**Revue de Traduction et Langues Volume 22 Numéro 01/2023**
**Journal of Translation Languages** **مجلة الترجمة واللغات**
**ISSN (Print): 1112-3974** **EISSN (Online): 2600-6235**

# De la notion de corpus dans les études contrastives en langues : choix, outils et exploitation

## *The Notion of Corpus in Contrastive Language Studies: Choices, tools and exploitation*

**Paul Fonkoua**
**Université de Yaoundé I - Cameroun**
**paulfonkoua@yahoo.fr**
**0009-0005-8540-4139**
**Auguste Bayiha**
**Université de Yaoundé I - Cameroun**
**cbillongbayiha@gmail.com**
**0009-0006-2097-7902**

**Comment citer cet article :**
Fonkoua, P., & Bayiha, A. (2023). De la notion de corpus dans les études contrastives en langues : choix, outils et exploitation. *Revue de Traduction et Langues 22 (1)*, pp-pp. 333-357.

**Reçu :** 14/03/2023 **; Accepté :** 05/ 06/2023**, Publié :** 30/06/2023

***Auteur correspondant : Auguste Bayiha***

**Keywords**

*Corpus, Contrastive Studies, Languages, Sub-corpus, Source Text, Target Text*

**Abstract**

*Since past decades, Contrastive Linguistics is in search of ways for its development. Such a situation results from the fact that the scientific context nowadays is overwhelmed with the advent of new disciplines; is highly saturated by needs for interdisciplinarity; and it imposes the search of paths which have to serve as guide for forthcoming studies. As far as Contrastive Linguistics is concerned, the ways leading to its development are methodological, theoretical, and epistemological in nature. Regarding this wide scope, this study mainly focusses on the methodological aspect of the discipline. Clearly, this paper examines the notion of "Corpus" within the field of Linguistics in general and Contrastive Linguistics in particular. As a new field of study, Contrastive Linguistics needs to be grounded on solid methodological foundations as well. First of all, it delves into addressing issues of methodological tools used while contrasting languages; then, it examines the issue of corpus selection for a specific study; and finally, it proposes ways for selecting and treating data which are often overlooked by professionals and students alike. To undertake the study hereof, it resorted to the Taxonomy approach in Natural Sciences elaborated by Carl Von Linné, Candole, Charles Darwin, Willi Hennig. The use of this methodological approach purposes at providing clear definitions of terms surrounding the notion of "Corpus"; at describing the various types of corpora, at underscoring their internal organisation; and at casting light on the useful tools for the constitution of a special kind of corpus relevant for studies in Contrastive Linguistics.The tasks mentioned earlier will provide a good grasp of how the notion of "corpus" is apprehended within the framework of contrastive analysis. This work shall equally equip young scholars looking forward to specialising in this highly technical scientific niche with abilities to better find markers in a methodology of proper and full selection, constitution and exploitation of a corpus in contrastive linguistics.*

***L'œuvre est sous la licence Creative Commons Attribution-NonCommercial 4.0 International***

**Mots clés**

*Corpus ; Études contrastives en langues ; Sous-corpus ; Texte source ; Texte cible*

**Résumé**

*Dans cet article, il sera question d'examiner la notion de « corpus » dans le champ de la linguistique en général, et dans le cadre des études de linguistique contrastive en particulier pour pallier aux problèmes de sélection et de constitution des données que rencontrent les praticiens et les étudiants dans ce domaine. Pour ce faire, l'approche taxonomique, empruntée aux sciences naturelles et initiée par Carl Von Linné, Candolle, Charles Darwin, Willi Hennig, sera convoquée et elle nous permettra d'éclaircir les questions terminologiques autour de la notion de « corpus » ; de décrire chaque type de corpus en exposant son organisation interne ; et de présenter les moyens exploitables pour la constitution d'un type particulier de corpus adéquat dans les études contrastives en langues. Ce travail pourra permettre aux jeunes chercheurs qui souhaitent s'investir dans ce domaine de recherche particulièrement technique de mieux se situer dans une méthodologie du choix, de la constitution et de l'exploitation judicieuse et optimale du corpus dans le champ de la linguistique contrastive. Pour ce faire, nous commençons par une mise au point terminologique et une clarification de la notion de corpus. Par la suite, nous présentons de manière illustrative les expériences asiatique, africaine, américaine et européenne en matière de constitution et de gestions des corpus pour enfin éclairer les stratégies méthodologiques pour le traitement des corpus.*

## 1. Introduction

Les études contrastives en langues acquièr ent toute leur substance, lorsque l'on met côte à côte au moins deux systèmes linguistiques donnés. Cependant, la comparaison exhaustive de deux systèmes linguistiques dans une seule étude semble impossible à cause du fait que la langue est un agglomérat à des degrés divers, c'est-à-dire dans ses ressources, dans son fonctionnement, dans ses rapports, dans ses mécanismes, dans la flexibilité de sa manipulation par les locuteurs. Face à cette limite, les linguistes en général et les contrastivistes en particulier ont très souvent recours à l'usage des corpus.

À ce propos, deux définitions peuvent être retenues : Pour Habert (2000, p.1), « un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques et extra-linguistiques explicites pour servir d'échantillon d'emplois déterminés d'une langue ».

Selon Dalbera (2002), un corpus est un ensemble d'éléments sur lequel se fonde l'étude d'un phénomène. Il est possible qu'un corpus contienne, lors de sa constitution, des données linguistiquement hétérogènes ou un ensemble d'éléments appartenant à plusieurs langues. Cela fait appel à la notion de « sous-corpus », connu en anglais sous le vocable pluriel « subcorpora ». De ce fait, un corpus est un ensemble des sous-corpus et inversement, un sous-corpus est un élément constitutif d'un corpus. La complexité présentée par la notion de « corpus » pose davantage le problème de sélection, de constitution et même de traitement, surtout dans le champ de la linguistique contrastive.

***L'œuvre est sous la licence Creative Commons Attribution-NonCommercial 4.0 International***

Cet article se donne alors pour objectif d'examiner la portée de la notion de « corpus » dans toutes ses ampliations au sein de la linguistique contrastive. Pour essayer de réaliser cette tâche, nous nous servirons de l'approche taxonomique, empruntée aux sciences naturelles et initiée par Carl Von Linné[1], Candolle[2], Charles Darwin[3], Willi Hennig[4]. Cette approche originellement consiste à décrire la diversité des organismes vivants et à les regrouper en entités appelées taxons afin de les identifier, les décrire, les nommer et les classer. Selon l'appropriation que nous ferons dans cette réflexion, cette approche nous permettra de distinguer la diversité typologique des corpus afin d'identifier chaque type, de le décrire, de nommer les sous-catégories et de les ranger sous un type particulier de corpus.

Cet article traitera des questions terminologiques en première articulation. En seconde articulation, les corpus pour les études contrastives seront examinés ; par la suite, il sera question de statuer sur le protocole général de la constitution d'un corpus inspiré des diverses expériences. Nous proposerons également quelques stratégies méthodologiques liées au traitement d'un corpus ; nous terminerons cette réflexion en abordant les exigences à remplir dans le processus de choix d'un type de corpus en fonction d'une étude donnée.

## 2. Questions terminologiques

Les questions d'ordre terminologique sont directement liées à la nomenclature des corpus. Il semble nécessaire d'aborder les différents points de vue des auteurs sur la terminologie en ce qui concerne les corpus parce qu'elle est abondante et confuse à la limite.

Dans le domaine de la traductologie, Baker (1993, 1995) fait la distinction entre les corpus parallèles et les corpus de traduction. Elle pense que les corpus parallèles font référence à plusieurs sous-corpus monolingues ayant un même cadre d'échantillonnage tandis que les corpus de traduction contiennent les textes sources et leur traduction.

Teubert (1996) pour sa part, fait la distinction entre trois appellations : les corpus comparables, les corpus parallèles et les corpus de traduction. Selon lui, les corpus comparables regroupent les textes originaux d'une langue A et les textes authentiques dans une langue B ; les corpus parallèles contiennent les textes originaux dans une langue A et

[1] Naturaliste suédois né le 23 mai 1707 à Rashult et décède le 10 janvier 1778. Il est bien connu par ses ouvrages entre autres *Animalium specierum* (1759), *Flora suecica* (1745), *Fauna suecica* (1745), *Fundamenta Botanica* (1735).

[2] Botaniste suisse ne le 4 février 1778 à Genève et décède le 9 septembre 1841 dans la même ville. Sa contribution théorique est vue à travers l'un de ses ouvrages intitulé : *Naturalis Regni Vegetabilis : sive, enumeratio contracta ordinum generum specierumque plantarum huc usque cognitarum, juxta methodi naturalis normas digesta* (1824-1873).

