Lemmatiser les noms hébreux dans les textes antiques et médiévaux grecs et latins

Jules Nuguet, Alice Leflaëc

Lemmatiser les noms hébreux en grec et en latin

Alice Leflaëc et Jules Nuguet

13 décembre 2023

1 Introduction

Toute notre réflexion sur la lemmatisation des noms propres en grec et en latin trouve sa source dans notre travail pour le projet JERIHNA, aussi allons-nous brièvement vous le présenter pour mettre en évidence les problèmes et les enjeux que nous rencontrons. Notre approche, en effet, n'est pas celle de linguistes mais de personnes qui ont besoin de lemmatiser un texte pour mener à bien leur projet.

Nous sommes tous deux chargés d'étude dans le projet JERIHNA qui prépare l'édition critique, à la fois numérique et papier, du *Livre d'interprétation des Noms hébreux* de Jérôme de Stridon. Ce texte latin de la fin du IV^{e} siècle se présente sous la forme d'un glossaire de noms hébreux tirés de la Bible, pour lesquels Jérôme offre une interprétation (plus ou moins discutable) de l'étymologie.

Cette édition critique est assortie d'une édition diplomatique des quinze manuscrits les plus anciens. Nous avons également souhaité exploiter, en quelque sorte, le glossaire réalisé par Jérôme en effectuant un travail documentaire plus approfondi sur chaque nom hébreu : c'est ce travail ontologique qui nous a conduit aux réflexions que nous exposons aujourd'hui sur la lemmatisation des noms propres.

2 Un projet sur les noms propres hébraïques

2.1 Spécificités du corpus

- Le texte des *Noms hébreux* contient un peu plus de 3000 noms rangés par livre biblique.
- Un même nom hébreu peut être traité dans plusieurs livres bibliques au sein des *Noms hébreux*, avec des variations possibles dans l'interprétation.
- Jérôme s'intéresse aux noms propres de la Bible, aussi bien de personnages que de lieux, mais il traite aussi certains noms communs ou certaines formes verbales comme des noms propres. Ex : *Ephphata* (ouvre-toi),

parole prononcée par Jésus au moment de la guérison d'un sourd-muet (cf. Mc 7:34).

- Une même entité peut être traitée sous deux orthographes différentes dans les *Noms hébreux*. Ex : Sina / Sinai

Ajoutons à cela que le texte des *Noms hébreux* est présent dans plus de 200 manuscrits avec de nombreuses variantes orthographiques pour un même nom. Si l'on souhaite utiliser ce texte comme ce qu'il est vraiment, un manuel que l'on consulte et non un ouvrage qu'on lit d'une traite, il apparaît nécessaire de relier entre eux les noms identiques et de distinguer les homonymes.

2.2 Les sources des *Noms hébreux*

Le *Livre d'interprétation des Noms hébreux* s'appuie sur des nombreuses sources :

- Le texte biblique rédigé en hébreu et en grec. Jérôme s'intéresse à des noms hébreux mais aussi à des noms grecs du Nouveau Testament pour lesquels il offre, fait surprenant et assez incompréhensible, une étymologie hébraïsante.
- Les traductions grecques de l'Ancien Testament proches de la LXX
- Des auteurs grecs : Philon d'Alexandrie, Origène et Eusèbe de Césarée ont pu être identifiés.

Le texte biblique est, bien souvent, médiatisé et les noms hébreux sont sans cesse traduits : d'origine hébraïque le plus souvent, ils sont traduits en grecs puis traduits en latin, la langue de notre glossaire.

Cet intérêt pour les sources de Jérôme est un aspect du travail documentaire que nous effectuons sur les noms hébreux traités par Jérôme.

