Une introduction linguistique pour les données de Machine learning?

Simon Gabay

Une introduction linguistique pour les données de *Machine learning*?

Simon Gabay

Universités de Neuchâtel et Genève (Suisse)

Nous le savons depuis longtemps, la transcription est un art plus complexe qu'il n'y paraît, notamment pour les textes anciens. Prenons l'exemple de cette ligne tirée de l'*Amour tyrannique* de Georges de Scudéry :

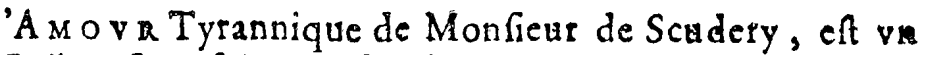

Figure 1 – Georges de Scudéry, *L'Amour tyrannique*, Paris : A.Courbé, 1640, p. 1.

Si l'on écarte l'alignement complet des anciennes formes avec les contemporaines (qui n'est probablement pas souhaitable [1]), certains éditeurs plus mesurés ont pris l'habitude de dissimiler ‹u› et ‹v› pour que cette dernière lettre note, comme aujourd'hui, le son [y] (*AMOVR*), d'abandonner les ligatures esthétiques (‹ﬅ›), de normaliser les variantes allographétiques (‹ſ›), *etc*. Si une telle pratique a d'abord eu pour objectif de « rendre service au lecteur » [2], la rigueur des règles [3] et autres conseils [4] des grands philologues ont eu pour effet secondaire de « puissamment contribuer aux études quantitatives » [5] tant « la qualité cumulative d'une science est étroitement liée au formatage et au statut de ses *data* » [6].

1. Simon Gabay, « Pourquoi moderniser l'orthographe ? Principes d'ecdotique et littérature du XVIIe siècle », *Vox Romanica*–73 (2014), p. 27-42, URL : http://periodicals.narr.de/index.php/vox_romanica/article/view/2254 ; Frédéric Duval, « Les éditions de textes du XVIIe siècle », dans *Manuel de la philologie de l'édition*, 2015, p. 369-394, URL : https://www.degruyter.com/view/books/9783110302608/9783110302608-017/9783110302608-017.xml.

2. Paul Meyer, « Instructions pour la publication des textes », *Bulletin de la Commission royale d'Histoire*, 86 (1922), p. 17-27, URL : https://www.persee.fr/doc/bcrh_0001-415x_1922_num_86_1_3970.

3. Mario Roques, « Règles pratiques pour l'édition des anciens textes français et provençaux », *Bibliothèque de l'École des chartes*, 87 (1926), p. 453-459, URL : https://www.persee.fr/doc/bec_0373-6237_1926_num_87_1_460551.

4. Françoise Vielliard et Olivier Guyotjeannin, *Conseils pour l'édition des textes médiévaux. Fascicule I, Conseils généraux*, Paris, 2014 (Orientations et méthodes).

5. F. Duval, « Transcrire le français médiéval : de l' "Instruction" de Paul Meyer à la description linguistique contemporaine », *Bibliothèque de l'École des chartes*–170-2 (2012), p. 321-342, URL : https://www.persee.fr/doc/bec_0373-6237_2012_num_170_2_464252.

6. Bernard Lacks, « Pour une phonologie de corpus », *French language studies*–18 (2008), p. 3-32.

Si une réflexion ecdotique rigoureuse a pu incidemment servir l'approche computationnelle, les deux méthodes conservent cependant des objectifs radicalement différents : la première tente d'appauvrir intelligemment le texte pour le rendre lisible, alors que la seconde souhaite enrichir au maximum le matériau textuel pour le rendre exploitable. Les normes ecdotiques sont donc par nature inaptes à accompagner correctement la *datafication* des données à l'époque de l'intelligence artificielle – soit celle de la mutualisation pour la constitution de grands ensembles. Or si un immense travail a bien été entrepris sur la standardisation des formats en sciences humaines [7], l'interopérabilité minimale des contenus, surtout linguistiques, constitue un écueil qu'il convient de surmonter.

La multiplication d'outils toujours plus puissants et plus simples à manipuler doit nous amener à repenser leur entraînement comme leur utilisation. En effet, la rapide démocratisation technique que connaissent les humanités numériques entraîne, voire encourage le développement de micro-solutions propres aux besoins de chacun qui entravent l'interopérabilité des données. Cette situation n'est évidemment pas satisfaisante.

