

Le Traitement Automatique des Langues (TAL) comme aide au traducteur : aperçu des outils

Adrien Dubied

Juin 2021

1 Introduction

Les "Humanités Numériques", ce concept entièrement nouveau pour moi il y a deux ans a immédiatement éveillé ma curiosité de par son approche pratique et interdisciplinaire, son caractère accessible et ses outils modernes. Il était question d'art, d'histoire de l'art, d'histoire, en bref, des Sciences Humaines. Sans être fin connaisseur, ni un spécialiste en la matière, j'ai toute fois toujours été très sensible à ces domaines. C'est en creusant d'avantage que j'ai réalisé toute l'ampleur des connaissances que cette nouvelle façon de faire de la recherche pouvait apporter, non seulement aux domaines en question, mais aussi à une multitude de disciplines diverses et variées. En effet, l'avenir des activités langagières ne se jouera pas sans les outils numériques qui permettent le TAL. Depuis Internet 2.0 et le Big Data, nous disposons d'une quantité infinie d'informations et notamment de textes disponibles en ligne, surtout en anglais. Depuis le spectre des Humanités Numériques, Franco Moretti (2005) [10] a exposé les avantages du Distant Reading (lecture en Corpus) par opposition au Close Reading (lecture classique). Pour traiter et étudier une telle quantité de textes le TAL est effectivement devenu une pratique incontournable. Mon travail consiste à présenter certains outils comme une aide pouvant servir aux langagiers souhaitant exploiter des corpus spécialisés, de domaines relativement nouveaux et méconnus du grand public, pour en tirer un lexique multilingue destiné à la transmission du savoir à des cultures linguistiques cibles.

Je suis le développement de la technologie Blockchain depuis fin 2017. Cette technologie disruptive amènera probablement la plus importante révolution technologique de ce début de XXI^e siècle. Selon le Service de Recherche du Parlement Européen (EPRS)[6], les différents concepts qui la composent comme la décentralisation, le pseudonymat, l'immuabilité des données, etc., ont le potentiel d'apporter des solutions à différentes problématiques sociétales et restent pourtant largement méconnus du grand public. La grande majorité des informations disponibles est en anglais et les concepts cryptographiques de la Blockchain sont difficilement appréhendables. Le partage multilingue des connaissances au travers de textes, mais aussi de ressources terminologiques spécialisées et vulgarisées constitue pour les publics ciblés un facteur clé à la compréhension et à l'adoption consciente de cette nouvelle technologie. Par mon travail basé sur l'étude d'un corpus multilingue aligné par phrases ou par segments (bi-texte), je souhaite présenter comment

la constitution et la publication d'un lexique multilingue pourraient être menée à bien en me servant de différents outils numériques puis, en abordant leurs avantages et leurs limites. Il ne s'agit donc pas d'un réel travail terminologique avec toutes les connaissances, la profondeur et la complexité que cela impliquerait. Le résultat final est très inabouti. En revanche, les outils présentés peuvent servir aux traducteurs désirants acquérir des connaissances dans un domaine nouveau et construire une mémoire de traduction qui pourra être utilisée dans un Outil de Traduction Assistée par Ordinateur (TAO).

2 Corpus

Un bi-texte est en réalité composé de deux corpus unilingues dont les phrases ou segments ont été alignés de sorte à ce qu'un segment en langue source soit systématiquement aligné à son correspondant en langue cible dans un tableur. Il est clair que pour disposer d'un tel corpus, il faut disposer de textes déjà traduits. La première étape pour la constitution de mon corpus consiste alors à sélectionner des textes disponibles en ligne gratuitement, en plusieurs langues et de source fiable. J'ai décidé de travailler sur les versions anglophones et francophones d'une étude publiée en 2017 par l'EPRS s'intitulant "How Blockchain technology could change our lives". Il s'agit d'un texte de près de 13'000 mots en anglais et 16'000 en français. Il s'agit d'un texte informatif et plutôt accessible au grand public, c'est à dire que les termes et les concepts sont bien expliqués. Il s'agit donc d'un texte parfait pour l'objectif de ce travail. Il est évident qu'il ne s'agit là que d'un échantillon de corpus et que pour arriver à un travail abouti, un corpus devrait contenir environ 1'000'000 de mots, selon les recommandations de Sinclair (2004)[12]. S'agissant d'un domaine de spécialité, quelques centaines de milliers de mots devraient suffire. Je n'ai donc pas eu besoin d'employer les techniques de "Scraping", développées par Marres (2013)[8] permettant d'aspirer automatiquement des contenus en ligne.

