Méthodologie pour la structuration semi-automatique d'un corpus lexicographique bilingue : le cas du dictionnaire français-kabyle.

Mahfoud MAHTOUT
Laboratoire DySoLa (Dynamiques Sociales et Langagières), Université de Rouen
mahfoud.mahtout@yahoo.fr

Résumé

L'objectif de cette contribution est de proposer une méthodologie de structuration de corpus à l'aide d'outils informatiques récents permettant aux linguistes non-spécialistes en informatique de constituer des bases de données lexicales permettant de tirer parti de toutes les informations contenues dans un corpus lexicographique. Il s'agit, plus exactement, de présenter le processus d'informatisation du *Dictionnaire français-kabyle* (1902-1903) de Huyghe et ce depuis sa numérisation, en passant par sa structuration, à la constitution d'une base de données lexicales interrogeables en ligne[1]. Cette méthodologie a le mérite d'être plus économe en temps de travail qualifié que la solution qui consiste à tout structurer manuellement. En outre, cette approche à l'avantage de donner des résultats probants en termes de structuration d'un corpus numérique bilingue facile à enrichir et à partager.

Mots-clés : dictionnaires bilingues, méthodologie, structuration, semi-automatique, informatisation, corpus, TAL, langues africaines.

1. Introduction

À l'ère des technologies numériques, l'informatisation de ressources lexicographiques anciennes constitue une alternative pour la sauvegarde, la valorisation et à terme l'équipement linguistique des langues à tradition orale. D'ailleurs, les langues africaines accusent un retard important dans ce domaine par rapport aux langues de l'Europe pour lesquelles des travaux d'informatisation de dictionnaires anciens ont été initiés depuis les années 1980. Bien que le patrimoine lexicographique des langues africaines soit particulièrement riche, son exploitation demeure très limitée faute notamment de sa disponibilité. En effet, la rareté et la fragilité des dictionnaires anciens réduisent l'accessibilité à un large public et menacent la pérennité des textes. Aussi, les outils informatiques actuels offrent des possibilités de conservation et de diffusion de ressources lexicales inestimables et permettent leur exploitation et leur valorisation.

Nous proposons, dans cet article, de présenter une méthodologie mise en œuvre pour l'informatisation du *Dictionnaire français-kabyle* (1902-1903) de Gustave Huyghe. Ce projet découle d'un constat simple mais important : il n'existe aucune tentative d'informatisation de dictionnaires bilingues anciens notamment ceux de la période coloniale en Algérie ; de plus, dans le domaine de la lexicographie français-langues d'Algérie, rares sont les sites internet qui proposent une ressource lexicale exploitable en ligne. La méthodologie que nous proposons apporte une réponse adaptée à ce type d'ouvrages à caractère non systématique dont le contenu est assez disparate et souvent peu structuré. De ce fait, la structuration du corpus ne peut être envisagée que de façon semi-automatique. Il convient pendant cette phase de se poser les questions suivantes : quelles sont les rubriques récurrentes ? Celles-ci suivent-elles une organisation constante ? Quelles sont les informations non systématiques ? Comment sont-elles disposées dans l'article ? Quelles sont les solutions techniques les plus innovantes et les moins coûteuses en temps à envisager ?

Avant tout, nous devons souligner que l'informatisation du *Dictionnaire français-kabyle* (1902-1903) a fait l'objet d'un partenariat avec le Département de Génie mathématique de l'Institut National des Sciences appliquées (INSA) de Rouen. Cette collaboration a donné lieu à l'informatisation d'un échantillon du dictionnaire en question. Notre réflexion a évolué au cours de la recherche et a été élargie à l'ensemble du corpus lexicographique.

Nous nous proposons dans un premier temps d'exposer les différentes étapes que nous avons suivies pour informatiser le *Dictionnaire français-kabyle*, de sa numérisation à son informatisation. Nous présenterons ensuite l'outil Adobe FrameMaker qui permet de structurer d'une façon logique les données d'un corpus et de

[1] Cette base de données est en cours de réalisation.

les rendre disponibles sous forme d'une base de données accessible en ligne[2]. Nous conclurons en exposant les différents modes de consultation du dictionnaire et les possibilités d'exploitation qu'offre la version informatisée.

2. Le corpus lexicographique

Le corpus lexicographique est constitué du *Dictionnaire français-kabyle, Qamus Rumi-Qbaili,* publié en 1902-1903, à Malines, en Belgique, chez Godenne, par le missionnaire berbérisant Gustave Huyghe. Cet ouvrage de 893 pages contient une nomenclature étendue et très détaillée, riche de plus de 15 000 entrées organisées par ordre alphabétique latin. Sur le plan matériel et formel, les articles sont disposés sur une colonne par page et les entrées sont typographiées en minuscule et en caractères gras. Les entrées polysémiques sont le plus souvent suivies de tournures pour en préciser le sens. Les équivalents kabyles notés en italique-gras rendent les différentes acceptions de la vedette française. Les périphrases prennent souvent la place de l'équivalent pour exprimer l'idée du terme français. Le dictionnaire du Père Huyghe foisonne d'exemples, de tournures et d'expressions (plus de 4000) qui rendent au mieux les différentes acceptions des vedettes. Les informations grammaticales sont mentionnées pour les mots kabyles, dans le corps de l'article : les termes indiquant l'aspect d'habitude, le parfait et le pluriel sont souvent notés.

