REY (Ch.) 2003, « Une expérience d'informatisation des dictionnaires anciens : les articles portant sur les sons dans l'*Encyclopédie méthodique* (1782-1832) », présentation-poster, Université de Provence. (Publication en ligne http://www.u-picardie.fr/LESCLaP/rey/presentation_poster_aix.pdf) (consultation : janvier 2013)

REY (Ch.) 2004, « Le balisage souple ou flottant : une piste pour l'informatisation des données encyclopédiques anciennes », Site du centre de recherches METAlexicographiques et DIctionnairiques Francophones (Metadif). (publication en ligne : http://www.u-picardie.fr/LESCLaP/rey/Reyc_balisage-souple.pdf) (consultation : janvier 2013)

REY (Ch.) et ZAOUI (C.) 2004, « Balisage XML "ciblé" : Une nouvelle approche dans l'informatisation des corpus », in *Actes de la conférence internationale sur la fouille de texte*, M.-H. Antoni et Fr. Yvon éds., Paris, École nationale supérieure des télécommunications (ENST), pp. 121–133.

REY-DEBOVE (J.) 1971, *Étude linguistique et sémiotique des dictionnaires français contemporains*, La Haye/ Paris, Mouton.

TEI Consortium, TEI P5 : *Guidelines for electronic text encoding and interchange (2.3.0.)*, http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html. (consultation : janvier 2013)

TUTIN (A.) et WIONET (Ch.) 2001, *Pour informatiser le* Dictionnaire Universel *de Basnage (1702) et Trévoux (1704) ; approche théorique et pratique*, Paris, Honoré Champion.

WOOLDRIDGE (T. R.) 1993, « Le flou en informatique textuelle », in *Texte*, Toronto, 13/14, pp. 275-289.

WOOLDRIDGE (T. R.) 1997, « Baliser un texte, c'est le penser : le cas du Dictionnaire de l'Académie française ». [Version révisée d'une communication faite à Paris, le 23 mai 1997, dans le cadre d'une journée d'études organisée par le Groupe d'Études en Histoire de la Langue] (Texte publié en ligne http://homes.chass.utoronto.ca/~wulfric/articles/gehlf597/) (consultation : janvier 2013)

WOOLDRIDGE (T. R.) 1999, « L'informatisation du Dictionnaire de l'Académie française », in *L'informatisation des dictionnaires anciens* (Actes du colloque-atelier international DictA1998), T. R. Wooldridge (éd.), Toronto, Siehlda. (Texte publié en ligne http://homes.chass.utoronto.ca/~wulfric/siehlda/dicta1998/trw_acad.htm) (consultation : janvier 2013)

Le corpus Babeliris : d'une méthodologie de constitution aux principales caractéristiques linguistiques

Ward VAN de VELDE
Centre de recherche Termisti
Institut supérieur de traducteurs et interprètes
Haute École de Bruxelles

Résumé

Cet article décrit le processus d'encodage du corpus *Babeliris*, réunissant des textes médico-administratifs du réseau hospitalier *IRIS*. Concrètement et sans revenir sur les modalités de collecte des textes, il présente les balises *TEI-P5* qui ont servi à encadrer les données textuelles variables, à partir desquelles peuvent naître des sous-corpus. Dans ce cadre, il explique le fonctionnement de l'interface créée à cet effet. Par ailleurs, il soulève la problématique du choix d'un aligneur multilingue *ad hoc* et, partant, d'un gestionnaire de mémoire de traduction, en exposant les avantages et les inconvénients de trois logiciels testés. Enfin, il décrit les principales caractéristiques linguistiques du pan français du corpus, dans le but de pouvoir juger de la représentativité de celui-ci dans le vaste monde des corpus de langue spécialisée et de langue générale.

Mots-clés :
linguistique de corpus, médecine, *TEI-P5*, alignement, concordance, mémoire de traduction.

1. Introduction : de la recherche *Babeliris*[1] à la naissance d'un corpus original

S'agissant de l'accès aux soins de santé, la législation belge veut que *la communication avec le patient se déroule dans une langue claire* (Ministère 2002). Ce principe présuppose notamment que le médecin s'exprime à sa patientèle dans une langue simplifiée. Le respect de cette disposition légale, inspirée de l'*Evidence based medicine* (Sackett *et al.* 1996), devrait particulièrement s'imposer au contact d'une patientèle allophone et/ou étrangère. Or, dans les faits, on constate que la loi ne prévaut pas toujours. Ainsi, à Bruxelles – terrain de nos recherches –, tous les patients ne sont pas égaux devant la démarche d'information médicale, et ce, en dépit des nombreuses actions menées dans le but d'aplanir les inégalités sociales en matière de santé (Boïteké 2011 : 9-10).

Les interactions orales jouent certes un rôle crucial dans la relation patient-médecin, mais une mauvaise compréhension de l'écrit médical peut aussi affecter la santé du patient, même si *l'écrit* ne peut détrôner *l'oral* (HAS 2008 : 6-7). Le centre de recherche TERMISTI, auquel nous ressortons, et le *Centrum voor Vaktaal en Communicatie* ont clairement compris ce double enjeu et se sont associés pour mener le projet Babeliris. Celui-ci a pour objectif d'améliorer la communication interculturelle en milieu hospitalier, appliquée plus spécifiquement aux hôpitaux publics bruxellois du réseau IRIS. Il se ramifie en deux volets complémentaires dans le but d'embrasser les facettes orale

1. Plus d'informations sur http://babeliris.be.

et écrite de la communication, respectivement traitées par les deux institutions susmentionnées.

Très concrètement, s'agissant du volet écrit, nous avons l'ambition de produire trois outils à valeurs formative et éducative destinés aux membres du réseau IRIS. Ce matériel didactique et pédagogique d'accompagnement du personnel soignant comprendra un *Guide de bonnes pratiques de rédaction* en français et en néerlandais destiné aux rédacteurs et aux traducteurs des hôpitaux, une banque de textes français-néerlandais alignés et une base de données terminologique multilingue, qui clarifiera le vocabulaire médico-administratif susceptible de poser problème à une patientèle ne maîtrisant ni les langues de la Région, ni l'actuelle *lingua franca* : l'anglais.