3 Choix des outils

Afin de sélectionner les outils nécessaires il faut d'abord identifier les différentes étapes qui composent le travail :

- Nettoyer le corpus
- Aligner le corpus
- Exploiter le corpus
- Structurer l'information tirée du Corpus
- Partager l'information

3.1 Nettoyer et aligner le Corpus avec Notepad++, RegEx et Excel

Les documents étant disponibles sur internet au format PDF, j'ai utilisé un convertisseur gratuit en ligne pour les transformer en format txt. Ainsi l'étape consistant à

En effet, pour créer un bi-texte qui puisse être lu comme tel par un Corpus Management System[11], tel que Sketch Engine, il faudra copier et coller toutes les lignes du texte source et cible dans une colonne d'un tableur en s'assurant que chaque phrase ou chaque

segment de mot soit aligné. Cette opération d'alignement est très chronophage et nécessite l'intervention humaine. Il s'agit de fusionner les différentes cellules d'une langue qui ne correspondent qu'à une seule cellule en langue correspondante. Dans mon cas, cela n'a pas duré plus d'une heure car le corpus ne comporte que 13'000, respectivement 16'000 mots. Pour mener à bien mon travail, étant donné qu'il s'agit d'un domaine de spécialité, j'estime qu'un corpus dix fois plus large aurait été suffisant pour créer une liste de termes qui soit représentative. Il faut évidemment que les textes sélectionnés proviennent d'instituts reconnus dans le domaine et qu'ils couvrent tous les sujets impliquant la technologie en question. Il faut également noter que plus le domaine évolue, plus il faudra mettre à jour le corpus en y ajoutant des textes. En dehors de ces contraintes chronophages cet exercice ne requiert aucune connaissance approfondie des fonctionnalités d'un tableur comme Microsoft Excel. Il s'agit simplement de deux colonnes avec sur la première ligne, la langue, en général exprimé en anglais. Dans chaque colonne, les segments sont introduits dans l'ordre, en face du segment correspondant :

A	B
English	French
How blockchain technology could change our lives	Comment la technologie de la chaîne de bloc pourrait changer nos vies
In-depth Analysis	Analyse approfondie
February 2017	févr. 17
PE 581.948	PE 581.948
STOA - Science and Technology Options Assessment	STOA - Évaluation des choix scientifiques et technologiques
AUTHORS	AUTEURS
Philip Boucher, Scientific Foresight Unit (STOA), DG EPRS, European Parliament	Philip Boucher, Unité de la prospective scientifique (STOA), DG EPRS, Parlement Européen
Susana Nascimento, Foresight, Behavioural Insights and Design for Policy Unit, DG JRC, European Commission (Chapters 6-8)	Susana Nascimento, Unité de la prospective, des études comportementales et de la conception déolotiques, DG JRC, Commission Européenne (parties 6 à 8)
Mihailis Kritikos, Scientific Foresight Unit (STOA), DG EPRS, European Parliament(Anticipatory Policy-Making sections)	Mihailis Kritikos, Unité de la prospective scientifique (STOA), DG EPRS, Parlement Européenparties sur l'élaboration de politiques d'anticipation)
ABOUT THE PUBLISHER	À PROPOS DE L'ÉDITEUR
To contact STOA or to subscribe to its newsletter please write to: STOA@ep.europa.eu	Pour contacter la STOA ou pour vous abonner à sa lettre d'information, veuillez écrire à l'adresssuivante: STOA@ep.europa.eu
This document is available on the Internet at: http://www.ep.europa.eu/stoa/	Ce document est disponible sur Internet à l'adresse suivante: http://www.ep.europa.eu/stoa/
Manuscript completed in February 2017	Rédaction achevée en février 2017
Brussels, © European Union, 2017	Bruxelles, © Union européenne, 2017
DISCLAIMER	CLAUDE DE NON-RESPONSABILITÉ
This document is prepared for, and addressed to, the Members and staff of the European Parliament as background material to assist them in their parliamentary work.	Le présent document est rédigé à l'attention des membres et du personnel du Parlement européen dane but de les aider dans leur travail parlementaire.
The content of the document is the sole responsibility of its author(s) and any opinions expressed herein should not be taken to represent an official position of the Parliament.	Le contenu de ce document relève de la responsabilité exclusive des auteurs et les avis qui y sont exprimés ne reflètent pas nécessairement la position officiellu Parlement européen.
Reproduction and translation for non-commercial purposes are authorised, provided the source is acknowledged and the European Parliament is given prior notice and sent a copy.	La reproduction et la traduction sont autorisées, sauf à des fins commerciales, moyennant mention de la source, information préalable du Parlement européen et transmission d'un exemplaire à celui-ci.
How blockchain technology could change our lives	Comment la technologie de la chaîne de blocs pourrait changer nos vies
Table of contents	Sommaire
How does blockchain technology work? 4	Comment la technologie de la chaîne de blocs pourrait changer nos vies..... 4
How does blockchain technology work? 5	Comment les chaînes de blocs fonctionnent-elles?..... 5

Fig. 2 - Cette approche simpliste et intuitive de données tabulaires organisées en colonnes (formalisées) peut être créée sous format .xlsx (tableaux) ou .csv (comma seperated value)

3.2 Exploiter et analyser le Corpus avec SketchEngine

Une fois le bi-texte terminé, l'outils SketchEngine permet son importation facile et intuitive. Il faut indiquer au programme qu'il s'agit d'un corpus bilingue. Il reconnaît dans la majorité des cas la paire de langue automatiquement grâce à la première ligne du tableur. Dans le cas contraire on peut simplement l'indiquer manuellement à l'aide de listes déroulantes. SketchEngine enregistre les corpus séparément. Nous avons donc un corpus en anglais et un autre an français qui peuvent être consultés séparément.

