

Vers un corpus numérique aljamiado: quelques exemples d'encodage XML-TEI.

Pablo Roza Candás
INALCO-Paris / SEAR-Oviedo

Pendant les dernières décennies, les études aljamiadas¹ ont vu une croissance significative en ce qui concerne le nombre de monographies, de dissertations et d'articles consacrés à l'édition de la production manuscrite morisque. Parmi ces études, deux projets développés à l'Université d'Oviedo et lancés par Álvaro Galmés de Fuentes ont été décisifs pour l'avancée scientifique dans le domaine: la *Colección de literatura española aljamiado-morisca* (CLEAM),² aujourd'hui une plate-forme bien consolidée pour l'édition de nouvelles compilations et d'études, et en ce qui concerne la lexicographie aljamiada, le *Glosario de voces aljamiado-moriscas* (GVAM),³ édité par la Bibliotheca Arabo-Romanica et Islamica de la même université asturienne.

On doit préciser néanmoins que la production aljamiada, malgré son grand intérêt du point de vue linguistique, continue toujours à être un domaine marginal dans les études hispaniques, cela étant dû en partie aux solutions ecdotiques, parfois très complexes, appliquées à ce corpus littéraire.⁴ L'un des aspects principaux de l'exploitation du corpus aljamiado (au-delà de l'intérêt littéraire) est, sans aucun doute, la connaissance approfondie de la langue quotidienne et de la variation dialectale des XVI^e et XVII^e siècles, puisque ces codex «marginiaux» reflètent souvent des aspects linguistiques qui apparaissent rarement dans les textes «chrétiens» du Siècle d'Or.

À cet égard, le projet que je présente ici essaie de rendre plus accessible cette production littéraire en utilisant les nouvelles ressources des humanités numériques. Nous avons besoin d'outils précis qui nous permettent d'effectuer des éditions fiables ainsi que d'accéder rapidement aux textes. Des supports solides et une classification correcte de ces matériaux nous permettront de répondre à des questions jusqu'à présent étudiées d'une manière partielle comme celles relatives à la dimension variationnelle de la modalité aljamiada ou la filiation textuelle des différents manuscrits du corpus.

¹ Ce travail a été réalisé dans le cadre du programme postdoctoral Marie Curie – Clarín Cofund (2014-2016) dans le Centre de Recherches Moyen-Orient Méditerranée (CERMOM EA 4091) de l'Institut National des Langues et Civilisations Orientales (INALCO) de Paris.

² Voir les volumes publiés entre 1970-2005: <http://www.arabicaetromanica.com/cleam/>

³ GÁLMÉS de FUENTES, ÁLVARO, M. SÁNCHEZ ÁLVAREZ, A. VESPERTINO RODRÍGUEZ, J. C. VILLAYERDE AMIEVA, *Glosario de voces aljamiado-moriscas*, Oviedo (Universidad de Oviedo, Departamento de Filología Clásica y Románica, Biblioteca Árabo-Románica), 1994, réimp. 2015, Gijón (Trea).

⁴ Sur ce point, nous avons choisi comme base de notre travail le système de translittération CLEAM (Colección de Literatura Española Aljamiado-Morisca) de l'Université d'Oviedo, dans lequel on combine parfaitement la lisibilité avec une fidélité maximale aux caractéristiques graphématiques des manuscrits. De cette façon, nous obtenons une translittération semi-paléographique qui nous permettra de restituer à tout moment le texte en caractères arabes

Jusqu'à présent, la seule expérience de recherche dans le domaine des humanités numériques appliquées au corpus aljamiado se réduit au projet⁵ que, avec Raquel Suárez García (SEAR- Université d'Oviedo), nous développons depuis 2012 dans le cadre du «Portal Andrés de Poza. Multiversion Texts and Digital Edition»⁶ de l'Université de Deusto. Cette équipe de recherche, menée par Carmen Isasi, s'est consacrée à l'édition et à la publication numérique de textes multi-version, y compris des traductions dans certaines langues européennes. Dans ce projet, nous avons exploré les possibilités d'une édition numérique des textes aljamiados sur la base d'une étude des différents récits de la tradition narrative morisque. Comme résultat de cette première expérience, nous avons pu établir les fondements pour une édition numérique des textes romans en caractères arabes en utilisant une méthodologie adaptée et simple mais, à notre avis, très efficace pour ce type de documents.

