ⴰⵙⵉⵏⴰⴳ-*Asinag,* 9, 2014, p. 75-90

La réalisation de grands corpus linguistiques berbères normalisés interopérables : enjeux culturels et enjeux d'ingénierie linguistique

Noura Tigziri (1) & Henri Hudrisier (2)

(1) Université de Tizi Ouzou
(2) Laboratoire Paragraphe, Université de Paris 8

We describe our involvement in projects aimed at the production of French and Franco arabo berber digital resources : the BNFB (a project of the OIF FFI [1]) and HumanitéDigitMaghreb (a project of the CNRS ISCC).

In this paper, we focus particularly on the methods used in HumanitéDigitMaghreb (the TEI, specifically applied to the structuration of speech corpora and corpora of poetry and folk tales). The link with the ethnomusicological TEI markup is expected but will be considered later.

We will also examine the practical and future issues of very large corpora, linguistically annotated in accordance with a common standard and designed to constitute, for the linguistic community (for us, the Berber world), the context necessary to interact with the future tools translation and e-semantics

On this last point, for written or oral (audio signal or transcription) corpora, it is essential that the research community about Berber cooperate to promptly equip Berber languages of modern tools for digital processing.

Introduction

La mise en place de corpus numériques est devenue une exigence si on veut sauvegarder notre patrimoine culturel et identitaire mais la création de corpus oraux, par exemple, avec leur transcription, leurs traductions, leur annotation nécessite des méthodes modernes qui puissent faciliter leur exploitation et leur accessibilité. Aussi dans notre article, présentons- nous l'une de ces méthodes, la TEI (Text Encoding Initiative) appliquée à la structuration d'un corpus en kabyle.

En lien avec d'autres collègues, nous sommes impliqués à des niveaux différents dans des projets de bibliothèques numériques, de production d'e-learning, de normalisation des TIC, de recueil linguistique et d'organisation de ressources de documents dans la dynamique des Humanités digitales. Sans pour autant en tirer

prétention, nous considérons que la sélection et le financement conséquents de ces actions de recherches dans des appels d'offres proposés par des instances comme l'OIF ou le CNRS, confortent la pertinence de nos choix méthodologiques. La large assiette des partenariats rassemblés [1] nous permet aussi de disposer de compétences interdisciplinaires (recherche littéraire, ethnologie, linguistique, bibliothéconomie, musicologie, ingénierie linguistique du document et des réseaux, expertise en normalisation) mais aussi de diversité géographique et multilingue. Il nous semble, en effet, primordial que cette diversité des points de vue, des langues, des modes d'expressions, des médias, des genres (écrit, oral, image, musique, théâtre, contes, poésie, etc.) soit pris en compte dans une collégialité numérique véritablement communicante.

Telles sont, en effet, les ambitions primordiales des travaux dans lesquels nous sommes engagés qui nécessitent cette large palette de disciplines, de langues, de métiers et de diversité internationale et institutionnelle :

- Faire communiquer des langues entre elles (la famille des langues berbères, les langues du Maghreb mais aussi à termes les langues mortes qui fondent son patrimoine mais encore rendre accessible les ressources que nous rassemblons de façon mondiale)
- Construire en synergie des corpus de documents numériques en prenant la précaution de négocier leur intercompatibilité normative de façon cohérente et concertée pour que quantité d'utilisateurs [2] (mais aussi de créateurs) de ressources puisse les réutiliser selon des diversités de facettes d'approche. Nous pensons, en effet, que rassembler des ressources numériques doit obligatoirement se faire en ayant le souci de déployer un maximum de scientificité mais en ayant le souci constant que ces travaux soient utilisables par d'autres disciplines, mais aussi puissent participer de prospérité numérique des communautés concernées par ces patrimoines[3].

Si nous affichons ces ambitions c'est parce que nous savons qu'à l'égal de la mutation de la « Galaxie Gutenberg[4] », la mutation de la « Galaxie Digitale » nous impose de prendre en compte les recompositions de collégialité interdisciplinaire et bien sûr la globalisation internationale et interlinguistique.

La Galaxie Gutenberg avait refondé les sciences, l'industrie et l'économie. La Galaxie Digitale nous impose elle aussi de revoir fondamentalement nos méthodes

[1] AUF, ISO, Télécom Paris Sud, Alliance Cartago, EHESS, Université de Paris 8, de Bordeaux 3, de Paris 10, d'Evry, de Tunis, de Tizi–Ouzou, d'Oujda, d'Agadir, de Niamey, Conservatoire de Rimouski-Quebec.

[2] Non obligatoirement prévu à l'origine des projets.

[3] Si le monde académique a d'année en année besoin de plus d'outils, d'équipement, de missions internationales et s'il ne peut raisonnablement augmenter « en pourcentage » sa part du budget national cela impose obligatoirement à prêter attention et à s'inscrire dans des synergies interdisciplinaires, des synergies internationales, des coopérations sciences–industrie et à être attentif aux retombées d'usage pour la prospérité de la société civile.

