Enjeux des corpus bilingues en diachronie longue : l'exemple du projet MICLE

Mathieu Goux

CORPUS

Corpus

25 | 2023
La constitution de corpus en diachronie longue

Enjeux des corpus bilingues en diachronie longue : l'exemple du projet MICLE

Bilingual historical corpora in the depth of time: the case of the MICLE project

Mathieu Goux

OpenEdition Journals

Electronic version
URL: https://journals.openedition.org/corpus/8468
DOI: 10.4000/corpus.8468
ISSN: 1765-3126

Publisher
Bases ; corpus et langage - UMR 6039

Electronic reference
Mathieu Goux, "Enjeux des corpus bilingues en diachronie longue : l'exemple du projet MICLE", *Corpus* [Online], 25 | 2023, Online since 18 January 2024, connection on 31 January 2024. URL: http://journals.openedition.org/corpus/8468 ; DOI: https://doi.org/10.4000/corpus.8468

Enjeux des corpus bilingues en diachronie longue : l'exemple du projet MICLE

Bilingual historical corpora in the depth of time: the case of the MICLE project

Mathieu Goux

1 Les enjeux des métadonnées dans les sciences du langage, et particulièrement dans le cadre de la linguistique de corpus ou de la linguistique outillée, sont à présent assez connus (Reppen 2010, Caron *et al.* 2019, Pincemin 2022). Les discussions les concernant se polarisent généralement sur deux sujets : d'une part, la question de la tokenisation (Lavrentiev *et al.* 2021), d'autre part, celle des métadonnées, notamment les jeux d'étiquettes morphosyntaxiques et les référentiels de lemmatisation, surtout pour les états de langue anciens (Souvay & Pierrel 2009) ou les langues peu ou non dotées (Madhavi 2018).

2 En revanche, il nous semble que peu de cas est encore fait d'autres éléments techniques qui influencent la préparation et l'accès aux données textuelles. Nous pensons notamment à la découpe des textes en propositions ou en phrases, à la conservation des informations philologiques ou la question des formats numériques des données. Ces éléments, considérés souvent comme relevant de l'ingénierie et comme à part des problématiques linguistiques, sont pourtant au cœur de celles-ci (Camps *et al.* 2020, Clérice 2022). Notre contribution se propose dès lors de faire un panorama de ces différents enjeux en les illustrant par les problématiques rencontrées, et les solutions apportées, par le projet ANR-DFG MICLE.

3 Le projet MICLE, qui a débuté en juin 2021[1], se dédie aux micro-indices du changement linguistique et notamment à l'évolution du caractère V2 de l'ancienne langue française et de l'ancien vénitien. Ces langues ont été choisies pour leur proximité typologique : ce sont non seulement deux langues romanes, mais également deux langues qui ont rapidement perdu leur caractère *pro-drop* à date ancienne (Poletto 2020, Wolfe 2020). L'une des pistes de recherche consiste dès lors à mesurer l'influence de cette propriété,

parmi d'autres, sur l'évolution générale de la syntaxe propositionnelle. Pour explorer cette hypothèse, un double corpus calibré génériquement et diachroniquement, composé de textes non-littéraires sur une période allant du 13e au 17e siècle, a été constitué. Les textes sélectionnés (9 pour la partie française, 8 pour la partie vénitienne) sont des minutes de procès et des styles de procédure, espacés autant que faire se pouvait d'une cinquante d'années. Quelques textes satellites, dont des correspondances privées, viennent compléter ce corpus noyau. La plupart de ces textes font l'objet d'une exhumation et ont été numérisés pour la toute première fois. Le corpus, dont la partie française a été publiée en avril 2023 (cf. fig. 1), comptera pour sa première version un peu moins de 500 000 *tokens*[2]. Il comportera un encodage fin tant en parties du discours qu'en informations syntaxiques, en dépendance et en constituants.

Figure 1. Textes de la partie française du corpus MICLE (version 0.9, avril 2023)

Corpus noyau :

Titre court	Date	Langue	Nombre de tokens	Étiquettes disponibles			Disponibilité du texte
				Universal Dependencies	UPenn	Presto	
Assises de Normandie	1207	Ancien français	4 766	✓	✓	✓	✓
ANYB 1292	1292	Anglo-Normand	33 557	✓	✓	✓	✘
Atirements et Jugiés	1314	Ancien français	15 232	✓	✓	✓	✓
ANYB 1340	1340	Anglo-Normand	49 203	✓	✓	✓	✘
Style et usage de l'échiquier	1425	Moyen français	71 903	✓	✓	✓	✓
Procès Jeanne d'Arc	1431	Moyen français	41 967	✓	✓	✓	✘
Réhabilitation Jeanne d'Arc	1450	Moyen français	8 716	✓	✓	✓	✓
Sorcellerie Guernesey	1563	Moyen français	12 190	✓	✓	✓	✓
Procès Bavent	1643	Français classique	3 343	✓	✓	✓	✓