Afin d'appréhender la langue écrite en usage dans les hôpitaux IRIS au départ de laquelle nous concevrons ces instruments, nous avons appliqué les principes défendus par la linguistique de corpus, souscrivant à une approche résolument *corpus driven* (Tognini-Bonelli 2001 : 84-99). Dans ce cadre, il nous paraît essentiel d'exposer la méthodologie qui a présidé à l'élaboration de notre collection de textes, dans le but notamment d'assurer une certaine transparence pour les futurs utilisateurs de notre corpus (Tymoczko 1998 : 655). Par ailleurs, même si nous ne prétendons pas à l'exemplarité, il n'en demeure pas moins que nous espérons que le *modus operandi* exposé ici pourra inspirer d'autres chercheurs conduisant une recherche similaire.

2. Vers une méthodologie d'encodage des documents

2.1. Du traitement primaire des textes à leur XMLisation : bref aperçu

Dans cette section, nous ne reviendrons ni sur les modalités de collecte documentaire ni sur les contraintes juridiques et/ou déontologiques que nous avons observées dans ce cadre (Van de Velde *à paraître*). Rappelons toutefois que les textes colligés obéissent à des critères bien définis au préalable : tous traitent d'une thématique médicale et/ou administrative, ont été rédigés par des membres du réseau IRIS, sont adressés au patient et sont encore en usage. Concrètement, nous avons réuni un vaste ensemble de documents médico-administratifs diffusés en toutes langues sur divers supports, auxquels risque d'être confrontée la patientèle IRIS : brochures, dépliants, affiches, panneaux, guides, formulaires, pages *Web*, etc. Enfin, nos textes étant relativement courts, nous avons retenu les écrits dans leur totalité : ce faisant, nous nous sommes inscrit dans la tradition de Sinclair (1991 : 19) et écarté de celle de Biber, théoricien du *stratified sampling* et du *proportional sampling* (1993 : 243-248).

Confronté à des écrits hétéroclites codés en divers formats, nous avons rapidement été amené à soulever la question de l'encodage textuel et à opter pour un langage de balisage adopté par une large communauté de spécialistes et de chercheurs. Celui qui a retenu notre intérêt, en raison du caractère pérenne et interopérable conféré aux écrits qu'il code, est le langage *XML* et, plus particulièrement, une application de celui-ci : la spécification *TEI*, dans sa version *P5*

(Bauman et Burnard : 2012), plus adaptée à des exploitations linguistiques et, selon Jacquemart (*s.d.* : 8), très souvent mise en œuvre dans la sphère médicale. Néanmoins, il y a loin d'un écrit papier ou d'un texte sauvegardé en .pdf à sa version *TEI*, sans compter le coût en temps et en moyens de cette entreprise de conversion (Popescu-Belis 2002 : 60-61). C'est de ce constat que rend compte ce schéma, inspiré des travaux de Hareide et Hofland (2012 : 89-90) et retraçant le traitement primaire des documents jusqu'à leur *XMLisation* :

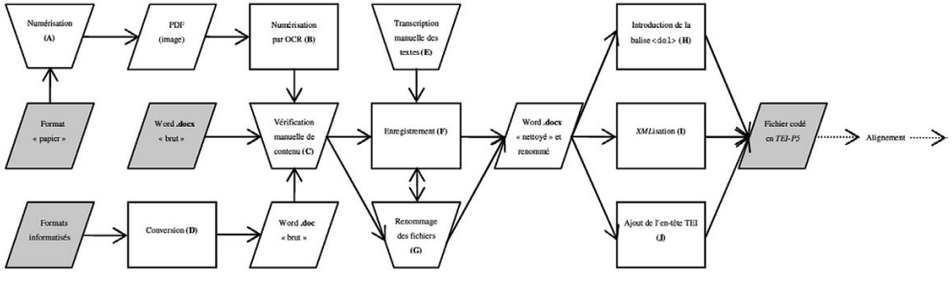

Figure 1 – Processus d'encodage documentaire : d'un texte brut à sa version structurée

Sur fond grisé apparaissent les trois formats textuels d'entrée – c'est-à-dire ceux des documents bruts obtenus tels quels par les hôpitaux IRIS – et celui de sortie – autrement dit celui des versions textuelles codées en *TEI-P5* –, et entre ces extrêmes figure le processus analytique de traitement documentaire. Celui-ci diffère selon le support textuel, mais converge vers un même but : soumettre le corpus à de puissants algorithmes de recherche.

S'agissant des écrits papier disponibles uniquement dans ce format, nous les avons convertis en des documents électroniques exploitables. Pour ce faire, nous les avons scannés (A) – afin d'obtenir des fac-similés en format .pdf (image) – puis numérisés à l'aide d'un logiciel de reconnaissance optique de caractères (OCR) : *Readiris*™ (B). À ce travail de numérisation, long de sept mois, s'est ajoutée une minutieuse tâche de révision (C), qui avait pour objectifs de transformer les signes mal interprétés par l'OCR en des chaînes de caractères lisibles, d'une part, et d'assurer une adéquation de mise en forme entre les versions originales et numérisées, d'autre part. On sait qu'une mauvaise transcription ou un découpage textuel inexact peut biaiser les résultats d'une quelconque analyse linguistique par logiciel (Véronis 2000 : 163). Quant aux documents *Word* sous format .docx, nous en avons examiné le contenu, veillant à les dépouiller de leurs images, mais à conserver le texte qui accompagnait celles-ci (C). Les autres écrits informatisés ont tous été convertis automatiquement en fichiers *Word* .doc (D) avant de subir un traitement similaire (C). Enfin, quelques publications ont été retranscrites manuellement (E). Il convient de préciser que, pour tous les textes, nous avons maintenu à dessein les erreurs syntaxiques et orthographiques, car celles-ci participent de la difficulté à comprendre le message.