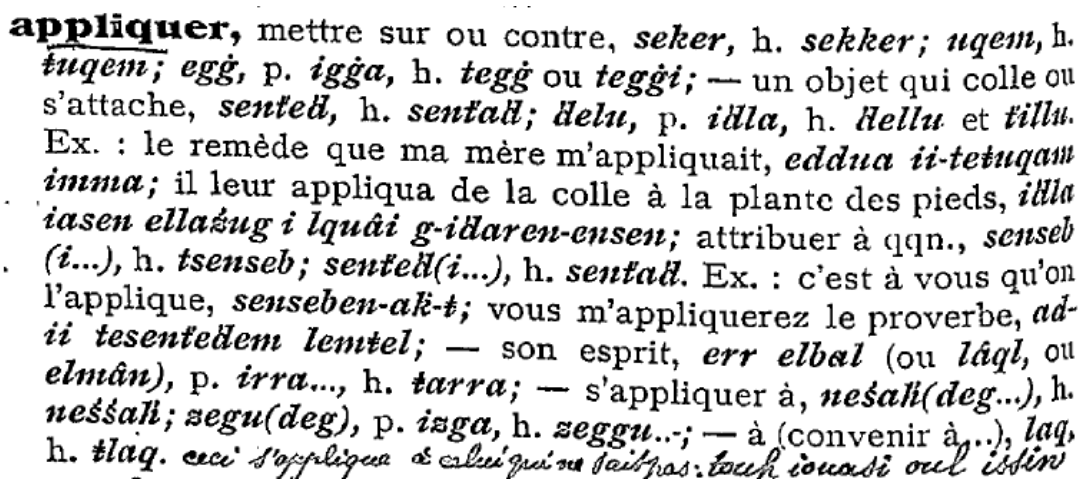
appliquer, mettre sur ou contre, ***seker,*** h. ***sekker; uqem,*** h. ***ṭuqem; egg,*** p. ***igga,*** h. ***tegg*** ou ***teggi;*** — un objet qui colle ou s'attache, ***senṭeḷḷ,*** h. ***senṭaḷḷ; ḍelu,*** p. ***iḍḍa,*** h. ***ḍellu*** et ***ṭillu.*** Ex. : le remède que ma mère m'appliquait, ***eḍḍua ii-tetuqam imma;*** il leur appliqua de la colle à la plante des pieds, ***iḍḍa iasen ellaẓug i lquâi g-iḍḍaren-ensen;*** attribuer à qqn., ***senseb (i...),*** h. ***tsenseb; senṭeḷḷ(i...),*** h. ***senṭaḷḷ.*** Ex. : c'est à vous qu'on l'applique, ***senseben-aḳ-ṭ;*** vous m'appliquerez le proverbe, ***aḍ-ii tesenṭeḍḍem lemṭel;*** — son esprit, ***err elbal*** (ou ***lâql,*** ou ***elmân),*** p. ***irra...,*** h. ***tarra;*** — s'appliquer à, ***neṡaḷḷ(deg...),*** h. ***neṡṡaḷḷ; zegu(deg),*** p. ***izga,*** h. ***zeggu...;*** — à (convenir à...), ***laq,*** h. ***ṭlaq.*** ceci s'applique à celui qui ne fait pas : touh iouasi ouel issin

Figure 1. Exemple d'organisation matérielle et formelle d'un article.

Cette figure illustre l'organisation microstructurelle d'un article type et permet de mettre en évidence le mode de transcription de la langue kabyle. Comme nous pouvons le constater, l'auteur a opté pour une transcription uniquement en caractères latins y compris pour le kabyle. L'auteur suit une règle simple représentant chaque phonème par une seule lettre. Pour ce faire, il recourt à des lettres conventionnelles suscrites, souscrites ou barrées pour transcrire les mots kabyles. Par exemple, Huyghe emploie le point pour indiquer les lettres faibles (neutres) et l'apostrophe pour désigner les lettres fortes (emphatiques). Bien qu'original, ce mode de transcription demeure d'une part, insuffisant pour rendre avec précision la prononciation kabyle et constitue, d'autre part, une première difficulté pour le traitement informatique du corpus (nous reviendrons sur ce point). Une seconde difficulté apparaît dans la figure 1 ; la présence de notes manuscrites dans le corps de l'article ne facilite pas la reconnaissance optique des caractères.

2.1 Contexte de rédaction et caractéristiques du *Dictionnaire français-kabyle*

Le *Dictionnaire français-kabyle* ne fut pas la première œuvre de Gustave Huyghe. En 1896, il compose le premier dictionnaire ayant le kabyle avant le français, ouvrage manuscrit, lithographié, puis imprimé en caractères typographiques en 1901 à Paris par l'Imprimerie nationale. Né en 1861 dans la commune de Morbecque, située dans le département du Nord, Gustave Huyghe est ordonné prêtre le 8 septembre 1884 et choisit de servir dans la Société des Missionnaires d'Afrique. Il sera envoyé, le 17 novembre de la même année, à la station Djamâa Saharidj, auprès de ses confrères de Kabylie. Très actif, le Père Huyghe fait la classe aux enfants, parcourt les villages de la Haute et de la Basse-Kabylie, soigne les malades et profite de ses visites pour consolider sa maîtrise de la langue kabyle. Et, tout ceci dans des conditions d'existence très difficile au milieu

[2] Le site fonctionne actuellement en "local". Il sera prochainement mis en ligne.

d'habitation rudimentaire : il faut souvent écrire à la lumière d'une lampe à pétrole, qui éclaire et chauffe la chambre. En octobre 1885, il est appelé au poste d'Ath-Menguellet, où il ne reste que quelques mois, puis affecté, en janvier 1886, à la station de Beni Smaïl. Ce sera le dernier poste dans lequel il exercera en Kabylie avant d'être appelé en Belgique (1887), puis envoyé en Tunisie (1897) où il reste plus de deux ans avant de retrouver l'Algérie en 1899 mais cette fois dans les Aurès, plus précisément à Arris, chez les Chaouis. Il confectionne alors le *Dictionnaire français-chaouia,* qu'il publie en 1906 à Alger chez Jourdan. Gustave Huyghe meurt le 01 décembre 1912, à l'âge de cinquante ans.

Le *Dictionnaire français-kabyle* réunit un matériau composé de plusieurs parlers kabyles recueilli principalement en Haute-Kabylie et dans certaines localités de la Basse-Kabylie où sont implantés des postes de mission. L'ouvrage du Père Huyghe contient une nomenclature étendue et très détaillée, riche de plus de 15000 entrées. Huyghe introduit dans sa nomenclature des mots référant, par exemple, à l'organisation sociopolitique de la société kabyle, *session* (*tajmâat*, « conseil des sages de la tribu »), *séminaire* (*timâmmert*, « établissement scolaire ») ; aux instruments d'agriculture ou de jardinage, *sarcloir*, *serpe*, *sécateur* ; à l'habillement, *savate*, *socque*, *saroual* ; etc. Son dictionnaire se caractérise par une accumulation de parlers kabyles qui varient suivant les tribus et les villages. L'ouvrage se termine par un appendice dans lequel le lexicographe nous donne une liste de mots tirés des expressions argotiques les plus répandues parmi les Kabyles. En 1904, ce dictionnaire est récompensé par le prix Volney, mais ne bénéficie d'aucune réédition.