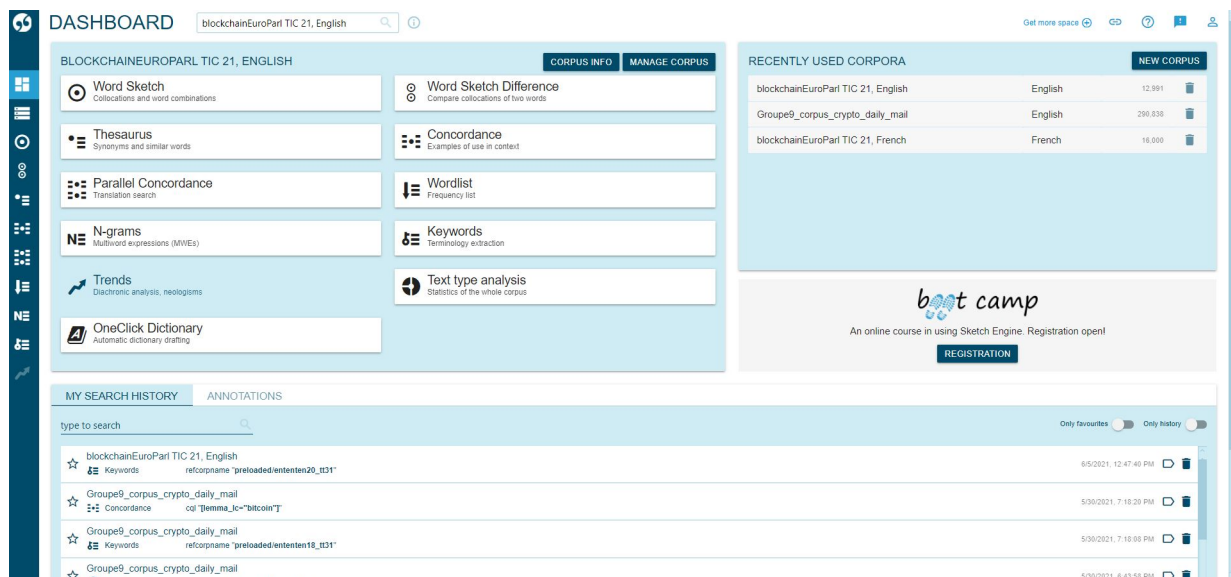


Fig. 3 - En haut à droite, les deux corpus blockchainEuroParl TIC 21. En haut à gauche, les différentes fonctionnalités d'exploration du corpus.

En se basant sur la fonction basique Keywords (en français, mots-clés) on peut facilement observer les mots simples ou composés qui apparaissent avec une fréquence élevée dans le corpus. On peut y sélectionner une liste de termes. Pour ce travail, j'ai repéré 19 termes spécifiques à la blockchain et essentiels à sa compréhension. Je rappelle qu'il ne s'agit pas d'un réel travail terminologique mais uniquement d'une exposition d'utilisation des outils numériques permettant de l'accomplir. Raison pour laquelle, les critères de sélection des termes ne sont pas développés dans ce travail.

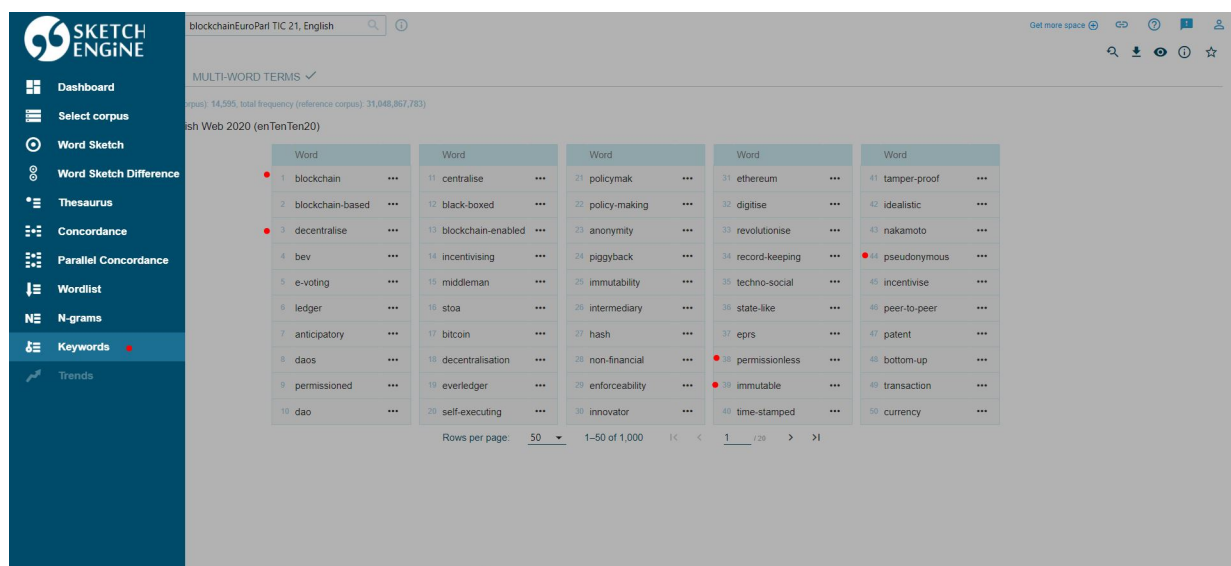


Fig. 4 - Les points rouges indiquent la fonction et les différents termes pouvant être explorés.