Corpus satellite :

Titre court	Date	Langue	Nombre de tokens	Étiquettes disponibles			Disponibilité du texte
				Universal Dependencies	UPenn	Presto	
Roche-Guyon	1502	Moyen français	41 149	✓	✓	✓	✘
Style Rouillé	1539	Moyen français	35 134	✓	✓	✓	✓
Fille possédée	1591	Moyen français	30 773	✓	✓	✓	✓

4 Notre contribution reviendra sur deux aspects particuliers, qui conditionneront la progression de notre réflexion et qui ont été au cœur du travail de constitution du corpus MICLE. Dans un premier temps, nous parlerons des enjeux d'étiquetage, à la fois en parties du discours mais également concernant le niveau de la proposition ou de la phrase, crucial pour l'analyse syntaxique. Dans un second temps, nous nous concentrerons sur la question de la chaîne de traitement, en évoquant d'une part la fidélité de la transcription et les informations philologiques que l'on doit ou peut conserver des sources originales, et d'autre part les formats numériques de données, qui conditionnent leur pertinence, leur interopérabilité et leur partage au sein de la communauté scientifique internationale.

1. Niveau du *token* et de la phrase

1.1. Étiquetages morphosyntaxiques

5 Nous ne reviendrons pas ici précisément sur les problèmes posés par la discrétisation des unités linguistiques à des fins d'étiquetage en parties du discours, ou sur les problématiques regardant la tokenisation ou la lemmatisation des corpus. Ces sujets ont été abondamment traités par la littérature scientifique, et les discussions se

poursuivent encore[3]. Nous voudrions cependant attirer l'attention sur trois observations, au cœur des problèmes rencontrés par le projet MICLE :

i. Les tentatives de produire des étiquettes universelles à l'instar de celles proposées par *Universal Dependencies* (UD)[4] conduisent à un grain grossier ou inadapté pour certaines langues, qui rend difficiles les analyses de détail.

ii. Dans le cadre des grands corpus en diachronie, les faits de grammaticalisation et de pragmaticalisation rentrent en contradiction directe avec un étiquetage en *token*, qui n'est pas une annotation dynamique susceptible de rendre compte des mécanismes de variation à l'œuvre (Lavrentiev *et al.* 2021).

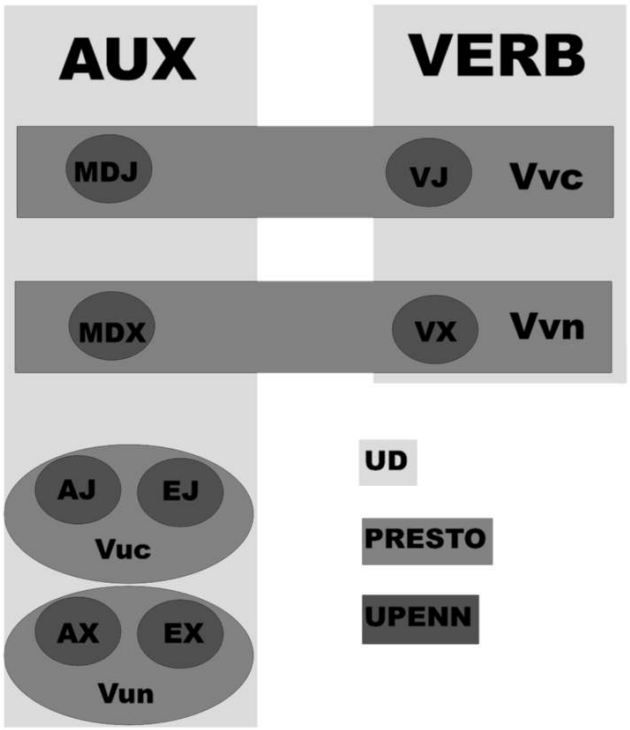
iii. L'ambiguïté ou l'ambivalence de certaines analyses oblige à situer l'annotation au sein d'une certaine école (structuraliste, générative, constructionnelle, etc.), et donc à opérer des choix qui engagent une certaine représentation des faits de langue. Par exemple, le corpus MCVF[5] utilise la notion de « position vide » propre à la grammaire générative pour ses annotations syntaxiques.