— Avec la transition de systèmes à base de règles à l'apprentissage profond, des corpus de plus en plus grands vont devenir nécessaires pour l'entraînement de modèles toujours plus efficaces et toujours plus généraux. La mutualisation des ressources exige donc de manière urgente la standardisation des données.
— Avec le développement de solutions spécifiques pour l'OCRisation, la segmentation, la lemmatisation, la normalisation, la reconnaissance d'entités nommées, la collation. . ., les chaînes de traitement décomposent de plus en plus le travail philologique [8] : il devient donc tout aussi important de pouvoir mutualiser des données pour une même tâche que de pouvoir exploiter les résultats d'un outil avec un autre.

En effet, si certains entraînent un modèle d'OCR à reconnaître le glyphe ‹ſ› comme un *s* long (‹ſ›) et d'autres comme un *s* rond (‹s›), les jeux de données perdent leur compatibilité. De plus, un lemmatiseur ne reconnaîtra pas (ou mal) le token *eſt* si les données d'entraînement ne contiennent que la forme *est*, et un système de NER aura plus de difficultés à reconnaître *Monſieur de Scudery* s'il a été entraîné sur sa forme normalisée *Monsieur de Scudéry*.

Il est possible d'imaginer que le coût de construction des modèles entraînera nécessairement leur faible nombre, et que tous les utilisateurs d'un même modèle produiront donc logiquement des données standardisées en très grande quantité. Mais ne nous leurrons pas : contrairement aux sciences de l'information, l'imposition d'une norme unique n'est ni pensable, ni souhaitable en philologie. La présence de données standardisées au départ n'implique en rien l'existence de données standardisées à l'arrivée, car l'ecdotique reste une science pragmatique, sujette à des ajustements en fonction des traditions textuelles ou des documents. Alors comment faire ?

Comme dans toutes les situations conflictuelles, à défaut de pouvoir résorber le nœud, il peut être utile de l'exposer afin d'en discuter. Plutôt que de tenter d'imposer un standard,

7. Lou Burnard, *What is the Text Encoding Initiative ? : How to add intelligent markup to digital resources*, 2014, URL : http://books.openedition.org/oep/426.

8. Ariane Pinche, Jean-Baptiste Camps et Thibault Clérice, « Stylometry for Noisy Medieval Data : Evaluating Paul Meyer's Hagiographic Hypothesis », dans *Proceedings of DH2019*, Utrecht, The Netherlands, 2019, URL : https://hal.archives-ouvertes.fr/hal-02182737.

il nous paraît plus pertinent d'inciter les acteurs à être transparents sur leur méthode. Une telle réflexion rejoint d'ailleurs les préoccupations de la philologie « traditionnelle », dont les praticiens doutent de plus en plus de la pertinence des informations qu'ils fournissent dans les introductions scientifiques de leurs éditions savantes [9]. Au XXI^{ème} s., afin de valoriser le travail effectué, quelles informations l'éditeur doit-il communiquer à ses pairs concernant le contenu de l'œuvre sur laquelle il travaille, et dans quel format doit-il le faire ?

Depuis plusieurs années, voire même des décennies, de nombreux formats d'échange et de description de données ont vu le jour : DCMI [10], OLAC [11], EAD [12], *etc*. La complexité des données allant en s'accroissant, des propositions d'amélioration continuent d'être faites, comme l'augmentation des métadonnées en METS [13] pour les *Ground truth* d'OCR [14] afin de décrire la structure de la page, le nom de la police, la taille des caractères, *etc.* :

```
<mets:xmlData>
    <gt:gt  xmlns:gt="http://www.ocr-d.de/GT/">
        <gt:state prop="acquisition"/>
        <gt:state prop="granularity/physical/document-related/page"/>
        <gt:state prop="granularity/physical/document-related/region"/>
        <gt:state prop="granularity/physical/document-related/text-line"/>
        <gt:state prop="granularity/physical/document-related/word"/>
        <gt:state prop="condition/production-related/document-faults/ink-from-facing"/>
        <gt:state prop="condition/wear/additions/informative/stamps"/>
        <gt:state prop="data-attributes/document-related/visual/text/font/typeface/blackletter"/>
        <gt:state prop="data-attributes/document-related/visual/text/font/multi-font/font-sizes"/>
        <gt:state prop="data-attributes/document-related/visual/text/font/multi-font/typefaces"/>
    </gt:gt>
</mets:xmlData>
```

FIGURE 2 – Extension des métadonnées METS pour OCR-D.