Sur la base de cette expérience pionnière, je mène actuellement dans le cadre du Centre de Recherches Moyen-Orient Méditerranée (CERMOM EA 4091) de l'Institut National des Langues et Civilisations Orientales (INALCO) de Paris, le projet de recherche «Digital Aljamiado: Development and Application of Digital Humanities Methodology to Spanish Aljamiado Corpus» (2014-2016). Ce nouveau projet se focalise sur les problèmes théoriques et pratiques liés à l'édition numérique XML (*Extensible Markup Language*), tels que la segmentation et l'alignement de versions ou le balisage pour lequel je suis les standards généralement reconnus du TEI (*Text Encoding Initiative*),⁷ ainsi que les critères fixés par le réseau international CHARTA (*Corpus hispánico y americano en la red: Textos antiguos*)⁸ dans lequel sont rassemblées plusieurs équipes de recherche dans le domaine des humanités numériques.

Dans ce travail je présenterai quelques cas concrets d'encodage XML-TEI appliqués au corpus aljamiado comme un exemple de l'utilisation et le rendement des nouveaux outils numériques. Sur la base de la proposition CHARTA-TEI⁹ fixée pour l'édition des textes espagnols en caractères latins, j'analyserai ici les possibilités d'adaptation de certaines de ces solutions ecdotiques aux textes en caractères arabes.

L'édition numérique XML et le balisage TEI

A grands traits, le XML (*Extensible Markup Language*) est un langage de codification des documents qui, avec le texte, incorpore des balises qui contiennent des informations sur la structure et le contenu du texte. De cette manière, un fichier XML est un fichier de texte normal mais avec un certain nombre d'éléments spécifiques. Un élément commence par une balise de début entre chevrons (<>) et finit toujours par une balise de fin (</>), placée aussi entre chevrons. Tous les éléments peuvent contenir zéro, un ou plusieurs attributs. Un attribut donne des précisions sur les éléments et leur contenu et il est composé d'un nom et d'une valeur.

⁵ Une édition numérique alignée des différentes versions du récit morisque de la *Donzella Arcayona* (ms. J 57 et J 3 du Consejo Superior de Investigaciones Científicas, ms. 5313 de la Biblioteca Nacional de España, ms. aljamiado de la Biblioteca Pública de Lleida et ms. 1944 de la Bibliothèque nationale d'Algérie).

⁶ En ligne: <http://andresdepoza.com>

⁷ En ligne: <http://www.tei-c.org>

⁸ En ligne: <http://www.charta.es>

⁹ CARMEN ISASI MARTÍNEZ (coord.), ANA LOBO PUGA, LEYRE MARTÍN AIZPURU, SANTIAGO PÉREZ ISASI, ELENA PIERAZZO y PAUL SPENCE (coord.), *Guía para editar textos CHARTA según el estándar TEI: una propuesta* (2014). En ligne: <<http://www.charta.es/investigacion/charta-tei/>>.

En ce qui concerne le balisage des textes on suit la proposition TEI (*Text Encoding Initiative*) dans sa version P5 (2007), résultant d'un projet interdisciplinaire des plusieurs universités européennes et américaines.¹⁰ Le standard TEI est sans doute le schéma le plus étendu au niveau international dans le balisage de textes. Ceci nous assure qu'à la fin les documents pourront être reconnus, réutilisés et partagés, ce qui est l'objectif de notre projet en fin de compte.

L'édition XML et le balisage TEI, sont deux processus complémentaires et souvent simultanés; en même temps que l'on établit le texte, on ajoute une série de balises relatives, par exemple à des ratures du copiste, des détériorations du manuscrit, etc. Dans une seconde phase, on ajoute des nouvelles balises concernant la classification du lexique, les toponymes et anthroponymes, les expressions arabes, etc.

Puisque notre intérêt est fondamentalement linguistique, le balisage qu'on a initialement considéré s'est focalisé, d'une part, sur la notation d'éléments de la langue, comme les traits dialectaux notables ou les phénomènes d'hybridation arabo-romane; d'autre part, nous classons les aspects codicologiques, comme les déchirements du papier, ou quelques autres dommages physiques des manuscrits qui peuvent conditionner l'édition d'une façon ou d'une autre.

1) Les marques de segmentation.