6 Ces problèmes sont d'autant plus importants dans le cadre d'un corpus bilingue comme celui du projet MICLE, puisque les deux langues analysées ont leur lot de spécificités qu'il convient de resituer dans ces enjeux d'annotation générale. Une solution, ou plutôt un contournement, a été trouvée : l'emploi de différents jeux d'étiquettes simultanément (UD, UPENN[6] et Presto[7]). Cette solution a deux avantages majeurs :

- Tout d'abord, deux des trois jeux sélectionnés (Presto et UPENN) ont été particulièrement pensés pour l'annotation de corpus en diachronie. Ils s'écartent, en ce sens, des annotations que l'on rencontre souvent dans les grands corpus de langues contemporaines au profit d'étiquettes d'un emploi plus souple, s'accommodant d'une approche dynamique des faits de langue.
- Ensuite, la confrontation de ces étiquettes permet de dépasser les problèmes de grain que l'on rencontre nécessairement dans un jeu particulier. Bien qu'un certain nombre d'étiquettes finissent par se rejoindre de jeu en jeu (notamment, les étiquettes renvoyant aux « mots grammaticaux », ou faisant partie de classes fermées comme les conjonctions, les pronoms, etc.), les variantes autorisent une navigation entre différents niveaux d'analyse.

7 Pour illustrer cela, considérons l'étiquetage des formes verbales. Nous avons représenté dans le graphique suivant la façon dont les jeux d'étiquettes interagissent (fig. 2) :

Figure 2. Interaction des étiquettes verbales pour les jeux UD, Presto et UPENN



8 Le système UD ne propose qu'une opposition entre auxiliaires d'un côté (qui incluent, notamment, les verbes modaux *vouloir*, *pouvoir*, *devoir*), et « verbes pleins » de l'autre, mais sans distinguer les modes personnels et non-personnels. Le jeu Presto rajoute ici une strate d'informations supplémentaires, en distinguant non seulement les auxiliaires *avoir* et *être* entre eux, mais également les formes à l'infinitif des formes conjuguées à un mode personnel[8]. En revanche, il ne distingue pas les verbes modaux. Enfin, le jeu UPENN rajoute cette information complémentaire et distingue infinitif, participes et formes à un mode personnel.

9 Cette confrontation illustre les avantages de multiplier les jeux d'étiquettes. Non seulement une requête peut les mélanger, par exemple pour chercher tous les types de verbes derrière un verbe modal ou tel auxiliaire, mais en cas d'interprétation difficile d'une structure précise, par exemple entre un participe et une forme personnelle en l'absence d'indication graphique comme un accent dans les textes anciens, ou dans les cas de formes véritablement ambiguës, le jeu UD permet malgré tout de moissonner les données sans engager une lecture qui créerait du bruit, ou du silence, dans les relevés. Du point de vue technique, du reste, la conversion d'un jeu d'étiquettes à l'autre peut se faire très aisément à partir d'une table de correspondances qu'exploitent des scripts qui, mécaniquement, appliqueront la transformation avec une intervention minimale de l'annotateur[9]. En ce sens, le problème du choix est contourné en proposant simultanément plusieurs solutions : à l'analyste, par la suite, de sélectionner le ou les jeux qui lui semblent le plus approprié pour sa recherche grâce à la documentation fournie. Cette solution assure également que l'utilisateur est familier avec, au moins, l'un des jeux sélectionnés, ce qui facilite l'exploration des données.

1.2. Délimitation de l'unité d'analyse maximale

10 Les outils d'annotation contemporains permettent de conduire automatiquement une analyse syntaxique en dépendance, selon différentes méthodes statistiques (Grobol et Crabbé 2021). D'autres corpus proposent, quant à eux, une annotation en constituants généralement faite manuellement, à l'instar du corpus MCVF dont nous avons parlé *supra*. Notons qu'à ce jour, il ne semble pas y avoir de procédures de conversion simple entre ces deux types d'analyses[10]. Que l'on travaille, cependant, en dépendance ou en constituants, un même problème se pose : celui de la dimension et des limites de l'unité d'analyse maximale considérée.

11 L'analyse syntaxique, effectivement, n'a de sens qu'au sein d'une unité maximale dans laquelle les relations de dépendance, de rections et de valence trouveront à s'interpréter. Cette unité, la proposition ou la phrase, fait l'objet de définitions multiples selon les traditions analytiques, mais elle reste dans tous les cas un concept plutôt récent dans notre histoire métalinguistique, en français ne serait-ce (Siouffi 2020). Les textes anciens répondaient à d'autres modes de structuration syntaxique et textuelle, ce qui peut poser des difficultés lors de la préparation des corpus. Considérons, ainsi, les deux phrases graphiques suivantes, issues du corpus MICLE, en français et en vénitien :

