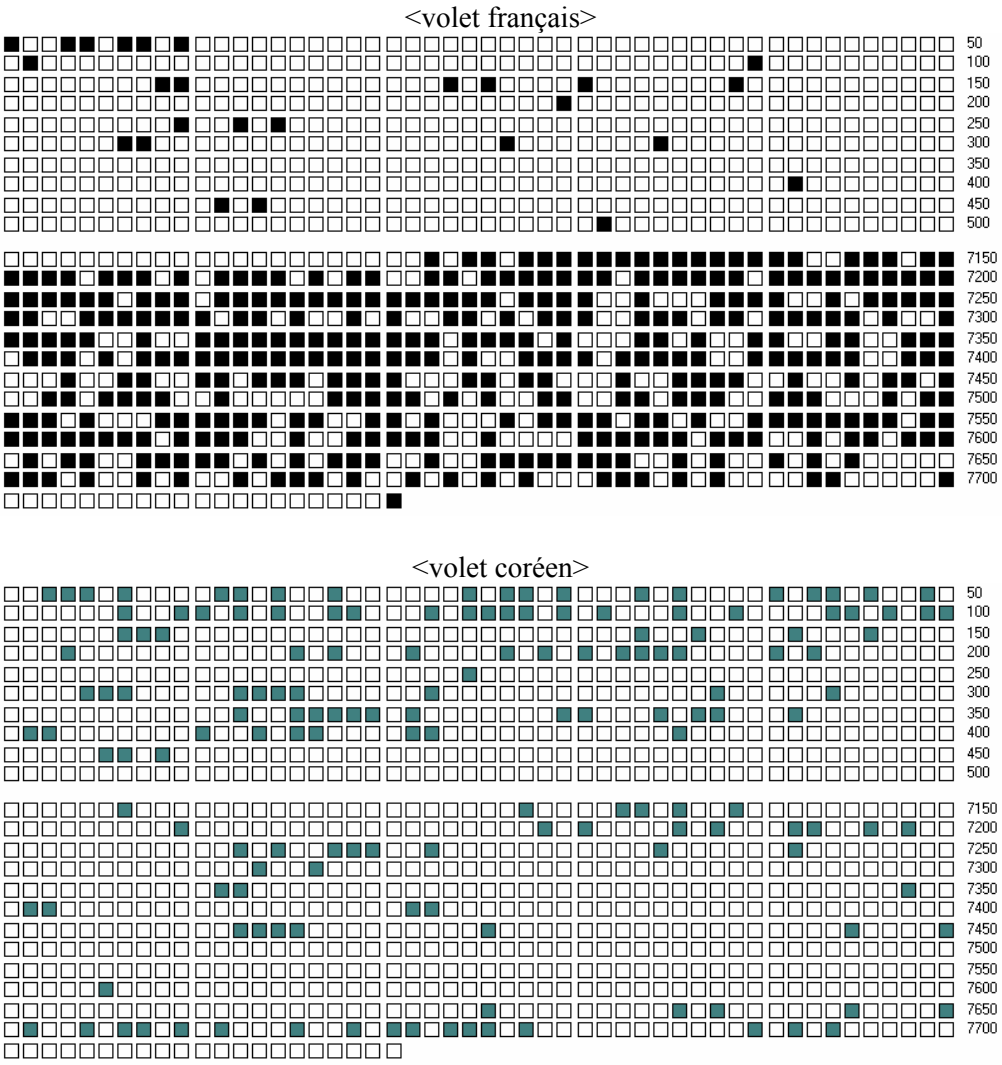


Figure 5 :
Extrait de la carte des sections ■ *homme_fr* et ■ *homme_co*
dans le corpus ***Droits***

La distribution du type *homme_co* ne s'accorde que très partiellement avec celle du type *homme_fr* (Figure 5). Une fréquence supérieure du TGen *homme_co* dans certaines parties

[13] Dans certains cas, un même carré peut contenir plus de deux phrases en fonction de la relation de correspondance avec l'autre volet.

nous amènera au constat que différentes expressions françaises : *êtres humains*, *individu*, *personne humaine* ; ainsi que des formes qui constituent des reprises anaphoriques de ces dernières, le pronom personnel *ils* ; et le pronom *chacun*, *tous* (Tableau 5) sont rendues en coréen par des formes relevant du TGen *homme_co*. L'écart des fréquences locales dans les parties 01 et 03 s'explique par la présence de ces équivalences traductionnelles.