<p xml:id="es35">Yā donzella, dīme de quién eres hija </p>

<p xml:id="es36">Dīšo ella:</p>

<p xml:id="es37">Yā el rey, mi nonbre es Arcayona i yo soy hija del rey Nağrab, señor de Al-Hindi (...)</p>

Par l'utilisation de l'élément <p>, nous fixons dans le document XML la structure du texte qu'on a préalablement établi. Par exemple, dans ce cas,¹¹ la structure dialogique du texte elle-même nous a servi de base pour la segmentation. À chacun de ces segments, l'éditeur XML assignera automatiquement un nombre dans le fichier (comme valeur d'un attribut), ce qui nous permettra ultérieurement de confronter dans la plate-forme de visualisation les différents passages comparables du texte.

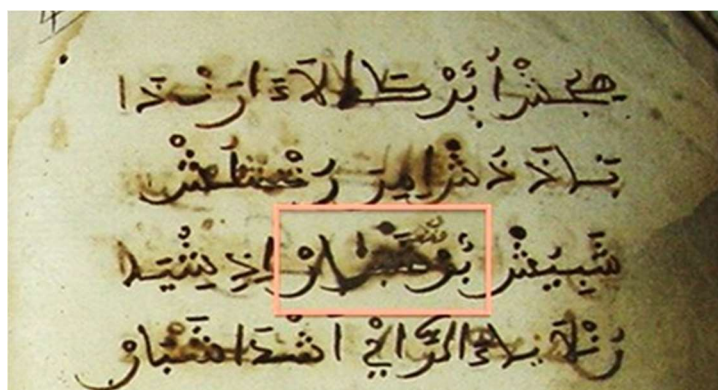
2) Les interventions du copiste.

¹⁰ En ligne: <http://www.tei-c.org>

¹¹ Ms. 1944 de la Bibliothèque nationale d'Algérie, fol. 55v.

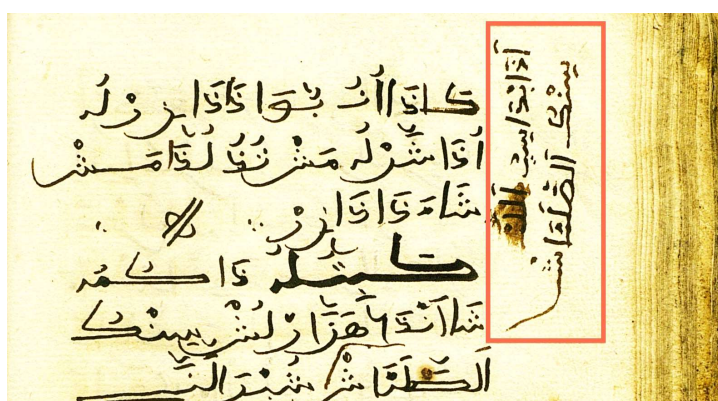
On considère ici toutes ces modifications effectuées par le copiste sur le document original pour corriger le texte écrit, comme des ratures, des annulations, des interlignés ou des notations en marge. En outre, on inclut dans ce paragraphe d'autres phénomènes en rapport aussi avec l'activité des copistes: comme les changements de main, les *lapsus calami* ou les espaces en blanc.

a) L'addition de texte.¹²



por <add place="arriba">su</add> saber

L'interligné est l'addition de texte de la même main du copiste ou d'une autre main ultérieure. On emploie l'élément <add> pour indiquer qu'il y a du texte ajouté, dans ce cas la forme de l'adjectif possessif "su". On utilise l'attribut (c'est-à-dire une valeur associée à l'élément), dans ce cas **place**, pour préciser le lieu où le texte apparaît : "**arriba**" (en haut) comme dans ce cas¹³ ou "**margen**" (en marge) comme dans le cas ci-dessous,¹⁴ etc.



<add place="margen">aḍebdeçió Allah çinco a'ssala'es</add>

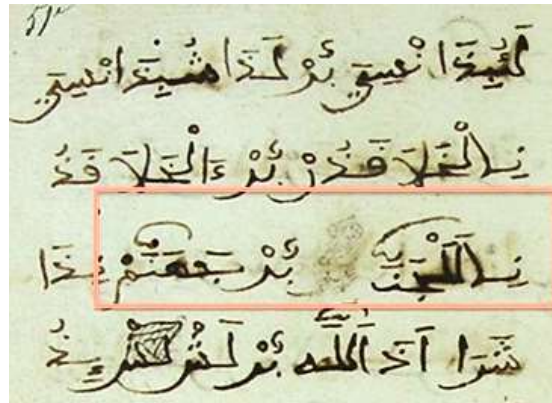
b) Les ratures faites par le copiste.¹⁵

¹² CARMEN ISASI MARTÍNEZ (coord.), ANA LOBO PUGA, LEYRE MARTÍN AIZPURU, SANTIAGO PÉREZ ISASI, ELENA PIERAZZO y PAUL SPENCE (coord.), *op. cit.*, pp. 60-62.