1. Lequel evesque leur exposa comment une femme nomme Jehenne, vulgairement appellee la Pucelle, avoit nagueres esté prinse et apprehendee en son dyocese, **laquelle**, tant a la requeste de et faicte par le tres chrestien et tres illustre prince le roy de France et d'Angleterre, de notres mere l' Université de Paris, a la sommacion de luy, et de venerable homme frere Martin Billon, vicaire general de l'inquisiteur de la foy en France, pour ce que ladicte femme estoit vehementement suspecte de crimes de heresie, luy avoit esté baillee et delivree, pour enquerir et informer sur les cas, crimes et malefices dont elle avoit esté accusee. (*Procès de Jeanne d'Arc*, 1431)
2. Ma se collui, de llo qual se dise ch' ell' abia habuto la rogandia o tramesso, serà morto fora de Venesia cença testamento, e lli beni de collui vignerà a lle mane de lli çùdisi a destribuir li entro li credori, sì se collui, **lo qual** domanda per rogandia o tramesso, o lli soi redi, o socedori, o comesarii, no pò questa causa plenamentre provar, ma presoncion à lli çùdisi provevelle, segondo che de sovra è conpreso, ço è che en lo quaterno del morto questa causa è scrita, o altra convignivelle, li çùdisi darà a collui lo qual domanda, o a lo rede, o socedor, o comesaro so, lo sacramento, si a li çùdisi parerà, veçuda, segondo ch' è dito de sovra, la bontade e ll' onestade de lle persone. (*Statuta Veneta*, 1290)

12 Dans ces transcriptions pour lesquelles la ponctuation du temps a été respectée, nous avons deux « phrases » graphiques de respectivement 119 et 155 *tokens*. Leur longueur les empêche, d'ores et déjà, de produire une structure en dépendance lisible, et rend plutôt compliquée une analyse en constituants. De plus, l'architecture même de ces unités graphiques, multipliant les compléments parenthétiques ou apposés, rend la perspective d'en faire une analyse syntaxique incertaine, voire douteuse. Aussi, l'annotateur du corpus doit trouver un moyen de résoudre ce problème et de créer des règles de découpe autorisant l'ajout de métadonnées syntaxiques qui, à défaut de rendre parfaitement compte de la réalité linguistique du temps, permettra la récolte d'occurrences. Ces règles de découpe, plus ou moins arbitraires, plus ou moins explicites, doivent du reste, dans le cadre d'un corpus bilingue, être homogènes et régulières.

13 L'intérêt de travailler par familles de langue permet cependant de simplifier ce travail : dans le cadre des exemples (1) et (2) précédents, nous avons pu repérer des outils anaphoriques au rôle approchant dans la dynamique textuelle (à l'instar du pronom relatif composé *lequel/lo qual*, mis en gras dans les exemples précédents, cf. Goux 2019). Nous avons dès lors considéré que ces indices marquaient le départ d'une nouvelle unité phrastique, justifiant une segmentation dans l'énoncé. Nous avons aussi considéré que certaines conjonctions, notamment *Et/È* et *Ou/O*, lorsque suivies par un nouveau référent et inaugurant un développement thématique, introduisaient une nouvelle unité d'analyse. Nous nous sommes également appuyés sur des critères issus moins du (méta)linguistique que du « périlinguistique », et de certaines conventions génériques, culturelles, typo-dispositionnelles… pour faciliter la préparation des données, comme la ponctuation (les points-virgules ou les deux-points marquent souvent des divisions fortes de la phrase graphique), les changements de paragraphe ou les manchettes (Pinzin et Goux 2022). Ces différentes stratégies d'alignement permettent, dès lors, de rechercher des occurrences croisées entre français et vénitien avec l'assurance d'une certaine homogénéité, et d'offrir des données comparables sur l'évolution historique de ces langues romanes.

14 Travailler en diachronie longue, du reste, invite à opérer ces choix d'alignement en accord avec un état de langue terminal dans lequel le modèle phrastique ou propositionnel est aligné avec les descriptions scientifiques contemporaines : les enjeux d'homogénéisation des données sont donc particulièrement forts, voire indépassables. Ils nécessitent une réflexion particulière, et surtout une articulation entre le projet de recherche et les données accessibles ou, comme le présentait Reppen (2010 : 31), « [*t*]*here must be a match between the language being examined and the type of material being collected* ».

15 Cette idée peut effectivement tomber sous le sens, mais elle engage une discussion entre pratique et théorie, entre constitution de corpus et objectifs de recherche, selon une synergie particulière qui n'est pas toujours explicitée. Surtout, ces choix qui s'imposent lors de la préparation du corpus ont deux conséquences directes et concrètes : d'une part, ils impliquent de déplacer l'unité d'analyse du niveau de la langue à celui du texte, et à prendre en compte la matérialité de cet objet (Caron & Dagenais 2019). Les deux approches ne sont pas nécessairement superposables, et si l'on peut se demander où se trouve le « sens du texte » (pour reprendre l'analyse de Pincemin 2022), reste qu'il faut toujours établir ici, et avant toutes choses, un « sens de corpus ». Il faut absolument que soit évidente la perspective de recherche à laquelle répond un corpus en particulier, comme elle détermine les choix d'étiquetage et, comme nous le verrons ci-après, ses aspects techniques. À l'analyste, une fois encore, de trier les données et de saisir ce qu'il cherche grâce à la documentation et aux possibilités de l'outil de visualisation qu'il utilise.