coréen	français
§$# 모든 **사람**은 태어날 때부터 자유롭고, 존엄성과 권리에 있어서 평등하다.	§$# *tous les **êtres humains** naissent libres et égaux en dignité et en droits.
$# **사람**은 이성과 양심을 부여받았으며 서로에게 형제의 정신으로 대하여야 한다.	$# ***ils** sont doués de raison et de conscience et doivent agir les uns envers les autres dans un esprit de fraternité.
§$# **모든 사람**은 인종, 피부색, 성, 언어, 종교, 정치적 /…/.	§$# ***chacun** peut se prévaloir de tous les droits et de toutes les libertés proclamés /…/.
§$# 모든 **사람**은 생명권과 신체의 자유와 안전을 누릴 권리가 있다.	§$# *tout **individu** a droit à la vie, à la liberté et à la sûreté de sa personne.
§$# 이러한 권리는 **인간**의 고유한 존엄성으로부터 유래함을 인정하며,	§$# *reconnaissant que ces droits découlent de la dignité inhérente à la **personne humaine**,
§$# 1. 성년에 이른 **남**녀는 인종, 국적 또는 종교에 따른 어떠한 제한도 받지 않고 혼인하여 가정을 이룰 권리를 가진다.	§$# 1. *à partir de l'âge nubile, l'**homme** et la femme, sans aucune restriction quant à la race, la nationalité ou la religion, ont le droit de se marier et de fonder une famille.
$# /…/, 고등교육도 능력에 따라 모든 **사람**에게 평등하게 개방되어야 한다.	$# /…/ # l'accès aux études supérieures doit être ouvert en pleine égalité à **tous** en fonction de leur mérite.
§$# 2. 교육은 인격의 완전한 발전과 **인권** 및 기본적 자유에 대한 존중의 강화를 목표로 하여야 한다.	§$# 2. *l'éducation doit viser au plein épanouissement de la personnalité humaine et au renforcement du respect des **droits de l'homme** et des libertés fondamentales.

Tableau 5 :
Exemple des expressions françaises correspondantes au type *homme_co*
dans le corpus ***Droits***

4.3. droits de l'homme/'인권ingwon'

Plusieurs méthodes (sélection des termes cooccurrents, calcul des segments répétés) permettent de constater que, dans notre corpus, la forme *homme* est en cooccurrence étroite avec la forme *droits*. Le segment *droits de l'homme* compte 986 occurrences dans le corpus. Cependant, on ne trouve aucune occurrence de la traduction littérale du segment français qui serait constituée par l'expression '인간의 권리inganui gwolli'. Le segment *droits de l'homme* est souvent traduit par la seule forme '인권ingwon' qui compte 1 244 occurrences. Si nous tentons de localiser ces occurrences à partir des pôles de recherche *saram* et/ou *ingan*, nous ne localiserons pas les occurrences de la forme *ingwon*. L'écart important des fréquences que l'on a constaté entre les types *homme_fr* et *homme_co* dans les parties 36-38, 40 tient bien fait que la majorité des occurrences qui relèvent de la forme *homme* apparaissent dans le corpus ***Droits*** en cooccurrence avec la forme *droits*, la plupart du temps

sous la forme *droits de l'homme*. On localise les occurrences correspondantes du type *ingwon*+ dans les dernières parties du corpus (cf. Figure 6).
Dans les cas ou le segment subit une inclusion, il est à nouveau rendu par *ingan*. Par exemple, les *droits fondamentaux de l'homme* est traduit par '**인간**의 (de l'homme) 기본 (fondamentaux) **권리** (droits)' et non plus '기본 **인권** gibonjeok *ingwon*'.

On vérifie, sur la figure 6, que les distributions dans les parties du corpus du couple *droits de l'homme/ingwon*+ sont assez similaires, à quelques expressions dues à la présence de segments comme *droits fondamentaux de l'homme, etc.*.