[3] Naturaliste et paléontologue anglais né le 12 février 1809 à Shrewsbury et trépasse le 19 avril 1882. Ses travaux sur l'évolution des espèces vivantes ont révolutionné la biologie avec son ouvrage *L'origine des espèces* (1859).

[4] Botaniste allemand né le 20 avril 1913 et décde le 5 novembre 1976. Il est célèbre pour avoir posé les fondements de la phylogénétique en développant le paradigme cladiste. Il publie l'un de ses ouvrages intitulé *Fondements d'une théorie de la systématique phylogénétique* (1950).

*L'œuvre est sous la licence Creative Commons Attribution-NonCommercial 4.0 International*

leur traduction dans une langue B ; et les corpus de traduction se composent des textes originaux dans une langue A et les différents textes traduits dans une langue B.

Aijmer et Altenberg (1996), Granger (1996), McEnery et Wilson (1996), Baker (1999), et Hunston (2002) corroborent le point de vue de Baker (1993, 1995) et partagent sa taxonomie. Selon Ebeling (1998), les corpus parallèles font référence à au moins deux sous-corpus qui présentent une sorte de parallélisme. Pour lui, les corpus parallèles globalisent les variétés comparables de corpus. Johansson (1998) conçoit les corpus parallèles comme l'ensemble constitué des textes sources et de leur traduction ou de plusieurs sous-corpus monolingues ayant un même cadre d'échantillonnage. Plus tard, Johansson (2007, p. 9) définit les corpus comparables comme regroupant les textes originaux comparables en différentes langues.

Pour Altenberg et Granger (2002, p. 8), les corpus de traduction comprennent les textes originaux et leur traduction vers une ou plusieurs langues pendant que les corpus comparables ne contiennent pas de traductions, mais des textes dans au moins deux langues qui sont comparables au niveau du genre, de la date de publication, etc.

Enfin, Nádvorníková (2010) considère les corpus parallèles comme l'ensemble constitué des textes originaux et de leur traduction et les corpus comparables comme étant uniquement composés de textes originaux regroupés sur la base d'un critère de comparabilité.

Dans le cadre de cet article, les diverses définitions des corpus comparables proposées par les auteurs ci-dessus mentionnés sont admises tandis que la nomenclature « corpus parallèles » est considérée comme synonyme à celle de « corpus de traduction » ; par conséquent, elles seront utilisées de manière interchangeable. À cet effet, l'équivalence terminologique peut être établie de la manière suivante :

**Tableau 1.**

*Équivalence terminologique sur la nomenclature des corpus*[5]

| **Point de vue de cet article sur les types de corpus** | **Teubert (1996)** | **Aijmer et Altenberg (1996)** | **Baker (1993)/McEnery et Wilson (1996)** | **Johansson (1998)** | **Ebeling (1998)** | **Altenberg et Granger (2002)** | **Nádvorníková (2010)** |
|---|---|---|---|---|---|---|---|
| *Comparables* | Comparable | Parallèle | Comparable | Parallèle | Parallèle | comparable | comparable |

[5] Ce tableau a pu être dressé lorsque nous avons fait le rapprochement entre les divers points de vue sur la terminologie liée à la nomenclature des corpus et à leur définition.

*L'œuvre est sous la licence Creative Commons Attribution-NonCommercial 4.0 International*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *De traduction* | Type comparable | De traduction | Parallèle<br>Parallèle | Parallèle<br>/ | /<br>/ | De traduction<br>/ | Parallèle |
| *ou parallèle* | de traduction | / | Parallèle | / | / | / | de traduction |

### 3. Les corpus pour les études contrastives

Les corpus pour les études contrastives sont usuels dans d'autres champs de recherche à l'instar de: l'enseignement des langues, la traduction, l'enseignement de la traduction, la traductologie[6], l'ingénierie des langues, la lexicographie bilingue, extraction terminologique. Ces corpus ont déjà été nommés selon leurs types et même selon leurs sous-types par plusieurs auteurs avant aujourd'hui. La quintessence des points de vue nous a permis de dresser un schéma résumant à des niveaux divers la conception des corpus en linguistique contrastive :

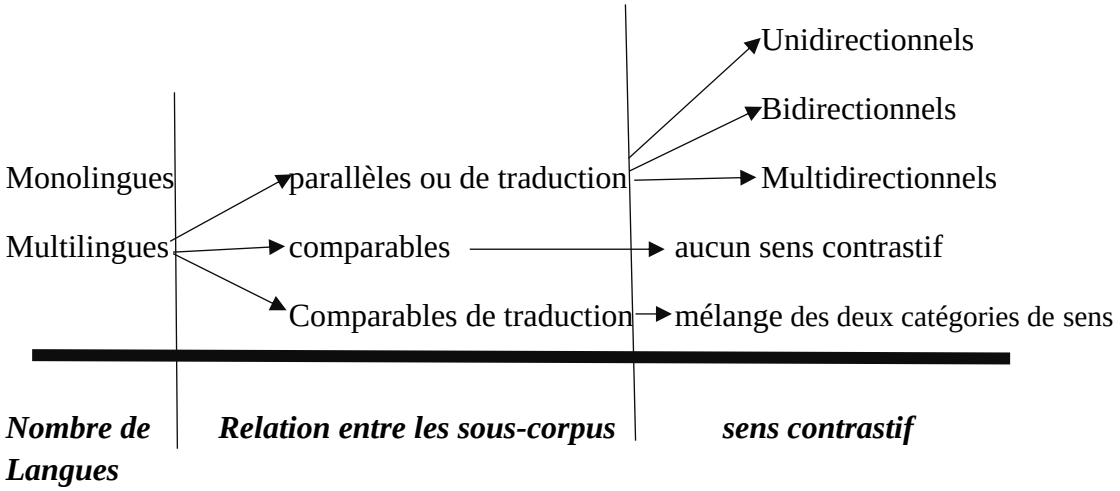



***Figure 1.*** Schéma synoptique sur la distinction des corpus[7]

[6] L'usage des corpus représente le point d'intersection entre la traductologie et la linguistique contrastive. L'ouvrage de Rudy Loock intitulé *La traductologie de corpus* (2016), publié dans *Presses Universitaires du Septentrion*, peut être consulté à ce sujet. Pour une consultation rapide et panoramique, Elizabeth C. Saint (2018) en donne le compte rendu. La distinction entre ces deux disciplines ont fait couler beaucoup d'encre à plusieurs auteurs à l'instar de Benaissa (2013) dans son article intitulé « Linguistique contrastive et traductologie : une relation dyadique ».

[7] Ce schéma a été inspiré d'Altenberg et Granger (2002), Rawoens (2008) et Nádvorníková (2017).

***L'œuvre est sous la licence Creative Commons Attribution-NonCommercial 4.0 International***

Au regard de ce schéma, trois niveaux de différences peuvent être recensés : le nombre de langues, la relation entre les sous-corpus et le sens contrastif.

### *3.1. Le nombre de langues en comparaison*

Le nombre de langues en comparaison permet la distinction entre les corpus monolingues et les corpus multilingues.

Les corpus monolingues sont des textes dans une seule langue. Par extension, ils peuvent renvoyer au produit de la traduction intralinguale. En d'autres termes, les corpus monolingues peuvent aussi être une réécriture d'un texte original dans la même langue que celle de l'original. Ce genre de corpus ne relève pas du domaine de la linguistique contrastive.

Les corpus multilingues par contre regroupent des sous-corpus dans plusieurs langues. Goeuriot (2009) rappelle qu'« il existe dans la littérature d'autres appellations pour les corpus multilingues. Fung et McKeown (1997) parlent de corpus non-parallèles ou de corpus parallèles bruités, et Rapp (1995) de corpus non-liés ». Ils possèdent plusieurs variantes : la variante bilingue, la variante trilingue, etc. Ce type de corpus est le domaine de prédilection des études contrastives en langues et la variante bilingue semble être la plus exploitée par les chercheurs. Par ailleurs, la structure interne du genre multilingue des corpus présente une certaine complexité causée par les relations diverses qui sont entretenues par ses sous-corpus de langues différentes.

### *3.2. La relation entre les sous-corpus*

La relation entre les sous-corpus d'un corpus multilingue révèle la distinction entre les corpus comparables, les corpus de traduction et les corpus comparables de traduction.

#### *3.2.1. Les corpus comparables*

Ils sont des textes originaux en plusieurs langues rassemblés sur la base d'une même variable qui peut être le genre, le type, le thème, l'auteur, la date de publication, le public cible, etc. Ebeling (1998, p. 2) fait ressortir 4 variétés comparables de corpus malgré le fait qu'il a une appellation différente des corpus comparables :

- Les sous-corpus représentant les langues ou les dialectes différents avec une *même* quantité de données, extraites des sources comparables ;
- Les sous-corpus exprimant le *même* contenu dans les langues ou dialectes différents ;
- Les sous-corpus exprimant le *même* effet dans les langues et les dialectes différents ;
- Un sous-corpus en tant que texte original et un autre sous-corpus en tant que texte traduit dans la *même* langue (la traduction intralinguale).