Il est ensuite possible de consulter ces termes grâce à la fonction Parallel Concordance qui va montrer d'un côté les occurrences du terme dans son contexte en langue source, de

l'autre dans son contexte en langue cible. Nous avons donc une vision alignée des apparitions en contexte. Cela permet non seulement d'observer comment le terme a été traduit, mais aussi de créer ses propres définitions dans les deux langues grâce à l'observation des Contextes Riches en Connaissances (CRC), selon Meyer (2001) [9]. Il s'agit d'une notion en terminologie qui désigne tous les éléments lexicaux et syntaxiques qui donnent une information sur un terme. Nous avons par exemple des mots ou phrases comme "est", "fonctionne", "il s'agit de", "fait partie de A", "signifie", "comprend A et B", "est une forme de", etc. Il y a des relations au niveau conceptuel entre les termes qui peuvent être synonymiques, hiérarchiques, hyperonymiques, méronymiques, etc.

① docR0	<s> How blockchain technology could change our lives </s>	<s> Comment la technologie de la chaîne de blocs pourrait changer nos vies </s>
① docR0	<s> Blockchains are a remarkably transparent and decentralised way of recording lists of transactions. </s>	<s> Les chaînes de blocs constituent un moyen remarquable de tenir un registre de transactions de manière transparente et décentralisée. </s>
① docR0	<s> Their best-known use is for digital currencies such as Bitcoin, which announced blockchain technology to the world with a headline-grabbing 1000% increase in value in the course of a single month in 2013. </s><s> This bubble quickly burst, but steady growth since 2015 means Bitcoins are currently valued higher than ever before. </s>	<s> Elles sont surtout connues pour être à la base de devises numériques comme le bitcoin, qui a mis la technologie de la chaîne de blocs sous le feu des projecteurs lorsqu'il s'espérait de 1 000 % en l'espace d'un seul mois en 2013. </s><s> Cette bulle avait alors éclaté rapidement, mais bitcoin connaît une croissance régulière depuis 2015 et son cours atteint aujourd'hui un niveau plus élevé que jamais. </s>
① docR0	<s> There are many different ways of using blockchains to create new currencies. </s>	<s> Il existe de nombreuses manières de créer de nouvelles monnaies à l'aide des chaînes de blocs. </s>
① docR0	<s> Blockchains are particularly well suited to situations where it is necessary to know ownership histories. </s>	<s> Les chaînes de blocs conviennent particulièrement dans les situations où il est nécessaire de conserver un historique de propriété. </s>
① docR0	<s> They also present opportunities in all kinds of public services such as health and welfare payments and, at the frontier of blockchain development, are self-executing contracts paving the way for companies that run themselves without human intervention. </s>	<s> Elles ouvrent également des possibilités dans toutes sortes de services publics comme le remboursement des frais de santé et le versement des prestations sociales. </s><s> À la pointe développement des chaînes de blocs se trouvent même les contrats intelligents, qui laissent entrevoir l'avenir où des entreprises autonomes fonctionneraient sans intervention humaine. </s>
① docR0	<s> Blockchains shift some control over daily interactions with technology away from central elites, redistributing it among users. </s>	<s> Les chaînes de blocs redistribuent aux utilisateurs une partie du contrôle qu'exercent les élites centralisées sur les interactions technologiques du quotidien. </s>
① docR0	<s> Indeed, the governments and industry giants investing heavily in blockchain research and development are not trying to make themselves obsolete, but to enhance their services. </s>	<s> En effet, les États et les géants du secteur qui investissent massivement dans la recherche et le développement en matière de chaînes de blocs n'ont pas pour objectif de se rendre eux-mêmes obsolètes, mais plutôt d'améliorer leurs services. </s>
① docR0	<s> For example, blockchain's transparency is fine for matters of public record such as land registries, but what about bank balances and other sensitive data? </s><s> It is possible (albeit only sometimes and with substantial effort), to identify the individuals associated with transactions. </s>	<s> Par exemple, la transparence des chaînes de blocs convient bien aux registres publics tels que les cadastres mais qu'en est-il des comptes bancaires et autres données sensibles? </s><s> Il est possible, quoique seulement parfois et au prix d'efforts substantiels, d'identifier les personnes associées aux transactions. </s>
① docR0	<s> While some blockchains do offer full anonymity, some sensitive information simply should not be distributed in this way. </s>	<s> Si certaines chaînes garantissent bien un anonymat total, certaines informations sensibles ne doivent tout simplement pas être distribuées de cette façon. </s>
① docR0	<s> Nevertheless, although blockchains are not the solution for every problem and even if they will not revolutionise every aspect of our lives, they could have a substantial impact in many areas and it is necessary to be prepared for the challenges and opportunities they present. </s>	<s> Néanmoins, même si les chaînes de blocs ne sont pas la solution à tous les problèmes et ne révolutionneront pas tous les aspects de nos vies, elles pourraient avoir des répercussions importantes dans de nombreux domaines, et il est nécessaire d'être préparé aux défis qu'elles présentent et aux opportunités qu'elles offrent. </s>
① docR0	<s> This report provides an accessible entry point for those in the European Parliament and beyond who are interested in learning more about blockchain development and its potential impacts. </s>	<s> La présente analyse informe, de manière accessible, celles et ceux, au Parlement européen et ailleurs qui souhaitent en savoir plus sur le développement des chaînes de blocs et leurs retombées potentielles. </s>
① docR0	<s> The section immediately below presents an introduction to how blockchain technology works. </s>	<s> La première partie présente un aperçu du mode de fonctionnement de la technologie de la chaîne de blocs. </s>
① docR0	<s> Finally, a concluding section presents some overall remarks and potential responses to blockchain development. </s>	<s> Enfin, l'inclusion formule quelques remarques générales et offre plusieurs réponses possibles au développement des chaînes de blocs. </s>
① docR0	<s> How does blockchain technology work? </s>	<s> Comment les chaînes de blocs fonctionnent-elles? </s>
① docR0	<s> Before attempting to understand how blockchain ledgers work, it is worth taking a look at traditional ledgers. </s>	<s> Avant de chercher à comprendre comment fonctionnent les registres reposant sur des chaînes de blocs il est utile de se pencher sur les registres traditionnels. </s>
① docR0	<s> Blockchain offers the same record-keeping functionality but without a centralised architecture. </s>	<s> Les chaînes de blocs offrent la même fonctionnalité de tenue de registres, mais sans architecture centralisée. </s>