3. Méthodologie

Il convient de souligner que le *Dictionnaire français-kabyle* de Huyghe, auquel un traitement informatique est appliqué, est conçu sous forme de base de données lexicales dont l'interface de consultation permet à l'utilisateur d'interroger et d'exploiter de façon personnalisée les données stockées. La méthodologie adoptée s'inspire des travaux menés dans le cadre de l'informatisation des dictionnaires français anciens (Leroy-Turcan et Wooldridge Russon : 1997 ; Dendien et Pierrel : 2003 ; Manuélian, Bruscand et al. : 2009, etc.) et ceux d'autres langues (De Tier et Van Keymeulen : 2010 ; Mazziotta et Renders : 2010 ; Enguehard et Mangeot : 2013).

3.1 De la numérisation à la récupération des données textuelles

La première étape a consisté à numériser la version papier du *Dictionnaire français-kabyle* (1902-1903). Pour ce faire, nous avons demandé sa numérisation à la bibliothèque universitaire de Grenoble (SICD 2)[3] qui propose au public un service gratuit de "numérisation à la demande". Une fois numérisé, le document est mis à notre disposition sous format PDF-image. Ensuite, nous avons procédé à son "océrisation", c'est-à-dire à la conversion du format PDF-image en format texte au moyen d'un logiciel OCR[4] qui permet la récupération des données textuelles. À l'issue de cette opération de conversion, nous avons procédé à la vérification et au contrôle du texte pour corriger les erreurs de reconnaissance[5]. En effet, le document source comporte des caractères accentués non pris en charge par le logiciel OCR et contient, à certains endroits, des annotations manuscrites qui encombrent le texte imprimé, ce qui rend le résultat de la reconnaissance de moindre qualité. Nous avons donc procédé à la révision des coquilles de toutes sortes en veillant particulièrement au respect du contenu linguistique du texte original et de ses caractéristiques typographiques. Une fois l'ensemble du texte lexicographique relu et corrigé, il était prêt pour la phase de la structuration des données.

3.2 Structuration des données

La deuxième étape consiste à analyser manuellement et minutieusement les différents types d'articles du dictionnaire afin de révéler leurs structures en caractérisant les différents éléments qui les constituent : spécifications typographiques (corps de l'entrée, corps de tous les autres éléments), environnement d'apparition de chaque élément, attributs, etc. Par exemple, les éléments suivants ont été annotés automatiquement : les entrées (gras, taille 14) les indications grammaticales (marques grammaticales), la traduction (italique), les marques d'usage (indications d'usage), les exemples (Ex :), les limites des articles, etc. Dans un second temps,

---

[3] Service interétablissements de coopération documentaire de Grenoble

[4] Optical Character Recognition (Reconnaissance optique des caractères). Le logiciel de reconnaissance utilisé est OmniPage 17. Ce logiciel possède une fonction d'apprentissage permettant de modifier les solutions d'OCR attribuées aux caractères mal interprétés. Cette fonction est utile pour notre document qui présente des caractères inhabituels.

[5] La correction a été effectuée par un groupe de bénévoles.

un balisage en norme XML (eXtensible Markup Language) est nécessaire pour obtenir une organisation des données de manière logique et hiérarchisée. L'intérêt de la structuration est multiple : elle permet l'exploitation des ressources du dictionnaire sur des supports différents, la constitution d'une base de données interrogeable, la gestion des mises à jour, la création des produits dérivés, etc. Le balisage des différents éléments devrait aboutir au résultat figurant dans le tableau suivant :

Structure profonde	Structure de surface
<ARTICLE><ZONE-ENTREE><ZONE-ADRESSE><**ADRESSE**>abattoir,</ADRESSE></ZONE-ADRESSE></ZONE-ENTREE><ZONE-GRAM><CATEGORIE-GRAMMATICALE><ZONE-TEXTE><ZONE-SEMANTIQUE><DIVISION-SEMANTIQUE><**TRADUCTION-KABYLE**>âric,<GENRE-NOMBRE><**PLURIEL**>;iâ-cen(B.A);&marquers-paranthèses;</PLURIEL></GENRE-NOMBRE></TRADUCTION-KABYLE><**TRADUCTION-KABYLE**>batuar (pris du français).&emprunt- franc;</TRADUCTION-KABYLE></DIVISION-SEMANTIQUE></ZONE-SEMANTIQUE></ZONE-TEXTE></CATEGORIE-	**abattoir** ***âric*** pluriel ***iâ-cen*** (B.A) ; ***batuar*** ♦(pris du français).

Tableau 1. Présentation de la structure d'un article et sa représentation en surface.

L'arborescence donne un aperçu de l'environnement de l'élément en question (présenté en gras dans le tableau, colonne de gauche). À un niveau supérieur de l'arborescence sont indiqués les éléments pouvant contenir l'élément analysé. Si celui-ci contient lui-même d'autres éléments, ces derniers sont listés dans un niveau inférieur. Toutefois, étant donné la non-systématicité des articles (leur structure varie considérablement à tel point que les différentes informations peuvent figurer à n'importe quelle position dans l'article), un traitement automatisé n'aurait donné que des résultats médiocres : les disparités existant entre les articles ne permettent pas une automatisation du découpage du texte lexicographique. Nous avons donc dû définir un schéma de codage suffisamment souple permettant de décrire les particularités de chaque article. Par rapport à cette difficulté, nous avons opté pour une solution alternative : celle d'utiliser un éditeur XML permettant l'insertion manuelle de balises aux endroits voulus pour compléter le fichier XML issu du premier traitement. Pour accomplir cette tâche, nous avons choisi la solution proposée par le logiciel Adobe FrameMaker.