Les corps des textes une fois validés, nous les avons enregistrés au format *Word* .docx (F) afin de permettre une XMLisation automatisée de nos fichiers. À ce stade, nous les avons aussi renommés sur le modèle *Titre du document_Hôpital concerné_Langue_Public cible_ Support textuel_Visée du document_Date de diffusion* (G). Faute de règles bien définies régissant l'appellation d'un document .xml

(Dufournaud et Gratsac-Legendre 2012 : 7), nous avons nommé nos fichiers de façon intuitive, veillant à définir au mieux les paramètres variables qui les caractérisent et tenant compte de l'exploitation informatique que nous envisageons de notre corpus.

La dernière étape du traitement liminaire des textes IRIS, dite d'XMLisation (I)[2], a consisté à convertir par lots notre banque de textes désormais codés en .docx en des documents .xml conformes à la TEI-P5, et ce, à l'aide de l'éditeur *Tei Vesta*. Dans le même temps, et toujours de manière automatisée, on a fait précéder chaque texte d'un en-tête *TEI-P5* (J), dans lequel on a injecté entre les balises *ad hoc* les sept métadonnées variables contenues dans le titre de nos documents préalablement enregistrés en .docx. Enfin, on a balisé les corps mêmes des messages – contenus entre les marques <body> – au niveau du paragraphe <p> et, à l'instar des travaux de Vihla (1998 : 75), de Sansonetti (2003 : 73-74) ou de Hareide et Hofland (2012 : 90), on y a neutralisé, à l'aide de la balise <del resp="Ward Van de Velde"> (Bauman et Burnard 2012 : 884-886 et Van de Velde *à paraître*), les données superflues susceptibles de biaiser toute analyse linguistique : adresses et courriels, contacts et numéros de téléphone, index et tables des matières, bibliographies et références bibliographiques… Manquant de précision, cette dernière tâche d'annotation intratextuelle mériterait toutefois une vérification manuelle, coûteuse en temps et en moyens.

2. Cette tâche a été confiée à M. Cédric Libert, jeune linguiste de l'Université catholique de Louvain-la-Neuve (Belgique) qui se spécialise en informatique.

8. <term>xxx</term> (dans <keywords>, inclus dans <profileDesc>) : service et unité auxquels se rattache l'écrit. Admettant le principe selon lequel toute classification s'avère délicate à établir en raison de son caractère intuitif (McEnery *et al.* 2006 : 21) et celui selon lequel « *les catégories ne s'excluent pas forcément […] ou semblent être en intersection* » (Habert 2000), nous avons opté pour une typologie de *genre* reposant sur le domaine, en raison de sa légitimité *per se* (Lee 2001 : 37-39, Biber 1993 : 245 et Sinclair 1996 : 7). Concrètement, la notion de *genre* s'appuie sur des critères extralinguistiques, externes à la production textuelle, autrement dit sur des paramètres situationnels et fonctionnels de l'écrit : contexte ou mode de production, situation d'énonciation, lecteur modèle, mode de publication, fonction ou visée et, dans le cas présent, thème ou domaine. Ayant pour ambition de nous rapprocher de la réalité du réseau IRIS et d'inclure des aspects tant administratifs, (para)médicaux, infirmiers que sociaux dans une seule taxinomie, nous avons considéré les domaines suivants (Van de Velde *à paraître*) :

TABLE 1. DOMAINES RETENUS POUR LE CORPUS *BABELIRIS*

SERVICES	UNITÉS
ADMINISTRATIF ET LOGISTIQUE	Accueil – admission – consultation ; archivage médical – bibliothèque ; droits, devoirs et protection du patient (service juridique et contentieux) ; hygiène hospitalière ; informatique – pages Web ; médiation – communication – relations publiques ; médiation interculturelle ; service social ; trésorerie – facturation ; non catégorisé
INFIRMIER	Soins infirmiers
MÉDICAL	Anatomie pathologique et cytogénétique ; anesthésiologie – réanimation ; biologie clinique et prélèvements ; cardiologie – chirurgie cardiaque ; chirurgie cervico-faciale et thoracique ; chirurgie mammaire et pelvienne ; chirurgie plastique, réparatrice et esthétique ; chirurgie vasculaire ; dermatologie ; endocrinologie ; gériatrie ; gynécologie – obstétrique – périnéologie – prénatale ; hémato-oncologie – dépistage et prévention du cancer – cancérologie mammaire ; hépato-gastro-entérologie – chirurgie digestive et abdominale ; hôpital de jour ; immuno-allergologie ; infectiologie – maladies infectieuses ; maladies tropicales ; médecine interne ; néonatalogie ; néphrologie – dialyse ; neurologie – neurochirurgie ; ophtalmologie ; ORL ; orthopédie – traumatologie – chirurgie orthopédique ; pédiatrie ; pneumologie – tabacologie – médecine du sommeil ; psychiatrie ; radio-isotopes – médecine nucléaire ; radiologie – imagerie médicale ; recherche hospitalière – éthique ; revalidation locomotrice, neurologique et cardio-pulmonaire ; rhumatologie ; sénologie ; sexologie – urologie ; soins supportatifs , continus et palliatifs ; soins intensifs – urgences ; stomatologie – dentisterie – orthodontie – chirurgie maxillo-faciale ; non catégorisé
PARAMÉDICAL	Diététique – nutrition ; ergothérapie ; kinésithérapie – ostéopathie ; logopédie ; pharmacie ; psychologie

Très concrètement, tout en-tête TEI-P5 de notre corpus se structure comme suit. Pour des raisons liées à l'impossibilité de développer ici la totalité de notre <teiHeader>, nous avons décidé de présenter uniquement le jeu de balises contenues dans la description du profil <profileDesc> :

```
<teiHeader>
 <fileDesc> […]
 </fileDesc>
 <encodingDesc> […]
 </encodingDesc>
 <profileDesc> […]
  <langUsage>
   <language ident="FR">français</language>
  </langUsage>
  <textClass>
   <keywords scheme="www.babeliris.org">
    <term>Service médical</term>
    <term>Infectiologie – maladies infectueuses</term>
   </keywords>
  </textClass>
  <textDesc>
   <channel>dépliant</channel>
   <constitution type="single"/>
   <derivation type="original"/>
   <domain type="medicine"/>
   <factuality type="fact"/>
   <interaction passive="group">non défini</interaction>
   <preparedness/>
   <purpose type="inform"/>
  </textDesc>
 </profileDesc>
 <revisionDesc status="approved">
 </revisionDesc>
</teiHeader>
```