*L'œuvre est sous la licence Creative Commons Attribution-NonCommercial 4.0 International*

Cette dernière variété est considérée par d'autres auteurs à l'instar de Zanettin, (1998), Culo et al. (2008) comme un corpus comparable monolingue (constitués de deux ensembles de textes, l'un composé de textes écrits dans une langue et l'autre composé de textes traduits dans cette même langue).

Une distinction supplémentaire est faite par Goeuriot (2009, p. 12) au sein des corpus comparables entre la catégorie généraliste et la catégorie spécialisée non dans le cadre du sens contrastif, mais dans leur contenu. Quelles que soient les taxonomies qu'on pourra élaborer au sein des corpus comparables, ils doivent satisfaire à la notion de « comparabilité » qui est fondamentale pour de tels corpus. À ce propos, Maia (2003) pense que la comparabilité est assurée par la forme et le contenu du corpus. La forme touche le nombre de textes et le nombre de mots de l'ensemble du corpus ainsi que la nature des différents textes (format, images, etc.).

Le contenu concerne la structure des textes, le registre de langue, les sujets abordés, etc. Goeuriot (2009, p. 19) perçoit les critères de comparabilité comme « les caractéristiques communes aux textes » et pense qu'ils permettent de garantir une certaine homogénéité dans le corpus. Selon Déjean et Gaussier (2002), ces critères sont qualitatifs et quantitatifs : le premier groupe intègre le genre, l'auteur, la période, le médium, etc. tandis que le second englobe les mesures de fréquence de certains traits linguistiques. Dans ce sens, Goeuriot (2009, p. 19) peut affirmer :

> Pour les corpus comparables de langue générale, le choix se porte souvent vers le genre, la période, le médium, tandis que pour les corpus spécialisés, le choix se porte plus souvent sur le thème, le genre, le type de discours…La comparabilité dépend des caractéristiques communes aux textes : plus ils ont des caractéristiques communes, plus ils sont comparables.

Il est aussi nécessaire de noter que la comparabilité peut être calculée. Nous nous gardons d'aborder cet aspect dans ce travail, mais pour plus d'éclairage sur cette question, cf. Déjean et Gaussier (2002). En guise d'exemple, LOB[8] et Brown[9] sont considérés comme des corpus comparables par Schlamberger (2002).

[8] Lancaster-Oslo/Bergen Corpus est constitué d'une collection d'un million de mots des textes écrits en anglais britannique, compilés dans les années 1970 avec la collaboration de l'Université de Lancaster, de l'Université de Oslo et de *Norwegian Computing Centre for the Humanities* à Bergen.

[9] The **Brown University Standard Corpus of Present-Day American English** (or just **Brown Corpus**) is an electronic collection of text samples of American English, the first major structured corpus of varied genres. This corpus first set the bar for the scientific study of the frequency and distribution of word categories in everyday language use. Compiled by Henry Kučera and W. Nelson Francis at Brown University, in Rhode Island, it is a general language corpus containing 500 samples of English, totaling roughly one million words, compiled from works published in the United States in 1961.

*L'œuvre est sous la licence Creative Commons Attribution-NonCommercial 4.0 International*

### *3.2.2. Les corpus parallèles*

Ils sont encore appelés les corpus de traduction. Il s'agit des textes sources dans une langue L1 et leurs traductions dans une autre langue L2. Pour Nádvorníková (2010 :10), les corpus serviront à des fins scientifiques (recherche linguistique contrastive, élaboration de livres de grammaire contrastive, rédaction de mémoires et de thèses) et pédagogiques (enseignement de langues, traductologie). Les corpus parallèles permettront également de créer des outils linguistiques (dictionnaires) sophistiqués ainsi que des outils logiciels (visant, par exemple, la promotion de la TAO, Traduction Assistée par Ordinateur).

Plusieurs critiques sont portées à l'endroit de ce type de corpus. Certains auteurs comme Teubert (1996), Déjean et Gaussier (2002), Maia (2003) pensent que les corpus parallèles sont inappropriés lorsqu'il s'agit des termes et des unités lexicales parce que le traducteur est souvent influencé par la langue source. Les textes traduits ne sont pas représentatifs d'une langue et par conséquent, ils ne doivent pas constituer un corpus. C'est la raison pour laquelle les textes traduits ne font pas partie de certains corpus nationaux des pays anglo-saxons. En guise d'illustration, nous pouvons citer le BNC (British National Corpus). La difficulté que ces auteurs posent au sujet des corpus parallèles est l'indisponibilité des traductions dans certaines langues. Ces critiques sur les corpus parallèles sont réfutées par d'autres auteurs à l'instar de Confiant (2003, p. 6) qui déclare que :

> La traduction semble être la voie royale pour que les langues accèdent à ce que Jean Bernabé appelle la souveraineté scripturale. C'est qu'elle oblige la langue à sortir de son cocon, des réalités qu'elle a l'habitude de désigner et la force à actualiser ses potentialités cachées.

Un autre avantage de ce type de corpus est que les textes sources et les textes traduits peuvent être alignés jusqu'au niveau du mot.

### *3.2.3. Les corpus comparables de traduction*

Ces corpus naissent de la combinaison ou de l'adjonction des deux précédents types. Selon Nádvornikóvá (2017, p. 15), ils *sont composés d'un sous-corpus de textes traduits et d'un autre sous-corpus (comparable quant à la taille et la composition) contenant des textes originaux de la même langue.* Ils représentent l'équilibre entre le pôle comparable et le pôle parallèle des corpus. Melnikova et al. (2009) présentent le fonctionnement interne des corpus comparables de traduction en ces termes :

> Tandis que les corpus parallèles permettent d'identifier, de manière directe des équivalents fonctionnels, entre unités et constructions, les hypothèses émises doivent être vérifiées sur des corpus comparables de grande taille, mieux à même de fournir des données sur le plan fréquentiel.

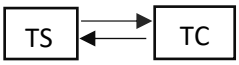
*L'œuvre est sous la licence Creative Commons Attribution-NonCommercial 4.0 International*

Les corpus comparables de traduction sont très rares. Mais selon McEnery et Xiao (2007), ENPC (English-Norwegian Parallel Corpus) et EMILLE corpus (Enabling Minority Language Engineering) en sont des exemples.

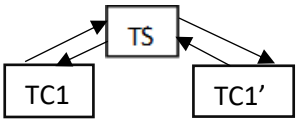
### *3.3. Le sens contrastif*

Le sens contrastif indique le positionnement d'un sous-corpus par rapport à l'autre. À ce niveau, il est nécessaire d'élaborer sur le type parallèle qui présente trois sens :

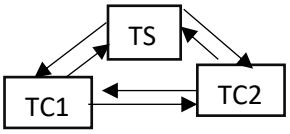
- Dans *le sens unidirectionnel*, un seul sous-corpus jouit du statut du « texte source » (le texte-directeur de la comparaison) alors que l'autre, du statut unique de « texte cible ».
- *Le sens bidirectionnel* fait à ce que les sous-corpus jouissent tour à tour du statut de « texte source » et de celui de « texte cible ». De plus, quelques degrés bidirectionnels des corpus de traduction peuvent être remarqués :
- **1er degré** marque une réciprocité comparative entre le texte source (TS) et le texte cible (TC) comme l'indique le schéma suivant :




- **2e degré** met en exergue deux traductions en une même langue d'un même texte source. Le sens bidirectionnel ne sera plus entre les textes traduits dans la même langue, mais entre ces derniers et le texte source.




- **3e degré** indique le sens entre deux traductions en différentes langues d'un même texte source. Le sens bidirectionnel est vu à tous les niveaux, c'est-à-dire entre le « texte source » et les textes cibles ou entre les textes cibles (Hasselgard, 2010)[10]. Dans ce cas, les sous-corpus jouissent tour à tour des statuts de « texte source » et de « texte cible » par rapport aux autres.




- Le *sens multidirectionnel* des corpus de traduction suppose une « langue pivot » autour de laquelle rotent plusieurs autres langues. Une telle situation permet une étude contrastive entre la « langue pivot » et les multiples autres langues qui

[10] Fait la distinction suivante sur les possibilités de comparaison des corpus de traduction bidirectionnels: « *(1)* originals and their translations; *(2)* original texts of the same type in both languages; *(3)* translations in both languages; *(4)* original and translated texts in the same language ».