2.3 Les noms hébreux en contexte

Nous avons pour projet de créer, à partir du texte de Jérôme, un index des noms hébreux qu'il traite. Cet index présentera l'orthographe de chaque nom dans diverses sources antiques et contemporaines à la fois des dictionnaires et concordances de référence pour l'Antiquité et des bibles en langues diverses : l'*Onomasticon* de Forcellini, le dictionnaire ecclésiastique de Sleumer, les concordances de Strong et de Hatch-Redpath et, pour les Bibles, la Vulgate, la LXX et le Nouveau Testament Grec, la BHS, la TOB, la Bible de Jérusalem, la Nouvelle Bible Segond et la NRSA (traduction oecuménique anglaise). Nous présentons toute la liste pour donner une meilleure idée des données à relier. Cet index contiendra aussi un relevé des occurrences de ce nom dans la Bible et dans divers textes patristiques, en particulier ceux dont s'inspire Jérôme. Il présentera aussi des remarques sur l'étymologie du nom selon les données de la philologie hébraïque actuelle. L'index sera relié à l'édition critique, aux éditions

diplomatiques mais aussi à la base de données BiblIndex qui recense les citations bibliques dans les textes patristiques.

Nous avons mentionné le relevé des occurrences des noms hébreux dans les textes patristiques, notamment dans les sources de Jérôme : un projet de création d'une base en XML-TEI de textes patristiques est aussi en cours de réalisation. L'objectif est double : relier cette base à BiblIndex en présentant en contexte les citations bibliques mentionnées dans BiblIndex et relier cette base à l'index des Noms hébreux en donnant, là encore, le contexte d'apparition des noms hébreux dans les oeuvres.

Ces divers projets au sein de JERIHNA font apparaître une grande variété de sources mais aussi une volonté et un besoin fort de relier toutes les données entre elles. Se pose donc la question de la manière de faire et, dans le cas des noms propres, recourir à la lemmatisation nous a semblé, dans un premier temps du moins, l'approche la plus simple.

3 Les noms propres hébreux entre lemme et entité

Si associer des formes grâce au lemme apparaît comme une bonne idée dans le cas des noms propres, on se heurte toutefois rapidement à une série de problèmes et d'interrogations.

3.1 Le choix d'un lemmatiseur

Le premier choix à faire est celui du lemmatiseur. Plusieurs facteurs distinguent les outils de lemmatisation :

- La tokenisation (certains tokenisent la ponctuation, d'autres coupent des mots,...)
- La normalisation, notamment pour le grec, par exemple le traitement des diacritiques
- Le choix du référentiel

On peut distinguer deux grandes familles de lemmatiseurs :

- Ceux qui se basent sur un dictionnaire (Eulexis, Collatinus, treeTaggers, GLEM,...)
- Ceux qui se basent sur des données en contexte (les modèles Spacy,...)

La plupart des articles sur la lemmatisation (quelle que soit la langue traitée) précisent, sans surprise, que le traitement des noms propres n'est pas le point fort des lemmatiseurs. Le texte des *Noms hébreux* de Jérôme est lui-même un cas-limite : il s'agit de noms traduits (avec souvent un passage par le grec entre l'hébreu et le latin), de noms parfois rares et présentés en quelque sorte hors

contexte puisque l'ouvrage est un glossaire. Dans l'optique de cette présentation, nous avons décidé d'effectuer quelques tests en prenant des textes "moins difficiles" dont nous aurons besoin pour la suite de projet et qui insèrent les noms hébreux dans un contexte : le début du *Commentaire de Daniel* de Jérôme pour le latin et le début de la première *Homélie sur Jérémie* d'Origène pour le grec.

3.1.1 Lemmatiseurs latins

Nous avons testé trois lemmatiseurs pour le latin : Collatinus, Pie Extended et LatinCy. Collatinus se contentant de présenter les différentes possibilités mais refusant de choisir un lemme de référence, nous l'avons rapidement écarté. Pour les deux autres, précisons d'emblée que nous ne nous sommes pas intéressés à l'analyse morpho-syntaxique ; nous avons, en revanche, jeté un oeil à l'étiquetage de la partie du discours : pour les noms hébreux, LatinCy présentait un peu moins d'erreurs d'étiquetage que Pie Extended. Dans les deux cas, les noms hébreux présents dans le texte étaient plutôt bien lemmatisés mais un certain nombre d'erreurs demeurent.