[4] Cf M. Mac Luhan. Sa thèse impliquait des hypothèses de mutations similaires que l'auteur pointait avec les mass médias des années 60.

selon de multiples impératifs, notamment : interdisciplinarité, approche multimédia, multilinguisme, synergies sciences-industrie, pluralité des usages, mondialisation numérique des ressources (cloud computing), e-sémantique.

Nous ne développons pas dans cet article la facette Bibliothèque numérique de nos travaux. Cette facette sera exposée dans d'autres publications et a été d'ailleurs soumise au comité du TICAM 2012. Il est cependant indispensable de signaler que, bien sûr, la réalisation de grands corpus de documents impose de s'inscrire au minimum dans les recommandations et les bonnes pratiques proposées par l'OCLC[5] et notamment s'appuyer sur le Dublin Core[6] partagé par la plupart des bibliothèques numériques dans le monde. L'avantage majeur de ce respect des recommandations de l'OCLC et du DC étant que, si on en donne l'autorisation, toutes les bibliothèques compatibles dans le monde peuvent venir « moissonner » nos ressources berbères, et que réciproquement nous pouvons enrichir nos propres corpus en venant moissonner nous-mêmes automatiquement toutes les bibliothèques numériques du monde grâce aux mots-clefs ou aux termes « tagués » qui représentent les problématiques qui traversent nos corpus.

Baliser des documents sous plusieurs facettes

Les tâches spécifiques décrites par les auteurs dans ce papier se focalisent plus spécifiquement sur la TEI et son importance grandissante pour rendre disponibles, interopérables, réutilisables et normalisées des ressources linguistiques qui peuvent être indifféremment des corpus oraux, des chansons, ou de la littérature. L'avantage de la TEI pour des ressources numériques, c'est qu'elle autorise des traitements par balisages successifs, facette par facette, et permet ensuite leur alignement multifacette, multisupport, multidisciplinaire et multilingue. Concrètement, une ressource sonore chantée et parlée kabyle pourra être analysée linguistiquement et transcrite, elle pourra être liée et alignée avec sa transcription, son analyse ethnomusicologique, puis ses transcriptions (par ex. en d'autres langues berbères et en fr., ar., es., en. …). Le texte lui-même pourra être l'objet d'un balisage correspondant à des analyses littéraire et poétique, elles aussi, alignées avec les autres facettes.

[5] Le Online Computer Library Center (OCLC), fondé en 1967, nommé à l'origine *Ohio College Library Center*, est une organisation à but non lucratif mondiale au service des bibliothèques dont le but est d'offrir un meilleur accès public aux informations et d'en réduire le coût. Plus de 60 000 bibliothèques dans le monde utilisent les services de l'OCLC afin de trouver, de cataloguer ou de conserver leurs ouvrages. Les bureaux de l'organisation sont situés à Dublin, Ohio (USA).

[6] Le Dublin Core est un schéma de métadonnées générique qui permet de décrire des ressources numériques ou physiques et d'établir des relations avec d'autres ressources. Il comprend officiellement 15 éléments de description formels (titre, créateur, éditeur), intellectuels (sujet, description, langue, …) et relatifs à la propriété intellectuelle. Le Dublin Core fait l'objet de la norme internationale ISO 15836, disponible en anglais et en français depuis 2003. (6 pages, c'est donc une norme extrêmement concise et facile à s'approprier). Il est employé par l'Organisation mondiale de la santé (OMS), ainsi que dans de très nombreuses institutions, états et entreprises. Le Dublin Core a un statut officiel au sein du W3C et bien sûr de l'ISO.

C'est en cela que les exigences de multidisciplinarité, de multilinguisme, de synergie sciences-industrie (notamment industrie de langue) ne sont pas des vains mots. Le cœur de la mutation cognitive de la Galaxie numérique se situe là. Dans une opérabilité synergique numérique d'analyses scientifiques en sciences humaines qui s'additionnent, se recomposent, s'interfécondent de façon croisée. C'est dans ce but que nous nous inscrivons dans le projet HumanitéDigitMaghreb qui a précisément pour objet de pousser plus loin les ambitions du projet BNFB et d'aider les participants francophones, arabophones et berbérophones du projet à s'approprier des méthodes de la TEI. Le but final sera, donc, que nous ne disposions pas seulement de bibliothèques numériques uniquement référentielles (ce qui est déjà bien !), mais que nous mettions en œuvre graduellement un « balisage savant » des ressources (linguistique, littéraire, musicologique) qui donnera une véritable valeur ajoutée notamment aux ressources berbères.