*L'œuvre est sous la licence Creative Commons Attribution-NonCommercial 4.0 International*

gravitent autour d'elle. Le corpus tchèque *InterCorp* (Nádvorníková 2017) et la version étendue du ENPC (Johansson, Ebeling et Oksefjell 2002 : figure 2) nous offrent un tel aperçu avec le tchèque et l'angla is comme « langue pivot » respective.

## 4. Sur la constitution du corpus

Le questionnement sur la constitution des corpus ne date pas d'aujourd'hui. Il représente le résultat d'une longue expérience à travers plusieurs coins et recoins dans le monde avec des textes appartenant à des langues variées. L'on peut se référer à Xiao (2008) et au site *https://sites.psu.edu/calpercorpusportal/resources/l1-corpora* pour visualiser les détails sur les corpus célèbres. Pour trouver une voie de sortie, les expériences asiatique, européenne, américaine et africaine seront examinées. Ceci nous permettra de ressortir les éléments communs ou de faire une synthèse qui pourra servir pour la confection des corpus.

### *4.1. L'expérience asiatique*

Pour l'expérience asiatique, nous retiendrons l'expérience chinoise et l'expérience sur les langues indiennes.

#### *4.1.1. L'expérience chinoise*

Le PCMW (Parallel Corpus of Medical Works) anglais – chinois est un corpus parallèle constitué de 18 ouvrages médicaux en anglais et de leur traduction en chinois dont 15 sont publiés par *Shanghai Scientific and Technological Literature Publishing House* et les 3 autres par *People's Medical Publishing House*, *China Medico-Pharmaceutical Science &Technological Publishing House* et World *Book Publishing Corporation Beijing Company*. Trois (3) facteurs ont été pris en considération pour la collecte des données : la qualité des textes collectés, le but de la confection du corpus et la disponibilité des textes. Parallèlement, le PCMW présente certaines qualités à savoir sa capacité large qui assure l'adéquation des ressources et sa représentativité, c'est-à-dire qu'il couvre les principaux champs du domaine médical.

Les données ont été converties en format électronique à l'aide du scanner OCR[11] et ensuite sauvegardées en tant que document Microsoft Word 2003. Un prétraitement manuel est fait pour éliminer les informations inutiles sur le document. La préparation et le traitement du texte sont réalisés en quatre étapes : la relecture du texte à traiter, la séparation des sous-corpus, l'alignement automatique (par *Self-Coded Python Programs*) des phrases étiquetées par les marques XML, et la vérification de l'alignement manuel.

[11] La reconnaissance optique de caractères (OCR) est une opération qui consiste à extraire du texte d'une image de page. L'image de page est la représentation électronique d'un texte et, le cas échéant, d'autres éléments tels que des en-têtes et des illustrations, et elle est obtenue par numérisation d'un document papier ou par ouverture d'un fichier image électronique.

*L'œuvre est sous la licence Creative Commons Attribution-NonCommercial 4.0 International*

Pour finir, le PCMW possède 54522 paires de phrases et plus de 2,5 millions de caractères chinois et de mots anglais (XiaoXiao et Shili, 2011).

#### *4.1.2. L'expérience sur les langues indiennes*

Selon McEnery et al. (2000), le corpus EMILLE est un corpus comparable *Bengali – Gujarati – Hindou – Panjabi – Singhalese – Tamil – Urdu.* Il a été conçu dans le cadre de *The MILLE Project* financé par EPSRC pour 18 mois à l'Université de Lancaster. Le corpus EMILLE contient les sous-corpus en langues écrites d'au moins 9 millions de mots et les sous-corpus oraux d'au moins 500.000 mots par langues.

Parmi les multiples objectifs de ce projet se trouve celui du développement des corpus. En ce qui concerne le corpus EMILLE, les données monolingues en langue écrite ont été collectées de *Lake House Printers* à Sri Lanka, *The Dept. of Health, The Sikh Parliament* à Birmingham, des journaux communautaires dans le Royaume-Uni, des journaux indiens en ligne, des brochures de santé, des textes religieux et des romans. Celles-ci ont été annotées selon les recommandations standards de codage des corpus (CES/TEI). Par ailleurs, les données orales ont été collectées des communautés à travers le Royaume-Uni ; puis stockées dans des disquettes ; enfin transcrites par des spécialistes. Toutes les données ont été alignées à l'aide de l'algorithme d'alignement des phrases, version Gale&Church (1993) de McEnery & Oakes (1996). Les vérifications sont faites par les locuteurs natifs grâce aux logiciels d'alignement de Langlais et al. (1998a, b).

### *4.2. L'expérience européenne*

Pour l'expérience européenne, nous retiendrons l'expérience russe, l'expérience norvégienne et l'expérience tchèque.

#### *4.2.1. L'expérience russe*

Le RNC (Russian National Corpus), comme le présente Grishina (2007), est un corpus ouvert qui était initialement un corpus de nature comparable de 125 millions de mots environ, exploitable depuis 2003 et accessible à partir du site : *www.rescorpora.ru* . Il contenait tous les genres et types de textes du XIXe siècle au XXIe. Il a été réalisé par des chercheurs des instituts variés et des universités russes à Moscow et à Saint-Pétersbourg. Les données qui le constituent sont à la fois écrites et orales : les premières relèvent du genre fictionnel et des genres non-fictionnels alors que les dernières, des films russes entre autres sources.

Selon Kakhilina et al. (2014), le RNC est annoté à quatre niveaux majeurs : le niveau métatextuel, le niveau morphosyntaxique, le niveau des accents, le niveau lexico-sémantique. Selon Mikhailov et Cooper (2016), en 2015 les corpus parallèles (arménien, biélorusse, bulgare, anglais, français, allemand, italien, latvien, polonais, espagnol et ukrainien) d'environ 24 millions de mots y ont été inclus. Ces textes ont été alignés au niveau des phrases.

*L'œuvre est sous la licence Creative Commons Attribution-NonCommercial 4.0 International*

### *4.2.2. L'expérience norvégienne*

ENPC (English-Norwegian Parallel Corpus) est considéré comme parallèle pour certains auteurs et comme un corpus comparable de traduction par McEnery et Xiao (2007). Ce projet a été financé par la Faculté des Arts et le *Department of British and American Studies* de l'Université de Oslo. L'équipe chargée du pilotage du projet était constituée de Stig Johansson, Knut Hofland, Jarle Ebeling et Signe OKsefjell (1999/2002). ENPC est constitué de 100 textes originaux (fictionnel et non-fictionnel) en anglais et 100 autres en norvégien et de leurs traductions respectives en norvégien et en anglais avec un total général de 2,6 millions de mots.

Les textes anglais incluent les productions de plusieurs variétés d'anglais tandis que pour les textes traduits, la priorité est mise sur ceux qui ont été publiés. Ces textes ont été classifiés en fonction du genre et des domaines. Après la digitalisation du corpus, les informations inutiles sont supprimées et les textes sont alignés à l'aide du *Translation Corpus Aligner*, un programme pour l'alignement automatique des phrases. Les textes sont étiquetés, relus et corrigés pour être mis dans la base de données. Le caractère ouvert du ENPC permet à ses auteurs d'envisager son expansion linguistique en incluant le suédois, le hollandais, le finnois, le portugais, l'allemand. Il convient de noter que, ENPC est un projet sœur du ESPC (English-Swedish Parallel Corpus) selon Altenberg et Aijmer (2000).

### *4.2.3. L'expérience tchèque*

Le projet de corpus parallèles *Intercorp* (*http://www.korpus.cz/intercorp/)* a été lancé par l'Institut du Corpus national tchèque (*http://ucnk.ff.cuni.cz/)* en 2005. Il représente le prolongement naturel du large corpus unilingue, le *Corpus national tchèque*. Le projet *Intercorp* est subventionné par le Ministère de l'Éducation nationale de la République Tchèque et l'accès aux corpus est gratuit. En ce qui concerne la composition du corpus, les textes littéraires sont les plus nombreux (compte tenu de leur grande disponibilité en versions parallèles). Cependant, les coordonnateurs tâchent d'inclure dans le corpus également les textes des autres genres, principalement issus des domaines scientifique et journalistique [Nádvorníková (2010)]. *InterCorp*, créé par l'Institut du Corpus national tchèque, contient actuellement des textes en 39 langues, y compris des langues romanes (le français, le portugais, l'espagnol, l'italien, le catalan et le roumain), au total 1 597 462 625 mots. La langue tchèque constitue la langue pivot du corpus (chaque texte est aligné par rapport à la version tchèque). Mais dans les recherches sur corpus, il est possible de mettre le tchèque de côté et d'effectuer des recherches directement sur les langues choisies (par exemple sur le français et l'espagnol). Étant donné le caractère synchronique du corpus, seulement les textes écrits après la Seconde Guerre mondiale sont admis. Le corpus contient actuellement plus d'un milliard et demi de mots (au total pour les 39 langues), mais il ne cesse de croître et l'Institut du Corpus national tchèque lance chaque année une nouvelle version élargie et améliorée (*voir http://ucnk.ff.cuni.cz/intercorp/?lang=en*) [Nádvorníková (2017)].