3.3.1 L'outil Adobe FrameMaker

Adobe FrameMaker[6] est un outil de publication automatisée multicanal intégrant un éditeur XML. Le mode de création XML fournit une interface utilisateur dotée d'un éditeur XML permettant de décrire le contenu d'un document d'une façon structurée et conformément aux normes d'échange de données ou de présentation sur le Web. L'interface utilisateur est simple et ne nécessite pas de connaissances approfondies de codage XML. Elle permet, entre autres, de créer des balises, de les appliquer aux éléments textuels sélectionnés en un seul clic. La fenêtre "Auteur" fournit une vue WYSIWYM (What You See Is What You Mean, ce que vous voyez est ce que vous voulez dire), affiche trois panneaux différents : le premier affiche le document de travail tel qu'il apparaîtra à la publication, le deuxième permet l'ajout ou la modification des marqueurs, balises et toutes autres variables et le troisième autorise l'application des opérations au document de travail. Ces outils visuels de création facilitent la structuration du texte lexicographique et permettent de voir simultanément la structure hiérarchisée et le contenu du document.

[6] Adobe FrameMaker version 12. Il existe d'autres éditeurs XML comme oXygen XML Editor ou Arbortext Editor qui présentent des fonctionnalités similaires.

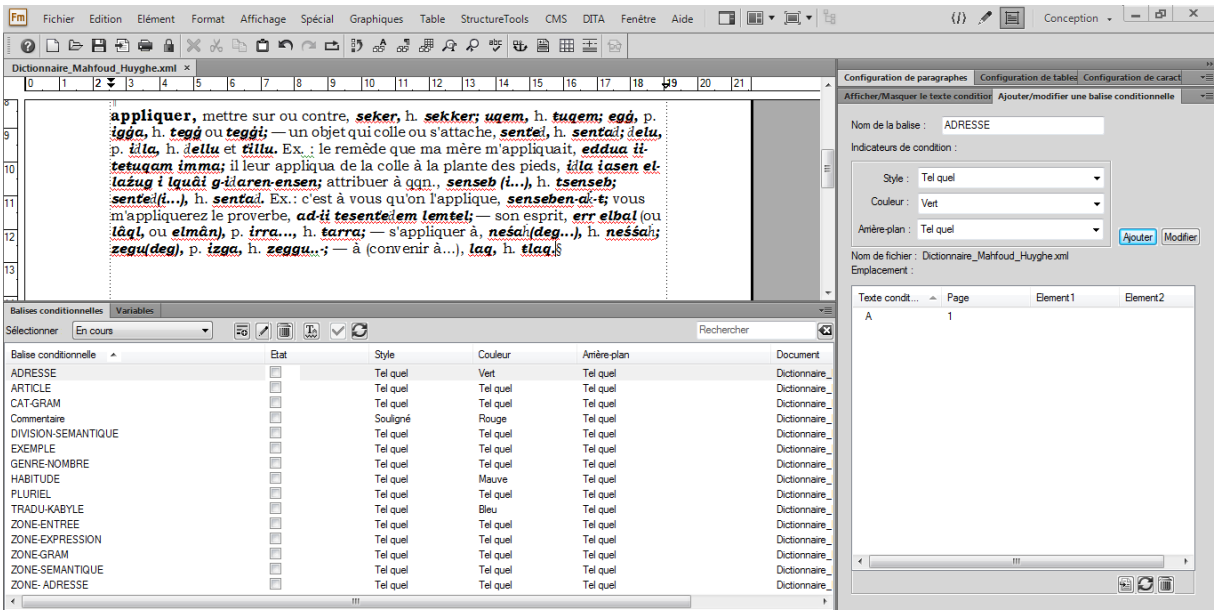

Figure 2. Interface utilisateur de l'éditeur XML Adobe FrameMaker.

Le conteneur de balise (en bas du la page) permet d'appliquer les balises choisies au texte sélectionné. Il est possible d'attribuer une couleur particulière pour chaque élément sélectionné du texte de manière à distinguer les balises d'arrière-plan. Ce mode d'affichage masque le balisage et ne montre que des zones de saisie correspondant aux différents éléments et attributs. Une seconde option permet d'afficher tout le balisage avec la structure hiérarchisée des balises. La configuration de cet outil fournit ainsi un moyen de contrôle semi-automatique suivant les règles relatives à la structure des articles : l'éditeur vérifie constamment les règles prédéfinies et ne permet d'insérer qu'un contenu conforme à ces règles.

L'outil Adobe FrameMaker constitue une aide précieuse pour l'informatisation des dictionnaires car il permet de décrire de manière précise et logique la structure formelle de chaque article. L'avantage qu'offre cet éditeur est d'être facile d'utilisation et permet surtout un gain de temps et d'effort par rapport à une structuration strictement manuelle du texte lexicographique.

3.3.2 Conception de la base de données SQL[7]

La conception de la base de données est fondée sur une analyse lexicographique minutieuse permettant de recenser tous les éléments qui composent le texte dictionnairique. L'utilisation d'outils conçus spécialement pour la modélisation de la base de données facilite l'expertise et l'identification des besoins tout en accélérant le processus de sa conception. Celle-ci doit répondre au moins à trois objectifs : structurer les données, les trier de manière à donner l'accès à un maximum d'informations utiles et les stocker dans la base.

Pour organiser les données, nous avons d'abord utilisé le modèle (ou technique) « Entité-Association » qui permet de construire des schémas théoriques de raisonnement, puis nous avons mis en œuvre un modèle de base de données relationnelle permettant de réaliser n'importe quelle requête : les Entités étant liées entre elles par des relations *de → à* (par exemple, *Entrée* conduit à *Traduction*). Mais avant d'aborder l'architecture de la base de données, il nous faut d'abord présenter le modèle *Entité-Association*.

3.3.3 Modèle Entité-Association

Le modèle *Entité-Association* est une représentation des données traitées sous forme de schéma logique. Pour construire ce schéma, nous avons utilisé un logiciel libre (AnalyseSi) qui permet de modéliser la base de données. Cet outil offre une grande souplesse en matière d'analyse, ce qui convient avantageusement à la structure non systématique du *Dictionnaire français-kabyle* de Huyghe.