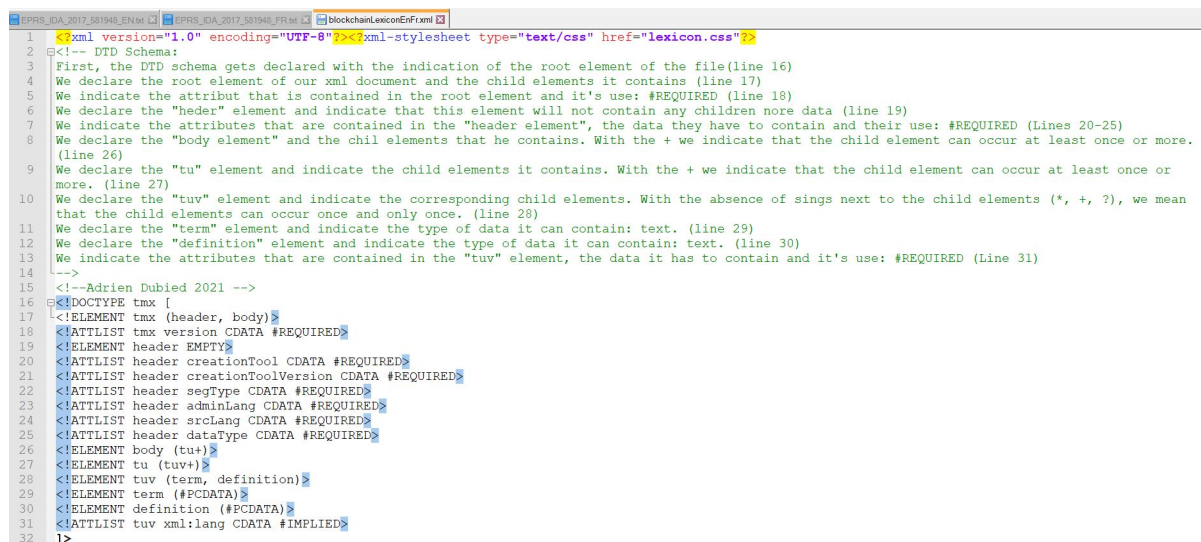
Fig. 5 - Le point rouge indique la fonction Parallel Concordance, surlignés en jaune, quelques exemples de CRC

Sketch Engine est un outil très intuitif. Il contient des vidéos tutoriels pour chaque fonctionnalité qui sont postées en ligne et qui permettent d'apprendre à utiliser l'outil. De plus, le service client est de qualité et rapide. Le seul inconvénient que je connais, est que le programme est payant si l'on n'y a pas accès depuis son compte universitaire.

3.3 Structurer et Modéliser l'information avec XML, XSLT, HTML et CSS

XML a été développé par le World Wide Web Consortium[1] en 1999 en tant que projet open source. Il s'agit d'un langage de balisage extensible, c'est à dire, adaptable à d'autres langages informatiques tels que HTML. Il permet de structurer l'information, de la manipuler et de la partager grâce à sa fonction d'interopérabilité. Il est caractérisé par ses <balises> et est basé sur des standards comme le Standard Generalized Markup Language et Unicode, également développé par le World Wide Web Consortium. Il permet notamment de gérer des données terminologiques et langagières structurées sous forme d'arbre ou de thésaurus. Il représente un avantage considérable pour le partage et la

