

INDEXATION D'UN QUOTIDIEN DE LANGUE ARABE

Défis et perspectives

Nohma Khayrallah

Enseignante à l'Université libanaise, faculté de journalisme et de documentation (section I, Unesco) – Chef du département d'indexation et des coupures de presse au Centre An-Nahar pour la recherche (Beyrouth, Liban).

Les articles, analyses et informations sont l'apanage d'un journal quotidien. Et pour en tirer un meilleur profit, un corpus judicieux s'impose. Cette accumulation, si elle est bien organisée, devient une importante et sérieuse source d'information. Et pour classer et réunir les informations de ces quotidiens, un index et une base de données précis et concis sont de rigueur. Au Moyen-Orient, un quotidien politique, le *Nahar*, a commencé en 1992 une expérience d'indexation sur des numéros scannés, fournissant une énorme base de données à tout chercheur, journaliste et étudiant pour approfondir la politique, l'économie, et la culture de cette partie du monde.

Mais comment indexer un quotidien qui a 65 ans d'âge quand on a affaire à :

- une langue qui a ses propres caractéristiques linguistiques et écrites ;
- un thesaurus traduit et qui doit s'adapter à un quotidien ;
- un contexte géopolitique complexe et difficile ?

Ainsi, quels sont les obstacles à un tel projet ? Quelles sont les difficultés auxquelles s'attendre ?

Pour répondre à ces questions, nous allons aborder dans cet article l'élaboration d'un index d'un journal politique libanais, les étapes qui l'ont marquée, les problèmes rencontrés et son adaptation à un environnement politique complexe. Nous aborderons de même l'utilisation d'un thesaurus en arabe pour un quotidien, cette expérience étant pionnière dans ce domaine. L'indexation est utile pour tout journal ou document qui ne peut être soumis à une indexation automatique à cause de son ancienneté ou d'une reconnaissance optique de caractères (OCR) imprécise.

Il s'agit dans cette étude d'établir un bref historique du journal puis d'aborder la démarche d'indexation et ses secrets.

HISTORIQUE DU NAHAR

Le premier numéro du *Nahar* date du 4 août 1933. Il a été fondé par Gibran Andraos Tuéni, avec la collaboration de ses compagnons de plume tels les grands journalistes Fouad Hbeich, André Tuéni, Tawfiq Youssef Awad et Afif Tibi, afin d'établir un nouveau modèle dans l'écriture journalistique loin de l'approche traditionnelle qui sévissait à l'époque.

Gibran Andraos Tuéni a quitté le Liban en 1947 pour l'Argentine, en laissant des recommandations quant à l'approche des politiques extérieures et intérieures du journal. Ce sera un « *journal arabe libanais, défendant et l'indépendance du Liban et sa propre indépendance financière, et qui, surtout, s'éloignera du sectarisme* ». Le document engage le journal à suivre la « *culture progressiste* », à ne pas rejeter ce qui vient de l'Occident, mais à l'assimiler « *dans notre propre culture et à l'ajouter à l'héritage arabe* ».

Tout journalisme n'est en fait qu'à l'image de la société qu'il représente. « *Le paysage médiatique libanais est le reflet de la structure politique du pays* » (Abou Assi, 2009). Pour décrire l'adaptation du journal à la complexité de la politique libanaise, il est essentiel de revenir à un article publié en page une du 7 octobre 1956 (El Hajj, 1956), article dans lequel Louis El Hajj, un grand journaliste de son époque et éditeur en chef du journal, dévoile les secrets de l'écriture journalistique propre au *Nahar*. Ces directives seront reprises par Ghassan Tuéni qui prônera certaines règles dans l'édition, continuant les directives de Tuéni père, et de Hajj :

- ne rien publier sans enquête préalable et sans s'assurer des points de vue de toutes les parties concernées ;
- dénicher les informations. Ne pas se limiter pas à celles qui s'imposent à nous, mais les chercher là où elles devraient être et où elles seront à venir ;
- militer pour la liberté, pour les courants de changements, lever la voix, et être les porte-parole des grandes et petites plaintes, des critiques, des rêves et des aspirations du peuple.

Le titre, une approche différente

Alors que les gros titres dans la presse arabe du début du siècle étaient formulés tels les titres des livres, le nouveau journalisme a adopté « le titre élargi », ce qui n'était pas d'un journalisme familier. Le titre est devenu ce résumé des faits qui contenait les éléments les plus importants de l'article, et alléchant le lecteur par les détails avant de commencer la lecture, comme le décrit le grand journaliste Farés El Khoury qui a dit que le titre de ce journal EST la nouvelle (El Hajj, 1956).