3. La genèse d'une interface génératrice de sous-corpus

Le choix des balises exposées *supra* laisse entrevoir que le langage *TEI* ne constitue pas un format *stricto sensu*, mais plutôt un vaste ensemble de recommandations destinées à coder un document

XML. En réalité, la *TEI* permet d'encoder une même information de diverses manières en fonction de l'interprétation donnée au balisage et de l'objectif de celui-ci en lien avec une étude donnée. Faute d'univocité des métadonnées textuelles, aucun logiciel développé à ce jour ne permet donc d'interpréter invariablement le jeu de (sous-)balises *XML/TEI*. Cette réalité a pour corollaire de devoir construire ou réaménager – à la lumière du travail d'analyse à effectuer – des outils intégrant la logique sémanticienne *TEI*.

En outre, malgré l'universalité supposée du langage TEI-P5, il n'en reste pas moins que certains concordanciers et outils d'analyse du discours posent des problèmes d'exploitation du balisage .**xml** et supportent uniquement des formats propriétaires (Daoust et Marcoux 2006 : 330-331). D'autres encore imposent l'usage de *DTD* particulières. Il serait toutefois ambitieux d'adapter constamment le marquage .xml à ces outils, sachant qu'ils s'accommodent du format de texte brut .txt. Comme l'atteste le tableau 2, qui met en correspondance certains des « linguisticiels » utilisés à TERMISTI avec leurs propres formats de représentation de données, le plus petit commun dénominateur des outils testés demeure le format .txt :

TABLE 2. FORMATS D'IMPORTATION DOCUMENTAIRE SUPPORTÉS PAR LES DIVERS LOGICIELS TESTÉS

	OUTIL LINGUISTIQUE	FORMAT D'IMPORTATION TEXTUELLE REQUIS
1	AntConc	.htm, .html, .txt et .xml
2	Cordial Analyseur	.cnr, .htm, .rtf, .sam, .txt et .wp
3	mkAlign	.htm, .html, .tmx, .txt et .xml
4	NooJ	.doc, .docx, .htm, .html, .pdf, .rtf, .txt, .xml...
5	TXM	**.tmx, .txt, .xml(-tei et dérivés)**

Fort de ces constatations, nous avons fait développer[8] un outil capable de générer des sous-corpus codés en **.txt** au départ de nos documents XMLisés. Donner naissance à ce genre d'outil est intéressant à plusieurs égards : pour comparer certains paramètres de rédaction de divers sites hospitaliers, pour étudier des écrits d'un seul service quel que soit l'hôpital dont ils émanent, etc. La constitution de ces sous-corpus repose bien entendu sur l'interrogation des *dimensions* variables de l'en-tête *TEI*, isolées en 2.1. (à l'exception du titre de l'écrit, unique par principe) : service, unité, site hospitalier, langue, lecteur modèle, support, visée, année de publication ou encore association de plusieurs de ces éléments. Ainsi, la *figure 2* indique que l'on recherche, dans notre corpus, tous les documents du service médical et, en particulier, de l'unité d'hémato-oncologie – dépistage et prévention du cancer – cancérologie mammaire de l'Institut Jules Bordet, rédigés en français, à visée injonctive, tirés de pages Web, ayant le patient pour lecteur modèle et composés à une date quelconque.

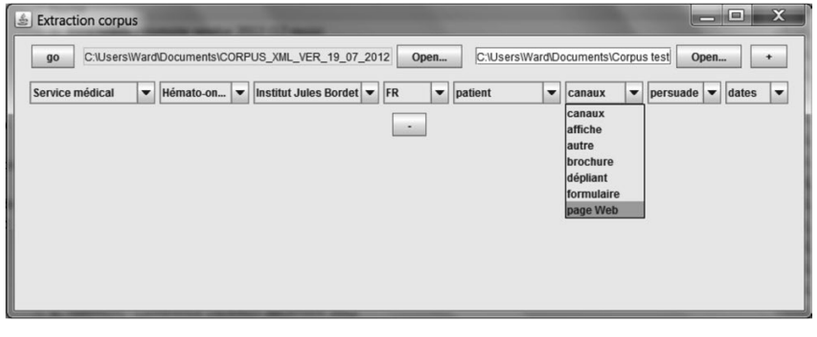

Figure 2 – Interface génératrice de sous-corpus

8. Toujours avec le soutien de C. Libert.

Enfin, suivant les conclusions de notre étude des concordanciers disponibles, nous avons veillé à ce que notre interface de sélection textuelle produise comme résultat de sortie un seul fichier regroupant tous les documents concernés par notre requête, codé au format .txt, et facilement exploitable par les outils d'analyse du discours présentés dans le tableau 2.

4. L'alignement du corpus *Babeliris*

Dans le contexte de notre projet *Babeliris*, nous avons aussi composé un corpus parallèle français-néerlandais au départ de l'ensemble de nos textes médico-administratifs bilingues[9]. En réalité, nous avons conçu un corpus spécialisé de traduction unidirectionnel, fondé sur un appariement de segments français-néerlandais, étant donné que la plupart des textes *IRIS* sont traduits du français vers le néerlandais comme le laisse supposer le tableau 3 (40,9 % de documents disponibles uniquement en français, contre 7,6 % pour le néerlandais) :

Tableau 3 – Textes unilingues et traduits du corpus *Babeliris*

	Langue	Nombre total d'écrits	Nombre d'écrits unilingues	Pourcentage d'écrits unilingues	Nombre d'écrits alignables	Pourcentage d'écrits alignables
1	Français	1 126	461	40,9 %	665	59,1 %
2	Néerlandais	720	55	7,6 %	665	92,4 %
	TOTAL	**846**	**516**	**28,0 %**	**1 330**	72,0 %

9. Obéissant au souhait de la structure faîtière de l'IRIS de penser l'aménagement linguistique prioritaire au sein des hôpitaux bruxellois, nous avons décidé d'aligner uniquement les documents disponibles en français et en néerlandais. La banque de textes ainsi alignés, de laquelle naîtront les outils mentionnés en 1, pourra également servir à alimenter les cours de langues sur objectifs spécifiques destinés au personnel (http://www.iris-hopitaal.eu/fr/).