Le courant des Humanités digitales et la TEI

Pour nous, le travail de formalisation des documents linguistiques s'opère sur deux versants complémentaires :

1. Celui de la numérisation des documents (Dublin Core) pour qu'ils puissent devenir disponibles, de façon normalisée et interopérable sur une plate-forme partagée en commun par les participants des projets (la plate-forme OMEKA[7]), mais qu'ils puissent aussi être moissonnés partout dans le monde sur des plates-formes répondant aux spécifications de l'OCLC et qu'à l'inverse, les participants des projets berbères et arabo-berbères précités puissent « moissonner eux aussi des documents dans toutes les bibliothèques numériques ».
2. Celui d'un balisage interne des documents déjà numérisés et référencés pour ce qui est de leur structure formelle, de leur morphologie, de leur signification, de l'ajout de gloses ou de notes, d'hypothèses explicatives ou encore du balisage de leur alignement avec des fichiers associés

[7] Omeka est un logiciel flexible et open source, conçu pour la publication sur le web de collections de documents numériques provenant de bibliothèques, de musées ou d'archives. Le logiciel est développé par le "Roy Rosenzweig Center for History and New Media". L'interface standard permet de parcourir la liste des documents (ou items), d'afficher les fichiers associés à chaque document, de filtrer par mot-clé, de parcourir les collections. Une recherche simple et avancée complète les possibilités de navigation. Des extensions permettent l'ajout de fonctionnalités, facilitant par exemple la création d'expositions électronique. L'administration du site, intuitive et fonctionnelle permet la gestion des collections, des documents et des fichiers associés à chaque document. Le type des documents peut être précisé : texte, image, son, vidéo, cours, histoire orale, email, site web, lien hypertexte, évènement ou personne. Les documents ont le statut public ou privé et peuvent être mis en avant sur la page d'accueil du site. Les informations descriptives de chaque document sont renseignées au format Dublin Core. Des métadonnées supplémentaires peuvent être ajoutées, dépendant du type du document, notamment la TEI. Les fichiers associés peuvent être du type texte (TXT, DOC, PDF, XML, JPG, TIFF), image (GIF, JPEG, PNG, TIFF), son (AIFF, MIDI, MP3, OGG, QT, RA, WAV) ou vidéo (AVI, MPEG, MP4, QT, SWF, WMV).

(transcription, traduction, interprétation ou autre versus médiatique comme des fichiers sonores ou vidéos associés à des textes, des partitions musicales, des photographies, des cartes, des réseaux ou des schémas).

C'est donc sur le point 2 que nous insistons dans cet article.

Le problème posé : la formalisation numérique normalisée des travaux en SHS

Les SHS travaillent globalement sur une matière plus « floue » que les sciences expérimentales et bien sûr que les sciences exactes : leur matériel principal est le document (souvent linguistique), leur outil d'analyse est le plus souvent l'argumentation textuelle et leurs résultats sont globalement des textes.

Evidemment certaines sciences humaines (notamment la linguistique) pratiquent depuis longtemps la formalisation d'un grand nombre de leur description (qu'elles soient morphologiques, syntaxiques, argumentatives, etc.). Cette pratique de formalisation a grandement facilité leur collaboration avec les informaticiens et explique pour partie les progrès en ingénierie linguistique. D'autres sciences humaines, les études littéraires par exemple ont été longtemps et sont aujourd'hui encore globalement rétives à la formalisation de leurs analyses. La recherche littéraire est de ce fait une science qui travaille sur le langage naturel, analyse, pose des hypothèses, les formalise sous forme d'énoncés en langage naturel et communique ses résultats sous la forme quasi exclusive de textes argumentés en langage naturel.

Cependant, notre objectif n'est pas de distribuer des bonnes ou mauvaises notes à telle ou telle catégorie de chercheurs en SHS mais de comprendre la mutation de méthode et d'habitus des chercheurs. De fait, la question qui devient récurrente est celle de la normalisation des pratiques face à une relative prolifération des outils d'aide informatique dans certains segments du travail linguistique.

Si nous nous appuyons sur une typologie grossière des travaux scientifiques et industriels sur le langage, nous pouvons distinguer :

→ Le travail sur les corpus oraux pour lequel il existe une relative prolifération des outils d'aide à la transcription, mais sur lequel il est urgent de s'entendre sur des standards.

→ Le travail d'analyse littéraire qui n'a longtemps connu que des outils très rustiques et limités comme les analyses statistiques de vocabulaire.

→ Les travaux terminologiques et lexicographiques dont les principes et méthodes ont été normalisés très tôt grâce notamment à Eugen Wüster qui a perçu très vite l'obligatoire nécessité de normaliser les pratiques en fondant dès 1937 ce qui allait devenir le Comité Technique 37 de l'ISO (ISO TC37).

→ Les travaux sur l'informatisation de l'écriture : la question est largement connue à l'IRCAM. Notons cependant que le scénario historique de ce qui s'est passé entre les années 1960 et aujourd'hui est une excellente leçon d'évolution technologique et de la longue durée d'appropriation technique, de l'impérieuse nécessité de

s'inscrire dans la normalisation et de la nécessité de comprendre le lien entre les progrès de l'environnement technique[8]. Pour des raisons historiques, l'écriture latine non accentuée a été dès le début prise en compte et normalisée. On connaît ensuite les normes successives et notamment la famille ISO 8859 qui prenait en compte les grandes écritures alphabétiques (latine, cyrilliques, arabe, grec, hébraïque…) mais qui ne pouvait coder ni les écritures idéographiques, ni les écritures sans intérêt industriel évident comme le tifinagh ou les écritures archéologiques (cunéiformes, hiéroglyphes). Sur ces derniers segments, on a bien sûr assisté à une relative prolifération de standards propriétaires « bricolés » par des laboratoires ou de petites sociétés informatiques. C'est ensuite grâce aux efforts des équipes des « chercheurs de terrain » (notamment à l'IRCAM) que la normalisation de ces technologies de transition a pu se faire.