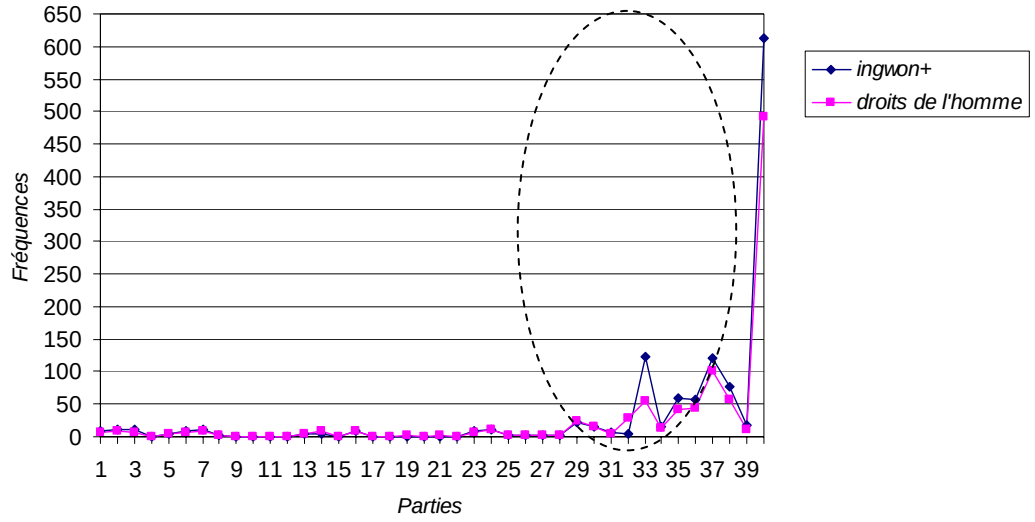


Figure 6 :
Les fréquences locales du couple *ingwon+/droits de l'homme*
dans les quarante parties du corpus *Droit*

L'écart constaté à propos de la partie 40 tient essentiellement au phénomène que nous venons de décrire plus haut. Cependant, après la prise en compte de ces variantes traductionnelles, les parties 33 et 40 montrent encore des écarts importants au plan fréquentiel dans la répartition des occurrences des deux *TGen* dont nous avons entrepris le rapprochement. Dans la partie 33, les types *ingwon*+ et *droits de l'homme* comptent respectivement 123 occurrences et 55 occurrences Dans le volet français, la fréquence du type *homme_fr* s'élève également à 55 occurrences, ce qui signifie que la forme *homme* n'apparaît dans cette partie que dans le contexte plus large *droits de l'homme*. Dans la partie correspondante du volet coréen, la fréquence locale du type *ingwon*+ dépasse largement celle de *droits de l'homme*. Cette différence provient du fait que le nom des organisations internationales contenant ce segment et leurs sigles respectifs sont fréquemment traduits en coréen par le même segment coréen.

Commission des ***droits de l'homme***	:	**인권**위원회
Commission Haute Commissariat des Nation Unies aux ***droits de l'homme*** HCDH	:	유엔**인권**고등판무관실

4.4. homme/'남자namja'

On peut fournir une explication du même type pour rendre compte de la fréquence nulle du ***TGen*** coréen *namja*+ (Voir Tableau 3). Le retour au texte permet de vérifier néanmoins la présence d'une opposition *homme/femme*. Dans les contextes où *homme* apparaît en cooccurrence avec *femme*, la plupart des occurrences coréennes apparaissent sous la forme : '남녀 namnyeo' (F=31), '남성 namseong' (F=40). *Namnyeo* est un mot composé indiquant « homme (남 nam) et femme (녀 nyeo) » et *namseong*, synonyme de *namja*, signifie, entre autres choses, un homme adulte.

La cartographie textuelle permet de représenter simultanément la localisation des occurrences du type *homme*+ et celle du type *femme*+ (*femme* 120 occurrences et *femmes* 55 occurrences). On compare ces résultats à la ventilation des occurrences du type coréen *namja*+ à partir du dépouillement de *namnyeo* et *namseong*.