Une chose nous a semblé intéressante à relever : que ce soit avec Pie Extended ou LatinCy, à l'une ou l'autre reprise, nous avons observé des cas de lemmatisation "logiques" (ex : Samson lemmatisé en Samso ou Esaia lemmatisé en Esaia) mais qui proposaient des formes que nous n'avons pas trouvées dans la Database of Latin Dictionaries qui recense pourtant un grand nombre de dictionnaires de référence en latin, dont le *Lexicon* de Forcellini sur lequel s'appuie le modèle LASLA de Pie Extended.

3.1.2 Lemmatiseurs grecs

Nous avons testé un plus grand nombre de lemmatiseurs grecs, huit en tout :

- Trois modèles Pie : le modèle grec présent dans Pyrrha, le modèle Pie avec les données de Perseus et celui utilisant les données de Proiel
- Deux modèles TreeTagger
- Le modèle GLEM
- Le modèle OdyCy de l'environnement spaCy
- Le modèle CLTK

Les modèles TreeTagger peuvent être écartés d'emblée, car ils ne lemmatisent pas et se contentent d'indiquer la POS. Quant au modèle GLEM, son fonctionnement par règles ne répond pas à nos attentes : la plupart des noms propres ne sont pas traités (mais de manière plus surprenante, des mots courants comme *theos* ou le pronom *autos* ne le sont pas non plus). Parmi les autres lemmatiseurs, les modèles CLTK, Pie Perseus et Proiel et OdyCy semblent offrir une meilleure lemmatisation des noms hébreux que le modèle grec de Pyrrha, ce qui peut s'expliquer par la grandeur des sets d'entraînement. Comme pour

le latin, un certain nombre d'erreurs subsiste toutefois. Le modèle Pie Proiel se distingue aussi des autres lemmatiseurs par une meilleure analyse de la POS et c'est probablement avec celui-ci que nous lemmatiserons les textes grecs de notre projet.

Il convient de préciser que nous n'offrons pas de véritable analyse comparatiste des lemmatiseurs : il aurait fallu aller beaucoup plus dans le détail et traiter non seulement de plus grandes portions de texte mais aussi une variété de textes. En revanche, cette réflexion sur le choix d'un lemmatiseur pour répondre aux besoins de notre projet a permis de mettre en évidence deux grands éléments :

- Plutôt que de créer un énième lemmatiseur, il serait préférable d'améliorer les modèles et, concrètement, JERIHNA et les Sources Chrétiennes pourront fournir, avec la base textuelle en préparation, un certain nombre de textes lemmatisés avec des données vérifiées pour les noms bibliques.

- Le choix des référentiels sur lequels s'appuient les lemmatiseurs n'est pas toujours très clair dans la documentation et on se demande comment le lemmatiseur identifie certains lemmes.

Mais avoir un lemme de référence (et lui-même bien référencé) ne résout pas tous les problèmes dans le cas des noms propres et dans l'objectif de relier ces noms dans différentes sources.

3.2 Le nom propre à la frontière entre le lexique et la pragmatique

Sans entrer dans un cours de linguistique (nous ne sommes absolument pas des linguistes !), nous pouvons mettre en évidence le cas particulier des noms propres : les homonymes sont bien plus fréquents que pour les noms communs et à la notion de lemme il convient d'ajouter celle d'entité nommée. Les noms hébreux traités par Jérôme présentent différents cas de figure qui mettent en évidence la nécessité de distinguer ces deux notions.

- Le cas des homonymes : un même nom propre peut renvoyer à deux entités ou plus. Ex : Beeri

- Le cas des pseudonymes ou des doubles noms : une même entité peut exister sous divers noms. Ex : Aulon / Haseroth

Alors que la lemmatisation est un processus avant tout lexical visant à ramener les différentes flexions ou variations orthographiques d'un mot à une forme de référence choisie de manière conventionnelle, la reconnaissance d'entités nommées relève davantage de la pragmatique en ce que la signification d'une forme donnée dépend de son contexte d'apparition. La tâche consiste à déterminer si cette forme correspond à une entité spécifique dans le monde réel. Ainsi, le contexte et la compréhension du sens dans un environnement particulier deviennent cruciaux.