2. Informations philologiques et formats numériques

2.1. Données philologiques

16 Les discussions concernant les parties du discours et les unités maximales d'analyse impliquent, en amont, une réflexion sur la préparation du matériau textuel, avant même que ne commence la phase d'annotation à proprement parler. Ces réflexions, qui

regardent le cheminement « du manuscrit à la base de données » (Pica 2022), ont été menées dans le cadre du projet MICLE du fait de l'hétérogénéité des sources textuelles mobilisées. Le corpus se compose de trois grandes familles de sources : des manuscrits et des imprimés du temps, qui n'avaient jamais été transcrits jusques là ; des imprimés du 19e siècle, qui opèrent une transcription des sources premières ; des éditions contemporaines, généralement parues dans des revues à comité de lecture et des ouvrages scientifiques. Ces trois familles de documents ont des caractères propres :

- Les manuscrits et imprimés du temps sont, évidemment, nos sources d'informations les plus sûres. La graphie et les informations typo-dispositionnelles sont de première main, et ces documents sont les plus à même de nous renseigner sur l'état de la langue du temps.
- Les transcriptions du 19e siècle ont des positions différentes quant à leurs sources. Certaines visent la fidélité de la transcription, tant en termes de graphie que de ponctuation. D'autres opèrent une série de transformations, voire de modifications des données, sans nécessairement documenter leurs choix.
- Enfin, les éditions modernes proposent généralement une transcription diplomatique ou semi-diplomatique de leurs sources. Elles ne conservent pas toutes les informations graphématiques (elles remplacent les *s* longs par des *s* ronds, ou régularisent les lettres ramistes par exemple), mais elles documentent avec grande précision leurs choix ainsi que les irrégularités ou curiosités de la source première.

17 La quantité de travail demandée pour la préparation du corpus, puis son exploitation scientifique, empêche généralement l'équipe de recherche de revenir méthodiquement aux sources manuscrites et d'en faire une transcription fidèle. Plus largement, que ce soit pour les transcriptions qui doivent être faites pour le projet ou pour les autres ressources textuelles, l'hétérogénéité des sources amène à réfléchir sur la quantité d'informations philologiques et médiales que nous voulons conserver dans les données. Si certaines de ces informations, à l'instar de la ponctuation ou des indications de début ou de fin de paragraphe, de section ou de partie, sont d'une importance capitale dans la mesure où elles engagent la représentation du sens et de la syntaxe du texte, d'autres informations peuvent être moins pertinentes dans le cadre d'une analyse linguistique. Les données graphématiques ou éditoriales (manchettes, effets de caractère, ratures…) peuvent effectivement ne pas être mobilisées pour une analyse particulière, mais leur documentation demeure, il nous semble, une nécessité.

18 Leur préservation doit s'envisager dans la continuité de la documentation générale du projet, et il convient de rendre accessibles les informations bibliographiques et philologiques pour permettre si besoin le retour à la source originale, à des fins de consultation, de vérification voire d'amélioration. Un corpus n'est pas qu'une simple base de données, élaborée à un moment donné du temps pour servir un projet de recherche circonscrit ; du moins, il ne devrait pas l'être. Il s'agit d'un objet scientifique, constitué selon des règles et des enjeux détaillés et documentés, pour être exploitable et interrogeable sur le temps long, voire enrichi au fur et à mesure de sa circulation dans le milieu académique (Galleron & Idmhand 2020, Poudat & Landragin 2017, Schreibman *et al.* 2004).

19 C'est la raison pour laquelle il nous a paru important, dans le cadre du projet MICLE, de ne pas nous contenter d'une transcription « modernisée », mais de mener une réflexion de fond sur les informations que nous retenons au sein de nos données pour conserver un équilibrage entre pertinence et efficacité, entre objectifs scientifiques propres à la thématique du projet et diffusion du corpus dans la communauté scientifique. Comme,

particulièrement, il n'y a pas de contraintes d'édition dans les livrables du projet, mais que les données doivent être accessibles d'une façon ou d'une autre, nous n'avons point modernisé la langue du temps et avons donc opéré une transcription « semi-diplomatique », ou intermédiaire. Nous avons dès lors choisi de ne pas conserver les informations graphématiques ou typo-dispositionnelles, au-delà de la ponctuation ou de l'architecture générale d'un texte particulier (en paragraphe, section, chapitre, etc.), mais nous avons cependant conservé certaines informations de correction des copistes et scribes du temps, lorsqu'un mot avait par exemple été barré, considérant que ces indices nous renseignaient sur le sentiment de langue des locuteurs.