---

7 Structured Query Language. Cette technique est, par exemple, mise en œuvre pour la structuration du dictionnaire Encarta

Après avoir défini les propriétés de chaque élément à intégrer dans le dictionnaire de données, nous avons procédé à la création des Entités et Associations pour obtenir le schéma MCD (Model Conceptuel de Données) ci-dessous.

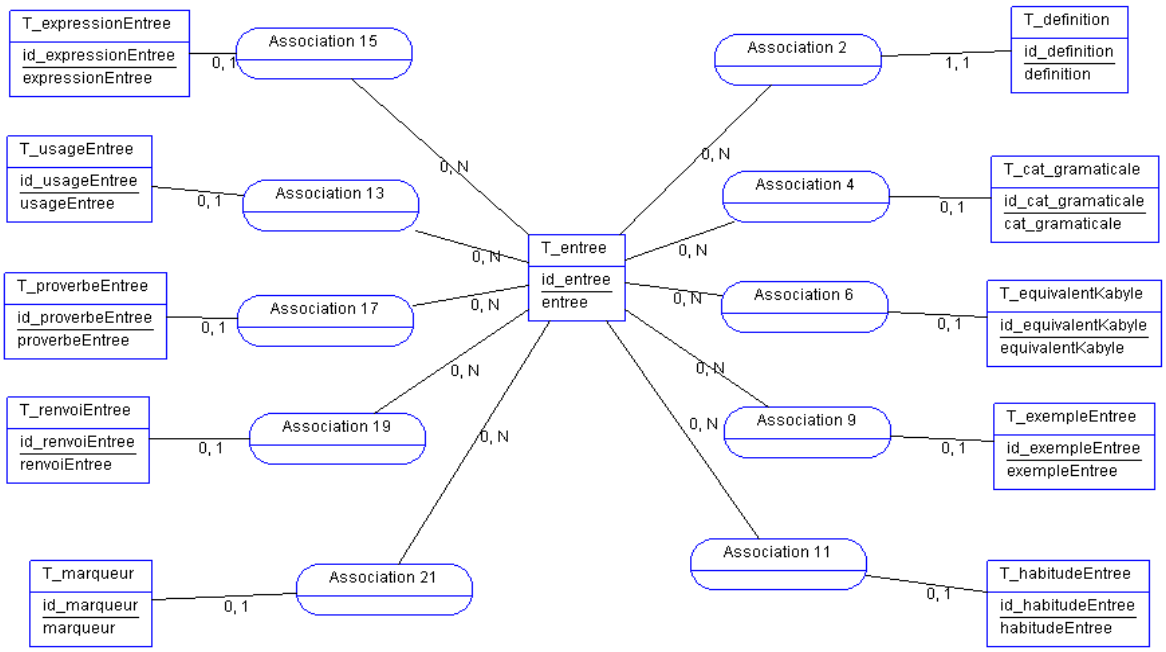

Figure 3. Schéma relationnel Entités-Associations (MCD).

Dans ce schéma, les Entités sont représentées par des cadres rectangulaires, les Associations par les formes ovales et les liens entre Entités et Associations sont symbolisés par des lignes marquant aussi la cardinalité de la relation (0, n) (0, 1). Notons que l'Entité « entrée » est centrale dans ce schéma dans la mesure où toutes les autres données convergent vers elle. C'est ce que représente le schéma MLD (Modèle Logique de Données) suivant :

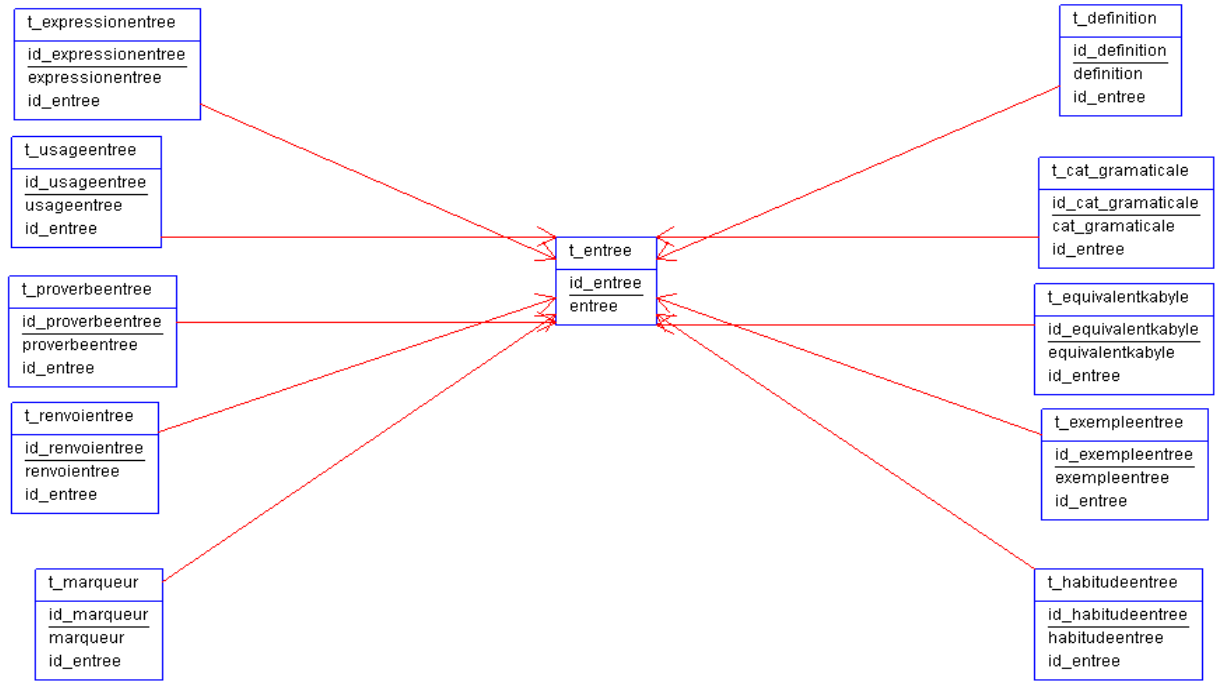

Figure 4. Schéma conceptuel MLD.