Dans le volet français de la carte des sections (Figures 7 et 8), les carrés noir indiquent la présence d'une occurrence du type *homme*+ ; un carré gris celles des occurrences du type *femme*+. Les carré bicolores (noir et gris) signalent la cooccurrence au sein d'une même section des types *homme*+/*femme*+. De manière symétrique, les carrés noirs de la carte des sections réalisée pour le volet coréen indiquent la présence des occurrences du type *namja*+. La cartographie révèle que le type ◨ *homme*+*femme* génère une représentation qui ressemble considérablement à celle établie à partir du type ■ *namja*+ pour le volet coréen. Le tableau 6 rassemble quelques cas qui font exception à cette règle et qui intéresseront le traducteur

coréen	français
/…/ 매춘행위를 목적으로 하는 **남녀**의 인신매매를 방지하기 위하여 본 협약에 의하여 그들의 의무로서 요구되는 조치를 채택하거나 /…/	/…/ les mesures destinées à combattre la traite des personnes de **l'un ou de l'autre sexe** aux fins de prostitution.
/…/ 유엔**인권**위원회와 **여성**의 지위에 관한 위원회의 진정 절차에 제출된 /…/	/…/ selon des procédures spéciales devant la * commission des **droits de l'homme** et la * commission de la condition de la **femme**.

Tableau 6 :

Exemples de cooccurrences *homme & femme* ne correspondant pas au *TGen namja*+ dans le corpus ***Droits***

La différence de fréquence constatée dans la partie 14 (Tableau 4) s'explique bien par la relation de cooccurrence du couple *homme/femme*. Le retour au contexte nous montre quelques segments comme *droits de l'homme et de la femme*, *entre l'homme et la femme*, *égalité de l'homme et de la femme*. La fréquence locale du type *namja*+ dans la partie 14 est effectivement beaucoup plus élevée que dans les autres parties (F=38).

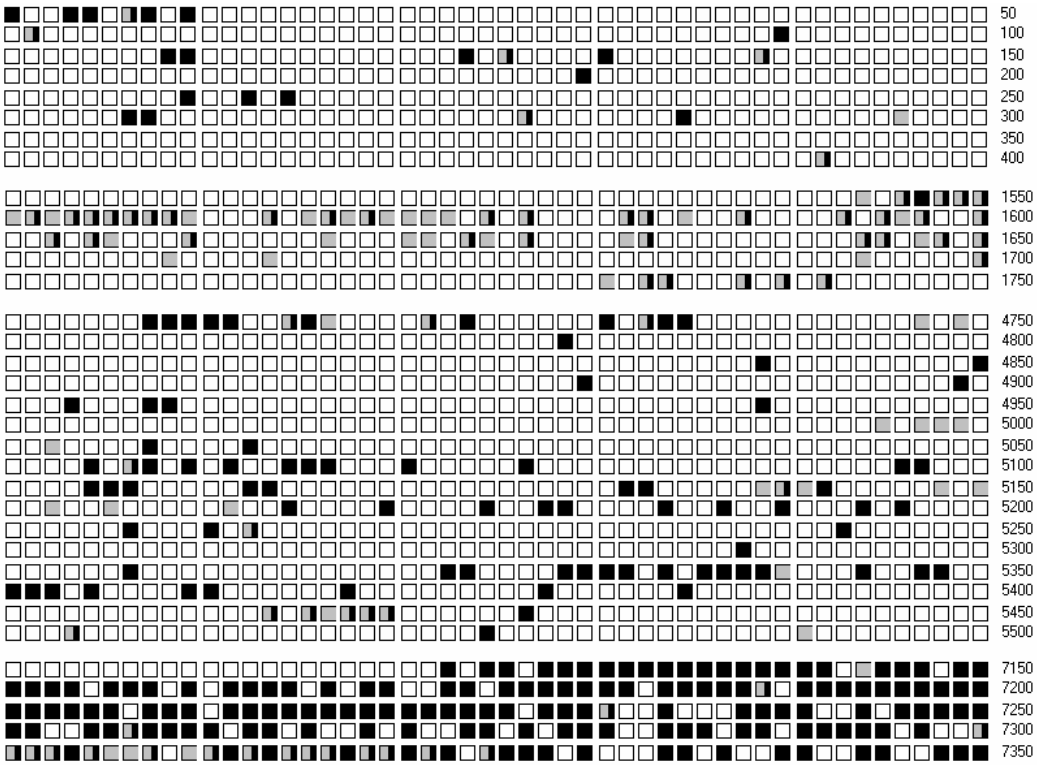


Figure 7 :

Extrait de la carte des sections ■ *homme_fr*, ■ *femme_fr* et ◨ *homme+femme* dans le volet français