20 Toutes choses égales par ailleurs, la quantité d'informations que nous préservons dans un processus de numérisation et d'outillage de corpus est toujours arbitraire, et peut toujours être motivée par une certaine perspective scientifique. Quelque choix que nous opérons, une sélection doit être faite, conséquence du changement de support et de la transformation de l'objet textuel. Notre perspective générale a donc été, d'une part, de documenter finement non seulement l'origine des ressources textuelles numérisées, mais également nos choix de transcription. D'autre part, de préciser au sein des données elles-mêmes les modifications effectuées « à la volée », et de conserver un maximum d'informations philologiques, de la ponctuation aux effets de caractère. Nous ne les exploitons pas nécessairement au sein de notre projet mais nous en conservons trace, pour permettre à de futures équipes de les utiliser à d'autres fins scientifiques.

2.2. Formats de données

21 La circulation des données au sein de la communauté scientifique internationale engage également une réflexion sur leur format d'encodage. Au regard des autres sujets traités précédemment, cette question semble encore peu abordée en sciences du langage, comparé à d'autres disciplines relevant des sciences humaines :

> [...] les formats qui participent de la production et de la circulation des écrits ne sont pas perçus par la plupart de leurs utilisateurs. Cependant, [...] ils constituent des acteurs importants dans la constitution des expériences sensibles des contributions scientifiques, et des communautés savantes contemporaines. (Mourat 2018 : 36)

22 Si l'on reprend la discussion de Clérice (2022 : 48-54), nous pouvons mettre en avant trois critères assurant une bonne « hygiène de la *data* », en sciences du langage comme dans d'autres domaines scientifiques :

- Transmissibilité : indispensable dans le cadre de l'échange scientifique, et condition *sine qua non* dans le cadre d'un corpus bilingue comme MICLE qui n'appartient pas à une équipe en particulier, la transmissibilité assure la lisibilité et l'exploitation des données au-delà du projet pour lequel le corpus a été élaboré.
- Visualisation : le format doit être aisément visualisable, quelle que soit la machine, et ce sans perte d'informations. En ce sens, il convient de s'orienter vers des formats libres, non-propriétaires et pérennes, avec un accès direct au « code source ».
- Édition et correction : dans le cadre de la circulation de la *data*, et de son possible enrichissement par d'autres projets, le format doit pouvoir être édité aisément. Les corrections, ajouts, modifications doivent pouvoir être faits avec une documentation minimale, et sans engager l'intégrité du fichier lui-même.

23 Dans l'histoire de la linguistique de corpus et des bases de données textuelles, plusieurs formats répondant à ces critères ont été employés : citons le format CONLL-U, l'XML-TEI et le PSD[11]. Cependant, là encore, les spécificités et l'échelle de ces formats engagent plusieurs représentations du fait de langue voire du fait de texte. Particulièrement, le CONLL-U et le PSD sont pensés pour encoder diverses informations métalinguistiques, mais ne peuvent gérer les informations philologiques, les effets de caractère ou l'architecture générale du texte. Ce sont pourtant des formats populaires, exploités par un très grand nombre de corpus tant en synchronie qu'en diachronie, et qui autorisent une discussion et une confrontation des phénomènes linguistiques entre plusieurs langues. En ce sens, leur exploitation nous semble être une évidence tant ils remplissent parfaitement leur office en termes de transmissibilité, de visualisation et d'édition, du moins dans le domaine des sciences du langage.

24 Dans la mesure, cependant, où ils ne peuvent parfaitement inclure l'intégralité des informations médiales que nous avons choisi de conserver de nos sources textuelles, nous avons élu le format XML-TEI comme format « de base » duquel peuvent être dérivés les autres. La TEI est certes orientée davantage vers l'édition textuelle plutôt que vers l'analyse linguistique, mais elle s'est dotée avec le temps d'un nombre impressionnant d'éléments destinés à cet office. Du reste, la possibilité de mélanger plusieurs standards au sein d'un même document, et de documenter leur arborescence, rend l'XML particulièrement propice à l'extension de l'annotation. Il s'agit, par ailleurs, d'un format hautement standardisé, lisible et éditable, qui le rend incontournable et facilite la circulation et l'exploitation des données linguistiques.

25 En ce sens, un fichier XML-TEI peut encapsuler toutes les informations présentes dans un CONLL-U et un PSD, et rajouter des informations philologiques et bibliographiques qui peuvent cruellement manquer des bases de données. Sa flexibilité lui permet, en outre, d'être facilement éditable et transformable par plusieurs langages de programmation comme Python, et de se prêter aisément à plusieurs procédures de visualisation, qu'il s'agisse d'une transformation en HTML pour un affichage web ou nativement par l'intermédiaire de logiciels de textométrie comme TXM[12]. Il s'agit, à l'heure actuelle tout du moins, du format qui réunit le mieux les qualités attendues pour un corpus linguistique, quels que soient ses objectifs scientifiques.