*L'œuvre est sous la licence Creative Commons Attribution-NonCommercial 4.0 International*

### *4.3. L'expérience américaine*

Pour l'expérience américaine, l'expérience des États-Unis d'Amérique et l'expérience canadienne seront retenues.

#### *4.3.1. L'expérience aux États-Unis d'Amérique*

ANC (American National Corpus) est un corpus comparable qui s'apparente au BNC (British National Corpus) selon Reppen et Ide (2004). Le projet ANC a été proposé par Fillmore, Ide, Jurafsky et Macleod en 1998. Il a été entrepris en coopération avec les maisons de publication, les organisations et les institutions académiques aux États-Unis. ANC contient les textes à partir de 1990. Ils sont composés de 55% d'ouvrages (non-fictionnels 41% et fictionnels 14%), 20% de journaux et magazines, 10% de textes électroniques et 5% de textes divers. De plus, certains corpus de spécialité taxés de « satellite » y sont aussi inclus si bien que, ANC contenait 11,5 millions de mots en octobre 2003. ANC est codé par XML conformément au codage standard de corpus XML ; l'annotation de celui-ci est faite à plusieurs niveaux parmi lesquels les parties du discours et les traits rhétoriques. Pour les parties du discours, elle a été faite spécialement à l'aide de « Bibber tagger ». ANC est disponible à partir du site : *http://americannationalcorpus.org*.

#### *4.3.2. L'expérience canadienne*

Le corpus *Hansard* canadien est dit parallèle parce qu'il est constitué des débats du parlement canadien publiés en langues officielles français et anglais. Selon Xiao (2008), Il contient des thématiques et des styles divers. Ce corpus existe en plusieurs versions : « USC version » comprend 1,3 million paires de fragments de texte alignés du 36e parlement canadien (1997-2000) avec 2millions mots pour chaque langue ; LDC (Language Data Consortuim) est une collection de textes parallèles Hansard publiée en 1995 ayant une portée temporelle de mi-1970 à 1988.

### *4.4. L'expérience africaine*

Le corpus SAWA est un corpus parallèle anglais-swahili élaboré en 2009 par De Pauw, Waiganjo Wagacha et De Schryver dans le cadre d'EACL[12]pendant l'atelier sur les technologies linguistiques pour les langues africaines. Les données (environ un demi-million de mots pour chaque langue) ont été collectées de plusieurs sources : le Nouveau Testament, Coran, la Déclaration des Droits de l'Homme, Kumasi.org, le sous-titrage des films, les rapports d'investissement, un traducteur local. Puis, la projection d'annotation a été initiée pour vérifier si la relation entre deux mots dans la langue source est similaire entre leur correspondance dans la langue cible : ceci étant une vérification préalable pour l'alignement. Ensuite, l'alignement est fait manuellement au niveau des phrases sur

[12] The European Chapter of the Association for Computational Linguistics (EACL).

*L'œuvre est sous la licence Creative Commons Attribution-NonCommercial 4.0 International*

l'interface UMIACS[13] ; un post-traitement manuel s'est imposé pour l'alignement au niveau des paragraphes parce que celui-ci a présenté une difficulté d'exportation des données des brochures colorées en PDF ; le marquage en signe du corpus, la conversion en UTF-8 et la séparation entre la ponctuation et les formes de mots ont présidé l'alignement au niveau des mots. Après cette tâche, les limites phrastiques ont été scannées dans le texte aligné en paragraphes servant à nouveau à l'annotation d'alignement manuel des mots sur l'interface UMIACS. Le caractère agglutinant du swahili a obligé une déconstruction morphologique des mots pour permettre l'application efficiente de GIZA++ qui est la méthode employée pour l'alignement des mots mise en œuvre par les méthodes comme de IBM1 à IBM5 et HMM.

### *4.5. Synthèse expérientielle*

Au regard de toutes ces expériences, il est à noter que la constitution d'un corpus, qu'il soit parallèle ou comparable, pour une étude linguistique doit obéir à certaines conditions qui sont classables en trois stations : les données, leur digitalisation et la logistique.

Pour la première station, *(1)* **la nature des données** doit être bien spécifiée et bien circonscrite. Ceci répond par exemple à la question sur le type de texte, sur le fait de langue, la variété linguistique, le type d'unité linguistique. *(2)* **La source des données** se doit être fiable, vérifiée et vérifiable. Il est toujours judicieux de la spécifier pour permettre la fixation de l'étude que l'on veut mener. C'est la raison pour laquelle les auteurs qui ont œuvré pour les corpus déjà confectionnés préfèrent entre autres les maisons de publications, les traductions éditées, les locuteurs natifs, les ouvrages publiés comme sources de collecte des données. *(3)* **la quantité des données** est très déterminante lors de la constitution d'un corpus. L'unité de quantification des données d'un corpus est le mot. C'est pourquoi pour chaque corpus ci-dessus mentionné, les auteurs prennent la peine de toujours préciser le nombre de mots.

Pour les corpus parallèles, *(4)* **la nature linguistique des sous-corpus** doit toujours être mentionnée parce que la plupart des corpus de traduction portent les noms des langues des sous-corpus qu'ils regroupent (BFSU English-Chinese Parallel Corpus, Babel Chinese-English Parallel Corpus, English-Swedish Parallel corpus, etc.) et lorsqu'il s'agit des corpus comparables, le domaine auquel appartiennent les types de texte doit être mentionné comme cela a été fait pour ANC, BNC, RNC.

La digitalisation des données fait référence à la conversion de celles-ci en version électronique pour permettre l'application efficiente des logiciels. Elle prend pour point de départ **l'usage des variétés de scanner** surtout lorsque les données sont en version papier ; ensuite **l'annotation** (concerne tant les corpus parallèles que les corpus comparables) que Mikhailov et Cooper (2016, p. 216) définissent comme :

[13] issu de l'intranet de "University of Maryland Institute for Advanced Computer Studies" (UMIACS).

*L'œuvre est sous la licence Creative Commons Attribution-NonCommercial 4.0 International*

> The addition of special markers to a text to denote text structure, grammatical classes, syntactic functions, etc. The purpose of annotation is to make it possible to perform searches and produce statistics for abstract level features. Annotation is also called markup. The term tagging is closely related, but in corpus linguistics this usually means grammatical annotation. Annotation is carried out in such a way that it is not shown in the search results (because it makes the output difficult to read). Encoding is a very closely related term and is sometimes used in a broader sense to mean annotation.

Le point d'achèvement de la digitalisation est **l'alignement** des sous-corpus entendu par Kraif (2015 : 16) comme *l'opération consistant à mettre en correspondance les segments équivalents*. Selon Mikhailov et Cooper (2016, p. 216), il désigne:

> Annotation that is used to link the corresponding segments of parallel texts. With an aligned corpus it is possible to collect contrastive data on the connections between the source and target texts, e.g. in the form of parallel concordances. Aligning can be performed manually, but for larger texts alignment software is used.

Il peut se faire manuellement ou à l'aide des logiciels. Que ce soit dans l'une des alternatives, l'alignement se fait progressivement, c'est-à-dire de l'unité supérieure à celle inférieure (du paragraphe au mot). Kraif (2015) propose une variété d'outils électroniques d'alignement : *Alinea, TXM, Lexico3, Plug aligner, K-Vec++, GIZA++, Mindo, Unitex MulItal, JAM, ParaConc*, etc. il est réalisé exclusivement sur les corpus parallèles de sorte que les enjeux de l'alignement sont liés à la distance filiale entre les langues des sous-corpus et la typologie de celles-ci.

La dernière station porte sur la logistique qui se réfère à l'ensemble des moyens à mettre en œuvre pour la réalisation d'un projet de constitution d'un corpus. Ces moyens sont **le financement** et **les droits d'auteurs**[14]. Les projets de grande envergure ont été financés par les gouvernements ou les institutions étatiques. En guise d'illustration, on peut citer Intercorp, ANC, RNC, ENPC, etc.

## 5. Stratégie méthodologique pour le traitement des corpus

Une stratégie méthodologique est l'art de combiner des opérations de façon ordonnée pour atteindre un but. Certaines opérations seront proposées ci-dessous et pourront être adoptées pour les recherches en linguistique contrastive. De manière non-exhaustive, nous proposons huit (8) considérations ou exigences :

[14] Voir Lucas A., et al. (2017), *Guide de droit d'auteur*, 3e édition.