Nous avons insisté sur cette question de la numérisation des écritures (qui pose désormais peu de problèmes) parce qu'elle est emblématique de l'obligatoire normalisation pour passer du foisonnement antiproductif des « standards propriétaires » comme c'était le cas avant Unicode et comme c'est encore le cas pour la transcription des corpus oraux.

En effet, comme le signale Thomas Schmidt[9], il existe aujourd'hui un choix relatif pour des outils d'aide à la transcription (de transcription informatique des corpus oraux[10]) et complémentairement une relative profusion de standards de formats pour coder de façon inutilement distincte les objets, les évènements, le contexte et les textes résultant de la transcription de ces corpus oraux. Certes, ces outils et formats présentent des différences mineures dues au contexte de leur développement et de leur production.

Par exemple, CLAN/CHAT a été développé pour transcrire et coder des corpus oraux d'enfants dans la base CHILDES alors que EXMARaLDA Partitur-Editor a été développé dans un contexte d'étude du multilinguisme et de la dialectologie. Tous ces outils ont des fonctionnalités similaires qui leur permettent simultanément de disposer d'un « player son » visualisant « l'enveloppe des productions sonores » et de zones de capture textuelle présentée en lignes parallèles (voir ci-dessous figure 1, une capture d'écran d'EXMARaLDA).

Pour compliquer encore la situation, les grands corpus mondiaux de transcription orale, ont bien naturellement développé des « conventions propriétaires » de codage des résultats dans leurs bases (voir ci-dessous figure 2, un tableau récapitulatif selon Thomas Schmidt).

---

[8] Dans ce cas particulier la nécessité d'attendre les progrès d'une informatique à 8 bits puis des processeurs à 16, 32, 64 bits et plus qui permettent maintenant de travailler directement en Unicode ce qui était plus problématique quand les processeurs étaient à 8 bits.

[9] Thomas Schmidt, « A TEI-based Approach to Standardising Spoken Language Transcription », *Journal of the Text Encoding Initiative* [Online], Issue 1 | June 2011, Online since 08 June 2011, connection on 08 July 2012. URL : http://jtei.revues.org/142 ; DOI : 10.4000/jtei.142

[10] ANVIL, CLAN/CHAT, ELAN, EXMARaLDA Partitur-Editor, FOLKER, Praat, Transcriber

File formats and transcription conventions for different spoken language corpora

Corpus (Language) [URL]	**File format**	**Transcription convention**
SBCSAE (American English) [http://projects.ldc.upenn.edu/ SBCSAE/]	SBCSAE text format	DT1 (DuBois et al. 1993)
BNC spoken (British English) [http://www.natcorp.ox.ac.uk/]	BNC XML (TEI variant 1)	BNC Guidelines (Crowdy 1995)
CallFriend (American English) [http://talkbank.org/]	CHAT text format	CA-CHAT (MacWhinney 2000)
METU Spoken Turkish Corpus (Turkish) [http://std.metu.edu.tr/ en]	EXMARaLDA (XML format)	HIAT (Rehbein et al. 2004)
Corpus Gesproken Nederlands (CGN, Dutch) [http://lands.let.kun.nl/ cgn/ ehome.htm]	Praat text format	CGN conventions (Goedertier et al. 2000)
Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK, German) [http://agd.ids-mannheim.de/ html/ folk.shtml]	FOLKER (XML format)	cGAT (Selting et al. 2009)
Corpus de Langues Parlées en Interaction (CLAPI, French) [http://clapi.univ-lyon2.fr/]	CLAPI XML (TEI variant 2)	ICOR (Groupe Icor 2007)
Swedish Spoken Language Corpus (Swedish) [http://www.ling.gu.se/ projekt/ old_tal/ SLcorpus.html]	Göteborg text format	GTS (Nivre et al. 1999)

Pour ce qui est du traitement numérique des corpus oraux, nous sommes donc confrontés à une situation tout à fait similaire à celle de la codification des écritures avant leur normalisation convergente avec Unicode. Il existe une relative anarchie des outils d'aide à la capture et à la transcription ainsi que des formats de codages. Comme le fait remarquer Lou Burnard[11], « Le constat est récurrent : à la variété des formats utilisés se superpose l'incohérence des pratiques conventionnelles de transcription des données orales. En dépit de plus de vingt années de pratiques convergentes, les communautés intéressées préfèrent travailler avec leurs propres

[11] Lou Burnard : « Encoder l'oral en TEI : démarches, avantages, défis…. » Conférence à la Bibliothèque Nationale de France, prononcée le 10 mai 2012, Publié le 19/06/2012 par Abigaël Pesses.

outils et conventions "maison". Pourtant, l'intérêt de se servir d'un format commun, voire pivot, est un sujet qui a été abordé dans la littérature académique à de multiples reprises : Edwards & Lampert (1993), MacWhinney (2007), Schmidt (2011). Ne serait-il pas finalement temps d'établir un format d'échange normalisé pour les données orales? …[La TEI grâce à son format fédérateur TEI transcription of speech et aussi grâce à son alliance en consortium avec l'ISO TC37 est actuellement en situation de devenir la norme[12]] ».