¹³ Ms. 1944 de la Bibliothèque nationale d'Algérie, fol. 2v.

¹⁴ Ms. L531 du Fondo Documental Histórico de las Cortes de Aragón (Saragosse), fol. 52v.

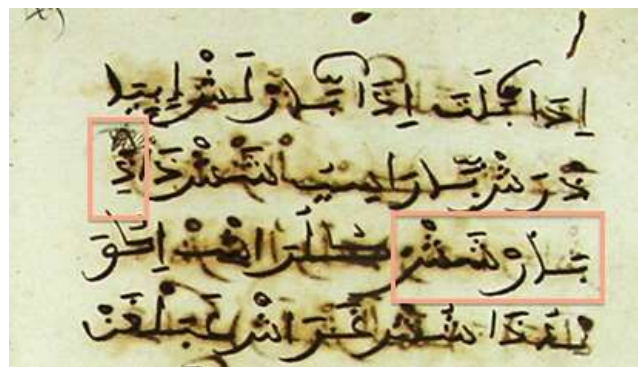
¹⁵ CARMEN ISASI MARTÍNEZ (coord.), ANA LOBO PUGA, LEYRE MARTÍN AIZPURU, SANTIAGO PÉREZ ISASI, ELENA PIERAZZO y PAUL SPENCE (coord.), *op. cit.*, pp. 58-59.



ni a alġanna ~~type="raspado"/>~~ por ġahannam

Le raturage suppose une façon de supprimer du texte par l'abrasion du support. Dans cette situation la transcription du passage éliminé est assez peu probable. On utilise l'élément **** pour indiquer qu'il y a du texte supprimé avec l'attribut **type**, pour préciser le type de raturage, dans ce cas¹⁶ **"raspado"** (érasé).

c) Les annulations de texte.¹⁷



d~~type="cancelado">a<add>i</add>~~versas

Les annulations sont définies comme un type spécial de rature dans laquelle le copiste souligne, avec par exemple un pointillé, les mots qu'il veut éliminer ou bien avec un simple trait sur le passage de texte à supprimer. Comme on peut voir dans ce cas,¹⁸ le copiste écrit *daversas*, possiblement une forme dialectale, et après il la corrige par *diversas*. Par conséquent, on emploie l'élément **** pour indiquer qu'il y a du texte annulé (la voyelle *fatha*) et on utilise l'attribut **type**, pour préciser le type d'annulation **"cancelado"** (annulé). De même, avec l'élément **<add>** on peut noter la correction faite par le copiste, dans ce cas l'addition de *kasra*.

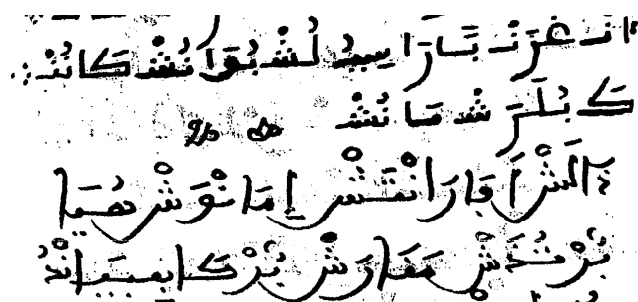
d) Les changements de main.¹⁹

¹⁶ Ms. 1944 de la Bibliothèque nationale d'Algérie, fol. 51r.

¹⁷ CARMEN ISASI MARTÍNEZ (coord.), ANA LOBO PUGA, LEYRE MARTÍN AIZPURU, SANTIAGO PÉREZ ISASI, ELENA PIERAZZO y PAUL SPENCE (coord.), *op. cit.*, pp. 56-57.

¹⁸ Ms. 1944 de la Bibliothèque nationale d'Algérie, fol. 25v.

¹⁹ CARMEN ISASI MARTÍNEZ (coord.), ANA LOBO PUGA, LEYRE MARTÍN AIZPURU, SANTIAGO PÉREZ ISASI, ELENA PIERAZZO y PAUL SPENCE (coord.), *op. cit.*, pp. 63-64.