L'alignement phrastique de segments et de sous-segments prend sens à la lumière des objectifs visés par le projet *Babeliris*, puisqu'il permet une extraction aisée de certains termes et de leurs équivalents vulgarisés, dégagés par la comparaison structurelle des langues des bi-textes[10]. Certains de ces termes vulgarisés viendront ensuite alimenter notre base de données terminologique. Par ailleurs, les textes parallèles alignés constitueront une banque de données bilingue, que nous convertirons en une vaste mémoire de traduction à faire valider par les traducteurs de la structure *IRIS* faîtière.

Afin de rencontrer au plus vite ces objectifs, nous avons dû choisir, pour nos écrits, un aligneur multilingue un gestionnaire de mémoires de traduction. Les trois outils testés dans ce cadre, dont les principales caractéristiques sont résumées dans le tableau 4, sont *WinAlign* de *SDL Trados*, *MultiTrans Prism* de *MultiCorpora* et *mkAlign*.

Tableau 4 – Comparaison des aligneurs testés

	Langue	Nombre total d'écrits	Nombre d'écrits unilingues	Pourcentage d'écrits alignables
1	Mise à jour du logiciel	2007	2011	2012
2	Nombre de langues alignables simultanément	2	2	2
3	Coût	Payant	Payant	Gratuit
4	Niveau d'appariement automatique le plus fin	Phrase	Phrase et sous-segment	Phrase
5	Niveau d'intervention manuelle autorisée	Mot	Sous-segment (encadré d'une ponctuation forte)	Mot
6	Méthode d'alignement	Flèches	Surlignage jaune	Tableau à double entrée
7	Format d'exportation échangeable	**.tmw**	**.tmx**	**.tmx**

10. Ainsi, pour un profane du langage médical, le terme *carence martiale* sera moins transparent que son équivalent néerlandais *ijzertekort*, qui littéralement se traduit par *manque de fer* et que nous retiendrons désormais sous cette forme. Par ailleurs, la confrontation des segments source et cible permettra aussi de relever certains cas de vulgarisation traductionnelle : ainsi le terme *vitamine liposoluble* renfermera davantage de renseignements que son équivalent néerlandais *oplosbare vitamine* en raison de la présence du préfixe d'origine grecque λίπος (graisse, corps gras), mais son sens sera plus complexe d'interprétation. C'est pourquoi nous préférerons finalement retenir la version *vitamine soluble* du terme.

Nous avons finalement retenu *Multi Trans Prism* dans sa version Expert. Même si ce dernier a son coût et ne permet pas une intervention manuelle au niveau du mot, il présente toutefois de nombreux avantages :

- Il produit censément des alignements automatisés de textes parallèles
- aussi 1 : N et N : 1[11] – avec une exactitude confirmée avoisinant les 95 %. Son algorithme sophistiqué est basé sur la longueur des phrases et des paragraphes, la mise en page et les marques de ponctuation. Dès lors, cet outil n'exploite ni lexiques bilingues ni cognats[12], si bien qu'en l'absence d'un ancrage lexical, il s'avère indépendant des langues traitées ;
- Il est doté d'une interface aussi ergonomique qu'intuitive et nous a semblé, en ce sens, plus convivial que les deux autres outils testés ;
- Il permet d'exporter tout appariement de segments vers le format universel d'échange de mémoires de traduction .tmx.

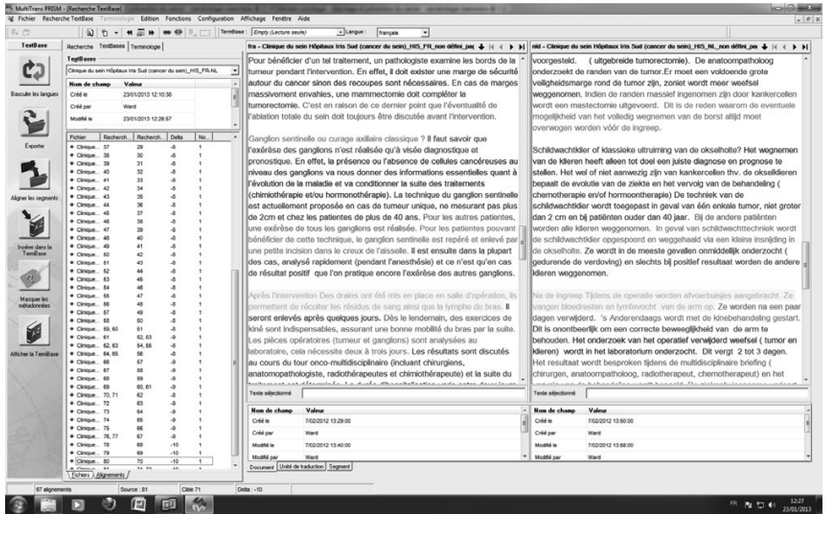

FIGURE 3 – INTERFACE DE L'ALIGNEUR *MULTITRANS PRISM*

11. Pour sa part, Véronis évoque l'idée de correspondances complexes m:n où m, n > 1 (2000 : 159).

12. La notion de *cognat* est traitée entre autres par Véronis (2000 : 159-160).

5. Les principales caractéristiques linguistiques du corpus français Babeliris

Dans cette section, nous exposerons quelques traits linguistiques du pan français de notre corpus (abrégé BAB., 519 393 mots), créé à l'aide de notre interface génératrice de sous-corpus, et les confronterons aux spécificités langagières de trois autres corpus, constitués à des fins de comparaison linguistique[13] :

- Corpus spécialisé de pédiatrie (abrégé PÉD., 154 864 mots) : élaboré au départ du cours polycopié du professeur P. Tounian (2002) de l'Université Pierre et Marie Curie ;
- Corpus spécialisé de notices pharmaceutiques (abrégé NOT., 335 431 mots) : conçu à partir des 141 notices[14] mises à jour entre le 15 et le 31 décembre 2012, glanées sur le site de l'Agence nationale de sécurité du médicament et des produits de santé (ANSM) ;
- Corpus de langue de presse généraliste et vulgarisée *L'Essentiel* (abrégé ess., 156 082 mots) : constitué au départ de 271 articles de journal rédigés en français simplifié et publiés en ligne[15] en Belgique entre le 1er mars 2009 et le 31 décembre 2012.