Dans les projets auxquels nous faisons références (BNBF et HumanitéDigitMaghreb) les participants se réclament globalement de ces deux disciplines, souvent des deux ensembles, mais aussi de la bibliothéconomie, l'histoire, la pétrographie, l'ethnologie, l'ethnolinguistique, la musicologie.

En nous inscrivant dans l'école de pensée des Humanités digitales et de la TEI qui est sa norme et son outil technique, nous voulons explicitement donner non seulement une réalité tangible et numérique à nos travaux, mais aussi les rendre facilement échangeables, cumulables, améliorables, modifiables partout dans le monde. Nous voulons que nos travaux linguistiques participent « pré-industriellement de l'ingénierie linguistique. Nous voulons que nos travaux de recherche littéraire soient non seulement visibles dans le monde entier, mais encore qu'ils s'inscrivent dans la synergie mondiale des études littéraires computorisables. Nous voulons aussi sur un plan plus spécifiquement pan berbère que nos travaux soient déjà facilement échangeables et cumulable entre nous et avec nos trois langues partenaires maghrébines (arabe, français, espagnol auxquelles il convient de rajouter l'anglais). C'est la raison primordiale de notre implication dans le projet HumanitéDigitMaghreb.

Qu'est-ce que HumanitéDigitMaghreb ?

HumanitéDigitMaghreb est un projet du CNRS-ISCC[13]. C'est une recherche-action dans laquelle sont engagés des acteurs de terrain (linguistes, chercheurs en littérature, culture, histoire tant française qu'arabe ou berbère) soucieux d'inscrire leurs pratiques dans l'organisation rationnelle de corpus numériques répondant aux recommandations mondiales des réseaux de bibliothèques numériques (OCLC) et des Humanités digitales. Ces acteurs de terrain sont associés avec des praticiens de la coopération francophone numérique, des spécialistes de l'information et de la communication, des fondateurs de la TEI et des Humanités digitales en France, des spécialistes de l'appropriation des usages du numérique et plus spécifiquement des patrimoines numériques. Parallèlement à son étude d'appropriation, cette recherche s'appuiera sur des réalisations en cours de structuration de corpus patrimoniaux franco-arabo-berbères.

[12] Cette conclusion entre crochets est d'Henri Hudrisier mais correspond à l'action des leaders de la TEI effectivement liés aux actions de l'iso TC37, notamment Laurent Romary convener de l'ISO TC37-SC4.

[13] Institut des sciences de la communication du CNRS (Centre National de la Recherche Scientifique).

Les partenaires historiques de la TEI

La TEI (Text Encoding Initiative) a été fondée à la suite d'une conférence sponsorisée par l'ACH (Association for Computers and the Humanities)[14] et financée par le NEH (U.S. National Endowment for the Humanities)[15]. Cette conférence avait lieu au Vassar College (Poughkeepsie, N.Y. - USA) en novembre 1987. Environ trente représentants du monde des bibliothèques, des sociétés savantes et de projets de recherche intéressés par le codage des textes et la recherche littéraire ainsi que d'informaticiens spécialisés en SGML étaient invités à cette conférence pour discuter la faisabilité d'un codage standard et élaborer des recommandations. Pendant la conférence, l'ACL (Association for Computational Linguistics)[16] et l'ALLC (Association for Literary and Linguistic Computing)[17] ont décidé de rejoindre l'ACH comme les sponsors d'un projet pour développer les Directives de la TEI (TEI Guidelines). En 1988, ils ont été rejoints par la Commission de la Communauté Européenne, l'Andrew W. Mellon Foundation[18] et le Social Science and Humanities Research Council of Canada[19].

[14] L'ACH (Association for Computers and the Humanities) a été fondée en 1978, une époque où la relation entre informatique et humanités, était encore très confidentielle. La plupart des grands universitaires du domaine jugeaient même qu'il s'agissait d'une alliance contre nature. En une trentaine d'année, le paysage a bien changé. L'ACH a mis en place un forum pour la recherche, des discussions et les explorations techniques qui ont alimenté cette transformation. L'ACH est devenue une association beaucoup plus vaste. Elle patronnait chaque année ; la conférence d'Humanités Numériques annuelle (maintenant patronnée par ADHO.

[15] Le NEH (U.S. National Endowment for the Humanities) est une agence fédérale américaine indépendante fondée en 1965 par le Président Lyndon Johnson. C'est le plus important organisme de financement dans le secteur des Humanités aux USA. Il intervient pour financer l'excellence culturelle, muséale, académique mais aussi la radio et la télévision, voire des bourses de recherches individuelles.