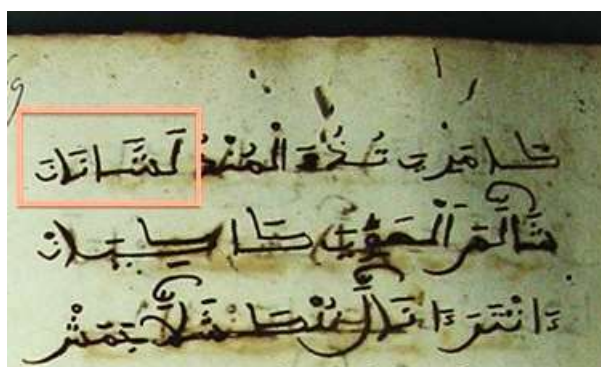


en gran preçio los buenos que nunca volarás menos **<handShift new="#h2"/>** de las afrentas i menwas huye por todas maneras porque viviendo

Par l'utilisation de l'élément **<handShift>** en combinaison avec l'attribut **new** on peut noter les différents copistes qui sont intervenus dans la confection du codex. De cette manière, dans le cas ci-dessus²⁰ on indique le changement de main avec l'annotation **"#h2"**, valeur que nous pouvons modifier ultérieurement selon les mains intervenant au cours de la copie, **"#h3"**, **"#h4"**, etc.

3) Les interventions de l'éditeur.

a) Les formes incorrectes.²¹



[la **<choice><sic>senena</sic><corr>setena</corr></choice>**]

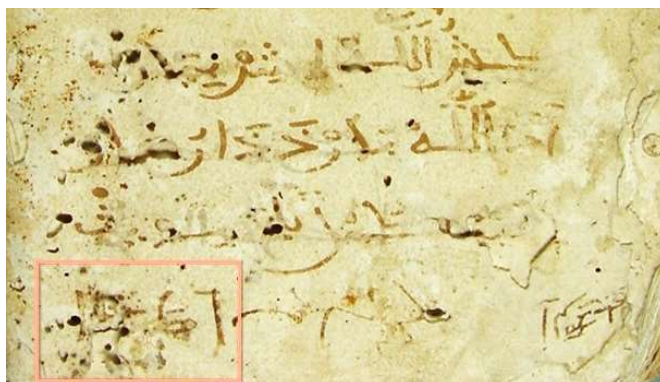
La prémisses dont nous partons est celle de l'intervention minimale, c'est-à-dire, on ne propose des corrections que quand nous sommes absolument sûrs que le texte original est erroné. Par exemple, dans ce passage²² on trouve la forme “la senena” par “la setena” (‘la septième’) dans laquelle le copiste a oublié le deuxième point sur la consonne *tā*. On utilise l'élément **<choice>** pour indiquer l'élection entre une forme erronée et une forme corrigée. Avec l'élément **<sic>** on peut noter la première “senena” et avec **<corr>** la deuxième, la forme que nous avons corrigée “setena”.

²⁰ Ms. 1163 de la Bibliothèque nationale de France, fol. 65r.

²¹ CARMEN ISASI MARTÍNEZ (coord.), ANA LOBO PUGA, LEYRE MARTÍN AIZPURU, SANTIAGO PÉREZ ISASI, ELENA PIERAZZO y PAUL SPENCE (coord.), *op. cit.*, pp. 65-66.

²² Ms. 1944 de la Bibliothèque nationale d'Algérie, fol. 35v.

b) *Les formes illisibles.*²³



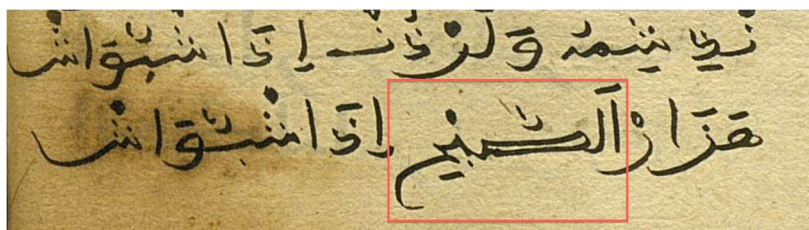
acá<subst><gap reason="ilegible"/><add>base</add></subst>

Un cas particulier est constitué par les passages illisibles en raison de la détérioration du support. Dans ce cas on a l'option de restituer le texte original en utilisant l'élément <subst> pour indiquer l'ajout de texte que l'on fait, en combinaison avec et <gap> qui nous indiquent qu'il y a un passage illisible. Avec l'attribut **reason** on indique que le texte est “ilegible” (illisible) dans l'original. Ensuite avec l'élément <add> on restitue le texte qui manque. Par exemple, dans ce passage²⁴ qui correspond à la fin d'un récit on peut restituer la forme “acábase”, c'est-à-dire ‘il s'achève’.