Ce schéma est obtenu par la transformation du modèle « MCD » en modèle « MLD » qui est directement exploitable par la base de données. Par ailleurs, l'insertion d'une clé étrangère (id_entree) dans toutes les autres tables en *relation* avec la table « *t_entree* » permet de faire le lien entre les informations contenues dans les autres tables et celles stockées dans la table entrée. Cette opération garantit ainsi l'intégrité des données pendant les différentes opérations de manipulation et de consultation des informations.

Une base de données est un fichier composé d'une ou de plusieurs tables. L'outil AnalyseSi nous offre la possibilité de construire et de générer un script de création des tables composant notre base de données relationnelle en respectant les contraintes fonctionnelles et référentielles de celles-ci. Une simple exécution du script SQL aboutit à la création des tables de la base de données.

3.3.4 Remplissage de la base de données

La méthode que nous avons adoptée est basée sur l'utilisation du document XML (Balises) pour créer un script SQL permettant de remplir les tables. En effet, le fichier XML contient les informations du dictionnaire regroupées entre des balises préalablement fixées. À partir de celui-ci, nous avons créé un fichier SQL contenant des requêtes d'insertion de données. Une fois le fichier importé et exécuté sous MySQL, la base de données est ainsi remplie. Voici un exemple de remplissage de la table *entrée*.

```
-- Contenu de la table `entree`

INSERT INTO `entree` (`id_entree`, `entree`) VALUES
(1, 'A,'),
(2, 'A,'),
(3, 'à, '),
(4, 'abaissant,'),
(60, 'abréger'),
(61, 'abreuver'),
(62, 'abreuvoir '),
(63, 'abréviation '),
(64, 'abri ');
```

Figure 5. Requêtes d'insertion dans la table *entrée*

3.3.5 L'interface

Une fois la base de données remplie, l'accès aux informations se fait par le biais d'une interface Web dynamique qui interprète les fichiers PHP (Hypertext Preprocessor) utilisés dans l'exploitation de la base de données. Ces fichiers sont une combinaison de script PHP, de requêtes SQL et du code HTML (Hypertext Markup language) qu'un navigateur Web interprète.

Lorsque l'utilisateur fait une requête, le PHP récupère les informations saisies et fait appel à la base de données pour récupérer les informations demandées et les afficher par la suite sous forme HTML qui gère la visualisation à l'écran. Voici une copie d'écran de l'interface d'interrogation du *Dictionnaire français-kabyle*.

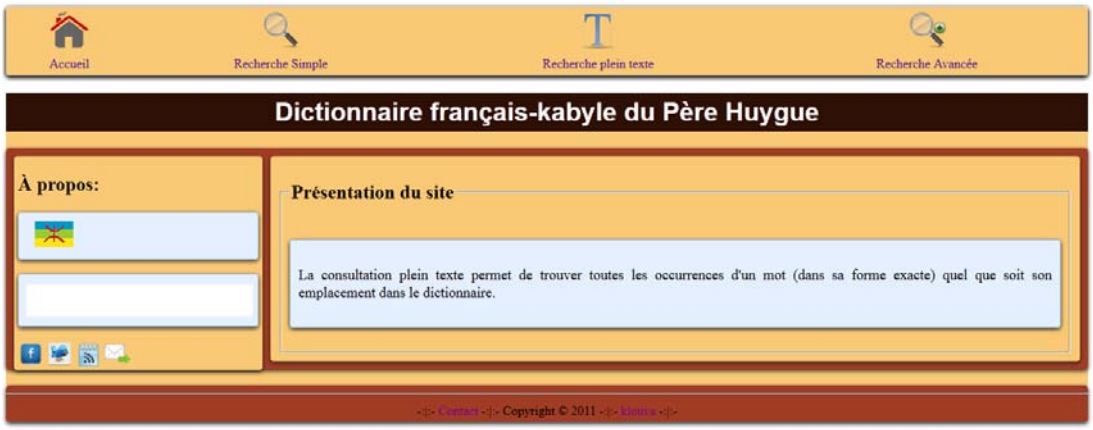

Figure 6. Interface d'interrogation du *Dictionnaire français-kabyle*.

3.3.6 La consultation du *Dictionnaire français-kabyle*

L'interface du *Dictionnaire français-kabyle* offre à l'utilisateur trois modes de consultation : une recherche simple, une recherche plein texte et une recherche avancée. Ces modes de recherche permettent à l'utilisateur d'avoir accès à un grand nombre d'informations impossible à exploiter dans toute version papier des dictionnaires.

3.3.6.1 La recherche simple

Ce mode d'interrogation, des plus traditionnels, permet un accès simple et rapide au contenu d'un article du dictionnaire. Cette consultation consiste à rechercher l'article concernant le mot saisi par l'utilisateur dans le menu *recherche simple*. La recherche dite « simple » s'effectue sur une entrée de la nomenclature qui donne accès à l'article lui correspondant : l'utilisateur saisit un mot sur lequel il souhaite obtenir des informations pour accéder directement à l'article. Nous avons pris le soin de présenter les textes des articles de manière aérée pour faciliter leur consultation. Voici une copie d'écran illustrant le mode de *recherche simple*.

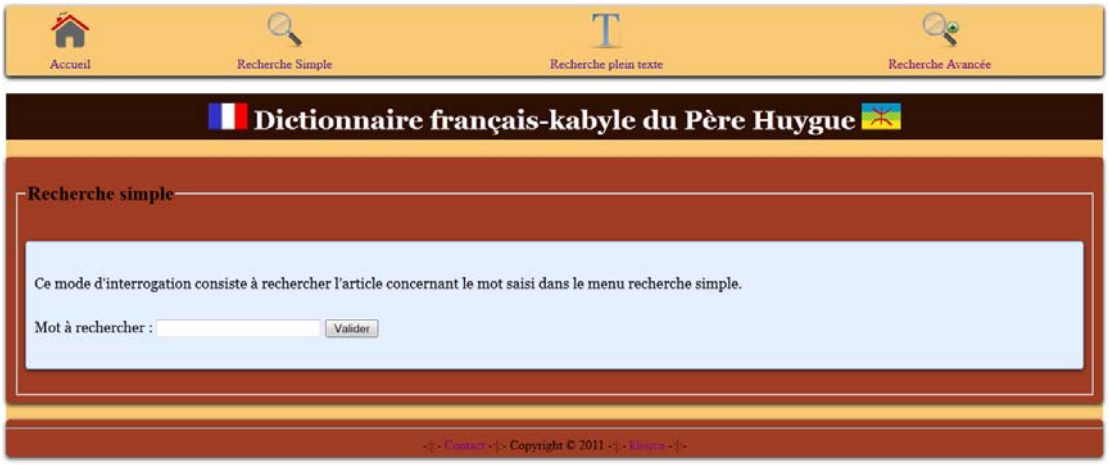

Figure 7. Résultat de la recherche en mode *recherche simple*.