Le contenu de `@prop` (*property*) reprend ici les valeurs proposées par l'*Ontology for Semantic Labelling of Document Data and Software Methods* [15], qui permet une description de grande qualité. Si l'idée est bonne, elle n'est pas parfaite : l'essentiel de l'information concerne en effet l'organisation des données visuelles et non textuelles, or la conflation des informations linguistiques (`<mods:languageTerm>`) et chronologiques (`<mods:dateIssued>`) contenues dans les métadonnées du fichier METS n'est pas suffisante pour décrire soigneusement le contenu. Malheureusement, s'il est possible de mélanger différents types d'image, il est bien plus compliqué de mélanger différents types de transcription (par exp. allographétique vs graphématique) [16] – le modèle ne

9. F. Duval, Céline Guillot-Barbance et Fabio Zinelli, « Introduction », dans, Paris, 2019 (Histoire et évolution du français), p. 7-32, DOI : 10.15122/isbn.978-2-406-08580-5.p.0007.
10. *DCMI Metadata Terms*, https ://www.dublincore.org.
11. *Open Language Archives Community*, https ://www.dublincore.org.
12. *Encoded Archival Description*, https ://www.loc.gov/ead.
13. *Metadata Encoding and Transmission Standard*, http ://www.loc.gov/standards/mets/.
14. Matthias Boenig, Konstantin Baierer, Volker Hartmann, Maria Federbusch et Clemens Neudecker, *Labelling OCR Ground Truth for Usage in Repositories*, de, 2019, DOI : 10.1145/3322905.3322916.
15. Christian Clausner et Apostolos Antonacopoulos, « Ontology and Framework for Semantic Labelling of Document Data and Software Methods », dans *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, 2018, p. 73-78, DOI : 10.1109/DAS.2018.46.
16. Dominique Stutzmann, « Paléographie statistique pour décrire, identifier, dater... Normaliser pour

pouvant à la fois apprendre à développer et maintenir une même abréviation, si les deux choix coexistent dans les données d'apprentissage.

Il est donc crucial d'intégrer des informations minimales sur la nature linguistique des données d'entraînement proposées. Afin de ne pas inventer de nouveau vocabulaire, nous proposons de reprendre celui de la TEI [17], et notamment les éléments de son modèle `editorialDeclPart` (<normalization>, <segmentation>...). Pour des équipes plus petites que celles des bibliothèques centrales ou universitaires, il pourrait être utile d'opter pour un encodage plus léger, par exemple en JSON avec quelques métadonnées simples en *Dublin Core* plutôt qu'en MODS – à la manière de métadonnées proposées par T. Clérice pour *Deucalion* [18].

```
{
    "dc:title": "Oraison funebre de Marie Terese d'Austriche",
    "dc:creator": "Jacques Bénigne Bossuet",
    "dc:publisher": "Sébastien Mabre-Cramoisy, Paris",
    "dc:date": "1683",
    "dc:language": "fr",
    "dc:identifier": "https://catalogue.bnf.fr/ark:/12148/cb36575655n",
    "dc:type": "Ground truth",
    "dc:description": [
        {
            "gt:prop": [
                "data-attributes/document-related/visual/text/script/latin",
                "data-attributes/document-related/visual/text/font/typeface/antiqua",
                "data-attributes/document-related/visual/columns/one"
            ],
            "tei:normalization": "none",
            "tei:segmentation": "original"
        }
    ]
}
```

FIGURE 3 – Métadonnées pour *ground truth* d'un imprimé.

Un tel système a l'avantage de se décliner simplement pour d'autres types de document, comme les données d'entraînement de lemmatisation ou d'étiquetage des parties du discours. Ces dernières nécessiteraient d'ailleurs probablement de préciser le référentiel utilisé pour la lemmatisation des tokens (*Morphalou* pour le français contemporain [19], *LGeRM* pour les états de langue plus anciens [20]), ainsi que le jeu d'étiquettes utilisé pour les POS (par ex. le standard international UD-Pos [21] ou la norme

coopérer et aller plus loin ? », *Codicology and Palaeography in the Digital Age–2* (2011), p. 247-277, URL : https://halshs.archives-ouvertes.fr/halshs-00596970.

17. L. Burnard, *What is the Text Encoding Initiative ? : How to add intelligent markup to digital resources...*

18. T. Clérice, « Deucalion et Pyrrha : Environnement pour la lemmatisation et la postcorrection à l'École des chartes », dans *Text Encoding : Latinists looking for new synergies*, Liège, Belgium, 2018, URL : https://halshs.archives-ouvertes.fr/halshs-02488858.