Un style journalistique particulier

Le journal est devenu une source importante pour de nombreux chercheurs parce qu'il offre à ses lecteurs des informations précises sur les affaires gouvernementales et la politique intérieure, ainsi que les déclarations et les discours des ministres et des députés de toutes tendances politiques confondues. Le journal publia en 1956 une feuille de route sur les secrets de la rédaction journalistique, lors d'un remaniement total du journal. Dans un article en page une du 16 octobre, une feuille détailla la répartition des nouvelles et articles sur les huit pages du journal. Ces divisions semblent de prime abord classique, mais c'est la qualité des nouvelles qui donna sa particularité au style journalistique adopté.

POURQUOI UN INDEX?

Une étude approfondie des articles du journal, ainsi que des demandes des utilisateurs, chercheurs, journalistes et étudiants, a révélé que les dossiers des coupures de presses n'étaient plus suffisants devant l'ampleur de l'information, et surtout devant les corrélations entre les sujets couverts par le journal.

Les coupures de presse classées dans des dossiers thématiques n'arrivaient plus à satisfaire la demande centrée sur les détails, les chiffres précis et quotidiennement renouvelés. Et surtout, il s'agissait d'avoir les informations rapidement et sur son poste, c'est-à-dire ne pas avoir à se déplacer pour consulter les dossiers papiers. Et devant l'ampleur des informations, il fallait passer à une recherche plus avancée qui « s'immisce » et « s'infiltrer » dans l'article pour en retirer profit. Un index s'imposait.

La spécificité des titres et la présentation des sujets facilitait l'approche. Et pour une bonne indexation d'un quotidien, cet index a dû répondre à ces questions :

Technique de l'indexation

- Quels types d'informations les utilisateurs chercheront-ils dans cinq ans ?
- Quelles sont les informations bibliographiques qui doivent être mentionnées et affichées pour une utilisation optimale ?

Pertinence de l'indexation

- Quel niveau d'indexation est prévu ?
- Quels articles ne seront pas sélectionnés pour l'indexation ?

Organisation de l'index

- Combien d'articles, ou de numéros du journal, seront indexés quotidiennement ?

Publication de l'index

- Quelles seront les utilisations futures de l'index ?
- L'index sera-t-il distribué ou vendu ?

POURQUOI UN THESAURUS?

La relation entre les index et les thesaurus est une relation fondamentale. Et le choix d'un thesaurus ne fut pas simple face à trois problèmes majeurs :

- l'ampleur de la tâche (cinquante neuf ans de journaux au moment où commença ce projet) ;
- la spécificité du style journalistique ;
- la langue arabe du quotidien.

Une étude faite sur un certain nombre de textes du journal, a montré qu'on pouvait établir une structure identifiable quoique particulière à ce style journalistique. En effet, dans la majorité des cas, les articles présentent une structure unifiée. Ils sont construits avec méthode pour faire passer un message, un fait, ou un évènement majeur : le premier paragraphe introduisant le sujet, le deuxième cernant les lieux, les interlocuteurs et résumant les faits, puis le corps de l'article donnant au lecteur une idée plus ou moins exhaustive du sujet qui a été traité. Ces constatations ont conduit à

admettre que la division de l'article en descripteurs était faisable et avec méthodologie, tout en préservant une unité globale du document par le scannage de la page entière, offrant ainsi une cohérence entre l'indexation et le document.

Le journal a entrepris son projet de digitalisation et d'indexation en 1992, et a adopté le thesaurus de l'Organisation de coopération et de développement économique (OCDE), traduit en arabe par Middle East Research and Studies (MERS).

Pourquoi le thesaurus de l'OCDE ?

Ce choix est justifié par des raisons majeures :

- Les thesaurus dans les sciences humaines et sociales sont moins nombreux et beaucoup moins précis que ceux pour les sciences pures et exactes (Khalifa, 2001). Il fut donc logique d'utiliser le thesaurus de l'OCDE dont le lexique est le mieux adapté à des informations d'ordre économique et social.
- Ce thesaurus existe en arabe dans le cadre des efforts de la Ligue arabe de traduire le thesaurus de l'OCDE. C'est donc dans ce contexte que parut en 1979 un thesaurus exhaustif des termes dans les domaines du développement socio-économique, et c'était peut-être la première tentative arabe d'aborder les thesaurus.
- Même si la version du thesaurus de l'OCDE traduite par la Ligue arabe ne fut pas une traduction précise, mais plutôt une « arabisation » et une adaptation du thesaurus étranger (Khalifa, 2001), le choix de la Ligue arabe de traduire ce thesaurus en particulier n'est pas à négliger.