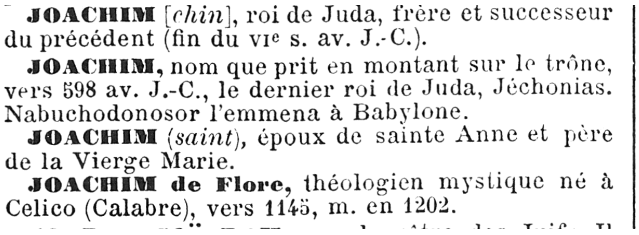
Dans le cadre de JERIHNA, les recherches sur l'orthographe des noms hébreux dans des sources plurilingues invitent à prendre en considération un troisième élément : la correspondance entre un élément et sa traduction. Comme le montre l'exemple avec Aseroth et Aulon, ce lien concerne bien plus la forme lexicale que l'entité nommée.

Il est intéressant de noter que, bien que la lemmatisation soit traditionnellement axée sur des considérations lexicales, la documentation sur les lemmatiseurs montre une évolution vers une prise en compte croissante du contexte et donc du sens. Ainsi, l'étiquetage morpho-syntaxique et le contexte d'apparition d'une forme gagnent en importance dans les approches modernes de la lemmatisation. Dans Pyrrha, par exemple, les homonymes sont parfois distingués à l'aide d'un chiffre qui renvoie aux entrées du dictionnaire de référence. Avec une telle distinction, la lemmatisation quitte le domaine purement lexical pour s'approcher de la pragmatique. Cela interroge la manière dont il faudrait, à terme, étiqueter le lemme des noms propres dans les divers lemmatiseurs : faut-il distinguer clairement le lemme de l'entité nommée, comme c'est le cas actuellement, puisque les lemmatiseurs n'indiquent que le lemme ou faudrait-il distinguer les homonymes, pour autant que cela soit possible, en s'appuyant sur des référentiels d'entités nommées (cf. exemple avec Beeri).

Les sens fournis par les dictionnaires et la lexicographie gagnent en importance pour désambiguïser les références entre différentes entités. Cette démarche entre dans le domaine de l'ontologie, où la création de liens entre divers systèmes de pensée facilite le repérage et la compréhension des entités dans leur contexte.

3.3 État de l'art des standards

Cette frontière entre ce qui relève du pragmatisme ou du monde lexical peut paraître abstraite. Les dictionnaires actuels, eux-mêmes, ne font que partiellement la distinction. Même pour des questions lexicales, le sens donné prend


JOACHIM [*chin*], roi de Juda, frère et successeur du précédent (fin du VIe s. av. J.-C.).

JOACHIM, nom que prit en montant sur le trône, vers 598 av. J.-C., le dernier roi de Juda, Jéchonias. Nabuchodonosor l'emmena à Babylone.

JOACHIM (*saint*), époux de sainte Anne et père de la Vierge Marie.

JOACHIM de Flore, théologien mystique né à Celico (Calabre), vers 1145, m. en 1202.

Figure 1: Joachim dans *Le Petit Larousse* 1924 (projet Nénufar)

plus d'importance.

Dans les textes, la question du lexique et du sens en contexte est encore plus liée et l'organisation de ces informations de nature différente devient un sujet important. Diverses modélisations existent comme la TEI. Dans le corps du

texte, on retrouve dans la balise <w> l'attribut @lemmaRef pour faire référence à une ontologie externe. Cependant c'est limité au lemme.

Exemple avec l'attribut @lemmaRef :

Nemo igitur putet eundem in Danielis principio esse Ioachim, qui in Hiezechielis exordio Ioiachin scribitur.

Que personne donc ne pense que ce Joachim du début du livre de Daniel est le même homme qui, au commencement du livre d'Ézéchiel, se trouve écrit sous la forme Joachin.