[16] L'ACL (Association for Computational Linguistics) est l'Association de référence mondiale pour les professionnels et les scientifiques travaillant sur les questions liant langages naturels et traitement informatique. L'ACL édite Computational Linguistics et organise des conférences annuelles (la 51ème conférence est prévue en 2013 à Sofia).

[17] L'ALLC (Association for Literary and Linguistic Computing) a été fondée en 1973 dans le but de favoriser des applications d'informatisation de l'étude du langage et de la littérature. L'ALLC s'intéresse à l'analyse des textes, aux corpus textuels, à l'histoire, l'histoire de l'Art, la musique, l'étude des manuscrits et à l'édition électronique.

[18] La Fondation Andrew W. Mellon de New-City et Princeton est une fondation privée, dotée de richesses accumulées par Andrew W. Mellon de la famille Mellon (Pittsburg, Pennsylvanie). C'est une fondation prestigieuse qui intervient dans l'enseignement supérieur, les bibliothèques et la communication savante, les musées et la conservation de l'Art, les arts de la scène, et les TIC. Plus précisément le développement de logiciels interréssant ses principaux champs d'intérêts ci-dessus.

[19] Le Social Science and Humanities Research Council of Canada en français Conseil de recherche en sciences humaines du Canada (SSHRCC- CRSHC) est un organisme du gouvernement fédéral canadien ayant pour mission d'appuyer la recherche et la formation avancée en milieu universitaire dans le secteur des sciences humaines.

On voit bien que ces associations fondatrices ne sont pas d'obscurs partenaires, ces diverses institutions opéraient une importante jonction synergique en fondant la TEI. Certes, la TEI a contribué à ce qu'un vaste public savant s'approprie des « standards bonnes pratiques » en matière de traitement et de communication pour l'étude savante des textes. Parallèlement, les institutions fondatrices n'oubliaient pas leurs objectifs fondateurs réciproques éminemment complémentaires : Humanités computationelles ; recherche littéraire par ordinateur ; linguistique computationelle ; recherche littéraire computationelle ; développement de logiciel pour le traitement de corpus culturels numériques et bibliothèques.

C'est d'ailleurs dans la perspective de ces objectifs que l'ALLC, en coopération avec l'ACH et la SDH-SEMI[20] ont préfiguré (dès 2002) puis fondé en 2005 l'ADHO[21].

On voit bien ainsi la synergie qui peut exister entre la TEI qui définit des standards et des bonnes pratiques et les Humanités digitales qui permettent que se socialisent ses usagers, qu'ils adaptent les outils (notamment ceux des bibliothèques numériques) à des besoins spécifiques, qu'ils échangent des méthodes, des modèles de structuration et de balisage de leurs corpus (en fait des TEI-DTD adaptées aux besoins de leurs corpus et de leurs pratiques d'analyse savantes et d'échanges).

Une synergie TEI, Humanités digitales et bibliothèques numériques

Toutes les universités héritières de ces premières universités européennes travaillant en latin utilisent le terme « Humanitas » pour désigner les disciplines de sciences humaines et sociales, ainsi que la recherche en Art et littéraire. Pour des raisons historiques, les institutions académiques anglophones gardent toujours vivant la désignation « Humanités » qui constitue toujours une sorte de métadiscipline recouvrant pratiquement ce que les francophones nommeraient « Arts et Lettres » parce qu'en français, l'expression «, ' les Humanités » est devenue un peu désuète. Quels que soient les termes, le monde anglo-saxon puis l'Europe du Nord et, avec un certain retard, la Francophonie s'emparent maintenant de l'expression Humanités digitales, ce qui redonne du sens à l'ancienne expression « les Humanités ».

En fait, on pourrait dire que de la rencontre de ces institutions et de leur convergence synergique sont nés deux axes de dynamique d'action fonctionnellement complémentaires qui rentrent en résonance avec une réalité de l'environnement technologique : les bibliothèques numériques. L'ADHO a adopté comme publication principale, le journal officiel de l'ALLC « Journal of Digital Scholarship in the Humanities » publié par les Oxford University Press. Deux

[20] Society for Digital Humanities-Société d'étude des médias interactifs (CAN).

[21] L'ADHO (Alliance of Digital Humanities Organizations) est donc une alliance internationale qui a pour objectif de soutenir les applications informatiques pour l'étude du langage et de la littérature : en fait les Humanités digitales. Elle le fait en soutenant des publications, des ateliers spécialisés (classes d'été), à travers aussi des groupes de travail thématiques répondant notamment à des disciplines et des sous-diciplines.

autres publications ont une portée mondiale en la matière : « DHQ, Digital Humanities Quartely » et « Digital Studies / Le champ numérique[22] tous deux publiés sous la responsabilité de l'ADHO.