19. ATILF-CNRS et Université de Lorraine, *Morphalou v3.1*, 2016, URL : https://www.ortolang.fr/market/lexicons/morphalou.

20. Id., *LGeRM*, ORTOLANG, 2017, URL : https://hdl.handle.net/11403/lgerm/v1.

21. Slav Petrov, Dipanjan Das et Ryan T. McDonald, « A Universal Part-of-Speech Tagset », *CoRR*, abs/1104.2086 (2011), URL : http://arxiv.org/abs/1104.2086.

CATTEX utilisée pour le français médiéval [22] et moderne [23]). Notons que les valeurs utilisées ne sont pas cette fois des éléments, mais des attributs du vocabulaire TEI (`@lemma`, `@pos`, voire `@msd` pour la morphologie), l'annotation linguistique portant nécessairement au niveau du mot (`<w>`).


```
{
    "dc:title": "LEM17",
    "dc:creator": [
        {
            "@id": "http://www.isni.org/isni/0000000419576448",
            "@label": "Simon Gabay"
        }
    ],
    "dc:publisher": [
        {
            "url": "https://github.com/e-ditiones",
            "@label": "E-ditiones project"
        }
    ],
    "dc:rights": [
        {
            "url": "https://creativecommons.org/licenses/by-sa/4.0",
            "@label": "CC-BY-SA"
        }
    ],
    "dc:date": "2020-05-15",
    "dc:language": "fr",
    "dc:coverage": [
        {
            "dc:issued": "1600-1700",
            "tei:pubPlace": "France"
        }
    ],
    "dc:type": "lemma",
    "dc:description": [
        {
            "tei:normalization": "none",
            "tei:segmentation": "original",
            "tei:hyphenation": "original",
            "tei:pos": [
                {
                    "url": "http://bfm.ens-lyon.fr/spip.php?article323",
                    "@label": "CATTEX09"
                }
            ],
            "tei:lemma": [
                {
                    "url": "https://hdl.handle.net/11403/morphalou/v3.1",
                    "@label": "Morphalou"
                }
            ]
        }
    ]
}
```


FIGURE 4 – Métadonnées pour données d'entraînement de lemmatiseur.

22. Sophie Prévost, Céline Guillot, Alexei Lavrentiev et Serge Heiden, *Jeu d'étiquettes morphosyntaxiques CATTEX2009*, rapp. tech., version 2.0. `http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_2.0.pdf`, Lyon, École normale supérieure de Lyon, 2013.

23. S. Gabay, J.B. Camps et T. Clérice, *Manuel d'annotation linguistique pour le français moderne (XVIe -XVIIIe siècles)*, 2020, URL : `https://hal.archives-ouvertes.fr/hal-02571190`.

Comme on le voit avec notre dernier exemple, le modèle est facilement extensible. Le format JSON permet un emboîtement simple, efficace et modulaire : il est possible de proposer les métadonnées du corpus, mais aussi d'ajouter les métadonnées de chaque document avec `dc:source`, ou encore de cumuler les métadonnées de plusieurs types d'annotation (lemmes, POS, entités. . .). Il est en revanche crucial d'avoir recours à des vocabulaires contrôlés, comme le *Dublin core* ou celui de la *Text Encoding Initiative*, afin de standardiser au maximum l'information, et exploiter pleinement les potentialités offertes par des méthodes recommandées par le W3C comme le JSON-LD (*JavaScript Object Notation for Linked Data* [24]) que nous utilisons ici.

24. Manu Sporny, Dave Longley, Gregg Kellogg, Markus Lanthaler et Niklas Lindström, *JSON-LD 1.0 - A JSON-based Serialization for Linked Data*, 2013, URL : `https://www.w3.org/TR/2013/PR-json-ld-20131105`.

Références

ATILF-CNRS et Université de Lorraine, *LGeRM*, ORTOLANG, 2017, url : https://hdl.handle.net/11403/lgerm/v1.

— *Morphalou v3.1*, 2016, url : https://www.ortolang.fr/market/lexicons/morphalou.

Boenig (Matthias), Baierer (Konstantin), Hartmann (Volker), Federbusch (Maria) et Neudecker (Clemens), *Labelling OCR Ground Truth for Usage in Repositories*, de, 2019, doi : 10.1145/3322905.3322916.

Burnard (Lou), *What is the Text Encoding Initiative ? : How to add intelligent markup to digital resources*, 2014, url : http://books.openedition.org/oep/426.