*L'œuvre est sous la licence Creative Commons Attribution-NonCommercial 4.0 International*

- L'écart familial et filial entre les langues en comparaison doit être pris en compte. En d'autres termes, le contrastiviste doit connaître à quelle famille et à quelle sous-famille appartiennent les langues en comparaison. Leclerc (2008) en donne déjà un aperçu lorsqu'il fait la distinction de 16 familles de langues[15] à travers le monde et Leclerc (2010) avec son dénombrement de 9 sous-familles[16] de la famille indo-européenne. De ce fait, il semble qu'il existe une corrélation entre l'éloignement filial entre les langues et la quantité élevée de divergences. De sorte que plus deux langues sont distantes l'une de l'autre, plus les divergences entre elles sont nombreuses. Cette opération est autant plus nécessaire parce qu'elle permet au chercheur de mieux orienter son travail et d'économiser l'énergie.

- Le type des langues en comparaison doit être pris en considération sur le plan syntaxique, morphosyntaxique, sémantique, etc. Plusieurs auteurs ouvrent des pistes à ce sujet comme Greenberg (1963), cité par Nexmeyer (2003), qui distingue six types de langue[17] sur le plan syntaxique. Cette opération permet d'anticiper sur les résultats de l'alignement des corpus ou sur les raisons de la difficulté de l'alignement des textes issus de certaines langues.

- Les différences formelles des langues en comparaison (système d'écriture, les caractères utilisés, etc.), les différences stylistiques imposées aux écrivains/auteurs par les langues elles-mêmes, les différences culturelles de ces langues en comparaison et la tradition scientifique ou professionnelle ou artistique de chaque auteur ou chaque traducteur doivent être cernées. Cela permettra au chercheur d'avoir un regard global et objectif sur les productions qui ne sont rien d'autre que les textes de différentes langues qui composent le corpus du travail.

- L'identification des pratiques propres à la tradition scientifique ou professionnelle ou artistique de chaque auteur ou de chaque traducteur. Cette opération permet de scruter les contenus textuels et de comprendre les raisons de la disposition des éléments des contenus textuels.

- Le réglage des problèmes de comparabilité et de similarité : il faut noter que le besoin de comparabilité et de similarité entre les sous-corpus des corpus comparables est très aigu par rapport à celui entre les sous-corpus des corpus parallèles qui sont composés de traductions. Pour le premier cas, Déjean et

[15] Indo-européenne, eskimo-aléoute, I. amérindiennes, I. australiennes, ouralienne, altaïque, bantoue, nigéro-congolaise, sino-tibétaine, chamito-sémitique, dravidienne, khoïsane, austronésienne, austro-asiatique, nilo-saharienne, I.papoues.
[16] Germaniques, romanes, slaves, celtiques, grecques, baltes, albanaises, iraniennes, indiennes.
[17] Les langues VSO, SVO, SOV, VOS, OSV, OVS.

*L'œuvre est sous la licence Creative Commons Attribution-NonCommercial 4.0 International*

Gaussier (2002) posent une condition *sine qua non* par rapport à la comparabilité des textes :

> Deux corpus de langues L1 et L2 sont dits comparables s'il existe une sous-partie non négligeable du vocabulaire du corpus de langue L1, respectivement L2, dont la traduction se trouve dans le corpus de langue L2, respectivement L1.

- De plus, il faudra s'assurer de la similarité entre les sous-corpus au niveau de leur taille ou du genre textuel ou de leur époque ou de leur date de publication. Pour le second cas, la comparabilité et la similarité ne sont plus recherchées au niveau des sous-corpus, mais plutôt au niveau de l'existence du fait linguistique à étudier dans les langues en comparaison. C'est dans ce sens que Hasselgarde (2010) aborde la notion de « perceived similarity ».
  La prise de conscience des pièges méthodologiques liés aux corpus parallèles et l'intégration de la façon de les éviter (voir Nádvorníková, 2017) sont d'une importance capitale.

- Faire usage des moyens de traitement fiables pour le traitement proprement dit des corpus. À l'ère des nouvelles technologies, les logiciels sont conseillés. Le traitement des corpus ne doit pas se confondre à l'analyse. Le traitement est immanent au corpus et se fait généralement à l'aide des logiciels et cet ensemble produit des données traitées en statistiques et pourcentages d'occurrences, de cooccurrences, de correspondances, d'équivalence tandis que l'analyse consiste à porter un regard extérieur aux données traitées et à expliquer leur état à l'aide des théories.
- Procéder aux vérifications du traitement différent des vérifications faites lors de la constitution des corpus. Pour le traitement des corpus de traduction, il faut avoir recours aux corpus comparables de traduction et pour celui des corpus comparables, les locuteurs natifs des différentes langues représentent des outils fiables de vérification qu'on peut convoquer.

## 6. Quel corpus pour quelle étude ?

L'intention qui nous anime n'est pas de confiner la créativité et l'ingéniosité du chercheur, mais de proposer des pistes qui pourront l'aider à opérer des choix pour des résultats probants dans le champ de la linguistique contrastive[18]. C'est la raison pour

[18] Pour son histoire et son évolution, cf. YLLERA, A., (2014), « Linguistique contrastive, linguistique comparée ou linguistique tout court ? » in *Contrastes*, Centre de Recherche en Linguistique Contrastive de l'Univeristé de Paris-III-Sorbonne Nouvelle (CRELIC) et KURTEŠ, S., (2006), "Contrastive Analysis at Work : Theoretical Considerations and their Practical Applications", *Estu. Ling.* Londrina, n° 9/1, pp111-140

*L'œuvre est sous la licence Creative Commons Attribution-NonCommercial 4.0 International*

laquelle les arguments présentés dans ce travail sont de nature hypothétique. Nous partons de la composition interne de la linguistique contrastive qui nous servira de base pour établir des jonctions entre une sous-discipline de la linguistique contrastive et une variété de corpus.

### *6.1. Linguistique contrastive et ses démembrements*

L'unanimité des auteurs comme Fisiak (1980*a*) et Ping Ke (2019) sur l'affiliation directe de la linguistique contrastive à la linguistique fait en sorte qu'elle hérite aussi de la composition interne de cette dernière. C'est pour cela que les démembrements de la linguistique contrastive sont cernables à travers sa domestication ou son appropriation des niveaux d'analyse linguistique et des autres branches de la linguistique.

Dans un premier axe, l'on peut relever les grammaires contrastives dont Nádvorníkóvá (2010 :10) fait allusion, les études contrastives phonologiques en anglais « phonological contrastive studies » abordées par Fisiak (1980*b*), l'analyse contrastive lexicale présentée par Ping Ke (2019), la lexicographie contrastive/bilingue dans laquelle s'inscrit la thèse de Le Serrec (2012), la stylistique comparée du français et de l'anglais de Vinay et Darbelnet (1958), la stylistique contrastive qui est une partie intégrante du modèle de Hartmann selon Creed (1995), la sociolinguistique contrastive dont parle Janicki (1980), la lexicologie contrastive selon Paillard (2000), l'analyse du discours comparée ou contrastive avec Von Münchow (2014).

Dans le second axe, les niveaux d'analyse linguistique peuvent aussi revêtir la nature contrastive. On peut à cet effet parler de la phonétique contrastive selon Ping Ke (2019), la morphologie contrastive (ibid), la syntaxe comparée de Guillemin-Flescher (1981), la sémantique contrastive présentée par Krzyżanowska (2013) et la pragmatique contrastive selon Ping Ke (2019). Au regard de ce qui précède, nous nous rendons compte de la riche densité interne de la linguistique contrastive et cela représente pour les chercheurs dans ce domaine un vaste terrain d'investigation qui requiert des outils sophistiqués et adéquats.

### *6.2. Choix motivé et justifié d'un corpus*

Le choix d'un corpus particulier dépend de l'orientation de l'étude et de l'objectif du chercheur. Bien que la présente déclaration soit indéniable, nous notons néanmoins que les corpus comparables semblent être plus appropriés pour l'extraction des termes bilingues dans le cadre de la lexicographie contrastive puisque selon Fung et Yee (1998, pp. 414-415), les corpus comparables présentent un autre avantage lorsqu'on recherche les néologismes, par exemple, les nouveaux noms propres y figurent plus rapidement que dans les textes parallèles, tels les noms des personnalités dans les journaux. Ces corpus semblent être plus usuels pour une analyse contrastive lexicale dans une spécialité donnée parce qu'ils sont dépourvus du phénomène « translationese », qui est une traduction biaisée et imposée par la langue source. C'est la raison pour laquelle Sinclair (1996, p. 12)

*L'œuvre est sous la licence Creative Commons Attribution-NonCommercial 4.0 International*

peut dire: The possibilities of a comparable corpus are to compare different languages or varieties in similar circumstances of communication, but avoiding the inevitable distortion introduced by the translations of a parallel corpus.