4) **La classification du lexique.**

On peut noter et classer de manière systématique toutes ces données qui peuvent être intéressantes pour une exploitation ultérieure des textes. Nous procédons ainsi à l'extraction de correspondances lexicales en vue de créer des glossaires (avec et sans contextes), comme exemple des réutilisations possibles du corpus pour des études linguistiques ou littéraires.

a) *Les formes et expressions arabes.*²⁵

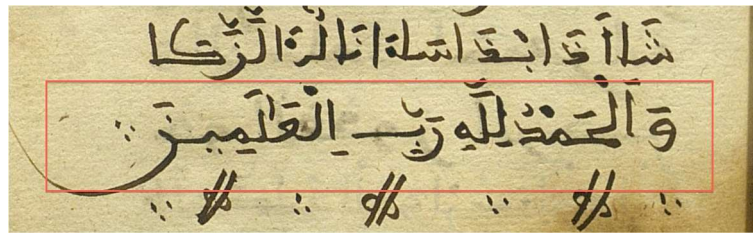


<foreign xml:lang="ar">a'ssubhi</foreign>

²³ CARMEN ISASI MARTÍNEZ (coord.), ANA LOBO PUGA, LEYRE MARTÍN AIZPURU, SANTIAGO PÉREZ ISASI, ELENA PIERAZZO y PAUL SPENCE (coord.), *op. cit.*, pp. 50-52.

²⁴ Ms. 1944 de la Bibliothèque nationale d'Algérie, fol. 65v.

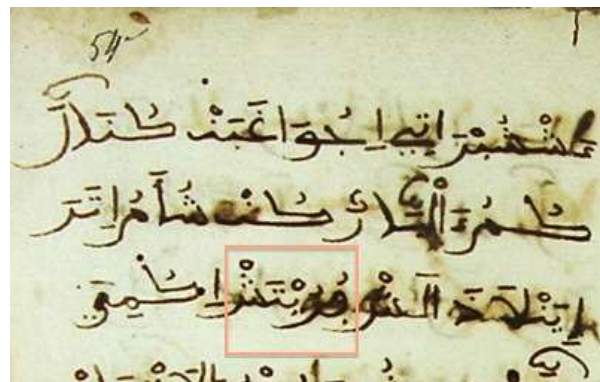
²⁵ CARMEN ISASI MARTÍNEZ (coord.), ANA LOBO PUGA, LEYRE MARTÍN AIZPURU, SANTIAGO PÉREZ ISASI, ELENA PIERAZZO y PAUL SPENCE (coord.), *op. cit.*, pp. 84-86.



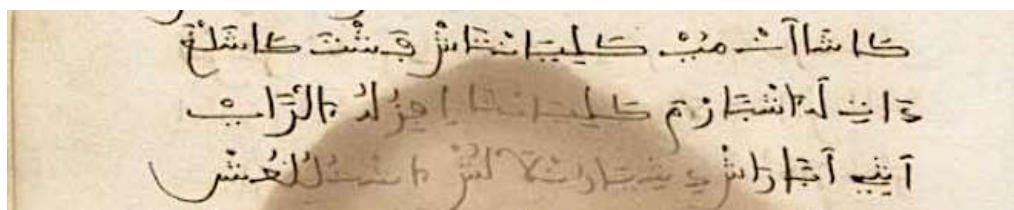
<foreign xml:lang="ar">wa-al-ḥamdu li-l-Lahi rabbi il-‘ālamīna</foreign>

C'est le cas, par exemple, des formes et des extraits en arabe pour lesquelles on utilise l'élément <foreign> avec l'attribut **lang** pour indiquer l'utilisation de l'arabe dans le passage. Dans ces cas,²⁶ on peut aussi ajouter sur la plate-forme de visualisation de l'édition une note dépliant avec la traduction pour faciliter la lecture des non-arabisants.

b) Les formes dialectales.



es sobre tī. I juegaban con-ella como el perro con su amo; i traíanle de las <seg type="dialectal">fruytas</seg> i comía.



que sean muy calientes fasta que salga de ti la esperma caliente. I fizolo el rey asī <seg type="dialectal">après</seg> dișiéronle los estlólogos.