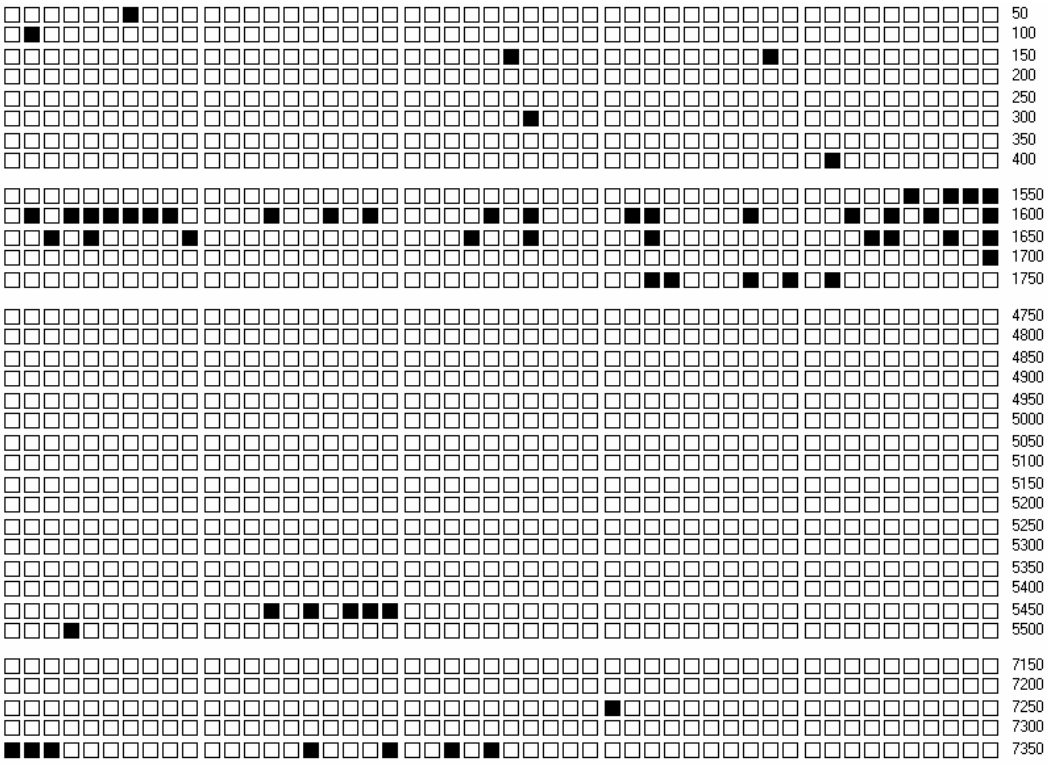


Figure 8 :

Extrait de la carte des sections ■ *namja+* dans le volet coréen

5. Conclusion

La traduction qui se donne pour objectif de transférer le sens d'un texte d'une langue à une autre mobilise des processus très complexes dans le cerveau humain. Lorsqu'il s'agit de langues n'ayant aucune parenté, la traduction des unités de la langue source vers des unités équivalentes dans la langue cible demande un travail encore plus complexe.
A partir de l'analyse lexicométrique du corpus ***Droits*** nous avons établi un certains nombre de rapports de correspondance pour le couple traductionnel *homme_fr/homme_co*. La complexité de ces rapports de traduction trouve sa source dans les différences profondes qui existent au plan linguistique et au plan culturel entre le français et le coréen. Cependant, l'observation des différences distributionnelles locales nous a permis d'établir un *schéma de traduction* du couple *homme_fr/homme_co* valable, pour le moins, à l'intérieur du corpus ***Droits***.

- homme → *saram*, *ingan*
  - si *homme* accompagne le mot *femme* → *namnyeo* ou *namseong*
- droits de l'homme → *ingwon*
  - si inclusion d'autres expressions lexicales → *ingan*
    *ex : droits **fondamentaux** de l'homme*
  - si il est suivi par le mot *femme* → *namja* ou *namseong*
    *ex : droits de l'homme et de la **femme***
- Autres expressions : *êtres humains*, *individu*, *personne humaine, chacun*, *tous* → *saram*, *ingan*

Dans cette étude, nous nous sommes attachés à la seule entité traductionnelle *homme_fr*/*homme_co* sans épuiser l'exploration des réseaux de cooccurrence autour de ces notions. Malgré ces limites, nous pensons avoir montré que l'analyse lexicométrique constitue désormais un outil extrêmement utile pour l'analyse des corpus parallèles qui concernent des langues sans parenté.