Jérôme, *Commentaire sur Daniel* (éd. et trad. R. Courtray, SC 602)

```
<p>
   <w lemmaRef="https://lila-erc.eu/data/id/lemma/113605
   ↪  " lemma="nemo" pos="pronoun">nemo</w>
   <w lemmaRef="https://lila-erc.eu/data/id/lemma⌋
   ↪  /106127" lemma="igitur"
   ↪  pos="adverb">igitur</w>
   <w lemmaRef="http://lila-erc.eu/data/id/lemma/120610"
   ↪  lemma="puto" pos="verb">putet</w>
   <w lemmaRef="https://lila-erc.eu/data/id/lemma⌋
   ↪  /109082" lemma="idem"
   ↪  pos="determiner">eundem</w>
   <w lemmaRef="http://lila-erc.eu/data/id/lemma/106748"
   ↪  lemma="in" pos="adposition">in</w>
   <persName
   ↪  ref="https://www.wikidata.org/wiki/Q171724">
         <w lemmaRef="https://lila-erc.eu/data/id/le⌋
         ↪  mma/4509" lemma="daniel"
         ↪  pos="proper_noun">danielis</w>
   </persName>
   <w lemmaRef="http://lila-erc.eu/data/id/lemma/119505"
   ↪  lemma="principium"
   ↪  pos="common_noun">principio</w>
   <w lemmaRef="http://lila-erc.eu/data/id/lemma/126689"
   ↪  lemma="sum" pos="verb">esse</w>
   <persName
   ↪  ref="https://www.wikidata.org/wiki/2307004">
         <w lemmaRef="https://lila-erc.eu/data/id/le⌋
         ↪  mma/10866" lemma="ioachim"
         ↪  pos="proper_noun">ioachim</w>
   </persName>
   <pc>,</pc>
   <w lemmaRef="http://lila-erc.eu/data/id/lemma/121354"
   ↪  lemma="qui" pos="pronoun">qui</w>
   <w lemmaRef="http://lila-erc.eu/data/id/lemma/106748"
   ↪  lemma="in" pos="adposition">in</w>
```

```
        <persName
        ↪  ref="https://www.wikidata.org/wiki/Q194064">
              <w lemmaRef="https://lila-erc.eu/data/id/le⌋
              ↪  mma/9227" lemma="hezechiel"
              ↪  pos="proper_noun">hiezechielis</w>
        </persName>
        <w lemmaRef="http://lila-erc.eu/data/id/lemma/102102"
        ↪  lemma="exordium" pos="common_noun">exordio</w>
        <persName
        ↪  ref="https://www.wikidata.org/wiki/Q319049">
              <w lemmaRef="https://lila-erc.eu/data/id/le⌋
              ↪  mma/10866" lemma="ioachim "
              ↪  pos="proper_noun">ioiachin</w>
        </persName>
        <w lemmaRef="http://lila-erc.eu/data/id/lemma/123840"
        ↪  lemma="scribo" pos="verb">scribitur</w>
     </p>
```