26 La question du format ne saurait être, en tous les cas, un point de détail lors de la constitution d'un corpus et elle doit être posée dès la préparation du projet tant elle implique à la fois une certaine relation aux données, et un temps de préparation et de formation souvent incompressible. Non seulement les chercheurs ne sont pas toujours formés à ces enjeux, et doivent dès lors acquérir des compétences leur permettant d'exploiter ces formats de données, mais les équipes de recherche doivent également incorporer des ingénieurs dont c'est le métier premier et qui ne sauraient être réduits au statut de techniciens. Cela suppose une répartition des tâches qui, du point de vue organisationnel, n'est pas toujours bien prise en compte, ou déléguée à des chercheurs sans formation *ad hoc*.

Conclusions

27 Qu'il s'agisse de la question des métadonnées, de la conservation des informations philologiques ou, encore, des formats, deux points nous semblent cruciaux :

- Tout d'abord, l'importance de la documentation. La moindre décision de transcription ou d'encodage est le lieu d'un choix, sur ce qui est conservé ou annoté, et ce qui ne l'est pas. Ces choix peuvent être motivés par une certaine perspective scientifique ou un objectif de recherche, mais ils ont toujours un certain degré d'arbitraire qu'il convient d'expliciter.
- Ensuite, la nécessité de réfléchir et de stabiliser des choix techniques liés aux formats d'encodage des données textuelles, leur visualisation, leur transmissibilité et leur édition. Ces choix techniques demandent une formation spécifique et une discussion constante avec les ingénieurs, qui participent pleinement au diagnostic scientifique présidant à l'annotation et à l'outillage du corpus et, finalement, à son analyse.

28 Comme nous le disions précédemment, un corpus n'est pas qu'une « simple » base de données, même si nous pouvons l'employer comme telle, pour vérifier une hypothèse de travail ou pour trouver des exemples illustratifs. Il s'agit avant tout d'un objet scientifique, avec ses limites, ses enjeux et ses perspectives propres, destiné à être partagé, à évoluer et à être augmenté, voire modifié. Un corpus qui ne peut se prêter à ces opérations aura un cycle de vie particulièrement court, et sombrera dans l'oubli : il faut viser, dès la conception, la postérité.

BIBLIOGRAPHY

Camps J.-B., Gabay S., Fièvre P., Clérice T. & Cafiero F. (2020). « Corpus and Models for Lemmatisation and POS-tagging of Classical French Theatre », halshs-02591388.

Caron P. & Dagenais L. (2019). « Entre plein texte et balisage fin : le Dictionnaire critique de la langue française de Jean-François Féraud en ligne sur le site du CNRTL », *in* P. Caron *et al.* (dir.), *L'enjeu des métadonnées dans les corpus textuels. Un défi pour les sciences humaines.* Rennes : Presses universitaires de Rennes, 149-170.

Caron P., Lay M.-H. & Defiolle R. (dir.). (2019). *L'enjeu des métadonnées dans les corpus textuels. Un défi pour les sciences humaines*. Rennes : Presses universitaires de Rennes.

Clérice T. (2022). *Détection d'isotopies par apprentissage profond : l'exemple de la sexualité en latin classique et tardif.* Thèse de doctorat de l'Université Lyon 3, sous la direction de Christian Nicolas, soutenue le 28 mars 2022.

Gabay S., Clérice T., Camps J.-B., Tanguy J.-B. & Gille-Levenson M. (2020). « Standardizing linguistic data: method and tools for annotating (pre orthographic) French », DDH '20, 15 17/10 2020, Hammamet. https://hal.science/hal-03018381.

Galleron I. & Idmhand F. (2020). « De l'interopérabilité à la réutilisabilité des éditions électroniques », *Humanités numériques* 1, https://doi.org/10.4000/revuehn.350.

Goux M. (2019). *Le pronom-déterminant LEQUEL en français préclassique et classique.* Paris : Classiques Garnier.

Grobol L. & Crabbé B. (2021). « Analyse en dépendances du français avec des plongements contextualisés », *TALN/RECITAL* 2021, Lille (virtuel), France.

Höfler S. (2002). *Link2Tree : A Dependency-Constituency Converter.* Thèse de doctorat de l'Université de Zürich. https://www.cl.uzh.ch/dam/jcr:00000000-6a77-a254-ffff-ffffa24f33e9/hoefler_liz_link2tree.pdf.

Lavrentiev A., Guillot-Barbance C. & Heiden S. (2021). « Enjeux philologiques, linguistiques et informatiques de la philologie numérique : l'exemple de la segmentation des mots », *Diachroniques* 8 : 76-102.

Mahdavi M. A. (2018). « Developing a Comprehensive Standard Persian Positional Tagset », *International Journal of Information Science and Management* 16(1) : 165-190.