gestion multilingue de données interopérables. Ses inconvénients sont liés au fait qu'il faille apprendre ce langage avant de pouvoir s'en servir, même s'il est plus simple que HTML, et que pour garantir l'interopérabilité des données, il faille se tenir strictement à un schéma prédéfini, par exemple dans un XML Schema. La moindre différence entre le document XML et son schéma peut entièrement compromettre les données traitées. J'ai placé le schéma XML dans le document XML mais il peut également se trouver sur un document à part. On y accède en y faisant référence sur le document XML. Ceci à l'avantage de pouvoir être appliqué à un nombre infini de documents. J'ai utilisé le programme Oxygen pour la création du Schéma, du document XML et la génération automatique du document HTML/CSS. Ce programme est spécifiquement conçu pour XML et ses différentes extensions possibles. XML peut également être utilisé sur un simple outil de traitement de texte comme la fonction Bloc-Note ou Notepad++. Le logiciel est payant si l'on n'y a pas accès avec son compte universitaire.



```

1 <?xml version="1.0" encoding="UTF-8"?><?xml-stylesheet type="text/css" href="lexicon.css"?>
2 <!-- DTD Schema:
3 First, the DTD schema gets declared with the indication of the root element of the file(line 16)
4 We declare the root element of our xml document and the child elements it contains (line 17)
5 We indicate the attribut that is contained in the root element and it's use: #REQUIRED (line 18)
6 We declare the "header" element and indicate that this element will not contain any children nore data (line 19)
7 We indicate the attributes that are contained in the "header element", the data they have to contain and their use: #REQUIRED (Lines 20-25)
8 We declare the "body element" and the chil elements that he contains. With the + we indicate that the child element can occur at least once or more.
  (line 26)
9 We declare the "tu" element and indicate the child elements it contains. With the + we indicate that the child element can occur at least once or
  more. (line 27)
10 We declare the "tuv" element and indicate the corresponding child elements. With the absence of sings next to the child elements (*, +, ?), we mean
  that the child elements can occur once and only once. (line 28)
11 We declare the "term" element and indicate the type of data it can contain: text. (line 29)
12 We declare the "definition" element and indicate the type of data it can contain: text. (line 30)
13 We indicate the attributes that are contained in the "tuv" element, the data it has to contain and it's use: #REQUIRED (Line 31)
14 -->
15 <!--Adrien Dubied 2021 -->
16 <!DOCTYPE tmx [
17 <!ELEMENT tmx (header, body)>
18 <ATTLIST tmx version CDATA #REQUIRED>
19 <ELEMENT header EMPTY>
20 <ATTLIST header creationTool CDATA #REQUIRED>
21 <ATTLIST header creationToolVersion CDATA #REQUIRED>
22 <ATTLIST header segType CDATA #REQUIRED>
23 <ATTLIST header adminLang CDATA #REQUIRED>
24 <ATTLIST header srcLang CDATA #REQUIRED>
25 <ATTLIST header dataType CDATA #REQUIRED>
26 <ELEMENT body (tu+)>
27 <ELEMENT tu (tuv+)>
28 <ELEMENT tuv (term, definition)>
29 <ELEMENT term (#PCDATA)>
30 <ELEMENT definition (#PCDATA)>
31 <ATTLIST tuv xml:lang CDATA #IMPLIED>
32 ]>

```

Fig. 6 - Voici la définition du schéma XML

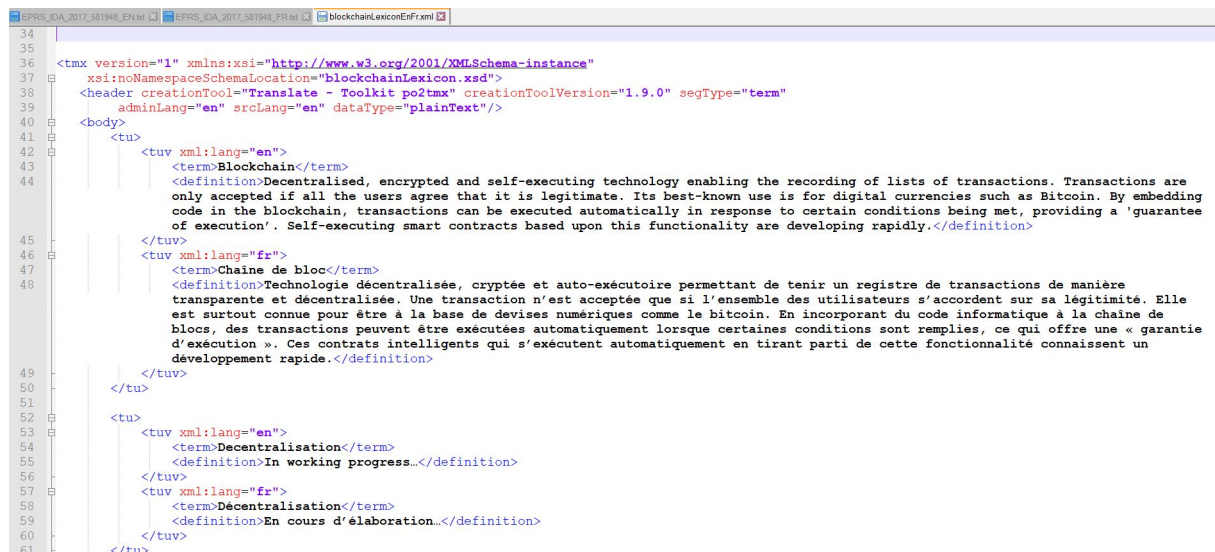


Fig. 7 - Voici la structure en Thésaurus de mon document XML

Comme le montre la figure 7, le document XML est structuré de la manière suivante :