Les corpus de traduction paraissent plus adéquats pour l'examen des rendus des notions et des procédés grammaticaux en vue de l'élaboration des grammaires contrastives. À titre illustratif, les corpus de traduction permettent de mieux cerner le rendu des différentes parties du discours d'une langue A à une langue B. De plus, ces corpus semblent être des outils appropriés pour l'étude des interférences linguistiques dans le cadre de la sociolinguistique contrastive de telle sorte que le discours d'un locuteur natif d'une langue A peut être examiné lorsqu'il parle une langue étrangère B.

Pour la stylistique contrastive, les corpus comparables et de traduction semblent intéressants dans la mesure où le mode de présentation d'un genre (lettre, CV, demande, carte d'invitation, etc.) dans une langue A peut être contrasté avec le genre équivalent dans une langue B ou alors les effets des figures de style actualisés dans un texte A peuvent être comparés au rendu de leurs effets dans un texte traduit B. De même, ces deux types de corpus peuvent servir efficacement pour les études dans le cadre de l'analyse du discours comparée. Les corpus de traduction et les corpus comparables peuvent dévoiler les particularités sur la cohérence, la cohésion des textes, sur les implicatures, etc.

Comme les études ne sont pas fortuites, les outils choisis par le chercheur ne doivent pas l'être aussi. Ils doivent au contraire se fonder sur une motivation objective et pouvoir être justifiés par celui qui les manipule. Dans cette perspective, il convient de mentionner que les faits étudiés en stylistique contrastive, en grammaire contrastive, en phonologie contrastive, en lexicographie contrastive, en analyse du discours comparée basées sur les corpus comparables ou de traduction, peuvent être abordés du point de vue phonétique, c'est-à-dire dans leurs variations de prononciation dans l'espace ; du point de vue morphologique, c'est-à-dire dans leurs formes ; du point de vue syntaxique, c'est-à-dire dans leurs agencements dans les structures plus ou moins étendues dans un texte ; du point de vue sémantique, c'est-à-dire dans leur significativité ou alors par rapport au sens qu'ils revêtent en contexte ; du point de vue pragmatique, c'est-à-dire dans les effets qu'ils produisent sur un auditoire ou sur un récepteur lorsqu'ils sont mis en œuvre. Tout ceci nécessite une certaine compétence de la part du contrastiviste.

### *6.3. Compétence minimalement requise*

D'entrée de jeu, on peut faire la distinction entre ceux qui constituent les corpus exploitables plus tard et ceux qui les constituent pour exploiter immédiatement ou ceux qui exploitent les corpus déjà constitués. Les indices de compétence que nous relèverons ici concernent ceux de la seconde catégorie. Les études contrastives en langues requièrent les éléments de compétence suivants :

La capacitation préalable sur les particularités du fait linguistique à étudier dans chaque langue en comparaison avant d'entamer le travail contrastif : le contrastiviste doit

*L'œuvre est sous la licence Creative Commons Attribution-NonCommercial 4.0 International*

pouvoir exposer le fait linguistique à étudier dans toutes ses subtilités dans chaque langue en comparaison :

- La connaissance profonde de son corpus, la parfaite maîtrise de sa ou ses source(s) et ses possibles influences : le contrastiviste doit pouvoir répondre à toutes les questions qu'on peut lui poser sur son corpus et sur sa source ;
- La maîtrise des procédés de la traduction lorsqu'on a affaire aux corpus de traduction : cela ne relève pas du hasard si les ouvrages (de référence) sont introduits par la présentation de ces procédés. On peut citer la *Stylistique comparée du français et de l'anglais* de Vinay et Darbelnet (1958), la *Stylistique comparée du français et de l'allemand* de Malblanc (1968), *Approche linguistique des problèmes de traduction anglais – français* de Chuquet et Paillard (1989) ;
- La présentation du travail contrastif doit se faire dans un style cohérent et digeste : on doit y voir une certaine linéarité.

## 7. Conclusion

Dans cette étude sur la notion de corpus dans les études contrastives en langues, nous avons d'abord abordé quelques questions terminologiques et nous avons réalisé que la linguistique contrastive possède une diversité de conceptions terminologiques sur la notion de corpus. Ensuite, les alternatives des types de corpus usuels en linguistique contrastive ont été amplement présentées. De plus, après avoir examiné la constitution du corpus au travers des différentes expériences, quelques fondamentaux ont été déduits. Ceux-ci ont permis de proposer certaines stratégies méthodologiques pour le traitement des corpus. Et enfin, nous avons essayé d'établir l'adéquation entre les différents types de corpus et les études possibles en linguistique contrastive.

En somme, les corpus sont les outils nécessaires pour faire de la recherche en linguistique contrastive. Leur usage nécessite au préalable la symbiose de plusieurs opérations au niveau de la sélection, de la constitution et du traitement. La délicatesse avec laquelle ces opérations doivent être menées interpelle alors le chercheur à adopter une démarche motivée, objective et justifiable. Ce travail pourra permettre aux jeunes chercheurs qui souhaitent s'investir dans ce domaine de recherche particulièrement technique de mieux se situer dans une méthodologie du choix, de la constitution et de l'exploitation judicieuse et optimale du corpus en vue des résultats prometteurs en linguistique contrastive.

***L'œuvre est sous la licence Creative Commons Attribution-NonCommercial 4.0 International***

## Références


[1] Aijmer, K., & Altenberg, B. (1996). Introduction. Aijmer, K. & Johansson, S. (eds). *Languages in Contrast. Text-based Cross-Linguistic Studies* (pp. 11-16). Lunds Universitets Publikationer.

[2] Altenberg, B., & Aijmer, K. (2000). The English-Swedish parallel corpus: A ressource for contrastive research and translation studies. *Corpus Linguistics and Linguistic Theory*.

[3] Altenberg, B., & Granger, S. (2002). Recent trends in cross-linguistic lexical studies. Altenberg, B., et Granger, S. (ed.). *Lexis in Contrast. Corpus-Based Approaches* (pp. 1-48). John Benjamin Publishing Company.

[4] Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. Baker, M., Francis, G. & Tognini-Bonelli, E., (eds). *Text and Technology: in Honour of John Sinclair* (pp. 233-252). John Benjamins.

[5] Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future Research. *Target*, (*7*), 223-243.

[6] Baker, M., (1999). The role of corpora in investigating the linguistic behaviour of professional translators. *International Journal of Corpus Linguistics*, (4), 281-298.

[7] Confiant, R. (2003). La traduction en milieu diglossique. *Atelier de recherche du sur l'enseignement du créole et du français dans l'espace américano-caraïbe (AREC-F)*, 1–17.

[8] Creed, M. (1995). *A contrastive analysis of French and English, social statistics texts*. Thesis, Dublin City University.

[9] Culo, O. et al. (2008). Empirical studies on language contrast using the English-German comparable and parallel croco corpus. *Proceedings of the LREC workshop on Comparable Corpora*, 47–51.

[10] Déjean, H. & Gaussier, E. (2002). Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica - Alignement lexical dans les corpus multilingues*.

[11] Delbera, J-P. (2002). Le corpus entre données, analyse et théorie. *Corpus*, (1).

[12] De Pauw., G. et al. (2009). The SAWA corpus: a parallel corpus English – Swahili. *EACL – workshop on Language Technologies for African Languages*, 9-16.

[13] Ebeling, J. (1998). Contrastive linguistics, translation, and parallel corpora. *Meta*, 43(4).

[14] Fisiak, J. (ed.). (1980*a*). *Theoretical Issues in Contrastive Linguistics*. John Benjamins B. V.

[15] Fisiak., J. (1980*b*). Contrastive analysis of phonogical systems. FISIAK., J., (ed.). *Theoretical Issues in Contrastive Linguistics*. John Benjamins B. V., 215-224.

[16] Fung., P. & Yee., L. Y. (1998). An IR approach for translating new words from nonparallel, comparable text. *Proceedings of the 17th international conference on Computational linguistics*, Montréal, 414– 420.