Très concrètement, nous analyserons notre corpus français *Babeliris* et ces trois autres corpus au regard de cinq caractéristiques

13. Nous sommes parti de l'hypothèse selon laquelle notre corpus *Babeliris* devait sans doute renfermer des caractéristiques linguistiques à la croisée du langage spécialisée et de la langue générale vulgarisée.

14. Sur son site *Internet*, cette agence française définit une notice comme étant une « a*nnexe de la décision d'autorisation comportant des informations notamment sur les indications thérapeutiques, contre-indications, modalités d'utilisation et les effets indésirables d'un médicament* » et précise que « *ces informations sont insérées dans le conditionnement (emballage) contenant le médicament [et sont plus] particulièrement destinée[s] au patient* » (http://agence-prd.ansm.sante.fr/php/ecodex/glossair/glossair.php#Notice).

15. Ces articles sont consultables sur http://www.journal-essentiel.be/.

que celles des textes de spécialité excèdent sans problème les 6,50 bits. Cet écart notable permet dès lors de distinguer un texte littéraire (LITT.) ou de langue générale (ESS.) d'un texte spécialisé (SPÉ.) Ainsi, avec ses 6,80 bits, notre corpus bab. se range du côté du discours spécialisé, marqué par des segments de phrase répétitifs (Van Campenhoudt 2011 : 15).

L'emploi des temps constitue une autre variable pertinente pour reconnaître les écrits spécialisés (*op. cit.* : 21-23). Kocourek nous apprend que dans le texte technoscientifique, *le présent de l'indicatif, à la voix active ou passive, est la forme verbale la plus fréquente, avec environ 85 %. À peu près dix pour cent sont réservés au passé* composé *et au futur, et le reste est réparti entre les autres temps ou modes* (1991 : 71). Le tableau 7 rend compte de la répartition de ces temps, ainsi que de l'imparfait et du passé simple, parmi l'ensemble des verbes (données obtenues à l'aide de la fonction *Étiquetage de textes de Cordial Analyseur*).

TABLEAU 7 – RÉPARTITION DES TEMPS VERBAUX DANS NOS SIX CORPUS

	TEMPS VERBAL	LITT.	SPÉ.	BAB.	PÉD.	NOT.	ESS.
1	Présent	18,60 % - 57,50 %	63,60 % - 77,00 %	60,80 %	78,40 %	51,60 %	63,10 %
2	Futur simple	1,70 % - 5,10 %	0,80 % - 9,80 %	6,70 %	2,50 %	2,90 %	3,70 %
3	Passé composé	3,20 % - 12,60 %	2,80 % - 12,80 %	4,60 %	3,00 %	6,70 %	11,50 %
4	Imparfait	4,80 % - 30,20 %	0,20 % - 3,10 %	0,70 %	0,60 %	0,40 %	4,20 %
5	Passé simple	0,30 % - 22,50 %	0,00 % - 1,70 %	0,20 %	0,10 %	0,00 %	0,10 %

Le temps qui domine dans BAB., comme dans PÉD., est le présent de l'indicatif. Quant aux emplois du futur simple et du passé composé (11,30 %), ils atteignent largement les dix pour cent évoqués par Kocourek (1991 : 71). Le futur simple est d'ailleurs très répandu dans bab., peut-être parce que « *dans les textes médicaux, [il] semble souvent utilisé pour donner des consignes en évitant la fonction conative* » (Van Campenhoudt 2011 : 22). En réalité, cette stratégie s'avère fréquente lorsqu'il s'agit de donner des instructions sous une forme modalisée. Enfin, l'imparfait et le passé simple semblent davantage réservés aux textes littéraires et sont presque absents de BAB., de PÉD., de NOT. et de ESS. rédigé en *français simplifié*.

Quatrième caractéristique d'un texte spécialisé : un faible embrayage pronominal. Van Campenhoudt évoque *une très faible présence des pronoms personnels des premières et deuxièmes personnes dans les textes spécialisés […] L'exception la plus notable est celle des modes d'emploi, instructions et autres notices vulgarisatrices* (*op. cit.* : 18-19). Pour sa part, Kocourek nous apprend que *le premier trait saillant de la syntaxe technoscientifique est la prédominance de la 3e personne* (1991 : 71). Ces traits langagiers ne semblent toutefois pas respectés dans BAB., comme l'indique le tableau 8 résumant le pourcentage de pronoms personnels embrayés par rapport à l'ensemble des pronoms personnels (mesures recueillies à l'aide de la fonction *Étiquetage de textes* de *Cordial Analyseur*).

Tableau 8 – Embrayage pronominal dans nos six corpus

	Pron. Embrayé	litt.	spé.	bab.	péd.	not.	ess.
1	1re pers. sg	24,30 % - 33,30 %	0,00 % - 6,00 %	5,40 %	0,00 %	0,00 %	3,50 %
2	2e pers. sg.	0,10 % - 1,60 %	0,00 % - 0,30 %	0,60 %	0,30 %	0,20 %	0,40 %
3	1re pers. pl.	7,20 % - 9,60 %	0,00 % - 6,50 %	7,90 %	0,60 %	0,00 %	4,00 %
4	2e pers. pl.	2,00 % - 6,20 %	0,00 % - 1,90 %	32,10 %	0,00 %	63,10 %	3,90 %

L'utilisation de la première personne, surtout au pluriel (1 358 occ. de *nous*), est très fréquente dans BAB. et désigne souvent le personnel *IRIS*. Quant à l'emploi du pronom de deuxième personne, il semble surreprésenté au pluriel de politesse (5 982 occ. de *vous*) et évoque la patientèle du réseau hospitalier. Cette distribution pronominale atypique nous empêche tout logiquement de rapprocher BAB. d'un autre corpus à notre disposition, même si celui-ci se démarque, tout comme dans NOT., par une forte présence de la 2e p. pl.