1. <tmx> Translation Memory eXchange. Le Root Element.
2. <header /> L'en-tête qui contient des informations telles que la langue du code, le type de données, etc.
3. <body> Marque l'ouverture du document qui sera visible. Comme pour un document HTML.
4. <tu> Pour Translation Unit.
5. <tuv xml:lang="en "> Pour Translation Unit Variant avec un attribut.
6. <tuv xml:lang="fr"> Pour Translation Unit Variant avec un attribut xml définissant la langue de cette variante.
7. <term></term> Balise contenant le terme.
8. <definition></definition> Balise contenant la définition.
9. </tuv> Balise fermante
10. <tuv xml:lang="fr"> Il en va ainsi de suite pour toutes les langues incluses dans le document.
11. <...></...>
12. </tmx>

Mes données étant structurés dans un document XML, j'ai utilisé le langage XSLT (eXtensible Stylesheet Language Transformations)[2] pour l'étape de la modélisation des données. En effet, il permet, en partant d'un document XML, d'appliquer un modèle de mise en page allant de la structure, par exemple l'ordre dans lequel les données apparaissent, au style appliqué. Pour le style, j'emploie le langage CSS dans une balise <style> du document XSLT. L'environnement de Développement Intégré (IDE) Oxygen permet la transformation du document XML en document HTML. Les données peuvent alors être mises à disposition en ligne.


```

37
38     border-style: solid;
39     border-width: 5px;
40     margin-bottom: 5px;
41   }
42   .term {
43     font-weight: bold;
44   }
45
46 </style>
47 </head>
48 <body>
49   <button><a href="blockchainLexiconFr.html">Français</a></button>
50   <h1 style="color: #FF797C;">Blockchain lexicon</h1>
51
52   <table>
53     <tr>
54       <th>Terms</th>
55       <th>Definitions</th>
56     </tr>
57     <xsl:for-each select="tmx/body/tu/tuv[@xml:lang='en']">
58       <xsl:sort select="term" order="ascending"/>
59       <tr>
60         <td class="term"><xsl:value-of select="term"/></td>
61         <td class="def"><xsl:value-of select="definition"/></td>
62       </tr>
63     </xsl:for-each>
64   </table>
65
66 </body>
67 </html>
68 </xsl:template>
69 </xsl:stylesheet>

```

Fig. 8 - Le langage XSLT peut être considéré comme un langage de programmation car il indique à l'ordinateur un modèle de disposition automatique des données xml.

Pour résumer, j'ai créé deux documents XSLT pour créer un document HTML par langue de sorte à pouvoir lier les deux documents par liens hypertexte et ainsi passer d'une page internet à l'autre.

```

35     margin-bottom: 5px;
36   }
37   .term {
38     font-weight: bold;
39   }
40
41 </style>
42 </head>
43 <body><button><a href="blockchainLexiconFr.html">Français</a></button><h1 style="color: #FF797C;">Blockchain lexicon</h1>
44
45   <table>
46     <tr>
47       <th>Terms</th>
48       <th>Definitions</th>
49     </tr>
50     <tr>
51       <td class="term">Bitcoin</td>
52       <td class="def">It is by far the most well-known digital currency. It announced blockchain technology
53       to the world. It has a feature whereby new bitcoins are generated and added to the
54       system, having an inflationary effect. It's distributed structure of the system coupled
55       with its cryptographic functionality make it incredibly robust. It's blockchain uses
56       resource-intensive algorithms.</td>
57     </tr>
58     <tr>
59       <td class="term">Block</td>
60       <td class="def">Bundle of new recorded transactions which is added as the latest link on a long 'chain'
61       of historic transactions forming a blockchain.</td>
62     </tr>
63     <tr>
64       <td class="term">Blockchain</td>
65       <td class="def">Decentralised, encrypted and self-executing technology enabling the recording of lists
66       of transactions. Transactions are only accepted if all the users agree that it is
67       legitimate. Its best-known use is for digital currencies such as Bitcoin. By embedding
68       code in the blockchain, transactions can be executed automatically in response to
69       certain conditions being met, providing a 'guarantee of execution'. Self-executing
70       smart contracts based upon this functionality are developing rapidly.</td>
71     </tr>
72   </table>
73
74 </body>
75 </html>
76 </xsl:template>
77 </xsl:stylesheet>

```

Fig. 9 - On peut aisément apercevoir le côté extensible de XML en observant la structure du document final HTML.

La maîtrise de ce langage est essentiellement avantageuse pour structurer des données qui grâce aux consignes données par le schéma permettent l'interopérabilité des données. Il faut cependant connaître ce langage et avoir des bases en HTML et CSS[4]. De plus, le format <tmx> permet l'importation des données dans un Outils de Traduction Assistée par Ordinateur offrant un gain de temps au traducteur dans son activité.