Il est également possible d'utiliser le système de StandOff pour lier le mot avec un référentiel extérieur :

```
<standOff>
   <superEntry type="lemma">
   <entry xml:id="lem_113605" >
      <lbl>nemo</lbl>
      <ref target="https://lila-erc.eu/data/id/lemma/113605"/>
   </entry>
   <entry xml:id="lem_106127" >
      <lbl>igitur</lbl>
      <ref target="https://lila-erc.eu/data/id/lemma/106127"/>
   </entry>
      <entry xml:id="lem_120610">
      <lbl>puto</lbl>
      <ref target="https://lila-erc.eu/data/id/lemma/120610"/>
   </entry>
      <entry xml:id="lem_109082">
      <lbl>idem</lbl>
      <ref target="https://lila-erc.eu/data/id/lemma/109082"/>
   </entry>
      <entry  xml:id="lem_106748">
      <lbl>in</lbl>
      <ref target="https://lila-erc.eu/data/id/lemma/106748"/>
   </entry>
      <entry xml:id="lem_4509" >
      <lbl>daniel</lbl>
```

```
      <ref target="https://lila-erc.eu/data/id/lemma/4509"/>
   </entry>

   </superEntry>
   <superEntry type="pos">
      <entry xml:id="pos_1">
         <lbl>pronoun</lbl>
         <ref  target="https://lila-erc.eu/lodview/ontologie
         ↪  s/lila/pronoun"/>
      </entry>
      <entry xml:id="pos_2">
         <lbl>adverb</lbl>
         <ref  target="https://lila-erc.eu/lodview/ontologie
         ↪  s/lila/adverb"/>
      </entry>
      <entry xml:id="pos_3">
         <lbl>verb</lbl>
         <ref target="https://lila-erc.eu/lodview/ontologie
         ↪  s/lila/verb"/>
      </entry>
      <entry xml:id="pos_4">
         <lbl>determiner</lbl>
         <ref target="https://lila-erc.eu/lodview/ontologie
         ↪  s/lila/determiner"/>
      </entry>
      <entry xml:id="pos_5">
         <lbl>adposition</lbl>
         <ref target="https://lila-erc.eu/lodview/ontologie
         ↪  s/lila/adposition"/>
      </entry>
      <entry xml:id="pos_6">
         <lbl>proper noun</lbl>
         <ref target="https://lila-erc.eu/lodview/ontologie
         ↪  s/lila/proper_noun"/>
      </entry>
   </superEntry>
   <listPerson>
      <person xml:id="pers_1" >
         <name>danielis</name>
         <idno type="wikidata">Q171724</idno>
      </person>
   </listPerson>
</standOff>
<text>
   <body>
      <p>
```

```
            <w ana="#lem_113605 #pos_1">nemo</w>
            <w ana="#lem_106127  #pos_2">igitur</w>
            <w ana="#lem_120610 #pos_3">putet</w>
            <w ana="#lem_109082 #pos_4">eundem</w>
            <w ana="#lem_106748  #pos_5">in</w>
            <w ana="#lem_4509 #pos_6 #pers_1">danielis</w>
        </p>
    </body>
</text>
```

Des outils davantage utilisés dans la sphère des TAL-istes ont également pris en considération cette question avec CONLL-RDF. Il existe donc des possibilités pour baliser des mots.

Figure 2: CONLL-RDF

Une fois les mot balisés au sein d'un texte et on peut les relier à une ontologie externe. Cette ontologie est un référentiel pouvant être aligné à d'autres référentiels. Par exemple : le Gaffiot aligné avec le Forcellini. Cela permet à plusieurs systèmes d'organisation des savoirs de coexister. Cette organisation des savoirs peut favoriser une meilleure compréhension du sens du mot en contexte. On peut également imaginer de faire des liens entre le domaine lexical et le domaine des entités nommées.

Pour organiser et construire ces systèmes de référentiels, il existe plusieurs schémas XML. La TEI est sûrement le schéma faisant le plus consensus. On retrouve un standard assez commun, TEI Lex-O, qui est le module dictionnaire de la TEI mais aussi le module de taxonomy. Mais dès qu'on souhaite lier des données entres elles, la TEI montre rapidement ses limites.

Le système de RDF s'avère davantage flexible pour connecter les référentiels. Des modèle de données tels qu'OntoLex peuvent se révéler utiles. Lorsqu'on regarde le schéma on voit bien que le sens prend une dimension centrale.

Le texte des *Noms hébreux* de Jérôme et le travail sur les sources, avec l'index et la base textuelle en préparation, posent d'autres questions de modélisation : à la distinction entre le lemme et l'entité nommée s'ajoute la traduction. Comment lier des ressources multi-langues ? Quelle granularité met-on en évidence

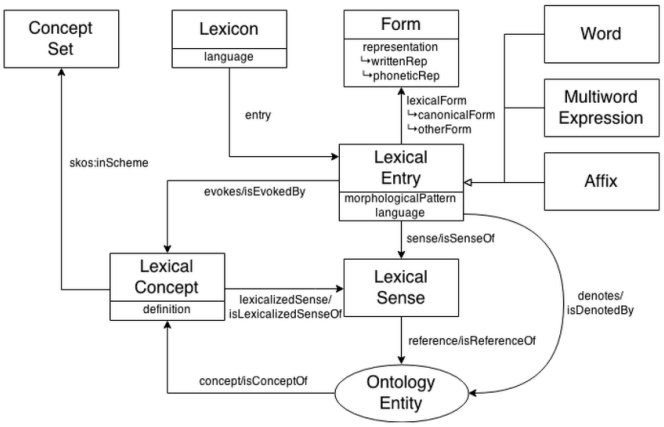


Figure 3: OntoLex

dans la relation entre un terme et son équivalent dans une autre langue : faut-il distinguer les dérivations de simples traductions ? Existe-t-il, en dehors du SKOS, une typologie des liens entre les traductions ?