Ce mode de recherche procède en quelque sorte de la même manière que la consultation manuelle dans les dictionnaires papier. En revanche, l'opération de recherche dans le dictionnaire informatisé offre l'avantage d'être nettement plus rapide.

3.3.6.2 La recherche plein texte

Ce mode d'exploitation donne à l'utilisateur la possibilité d'effectuer à travers l'intégralité du texte lexicographique des recherches *plein texte* qui permettent l'accès à des informations disséminées dans tout le texte du dictionnaire. Cette option de recherche permet à l'utilisateur de trouver toutes les occurrences de la

forme saisie, et ce quelle que soit sa position dans le texte. Ainsi, le texte du dictionnaire est parcouru dans son intégralité et les occurrences trouvées sont mises en évidence.

L'utilisateur saisit donc un mot ou une expression de son choix dans le menu *recherche avancée* et la liste des résultats trouvés s'affiche à l'écran. Ces résultats sont classés par degré de pertinence : la forme recherchée est considérée de "forte pertinence" lorsqu'elle correspond à une entrée, de "faible pertinence" lorsqu'elle figure dans le contenu de l'article. Dans ce dernier cas, les éléments trouvés sont affichés en fonction de la fréquence des formes occurrentes dans le texte de l'article. Il est à signaler que toute requête de recherche plein texte ne tiendra pas compte des mots de moins de trois caractères du fait de leur trop grande utilisation. De même que l'interface d'interrogation ne tiendra pas compte des accents. Ainsi, le mot « *abnégation* » peut être saisi sans accent sur le « e » (abnegation). Voici une copie d'écran illustrant le mode de *recherche plein texte*.

Figure 8. Résultat de la recherche en mode *recherche plein texte*.

L'avantage de ce mode de recherche est de fournir à l'utilisateur une grande quantité d'informations auxquelles il n'aurait pas eu accès lors d'une consultation traditionnelle dans le dictionnaire papier.

3.3.6.3 La recherche avancée

Outre les modes de recherche simple et plein texte, l'interface d'interrogation permet d'effectuer des recherches ciblées dans des sections spécifiques des articles du dictionnaire au moyen de critères multiples. La recherche avancée permet donc une interrogation plus fine de la base de données grâce une analyse minutieuse du dictionnaire. Elle offre à l'utilisateur la possibilité de mener sa recherche en utilisant un ou plusieurs critères situés sous la zone de saisie du menu *recherche avancée* ou alors de saisir librement un mot dans la zone de saisie. Ainsi, l'utilisateur peut limiter sa recherche à un seul critère ou en combiner plusieurs parmi les **dix** critères suivants : catégories grammaticales, exemples, expressions, équivalents kabyles, forme d'habitude, indicateurs sémantiques, marques d'usage, marqueurs entre parenthèses, proverbes et renvois.

L'une des caractéristiques du *Dictionnaire français-kabyle* de Huyghe consiste dans l'abondance des exemples et expressions qu'il propose. L'un de nos objectifs est de valoriser cette richesse pour le lecteur contemporain, avide de témoignages historiques et culturels. La fonction *recherche avancée* permet à l'utilisateur, par exemple, de chercher dans les exemples et les expressions un mot ou un groupe de mots donnés. Cependant, s'il est possible de croiser plusieurs critères de recherche, certaines combinaisons pourraient s'avérer non pertinentes. Par exemple, combiner le critère « *catégorie grammaticale* » avec « *marque d'usage* ». Des solutions comme le « *Système de Recherche Dynamique* (SRD) »[8] peut être une des solutions pour éviter de telles recherches. Voici une copie d'écran illustrant le mode de la *recherche avancée*.

[8] Le SRD permet de déterminer et de mettre en évidence les critères pertinents en même temps que l'utilisateur sélectionne les critères de sa recherche.

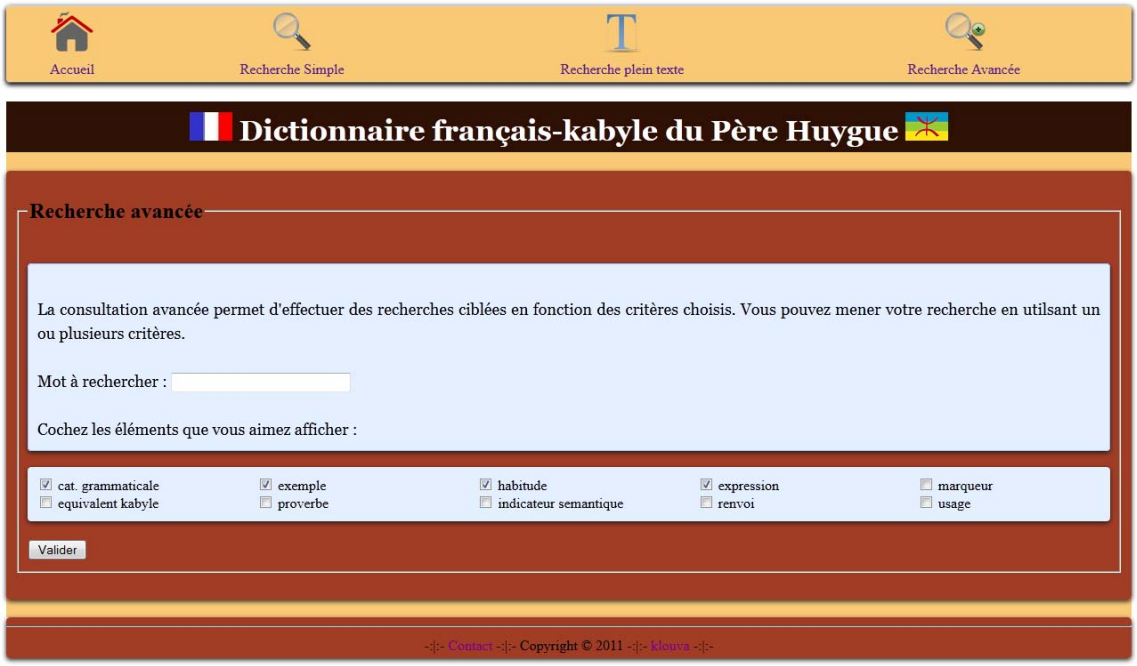

Figure 9. Résultat de la recherche en mode *recherche avancée*.