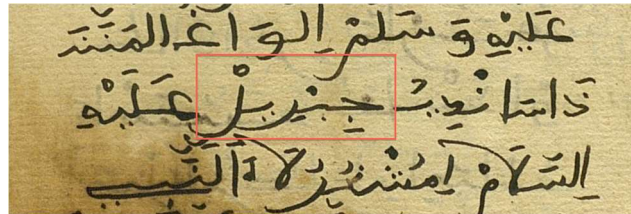
On traite de la même façon, les traits dialectaux, et particulièrement le lexique aragonais. Il est intéressant de procéder à une récolte systématique des éléments aragonais qui nous permettra de classer et d'analyser ultérieurement ce lexique depuis une perspective dialectologique plus précise que celle qui a été à l'œuvre jusqu'à présent.

De cette façon, afin de noter toutes ces données dialectales, on propose l'élément

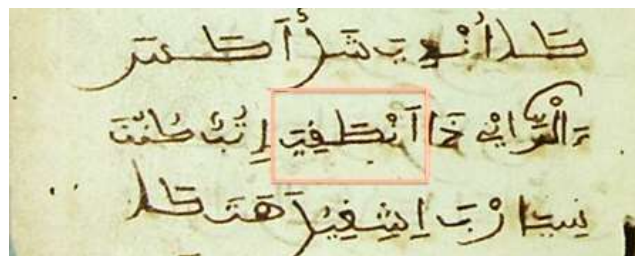
²⁶ Ms. L531 du Fondo Documental Histórico de las Cortes de Aragón (Saragosse), fols. 58v et 108r respectivement.

<seg> en combinaison avec l'attribut **type** avec la valeur **dialectal** pour noter, par exemple dans le premier passage²⁷ la filiation aragonaise de la voix *fruytas* 'les fruits' face au castillan *frutas*, ou dans le deuxième exemple²⁸ la forme *après* 'après' face au castillan 'después'.

c) *Les toponymes et anthroponymes.*²⁹



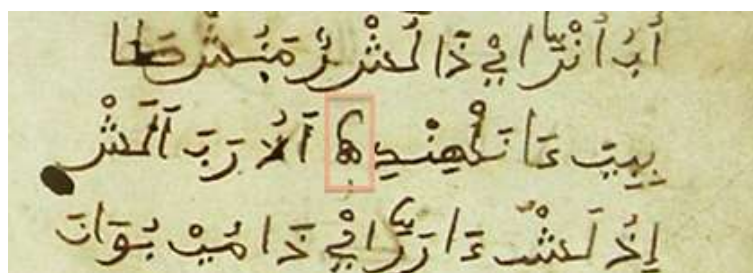
<persName key="Ġibrīl">Ġibrīl</persName>



<placeName key="Antākiya">Antāqia</placeName>

Par l'utilisation des éléments <persName> et <placeName> on peut noter les anthroponymes et toponymes respectivement. Dans ces cas³⁰ l'attribut **key** nous permette d'ajouter une forme régularisée (dans ces cas, "Ġibrīl" et "Antāqiya") sous laquelle on pourrait regrouper dans la plate-forme de visualisation les différentes variantes qui apparaissent dans le texte.

4) **Les caractères spéciaux.**



Ubo un rey de los romanos que vivía en-Al-Hindi <seg>

²⁷ Ms. 1944 de la Bibliothèque nationale d'Algérie, fol. 54r.

²⁸ Ms. J 57 du CSIC, fol. 33r.

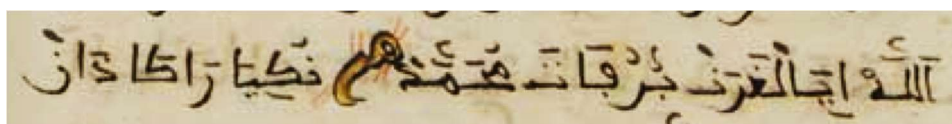
²⁹ CARMEN ISASI MARTÍNEZ (coord.), ANA LOBO PUGA, LEYRE MARTÍN AIZPURU, SANTIAGO PÉREZ ISASI, ELENA PIERAZZO y PAUL SPENCE (coord.), *op. cit.*, pp. 97-100.