Citons encore pour mémoire afin de survoler la problématique des Humanités digitales : *Humanist:* Un séminaire électronique sur les applications de l'informatique aux Humanités http://www.allc.org/publications/humanist

Mind Map of the Digital Humanities: Une cartographie conceptuelle de l'univers des Humanités digitales disponibles sur http://www.allc.org/publications/mind-map-digital-humanities

Cela permet d'avoir une vision synoptique et facile d'accès mise à jour en permanence par l'ensemble des communautés TEI, des publications, des outils disponibles. Notre ambition serait que TEI berbère y figure bientôt.

Dans cette partie, nous appliquons le codage TEI sur un corpus kabyle. Nous avons choisi de travailler sur un écrit parce que l'oral présuppose un certain nombre de décisions à prendre en ce qui concerne la définition de certains concepts tels par exemple la phrase, l'énoncé, le paragraphe... C'est pour cela que pour cette première application, nous avons jugé plus pratique de travailler sur un corpus écrit. Il s'agit de la traduction en kabyle de Kamal Bouamara de "Jours de Kabylie" de Mouloud Feraoun. L'oeuvre contient un certain nombre de parties. Nous avons travaillé sur deux parties pour montrer comment se fait le codage en TEI.

La première partie est "Taddart-iw" (mon village), la deuxième est "Tajmaɛt n At Flan" (la djemaa de Flan (un tel)).

Tout codage en TEI commence par la définition des éléments à mettre dans le <TeiHeader>. Pour notre part, nous avons le <TeiHeader>, comme ceci :

[22] *Digital Studies / Le champ numérique* (ISSN 1918-3666) est une publication universitaire spécialisée paraissant trois fois par an, destinée aux chercheurs dans le domaine des sciences sociales numériques et ayant pour objectif de leur offrir une ressource de niveau universitaire et de fournir un cadre formel à leurs activités de recherche. DS/CN est publiée par la Society for Digital Humanities / Société pour l'étude des médias interactifs (SDH/SEMI), un organisme affilié à l'Association for Computers and the Humanities (ACH) et à l'Association for Literary and Linguistic Computing (ALLC), via l'Alliance of Digital Humanities Organisations (ADHO).

```
<TeiHeader>
        <fileDesk>
                <
                <publicationStmt>
                        <publisher> ENAG </publisher>
                        <pubPlace> Alger </pubPlace>
                        <date>1998 </date>
                </publicationSmt>
        </fileDesk>
</teiHeader>
```

Avec deux attributs dans le <fileDesk>, le titre <titleStmt> qui précise tout ce qui est relatif au titre avec l'intitulé de l'ouvrage, l'auteur et on a ajouté la balise <editor> pour spécifier que c'est une traduction et la balise <relatedItemtype="translatedFrom"> pour donner la traduction.

```
        <titleStmt>
                        <title> Ussan di Tmurt </title>
                        <author> Mouloud Feraoun </author>
                        <editor role="translator"> Kamal Bouamara </editor>
                        <relatedItemtype="translatedFrom">
                          <bibl>
                                <author>Mouloud Feraoun</author>
                                 <title>Jours de Kabylie</title>.
                                   <date>1954</date>
                        </bibl>
                         </relatedItem
                </tileStmat>
```

et les données concernant la publication de cet ouvrage comme, l'éditeur la date, le lieu de publication.

```
                <publicationStmt>
                        <publisher> ENAG </publisher>
                        <pubPlace> Alger </pubPlace>
                        <date>1998 </date>
                </publicationSmt>
```

Une fois ces données introductives définies, nous passons au codage du texte lui-même.

```
<text>
        <body>
                <div n=1>
                        <head> taddart-iw </head>
```

Avec le corps du texte <body> comprenant un attribut qui lui-même est subdivisé en <head> (entête) et <p> (paragraphe).

Dans cette dernière balise nous définissons une balise <S> (phrase). Nous définissons la phrase au sens large du mot. Un segment compris entre deux points. Comme nous avons à prendre par moment des décisions par rapport à la définition du « mot » et du « mot composé » qui a deux parties reliées par un trait d'union, phénomène assez courant en kabyle, nous avons opté pour considérer que le mot composé de n éléments est un seul mot avec comme la définition du mot « tout élément compris entre deux blancs », dans ce cas, le codage en TEI se fait de cette manière

Exemple :

d win d-yettalsen (mot composé : d-yettalsen)

<w>d</w>

<w>win</w>

<w>d</w>

<hyphen>-</hyphen>

<w>yettalsen</w>

Si on considère le mot composé de n parties comme étant un seul mot, dans ce cas, le codage se fait ainsi :

Exemple : d win d-yettalsen (mot composé : d-yettalsen)

<w>d</w>

<w>win</w>

<w>d<hyphen>-</hyphen>yettalsen</w>

Ainsi le codage TEI du paragraphe et de la phrase est donné par ce qui suit :

```
<p>
    <s>
        <w>Ur</w>
        <w>lliɣ</w>
        <w>ara</w>
        <w>seg</w>
        <w>wid</w>
        <hyphen>-</hyphen>
        <w>nni</w>
        <w>yettɣuddun</w>
        <w>taddart</w>
        <hyphen>-</hyphen>
        <w>nsen</w>
        <pc>.</pc>
```

```
</s>
<s>
        <w>Mi</w>
        <w>ur</w>
        <w>fneṭzeɣ</w>
        <w>ara</w>
        <w>s</w>
        <w>waya</w>
        <w>nezzeh</w>
        <pc>,</pc>
        <w>ẓriɣ</w>
        w>acuɣer</w>
        <pc>.</pc>
</s>
```

Le deuxième corpus est un corpus oral. Il s'agit d'un meeting tenu par le président du parti du Rassemblement pour la culture et la démocratie, le docteur Saïd Saadi, lors des élections législatives de 2002.