Enfin, même s'il convient de les interpréter avec prudence en l'absence d'une analyse sémantique rigoureuse, les longueurs moyennes des mots et des phrases ainsi que l'indice de Flesch-De Landsheere permettent aussi de caractériser un texte spécialisé. Kocourek écrit ainsi que « *les phrases technoscientifiques se caractérisent par leur longueur […] moyenne […] entre 28 et 29 mots* » (1991 : 73). Les mots du discours scientifique semblent aussi suivre cette tendance à être longs (Van Campenhoudt 2011 : 10). Enfin, l'indice de Flesch-De Landsheere permet d'évaluer grossièrement la lisibilité d'un texte, qui sera généralement inférieure pour un écrit spécialisé (*ibid.* : 13). Le tableau 9 précise,

pour nos six corpus, les trois mesures lexicométriques susdécrites, définies avec *Cordial Analyseur* et *TextStat*.

Tableau 9 – Quelques relevés lexicométriques de nos six corpus

	Mesure lexicométrique	litt.	spé.	bab.	péd.	not.	ess.
1	Nombre moyen de lettres par mot	4,32 - 4,61	4,60 - 5,29	5,19	5,39	5,42	4,71
2	Nombre moyen de mots par phrase	16,29 - 23,16	27,09 - 46,45	10,95	15,22	10,53	11,77
3	Indice de Flesch-De Landsheere	41,86 - 48,32	19,00 - 42,66	17,63	4,24	14,55	32,35

La longueur moyenne des phrases de BAB. (mais aussi de NOT. et de PÉD.) étant biaisée par la présence de nombreux tableaux et de listes à puces, nous nous focaliserons uniquement sur les variables ***1*** et ***3***. S'agissant du nombre moyen de lettres par mot, on notera que les termes de BAB., de PÉD. et de NOT. sont relativement longs et rejoignent en ce sens les modalités d'écriture de SPÉ, à la différence de ceux de ESS. qui sont censés être accessibles à un large public. Enfin, des indices de Flesch-De Landscheere bas pour BAB., PÉD. et NOT. attestent d'une véritable complexité, d'autant plus marquée qu'ils sont largement inférieurs aux scores de spé. Les textes de BAB., de PÉD. et de NOT. pourraient donc être qualifiés de « *très difficiles* » (De Landsheere 1982).

Nous serions tenté d'en conclure pour l'instant que nous sommes en présence d'un corpus original, qui laisse apparaître des caractéristiques globales du langage spécialisé (parties du discours, emploi des temps, information mutuelle et mesures lexicométriques) et qui se démarque par une sur-représentativité de la fonction conative, typique des notices pharmaceutiques et des modes d'emploi.

6. Conclusion et perspectives

Aujourd'hui, de nombreuses études empiriques du langage appliquent les méthodes de la linguistique de corpus. L'exploitation d'une vaste collection de textes s'est ainsi imposée à nous comme une évidence.

Dans cette logique, il nous a fallu procéder à la *phase de préparation du corpus* (Sansonetti 2003 : 73) ou, pour le dire avec les mots d'Habert, à la *phase de dépouillement du corpus*, à savoir à cette « *exigence de standardisation des textes contenus dans un corpus […] destinée avant tout à les rendre comparables, à les stabiliser le temps d'une expérience* » (1997 : 194). C'est dans ce contexte que se sont inscrites nos réflexions sur les balises *TEI-P5* et sur un aligneur *ad hoc* pour nos textes bilingues. La genèse de notre interface génératrice de sous-corpus s'est, elle aussi, posée comme un préalable à l'étude linguistique, en l'espèce du pan français de notre corpus *Babeliris*. À l'aide de cet outil, nous avons facilement pu en conclure que le français des hôpitaux *IRIS* rejoignait celui des textes spécialisés marqués par une prédominance de la 2e p. pl.

Enfin, les perspectives d'exploitation pour notre corpus sont nombreuses. On pourrait imaginer générer des sous-corpus dans le but d'analyser, par exemple, les spécificités langagières propres à chaque hôpital *IRIS* ou à chaque unité. Il serait aussi intéressant de développer un outil qui croiserait interface de sous-corpus et aligneur, en vue d'étudier les procédés traductionnels spécifiques à un pan de la littérature *IRIS* et d'en envisager la vulgarisation. C'est que notre objectif in fine est, avant tout, d'encourager la lisibilité des écrits destinés au patient.

Bibliographie

BAUMAN (S.) et BURNARD (L.) 2012, TEI-P5: *Guidelines for Electronic Text Encoding*, Charlottesville, Text Encoding Initiative Consortium, version révisée 2.0.2. (www.tei-c.org/guidelines/p5)

BIBER (D.) 1993, « Representativeness in Corpus Design », in *Literacy and Linguistic Computing*, 8/4, Oxford, pp. 243-257.

BOÏTEKÉ (P.) 2011, *État de la question. Emploi des langues à Bruxelles : mieux tenir compte de la réalité sociologique bruxelloise*, rapport commandité par la Fédération Wallonie-Bruxelles et réalisé par l'Institut Emile Vandervelde, Bruxelles.

CHURCH (K.W.) et HANKS (P.) 1990, « Word Association Norms, Mutual Information, and Lexicography », in *Computational Linguistics*, Cambridge, MA, 16/1, pp. 22-29.