6 Références

Isabelle, P. et Warwick-Armstrong, S. (1993). « Les corpus bilingues : une nouvelle ressource pour le traducteur ». In P. Bouillon et A. Clas (Dir.), *La Traductique : études et recherches de traduction par ordinateur*, Les Presses de l'Université de Montréal, pp. 288-306.

Isahara, H. et Haruno, M. (2000). « Japanese-English aligned bilingual corpora ». In J. Véronis (Ed.), *Parallel Text Processing : Alignment and Use of Translation Corpora*, Dordrecht / Boston / London : Kluwer Academic Publishers, pp. 313-334.

Lamalle, C. et Salem, A. (2002). « Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels ». In *Actes des 6es Journées internationales d'Analyse statistique des Données Textuelles*, Saint-Malo, 2002, pp. 403-412.

Lebart, L. et Salem, A. (1994). *Statistique textuelle*. Paris : Dunod.

Martinez, W. et Zimina, M. (2002). « Utilisation de la méthode des cooccurrences pour l'alignement des mots de textes bilingues ». In *Actes des 6es Journées internationales d'Analyse statistique des Données Textuelles*, pp. 495-506.

Rastier, F. (2005). « Enjeux épistémologiques de la linguistique de corpus ». In G. Williams (Dir.), *La linguistique de corpus*, Rennes : Presses Universitaires de Rennes, pp. 31-45.

Salem, A. (1987). *Pratique des segments répétés, Essai de statistique textuelle*. Paris : Klincksieck.

Salem, A. (2004). « Introduction à la résonance textuelle ». In *Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles*, Louvain-la-Neuve, pp. 986-992.

Salkie, R. (2000). « Quelques questions méthodologiques dans l'exploitation des corpus multilingues », in M. Bilger (Ed.), Corpus : *Méthodologie et applications linguistiques*, Paris : Honoré Champion, pp. 180-195.

Shin, J. H., Han, Y. S. et Choi, K.-S. (1996). « Bilingual Knowledge Acquisition from Korean-English Parallel Corpus Using Alignment Method (Korean-English Alignment at Word and Phrase Level) ». In *Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen, Denmark, pp. 230-235.

Simard, M., Foster, G. et Isabelle, P. (1992). Using Cognates to Align Sentences in Bilingual Corpora. *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, Montreal, Canada, pp. 67-81.

Véronis, J. (2000). « From the Rosetta stone to the information society ». In J. Véronis (Ed.), *Parallel Text Processing : Alignment and Use of Translation Corpora*, Dordrecht / Boston / London : Kluwer Academic Publishers, pp. 1-24.

Zimina, M. (2000). « Alignement de textes bilingues par classification ascendante hiérarchique ». In *Actes des 5es Journées internationales d'Analyse statistique des Données Textuelles*, Lausanne, pp. 171-178.

Zimina, M. (2002). « Repérages lexicométriques des équivalences à basse fréquence dans les corpus bilingues ». In J. Véronis (Ed.), Revue électronique *Lexicometrica*, n. spécial « Corpus alignés ».

Zimina, M. (2004a). « Alignement textométrique des unités lexicales à correspondances multiples dans les corpus parallèles ». In *Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles*, Louvain-la-Neuve, pp. 1195-1202.

Zimina, M. (2004b). *Approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles*. Thèse de doctorat, Université Paris III.

7 Fonctionnalités *Lexico3* utilisées dans cette exploration

N°	*Fonctionnalité*	*Résultat*
5.5	Courbe d'accroissement du vocabulaire	*Figure 5*
5	Principales caractéristiques lexicométriques (PCLC)	*Tableau 2*
5.5	Courbe d'accroissement du vocabulaire	*Figure 1*
5.4	Diagramme de Pareto	*Figure 2*
6	Ventilation dans les parties	*Figure 4, 6*
8	Groupe de formes	*Figure 3*
7	Carte des sections	*Figure 5, 7, 8*