³⁰ Ms. L531 du Fondo Documental Histórico de las Cortes de Aragón (Saragosse), fol. 52v et ms. 1944 de la Bibliothèque nationale d'Algérie, fol. 54r.

type="trebol"/> ; adoraba a las ídolas. Era rey de muy buena

L'adaptation de la proposition CHARTA-TEI³¹ nous permette aussi de noter les caractères spéciaux souvent employés par les copistes pour souligner quelques mots clés (termes religieux, toponymes, anthroponymes, etc.) ou bien comme un simple élément décoratif du manuscrit. De cette manière, dans le cas ci-dessus³² on peut indiquer avec l'élément **<seg>** en combinaison avec l'attribut **type** l'apparition d'un **"trebol"** (trèfle) après le toponyme Al-Hindi.

De même, par la combinaison des éléments **<figure>** et **<graphic>** on peut créer un lien (dans le cas ci-dessous **url="trebol.jpg"**) vers l'image que nous avons préalablement balisé.³³



Allah y-el gran profeta Muḥammad **<seg type="trebol"><figure> <graphic url="trebol.jpg"> </figure></seg>** no quiere que den

Conclusion

En définitive, j'ai présenté ici quelques exemples des multiples possibilités que les nouvelles ressources des humanités numériques nous offrent pour l'édition et l'étude du corpus morisque. A cet égard, le but de l'encodage XML-TEI appliqué aux textes aljamiados est de classer et de mettre à disposition toute une série de matériaux qui seront d'une grande utilité pour des futures études, comme l'analyse de la variation linguistique (notamment en ce qui concerne les aspects diachroniques et diatopiques), la filiation et l'établissement des familles textuelles ou les études comparatives de la tradition aljamiado-morisque avec d'autres traditions littéraires.

L'édition numérique XML-TEI ajoute sans doute une nouvelle valeur à l'édition d'un témoignage littéraire concret dans le contexte du corpus aljamiado. De cette façon, on croit que le développement de cette méthodologie sera d'utilité pour l'étude de cette production textuelle caractérisée par la diversité de leurs versions.

Annexe

³¹ CARMEN ISASI MARTÍNEZ (coord.), ANA LOBO PUGA, LEYRE MARTÍN AIZPURU, SANTIAGO PÉREZ ISASI, ELENA PIERAZZO y PAUL SPENCE (coord.), *op. cit.*, pp. 66-78.

³² Ms. 1944 de la Bibliothèque nationale d'Algérie, fol. 1v.

³³ Ms. J 4 du CSIC, fol. 155v.

Exemple d'édition XML-TEI



Ms. 1944, Bibliothèque nationale d'Algérie, fols. 1v – 2 r.

<TEI>
 <teiHeader/>
 <text>
 <body>
 <p xml:id="es4">Yā <persName key="Ka'b al-Aḥbār">Ka'bā</persName>,
 cōntanos alguna cosa maravillosa.</p>
 <p xml:id="es5">Dīšo <persName key="Ka'b al-Aḥbār"><subst><gap
 reason="ilegible" unit="chars"
 quantity="5"/><add>Ka'bā</add></subst></persName>.</p>
 <p xml:id="es6">Plázeme, yā rey de los <subst><gap reason="ilegible" unit
 ="chars" quantity="3"/><add>cre</add></subst>yentes. As de saber que en los
 primeros del mundo ubo un rey de los romanos que vivía en-<placeName key="Al-
 Hind">el-Hindi</placeName><seg type="trebol"/>; adoraba a las ídolas, era de muy buena
 condiçión; <choice><sic>gbernaba</sic><corr>governaba</corr></choice> su reino
 <choice><sic>can</sic><corr>con</corr></choice> mucho amor i justiçiā. <pb n="2r"/> I
 nunca engendró criatura fasta que fue <choice><sic>di</sic><corr>de</corr></choice> çien
 años. I pensando el rey en cómo no ternía hichos, cayóle por-ello grande pienso. I-un día
 mandó llegar todos los sabioç de su reino. I llegados, dísoles el rey que mirasen en su saber i
 <choice><sic>senaçiā</sic><corr>sençia</corr></choice> si abía de tener</p>
 </body>
 </text>
 </TEI>