[17] Goeuriot., L. (2009). *Découverte et caractérisation des corpus comparables*


*L'œuvre est sous la licence Creative Commons Attribution-NonCommercial 4.0 International*

*spécialisés*. Thèse, Université de Nantes.
[18] Granger, S. (1996). From CA to CIA and Back: An Integrated Approach to Computerized Bilingual and Learner Corpora. Aijmer, K., Altenberg, B. & Johansson, M. (eds). *Languages in Contrast. Text-Based Cross-Linguistic Studies*, Lund University Press, 35-51.
[19] Grishina., E. (2007). Spoken Russian in the Russian National Corpus (RNC). *Institute of Russian Language*, Moscow.
[20] Guillemin-Flescher., J. (1981). *Syntaxe comparée du français et de l'anglais. Problèmes de traduction*. Gap, Ophrys.
[21] Habert., B. (2000). Des corpus représentatifs : De quoi, pour quoi, comment ? Bilger, M. (ed.). *Linguistique sur corpus. Études et réflexions*. Presses Universitaires de Perpignan, (31), 11–58.
[22] Hasselgard, H. (2010). Contrastive analysis/contrastive Linguistics. Malmkjær, K. (eds). *The Routledge Linguistics Encyclopedia*, Routledge, 3rd ed, 98-101.
[23] Hunston., S. (2002). *Corpora in applied Linguistics*. Cambridge University Press.
[24] Janicki., K. (1980). Contrastive sociolinguistics – Some methodological considerations. Fisiak., J. (ed.). *Theoretical issues in contrastive linguistics* (pp. 11-19). John Benjamins B. V.
[25] Johansson., S. (1998). On the role of corpora in cross-linguistic research. Johansson., S. & Oksefjell., S. (eds). *Corpora and cross-linguistic research: theory, method, and case studies* (pp. 3-25). Rodopi.
[26] Johansson., S. (2007). *Seeing through Multilingual Corpora*. John Benjamins.
[27] Johansson., S. Ebeling., J. & Oksefjell., S. (2002). English-Norwegian parallel corpus: Manual. *Department of British and American Studies*. University of Oslo.
[28] Kraif., O. (2015). *Corpus parallèles, corpus comparables : Quells contrastes* ? HDR, Université de Poitiers.
[29] Krzyzanowska., A. (2013). La sémantique contrastive aujourd'hui. *19th International Congress of linguists*, Genève.
[30] Leclerc., J. (2008). Les familles linguistiques dans le monde. *Ethnologue (16e ed.) du Summer Institute of Linguistics du Texas.*
[31] Leclerc., J. (2010). Les langues indo-européennes. *Ethnologue (16e ed.) du Summer Institute of Linguistics du Texas.*
[32] Le Serrec., A. (2012). *Analyse comparative de l'équivalence en corpus parallèle et en corpus comparable : application au domaine du changement climatique*. Thèse, Université de Montréal.
[33] Maia., B, (2003). What are comparable corpora? *Proceedings of the pre-conference workshop on multilingual corpora: Linguistic requirements and technical perspectives*. Grande Bretagne, 27-31.
[34] Mauranen., A. (2002). Will 'translationese' ruin a contrastive study? *Languages in Contrast*, 2(2), 161–185.
[35] McEnery., A. et al. (2000). EMILLE: Building a corpus of South Asian Languages.

*L'œuvre est sous la licence Creative Commons Attribution-NonCommercial 4.0 International*

*Disponible en ligne à https://www.asjp.cerist.dz/en/Articles/155*

*Epsrc-funded 18-month research project*, Lancaster University. https://www.mt-archive.info/BCS-2000-McEnery.

[36] McEnery., A. & Wilson., A. (1996). *Corpus Linguistics*. 1st ed. Edinburgh: Edinburgh University Press.

[37] McEnery., A. & Xiao., R. (2007). Parallel and Comparable Corpora: what are they up to? Anderman., G. & Rogers., M. *Incorporating Corpora: Translation and the Linguist*. Multilingual Matters édition.

[38] Melnikova., E., Novakova., I. & Kraif., O. (2009). Quels corpus pour l'analyse contrastive ? L'exemple des constructions verbo-nominales de sentiment en français et en russe. *Actes des 6èmes Journées de la Linguistique de Corpus* (disp. à l'adresse :http://www.licorn-ubs.com/jlc6/ACTES/Melnikova_etal_JLC09.pdf).

[39] Mikhailov., M. & Cooper., R. (2016). *Corpus Linguistics for Translation and Contrastive studies*. Routledge.

[40] Nádvorníková., O. (2010). Les corpus parallèles : L'espace pour l'analyse contrastive. *Études Romanes de BRNO.*

[41] Nádvorníková., O. (2017). Pièges méthodologiques des corpus parallèles et comment les éviter. *Corela*, 1-30.

[42] Newmeyer., F. (2003). Basic word order' in formal and functional linguistics and the typological status of 'canonical sentence types'. Willems., D. et al. (eds). *Contrastive analysis in language – identifying linguistic units of comparison*. Palgrave Macmillan, 69-88.

[43] Paillard., M. (2000). *Lexicologie contrastive anglais-français : formation des mots et construction du sens*. OPHRYS.

[44] Ping Ke. (2019). *Contrastive Linguistics*. Peking University Press.

[45] Rakhilina., E. et al. (2014). Russian National Corpus. *SCLC – RNC workshop*, University of Harvard.

[46] Rawoens., G. (2008). Les corpus bilingues et la linguistique contrastive. Une étude des constructions causatives basée sur un corpus parallèle néerlandais suédois. *9ème Journées internationales d'analyse statistiques des données textuelles (JADT)*, Lyon, ENS Lettres et sciences humaines, 971-980.

[47] Reppen., R. & Ide., N. (2004). The American National Corpus – Overall Goals and the First Release. *Journal of English Linguistics*.

[48] Schlamberger., B. M. (2020). La préparation des corpus parallèles et comparables – nouvelles bases pour la traduction entre le français et le slovène. *Cahiers du Plurilinguisme européen*. https://doi.org/10.57086/cpe.246

[49] Sinclair., J. (1996). Preliminary recommendations on corpus typology. *Rapport technique. EAGLES (Expert Advisory Group on Language Engineering Standards)* dans *http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html*

[50] Teubert., W. (1996). Comparable or parallel corpora? *International Journal of Lexicography*, 9(3), 238-264.

[51] Vinay., J-P. & DARBELNET., J. (1958). *Stylistique comparée du français et de*

*L'œuvre est sous la licence Creative Commons Attribution-NonCommercial 4.0 International*

*Disponible en ligne à https://www.asjp.cerist.dz/en/Articles/155*

*l'anglais.* Marcel Didier.
[52] Von Münchow., P. (2014). L'analyse du discours Contrastive : comparer des cultures discursives. Grezka., A. et al. (Eds). *Les Sciences du langage en Europe : tendances actuelles*. Paris : Education Discours Apprentissages, 75-92.
[53] Xiao., R. (2008). Well-known and Influential Corpora. Ludeling., A. & Kyto., M. (eds). *Corpus Linguistics: An International Handbook* (1). Mouton de Gruyter.
[54] Xiaoxiao., C. & Shili., G. (2011). The construction of English-Chinese parallel corpus of medical works based on self-coded python programs. *Procedia Engineering*, (24), 598-603.
[55] Zanettin., F. (1998). Bilingual corpora and the training of translators. *Meta*, 43(4), 616–630.

**Notices bio-bibliographiques**

**Fonkoua Paul** est enseignant-chercheur à l'Université de Yaoundé 1 (Chargé de Cours), à la Faculté des Arts, Lettres et Sciences Humaines. Il est titulaire d'un doctorat Ph. D en sociolinguistique et en analyse du discours. Il est auteur de plusieurs articles dans les revues nationales et internationales et co-auteur d'un ouvrage sur la langue et le style dans la chanson camerounaise. Il a notamment collaboré dans *ANADIS*, revue d'analyse du discours de l'Université Suceava – Roumanie et dans *Le français en Afrique*, revue de sociolinguistique, éditée par CNRS. Ses domaines de recherche sont entre autres la sociolinguistique (urbaine), l'analyse du discours (littéraire), le plurilinguisme et les contacts de langues. Il peut être contacté aux adresses électroniques suivantes : paul.fonkoua@univ-yaounde1.cm et paulfonkoua@yahoo.fr

**Bayiha Auguste** est doctorant au département d'Études Bilingues de l'Université de Yaoundé 1, spécialisé en linguistique contrastive et il est aussi enseignant vacataire dans le même département et au service de langues de la Faculté de Science dans la même université. Il est titulaire d'un Master en études contrastives en langues depuis 2019 à l'Université de Yaoundé 1 et d'un DIPES (Diplôme des Professeurs d'enseignement secondaire) à l'ENS (Ecole Normale Supérieure) de Yaoundé depuis 2018. Il peut être contacté à l'adresse électronique suivante : cbillongbayiha@gmail.com

## Déclaration de conflits d'intérêt

Les auteurs n'ont déclaré aucun conflit d'intérêt en ce qui concerne la recherche, la paternité et/ou la publication de l'article.

*L'œuvre est sous la licence Creative Commons Attribution-NonCommercial 4.0 International*