Clausner (Christian) et Antonacopoulos (Apostolos), « Ontology and Framework for Semantic Labelling of Document Data and Software Methods », dans *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, 2018, p. 73-78, doi : 10.1109/DAS.2018.46.

Clérice (Thibault), « Deucalion et Pyrrha : Environnement pour la lemmatisation et la postcorrection à l'École des chartes », dans *Text Encoding : Latinists looking for new synergies*, Liège, Belgium, 2018, url : https://halshs.archives-ouvertes.fr/halshs-02488858.

DCMI Metadata Terms, https ://www.dublincore.org.

Duval (Frédéric), « Les éditions de textes du XVIIe siècle », dans *Manuel de la philologie de l'édition*, 2015, p. 369-394, url : https://www.degruyter.com/view/books/9783110302608/9783110302608-017/9783110302608-017.xml.

— « Transcrire le français médiéval : de l' "Instruction" de Paul Meyer à la description linguistique contemporaine », *Bibliothèque de l'École des chartes*–170-2 (2012), p. 321-342, url : https://www.persee.fr/doc/bec_0373-6237_2012_num_170_2_464252.

Duval (Frédéric), Guillot-Barbance (Céline) et Zinelli (Fabio), « Introduction », dans, Paris, 2019 (Histoire et évolution du français), p. 7-32, doi : 10.15122/isbn.978-2-406-08580-5.p.0007.

Encoded Archival Description, https ://www.loc.gov/ead.

Gabay (Simon), « Pourquoi moderniser l'orthographe ? Principes d'ecdotique et littérature du XVIIe siècle », *Vox Romanica*–73 (2014), p. 27-42, url : http://periodicals.narr.de/index.php/vox_romanica/article/view/2254.

Gabay (Simon), Camps (Jean-Baptiste) et Clérice (Thibault), *Manuel d'annotation linguistique pour le français moderne (XVIe -XVIIIe siècles)*, 2020, url : https://hal.archives-ouvertes.fr/hal-02571190.

Lacks (Bernard), « Pour une phonologie de corpus », *French language studies*–18 (2008), p. 3-32.

Metadata Encoding and Transmission Standard, http ://www.loc.gov/standards/mets/.

Meyer (Paul), « Instructions pour la publication des textes », *Bulletin de la Commission royale d'Histoire*, 86 (1922), p. 17-27, url : https://www.persee.fr/doc/bcrh_0001-415x_1922_num_86_1_3970.

Open Language Archives Community, https ://www.dublincore.org.

PETROV (Slav), DAS (Dipanjan) et MCDONALD (Ryan T.), « A Universal Part-of-Speech Tagset », *CoRR*, abs/1104.2086 (2011), URL : http://arxiv.org/abs/1104.2086.

PINCHE (Ariane), CAMPS (Jean-Baptiste) et CLÉRICE (Thibault), « Stylometry for Noisy Medieval Data : Evaluating Paul Meyer's Hagiographic Hypothesis », dans *Proceedings of DH2019*, Utrecht, The Netherlands, 2019, URL : https://hal.archives-ouvertes.fr/hal-02182737.

PRÉVOST (Sophie), GUILLOT (Céline), LAVRENTIEV (Alexei) et HEIDEN (Serge), *Jeu d'étiquettes morphosyntaxiques CATTEX2009*, rapp. tech., version 2.0. http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_2.0.pdf, Lyon, École normale supérieure de Lyon, 2013.

ROQUES (Mario), « Règles pratiques pour l'édition des anciens textes français et provençaux », *Bibliothèque de l'École des chartes*, 87 (1926), p. 453-459, URL : https://www.persee.fr/doc/bec_0373-6237_1926_num_87_1_460551.

SPORNY (Manu), LONGLEY (Dave), KELLOGG (Gregg), LANTHALER (Markus) et LINDSTRÖM (Niklas), *JSON-LD 1.0 - A JSON-based Serialization for Linked Data*, 2013, URL : https://www.w3.org/TR/2013/PR-json-ld-20131105.

STUTZMANN (Dominique), « Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin ? », *Codicology and Palaeography in the Digital Age*–2 (2011), p. 247-277, URL : https://halshs.archives-ouvertes.fr/halshs-00596970.

VIELLIARD (Françoise) et GUYOTJEANNIN (Olivier), *Conseils pour l'édition des textes médiévaux. Fascicule I, Conseils généraux*, Paris, 2014 (Orientations et méthodes).