3.4 Partager l'information avec GitHub

Afin de partager son travail dans un esprit de transmission général de savoir mais aussi afin de pouvoir collaborer avec d'autres contributeurs (traducteurs, terminologues, informaticiens, etc.) désirants apporter leur pierre à l'édifice, la plateforme GitHub[3] est un outil extrêmement intuitif pour la panoplie d'activité qu'il propose. Il s'agit d'un véritable réseau social pour des développeurs mais aussi d'autre profils impliqués dans l'élaboration de projets numériques. Il permet effectivement d'ouvrir un dossier ouvert et partagé (repository) pouvant contenir un fichier README puis tous les fichiers et les documents en lien avec le projet numérique. Le fichier README sert de carte de visite pour les visiteurs du Repository. Il est possible de soigner sa mise en page grâce au langage Markdown qui est une sorte de HTML/CSS simplifié. Les visiteurs ont un degré d'accessibilité qui est défini par la licence que le créateur du projet choisit. Des contributions peuvent être soumises, révisées puis approuvées par les autres membres du projet avant d'être publiées. Le Repository garde une trace de toutes les modifications apportées et leur historique grâce au Système de contrôle de version Git. D'ailleurs, de nombreux projets open source dans le domaine de la Blockchain utilisent cette même plateforme comme base de travail. En plus de la gestion du projet, GitHub permet également de publier gratuitement, du contenu en ligne via ses propres serveurs. Voici les liens cliquables vers le Repository du projet ainsi que la page web avec le résultat de mon travail, le lexique En-Fr :

- [Blockchain Lexicon GitHub Repository](#)
- [Blockchain Lexicon EN/FR](#)

4 Conclusion

Si le Big Data montre aujourd'hui ses limites en termes de sauvegarde de la sphère privée et de la démocratie, il offre aussi, à ceux qui savent s'en servir, une nouvelle forme de liberté. La liberté de s'intéresser à tout en accédant à l'information à moindre coûts et ensuite partager ses connaissances ainsi que son travail à large échelle. Ce bref aperçu de certains outils de TAL, permet d'introduire les langagiers souhaitant profiter des opportunités technologiques à l'air de Internet 2.0. En effet, le TAL permet de gagner du temps en offrant la possibilité de traiter automatiquement une large quantité de données disponibles en ligne. De plus, la création d'un bagage numérique peut aujourd'hui servir de carte de visite pour capter l'intérêt de potentiels collaborateurs ou clients. On peut le voir comme l'équivalent d'une expérience personnelle indépendante servant de base pour façonner sa future vie professionnelle. Il faut tout de même garder à l'esprit que ce genre de travail n'est possible qu'en joignant des connaissances multilingues, terminologiques, terminographiques et informatiques. En général il s'effectuera en équipe mais ce travail montre qu'il peut également être mené à bien seul, moyennant du temps et la curiosité d'apprendre. Le domaine de l'informatique prend de plus en plus de place dans tous les métiers et la traduction n'est pas étrangère à ce phénomène. De plus, les travaux de Cabré (2007)[5] montrent bien l'importance pour les langagiers de pratiquer la terminologie et de mieux comprendre les concepts inhérents aux langues. On pourrait d'ailleurs améliorer la structure du recueil terminologique en l'intégrant à une base de données relationnelle, permettant l'indexation des concepts entre eux et la création d'un réseau de connaissances. Ce travail peut également servir de base à l'élaboration d'un guide pratique pour le traducteur couvrant le TAL de la recherche d'information à sa valorisation en passant par la modélisation et l'analyse des données. Il existe bien sûr une panoplie beaucoup plus large d'outils capables de mener à bien les différentes étapes exposées.

Comme le montre GitHub, le travail sur des projets ouverts et collaboratifs représente également une voie d'avenir pour les métiers de langagiers. En effet, les développements technologiques tendent vers la décentralisation des activités dans la majorité des domaines. Les agences de traduction ne seront-elles à l'avenir plus que des plateformes numériques décentralisées permettant à "n'importe qui" d'apporter sa contribution aux projets en cours ? Voici une question dont la réponse se trouvera peut-être du côté de la technologie Blockchain.

Références

- [1] Extensible Markup Language (XML) <https://www.w3.org/XML/>, 2021-06-05.
- [2] The Extensible Stylesheet Language Family (XSL) <https://www.w3.org/Style/XSL/>, 2021-06-05.
- [3] GitHub vs GitLab : avantages et inconvénients de ces plates-formes | À partir de Linux, <https://blog.desdelinux.net/fr/github-vs-gitlab/>, 2021-06-05.
- [4] HTML Standard, <https://html.spec.whatwg.org/multipage/>, 2021-06-05.
- [5] M Teresa Cabré. Constituer un corpus de textes de spécialité. *Cahiers du CIEL*, pages 37–56, 2007.
- [6] European Parliament. Directorate General for Parliamentary Research Services. *How blockchain technology could change our lives : in depth analysis*. Publications Office, LU, 2017.
- [7] Jeffrey EF Friedl. *Mastering regular expressions*. " O'Reilly Media, Inc.", 2006.
- [8] Noortje Marres and Esther Weltevrede. Scraping the social? issues in live social research. *Journal of cultural economy*, 6(3) :313–335, 2013.
- [9] Ingrid Meyer. Extracting knowledge-rich contexts for terminography. *Recent advances in computational terminology*, 2 :279, 2001.
- [10] Franco Moretti. *Graphs, maps, trees : abstract models for a literary history*. Verso, 2005.
- [11] O Nevzorova, D Mukhamedshin, and R Gataullin. Developing corpus management system : architecture of system and database. In *Proceedings of the International Conference on Information and Knowledge Engineering (IKE)*, pages 108–112. The Steering Committee of The World Congress in Computer Science, Computer ..., 2017.
- [12] John Sinclair and Ronald Carter. *Trust the text : Language, corpus and discourse*. Routledge, 2004.