Codage du TEI HEADER

```
<teiHeader>
   <fileDesc>
    <titleStmt>
       <title>Meeting politique</title>
          <author> Said Sadi="Président du  RCD"  </author>
    </titleStmt>
    <publicationStmt>
       <pubPlace>type="Tizi Ouzou">Stade Oukil Ramdane</pubPlace>
          <date> 2002.05.02</date>
    </publicationStmt>
    <sourceDesc>
       <recordingStmt>
        <recording type="audio" dur="P30M">
        <equipment>
           <p> audio tape, réalisé par B. S.</p>
         </equipment>
        </recording>
       </recordingStmt>
      </sourceDesc>
    </fileDesc>
 </teiHeader>
```

Codage du corps du texte

```
<text>
   <body>
     <incident>
       <desc>Applaudissement</desc>
     </incident>
     <u>who="# Saïd  Sadi ">
       <seg>Azul</seg>
     </u>
     <incident>
       <desc>Applaudissement</desc>
     </incident>
     <u>who="# Saïd  Saadi ">
       <seg>Azul <pause dur="PT10S"/> Azul d ameqran</seg>
     </u>
     <incident>
       <desc>Applaudissement</desc>
     </incident>
     <u>who="# Saïd  Saadi ">
       <seg>Tsellem-d deffir kunwi i  igubrentili</seg>
     </u>
     <incident>
       <desc>Applaudissement</desc>
     </incident>
     <u>who="# Saïd  Sadi ">
       <seg>
         Azul<pause dur="PT20S"></pause>
       </seg>
       <seg>Yidwen am yiḍelli am assa am uzekka wer ttagadut<pause
dur="PT20S"/></seg>
     </u>
     <incident>
       <desc>App</desc>
     </incident>
     <u>who="# Saïd  Saadi ">
       <seg>Qqaren- as imezwura- nneɣ<pause dur="PT10S"/> isers uḥeddad
tafḍist<pause dur="PT10S"/>irfed-itt mmi-s<pause dur="PT20S"/></seg>
     </u>
     <incident>
       <desc>Applaudissement</desc>
     </incident>
     <u>who="# Saïd  Sadi ">
       <seg>D ayen igellan di tiɣri n nouvembre rebɛa uxemsin<pause
dur="PT10S"/>Dayen igellan di la plate forme n la soumam<pause dur="PT05S"/>
ayen i gellan<pause dur="PT05S"/>deg dusyi –nni i d nexdem deg Σekkuren deg
seggasen n tmanyin<pause dur="PT05S"/> i gellan di la plate forme Lleqsar<pause
dur="PT20S"/> </seg>
```

```
    </u>
  </body>
</text>
```

Conclusion

La TEI a l'avantage de rendre disponibles, interopérables, réutilisables et normalisées des ressources linguistiques. Il est vrai que le travail de codage est fastidieux surtout quand il s'agit de travailler sur des corpus de grandes tailles mais vu les avantages que présente cette méthode, nous avons tout intérêt à l'exploiter et nous l'approprier à cause de l'un de tous ses avantages dont l'interopérabilité.

Quel que soit le corpus choisi, la disponibilité des balises de codage rend la tâche de balisage plus facile à appréhender. En effet, les membres du consortium qui ont établi la TEI ont prévu absolument toutes les balises utiles dans le codage de ressources linguistiques quelles qu'elles soient : corpus écrit, corpus oral et de quelle que soit la discipline : linguistique, littérature, musique

Bibliographie

Ben Henda M., (2012), *Vision historique, technique et prospective des systèmes d'information et de communication interopérabilité normative globalisée*. Mémoire de HDR sous la dir. De Roland Ducasse, Université Bordeaux III.

Ben Henda M. et Hudrissier H., (2000), Normalisation et terminologies multilingues pour les TICE in *Forum Terminologique International*, Université de Sousse 20 au 23 novembre 2000.

Gille B., (1978), *Histoire des techniques*, Encyclopédie de la Pléiade, Gallimard, Paris.

Hudrissier H. et Romary L., (2003), Le balisage normalisé des concepts et documents en liaison avec les normes de l'CAD. In *Normes et standards pour l'apprentissage en ligne*, Versailles, 18 mars 2003. http://www.initiatives.refer.org/inititaives, 2012.

Hudrissier. H. (2009), La nécessité d'adapter internet à la mondialisation linguistique, *in Critique de la société de l'information* (coordonné par J.P. Lafrance). Les Essentiels d'Hermès, CNRS éditions, Paris, p. 115-134.