Ce mode de recherche avancée offre l'avantage d'utiliser les différentes possibilités d'interrogation en multipliant la combinaison des critères. Ce type d'exploitation permet de mener des recherches expertes et ciblées dans le contenu du dictionnaire, ce qui facilite grandement la consultation. Notons enfin que ce mode de consultation inclut la recherche simple. Nous voyons que l'outil offre de nombreuses fonctionnalités qui peuvent être mises au service de la valorisation du patrimoine lexicographique.

Conclusion

La principale qualité de la version informatisée du *Dictionnaire français-kabyle* réside dans le respect de l'édition originale : le contenu textuel de la version papier correspond à celui de la version informatisée. La valeur ajoutée de la version informatisée réside dans les différentes possibilités d'exploitation des informations. Cela est rendu possible par le travail de structuration des données effectué en amont. La méthodologie décrite dans cette étude permet aux utilisateurs les moins familiarisés avec le système de balisage de travailler dans une interface simple d'utilisation. L'outil Adobe FrameMaker est une solution permettant aux linguistes ayant des connaissances suffisantes en informatique de procéder à la structuration de leurs corpus. Partant d'un fichier XML issu d'un premier traitement, le linguiste permet de compléter le balisage des données de manière souple : il peut choisir de nommer librement les balises et de les appliquer ensuite au texte sélectionné, le tout dans un format hiérarchisé exploitable et transférable dans d'autres applications. Une fois la phase de structuration terminée, le fichier XML traité par un script PHP permet d'implémenter la base de données. Enfin, une interface Web dynamique donne accès aux informations contenues dans la base de données et permet surtout de formuler des requêtes pour explorer d'une façon fine les informations contenues dans le texte lexicographique. Les trois modes de consultation sont fonction des centres d'intérêts des usagers qui, dans leurs recherches, choisissent tel ou tel mode selon les requêtes saisies desquelles dépendent directement des éléments de réponse.

L'avantage indéniable qu'apporte cette méthodologie tient dans le travail de structuration des données qui est une phase décisive dans l'informatisation des dictionnaires anciens. Les outils informatiques actuels rendent possible, selon une méthodologie simple et économe d'élaboration, l'accès à des ressources inestimables disponibles tant pour les spécialistes de langues que pour le grand public. Au-delà de la sauvegarde d'un patrimoine linguistique commun, la méthodologie décrite dans cette étude permet d'envisager à l'avenir l'actualisation de ces ressources lexicales anciennes. En rendant leur contenu pleinement exploitable, elles pourraient être enrichies facilement par l'ajout d'informations manquantes et contribuer avantageusement à la constitution d'outils lexicographiques modernes.

Bibliographie

DENDIEN J., PIERREL J. (2003), « Le trésor de la Langue Française informatisé. Un exemple d'informatisation d'un dictionnaire de langue de référence », dans *Traitement automatique des langues*, vol. 44, n° 2.

DE TIER V., VAN KEYMEULEN J. (2010), « Software Demonstration of the Dictionary of the Flemish Dialects and the Pilot Project Dictionary of the Dutch Dialects », dans *Proceedings of the 14th Euralex International Congress*, 6-10 juillet 2010, Leeuwarden, Ljouwert: Fryske Akademy, p. 620-627.

ENGUEHARD C., MANGEOT M. (2013), « LMF for a selection of African Languages », dans Gil Francopoulo (éd.), *LMF, Lexical Markup Framework*, p. 99-118.

GASIGLIA N., (2009), « Évolutions informatiques en lexicographie : ce qui a changé et ce qui pourrait émerger », dans *Changer les dictionnaires, Lexique*, n°19, Villeneuve d'Ascq : Presses universitaires du Septentrion, p. 235-298.

LEROY-TURCAN I., WOOLDRIDGE R. (1997), « L'informatisation des premiers dictionnaires de langue française : les difficultés propres à la première édition du *Dictionnaire de l'Académie française* », dans PRUVOST Jean (éd.), *Les Dictionnaires de langue française et l'informatique*, Université de Cergy-Pontoise : Centre de Recherche Texte/Histoire, p. 69-86.

MANUELIAN H., BRUSCAND A., CHOLEWKA N. et HETZEL A. (2009), « Le Petit Larousse Illustré de 1905 en ligne *:* Présentation *et* secrets de fabrication », dans *Études de Linguistique Appliquée* (ÉLA), n°156, p. 453 - 474.

MAHTOUT M. (2012), *Les dictionnaires bilingues en Algérie pendant la période coloniale, 1830-1930 : histoire, analyse et perspectives d'avenir*, Thèse de l'Université de Rouen, 2 vol.

MAZZIOTTA N. et RENDERS P. (2010), « Vers un enrichissement raisonné de la rétroconversion du Französisches Etymologisches Wörterbuch *(*FEW*)* », dans *Proceedings of the 14th Euralex International Congress*, 6-10 juillet 2010, Leeuwarden, Ljouwert: Fryske Akademy, p. 1026-1032.

WOOLDRIDGE R. (1994), « Projet d'informatisation du *Dictionnaire de l'Académie* (1694-1935) », dans QUEMADA B., et PRUVOST J. (éd.), *Actes du Colloque sur le* Dictionnaire de l'Académie française *et la lexicographie institutionnelle européenne*, Paris : Institut de France, p. 309-320.

Logiciels

Adobe FrameMaker version 12

Omnipage 17 (logiciel de Reconnaissance optique des caractères)

AnalyseSi