DAOUST (F.) et MARCOUX (Y.) 2006, « Logiciels d'analyse textuelle : vers un format XML-TEI pour l'échange de corpus annotés », in *Actes des 8es journées internationales d'analyse statistique des données textuelles (JADT) 2006*, Besançon, Presses universitaires de Franche-Comté, 1, pp. 327-340.

DE LANDSHEERE (G.) 19825, *Introduction à la recherche en éducation*, Liège, Thone.

DUFOURNAUD (N.) et GRATSAC-LEGENDRE (V.) 2012, *Manuel d'encodage XML-TEI – édition numérique de manuscrits baroques. Recommandations pour une application TEI ENBaCH, École des hautes études en sciences sociales* (EHESS), Paris, 60 pp.

HABERT (B.) 2000, « Des corpus représentatifs : de quoi, pour quoi, comment ? », in *Linguistique sur corpus. Études et réflexions*, M. Bilger, Perpignan, Presses Universitaires de Perpignan, pp. 11-58.

HABERT (B.), NAZARENKO (A.) et SALEM (A.) 1997, *Les linguistiques de corpus*, Paris, Masson et Armand Colin, 240 pp., Coll. U Linguistique.

HAREIDE (L.) et HOFLAND (K.) 2012, « Compiling a Norwegian-Spanish parallel corpus », in OAKES (M.P.) et JI (M.), *Quantitative Methods in Corpus-Based Translation Studies*, Studies in Corpus Linguistics, 51, Amsterdam, John Benjamins Publishing Company, pp. 75-113.

HAUTE AUTORITÉ DE LA SANTÉ 2008, *Élaboration d'un document écrit d'information à l'intention des patients et des usagers du système de santé*, Saint-Denis, Haute Autorité de Santé.

HUIGEN (M.) 2004, *Zelf brochures schrijven*, coll. TaalAnker: *Hoe formuleer ik het? Cahierreeks voor taalgebruikers*, 48, Alphen aan den Rijn, Kluwer.

ISO 639-1 2002, *Codes pour la représentation des noms de langue. Partie 1 : Code alpha-2*, Genève, Organisation internationale de normalisation.

JACQUEMART (P.) *s.d.*, *Conception d'un corpus de textes médicaux*, rapport de stage effectué dans le laboratoire DIAM–SIM/DSI /AP-HP sous la direction de P. Zweigenbaum, Paris (www.limsi.fr/npz/ftpapiers/jacquemartdea99.doc).

KOCOUREK (R.) 1991, *La langue française de la technique et de la science. Vers une linguistique de la langue savante*, Wiesbaden, Oscar Brandstetter Verlag & co.

LEE (D.) 2001, « Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle », in *Language Learning & Technology*, s.l., 5/3, pp. 37-72.

McENERY (T.), TONO (Y.) et XIAO (R.) 2006, *Corpus-Based Language Studies: An Advanced Resource Book*, Routledge Applied Linguistics, Londres, Routledge.

MINISTÈRE BELGE DES AFFAIRES SOCIALES, DE LA SANTÉ PUBLIQUE ET DE L'ENVIRONNEMENT 2002, *Loi relative aux droits du patient*, publiée au *Moniteur belge* le 26 septembre 2002, pp. 43719-43724.

POPESCU-BELIS (A.) 2002, « Constitution de banques de textes multilingues : un mécanisme fondé sur le standard XML », in *Cahiers du Rifal*, Paris, 23, pp. 56-60.

SACKETT (D.L.) *et al.* 1996, « Evidence based medicine: what it is and what it isn't », in *BMJ*, 312, pp. 71-72.

SANSONETTI (L.) 2003, « Approche lexicométrique de corpus d'interactions verbales entre un adulte et un enfant en cours d'acquisition du langage. Résultats d'expérience », in WILLIAMS (G.), *Textes et Corpus : Actes des Troisièmes Journées de la Linguistique de Corpus*, pp. 71-84 (web.univ-ubs.fr/corpus/jlc3/jlc2003.pdf).

SINCLAIR (J.) 1991, *Corpus, Concordance, Collocation. Describing English Language*, Oxford, University Press.

SINCLAIR (J.) 1996, *Preliminary Recommendations on Corpus Typology*, rapport technique, EAGLES (Expert Advisory Group on Language Engineering Standards), Birmingham University (www.ilc.cnr.it/eagle596/browse.html).

TOGNINI-BONELLI (E.) 2001, *Corpus Linguistics at Work*, Amsterdam, John Benjamins Publishing.

TOUNIAN (P.) 2002, Pédiatrie, *DCEM 3*, Paris, Université Pierre et Marie Curie, cours polycopié.

TYMOCZKO (M.) 1998, « Computerized Corpora and the Future of Translation Studies », in CLAS (A.), *Meta : journal des traducteurs*, Montréal, 43/4, pp. 652-660.

VAN CAMPENHOUDT (M.) 2011, *Linguistique française et exercices III. Le français spécialisé*, Bruxelles, Institut supérieur de traducteurs et interprètes, cours polycopié de 3e bachelier, document interne.

VAN DE VELDE (W.) *à paraître,* « La construction du corpus d'écrits médico-administratifs BABELIRIS : vers une méthodologie de compilation et d'encodage des textes », in *Actes de la conférence Traduction et Innovation*, Université Paris Diderot (Paris VII), 13–15 décembre 2012.

VAN OMMEN (H.) et VAN KUPPENVELD (E.) 19953, *Professionele bedrijfscommunicatie. Het handboek voor tekstschrijvers*, Groningen, Wolters-Noordhoff.

VÉRONIS (J.) 2000, « Alignement de corpus multilingues », in PIERREL (J.-M.), *Ingénierie des langues. Informatique et systèmes d'information*, chapitre 6, Paris, Hermès Sciences.

VIHLA (M.) 1998, « MEDICOR: a corpus of contemporany American medical texts », in *ICAME Journal*, Bergen, 22, pp. 73-80.

ZWEIGENBAUM (P.) et al. 2001, « Building a text corpus for representing the variety of medical language », in V. Patel *et al., Medinfo 2001,* Proceedings Corpus Linguistics, Lancaster, pp. 290-294.