Mourat R. de. (2018). « Le design fantomatique des communautés savantes : enjeux phénoménologiques, sociaux et politiques de trois formats de données en usage dans l'édition scientifique contemporaine », *Sciences du design* 8(2) : 34-44.

Pica M. (2022). « Harmoniser le corpus ConDÉ : de l'image à la ressource linguistique », *Studia Linguistica Romanica* 8 : 131-154.

Pincemin B. (2022). « Sémantique textométrique », *in* A. Biglari et D. Ducard (dir.), *La sémantique au pluriel*. Rennes : Presses universitaires de Rennes, 373-396.

Pinzin F. & Goux M. (2022). « The MICLE Project (Micro-Clues of Linguistic Evolution). Theoretical goals and methodological considerations », *Venise et la France : similitudes, spécificités, interrelations (1300-1800)*. Conference paper, may 2022.

Poletto C. (2020). « More than one way out. On the factors influencing the loss of V to C movement », *Linguistic Variation* 19(1) : 47-81.

Poudat C. & Landragin F. (2017). *Explorer un corpus textuel : Méthodes – pratiques – outils*. Louvain-la-Neuve : De Boeck.

Prévost S. (2015). « Diachronie du français et linguistique de corpus : une approche quantitative renouvelée », *Langages* 197 : 23-45.

Reppen R. (2010). « Building a corpus. What are the key considerations? », *in* A. O'Keeffe et M. McCarthy, *The Routledge Handbook of Corpus Linguistics.* Londres : Routledge, 31-37.

Schreibman S., Siemens R. & Unsworth J. (2004). *A Companion to Digital Humanities.* Hoboken : Blackwell Publishing.

Siouffi G. (dir.). (2020). *Une histoire de la phrase française. Des Serments de Strasbourg aux écritures numériques.* Arles : Actes Sud.

Souvay G. & Pierrel J.-M. (2009). « LGeRM : lemmatisation de mots en moyen français », *Traitement Automatique des Langues* 50(2).

Wolfe S. (2020). « Redefining the typology of V2 languages. The view from Medieval Romance and beyond », *Linguistic Variation* 19(1) : 16-46.

NOTES

1. https://www.unicaen.fr/projet_de_recherche/micle/

2. Le corpus intégral sera accessible via le portail txm-crisco.huma-num.fr/ et un dépôt GIT dédié. À la date d'écriture de cet article (juin 2023), seule la partie française du corpus est disponible sur le serveur *TXM-Crisco*. Les détails du corpus sont accessibles par la page d'accueil du serveur.
3. Voir, notamment, les article de Gabay *et al.* (2020), Lavrentiev *et al.* (2021) et, via l'exemple du perse, Mahdavi (2018).
4. universaldependencies.org/
5. https://github.com/beatrice57/mcvf-plus-ppchf
6. https://www.ling.upenn.edu/hist-corpora/annotation/index.html
7. presto.ens-lyon.fr/
8. Les participes et gérondifs ont une étiquette à part, non représentée sur le graphique.
9. Ces outils de conversion seront rendus disponibles sur le dépôt *Gitlab* du projet.
10. Voir, cependant, la thèse de Höfler (2002), bien que ce travail n'ait pas été suivi d'autres projets de cet ordre à notre connaissance.
11. Pour le format CONLL-U, voir https://universaldependencies.org/format.html ; pour le XML-TEI, voir https://tei-c.org/ ; pour le PSD, voir https://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-4/index.html.
12. https://txm.gitpages.huma-num.fr/textometrie/

ABSTRACTS

The increasing number of very large corpora in historical linguistics has led to numerous discussions on annotation procedures and associated metadata. Particularly the issues of part-of-speech tagging and tokenisation are often discussed in the literature. Other crucial topics, however, seem to be less talked about. We are thinking of the splitting of linguistic data into propositions or "sentences", the preservation of philological information, or the question of encoding and data formats. Our contribution explores these issues by taking the example of the MICLE corpus, which had to solve unprecedented difficulties during its constitution.

La multiplication des très grands corpus en linguistique historique a entraîné des discussions nombreuses sur les procédures d'annotation et les métadonnées associées, notamment concernant les questions relevant de l'étiquetage morphosyntaxique et de la tokenisation. D'autres sujets cruciaux, en revanche, semblent moins abordés, comme la question de la découpe en propositions ou en « phrases » des données linguistiques, la préservation des informations philologiques ou, encore, la question de l'encodage et des formats de données. Notre contribution explore ces thématiques en prenant exemple sur le corpus MICLE, qui a dû résoudre des difficultés inédites au long de sa constitution.

INDEX

Keywords: corpus linguistics, computational linguistics, historical linguistics, format, syntactic annotation
Mots-clés: linguistique de corpus, linguistique outillée, diachronie, format, annotation syntaxique

AUTHOR

MATHIEU GOUX

CRISCO (UR 4255), Université de Caen Normandie