Toutes ces questions de modélisation de données sont loin d'être entièrement résolues, mais notre travail peut s'insérer dans des problématiques plus larges.

4 Le Linked Open Data : la solution ?

4.1 Entité nommée et Linked Open Data

Aujourd'hui les bases prosopographiques sont assez présentes dans le LOD, que ce soit au niveau international ou national. Elles peuvent être collaboratives ou alimentées uniquement par des spécialistes : Wikidata, data-Biblissima, Viaf, IDref,... Quelle que soit leur granularité, les données sont présentes. De plus en plus, les données en contexte sont reliées à ces bases de référentiels dans un souci de désambiguïsation des homonymes. Cela favorise le développement d'outils entity-linking puisque de plus gros jeu de données permettent leur amélioration. Rien que pour cette tâche, Biblissima a besoin de développer, en son sein, une base de textes reliée a son propre référentiel data.biblissima (cela semble être un objectif du cluster 5b).

4.2 Linked Open data en linguistique des langues anciennes

Dans le même esprit que ce qui est fait pour les entités nommées, BIBLISSIMA aurait besoin d'un cadre de dépôt des différentes données linguistiques. Les approches peuvent être différentes : intérêt pour les dictionnaires, besoin de lemmatiser des textes par règle ou avec l'IA et sans doute d'autres approches

que nous oublions. Les données, en revanche, présentent un certain nombre de similitudes, voire sont les mêmes. Il y aurait donc un véritable intérêt à les aligner entre elles. Cela permettrait, par exemple, aux utilisateurs de mieux comprendre le choix des lemmes qui a pu être oépré dans certains cas.

Pour le latin, il existe déjà une ERC, Linked latin (LiLa), où des lemmes sont liés à divers référentiels linguistiques. Le site de l'ERC présente 3 interfaces : une interface de navigation (peu intuitive pour les novices) à travers les référentiels à la manières de data-biblissima, une iterface pour lancer des requêtes SPARQL et une interface de requête plus simple retournant une liste de lemmes. Si ce projet est un très bel apport à l'interconnexion des données linguistiques en latin, nous pouvons relever deux limites pour Biblissima :

- Les références de certains lemmes ne sont pas données.
- Les données traitées ne concernent que le latin (c'est déjà un gros apport !).

Concrètement, le travail que nous effectuons sur les *Noms hébreux* de Jérôme nous amène à traiter un certain nombre de données linguistiques qui pourraient venir enrichir un tel référentiel, que ce soit par l'ajout de lemmes sur des noms hébreux peu courants et par un travail sur les référentiels relatifs à ces noms. Nous pourrions entrer en contact avec les membres de cette ERC et leur proposer nos données, mais nous nous demandons s'il n'y aurait plutôt un intérêt à créer un dépôt linguistique de type LOD au sein de Biblissima. Nous y voyons plusieurs intérêts :

- Il ne s'agirait pas de partir de rien : il serait possible de récupérer les données de Collatinus et d'Eulexis par exemple.
- Une telle base permettrait de mettre en parallèle le latin et le grec et pourrait ouvrir la réflexion sur les questions de traductologie.
- Ce référentiel pourrait alimenter des outils comme Pyrrha .
- Si l'on se concentre, dans un premier temps au moins, sur les noms propres, on pourrait aussi envisager un travail d'alignement avec les entités nommées dans data.biblissima.

Nous sommes bien conscients de l'ampleur du chantier proposé... Dans le projet JERIHNA, nous sommes très régulièrement confrontés au besoin d'aligner les référentiels pour mieux traiter les données. Mais la tâche est telle qu'elle ne peut être réalisée qu'à plusieurs et sur du long terme. En proposant la création d'une telle base, nous cherchons avant tout à lancer la discussion sur le traitement des données linguistiques dans Biblissima.