Ce thesaurus fut donc un outil de contrôle pour introduire une analyse objective concernant tous les documents arabes dans le domaine socio-économique, en favorisant les sujets de développement industriel (Khalifa, 2001). Chaaban AbdelAziz Khalifa¹ qualifie la traduction de ce thesaurus comme un travail incomplet et primitif. Il lui reproche, entre autres, l'utilisation des termes au singulier plutôt qu'au pluriel (Khalifa, 2001).

MERS entreprit donc de faire sa propre traduction. Cette version fut utilisée par plusieurs journaux libanais pour indexer leurs numéros. Ce thesaurus traduit en arabe fournissait les concepts et les hiérarchies adéquates, et permettait d'incorporer les nouveaux sujets qui s'imposaient au fur et à mesure du travail, en étant plus adapté au lexique utilisé par les journaux libanais notamment.

1 Chaaban AbdelAziz Khalifa est professeur à la faculté de littérature à l'Université du Caire, président de l'Association égyptienne pour les bibliothèques et l'information. Il est l'auteur de *La liste des en-têtes arabes standardisées pour les bibliothèques, centres de recherche et bases de données* (qa'imat rou'ouss el mawdouat al arabia al qyassia lilmaktabat wa marakiz el maaloumat wa qawaed el bayanat), trois volumes, publié chez Dar El Shourouq en 2001.

L'INDEXATION D'UN JOURNAL ARABE

La méthodologie du travail

Une méthodologie stricte et particulière s'imposait. Elle est le résultat de l'expérience.

Le classement alphabétique se fait en ne tenant pas compte de l'article défini [AL] en début de mots.

- Tous les articles du journal sont inclus dans l'index à l'exception des nécrologies, des annonces et des photos.
- L'indexation des photos se passe en parallèle de l'indexation du journal, quoique la banque de données, le thesaurus et les notices bibliographiques soient quasiment les mêmes, à part quelques différences spécifiques aux photos.
- Les notices bibliographiques sont divisées en deux grandes catégories :
 - Les notices bibliographiques communes au *numéro* en entier, notamment la source, la date du numéro, le numéro de série.
 - Les notices bibliographiques qui appartiennent à *chaque page* séparément. Et ces notices lient les entrées et sont publiées sur cette même page, mais se différencient par le numéro de la colonne où débute chaque article.
- Certains événements majeurs nécessitent un grand nombre de descripteurs, ce qui charge l'index et rend la recherche très complexe. Pour résoudre ce problème, on a eu recours à des « descripteurs thématiques » tels que : « La Guerre de la Montagne », ou l'« Invasion Israélienne 1982 ».
- Ces « descripteurs thématiques » répondent à la terminologie utilisée par le journal, une objectivité stricte concernant le choix des descripteurs.
- Tous les signataires sont classés dans la case des « Auteurs », même s'ils sont des reporters ou des journalistes délégués dans les différentes régions.
- Les noms des personnes sont classés par ordre alphabétique du nom de famille, exceptés les noms des souverains, des princes, des héritiers, les saints, les pontifes, les patriarches, les cardinaux... classés par leur prénom.
- Pour les autres il était très difficile d'avoir le nom officiel, par manque de ressources et de références. Concernant les noms arabes anciens, il fallait avoir recours à plusieurs références en la matière, notamment le livre de Kheir ElDine El-Zarkali (1896–1976) (El-Zarkali, 1980) sur les *Biographies arabes*.

Le choix des descripteurs

Le rôle d'un thesaurus est, d'un côté, de faciliter le travail de l'indexeur et, d'un autre, de cerner les sujets. Et pour servir un quotidien politique libanais dont le style ne fait que couvrir la complexité du pays, il fallait faire en sorte qu'une liste de « descripteurs spécifiques » soit ajoutée de prime abord avant d'attaquer les sujets publiés dans le journal. Ces descripteurs concernent les personnes, les partis politiques, les associations, les syndicats, certains sujets divers.

Méthodologie de travail

Quelques règles essentielles sont nécessaires pour indexer les nouvelles afin de faciliter la recherche et bâtir des «dossiers» thématiques virtuels.

- Ainsi, une rencontre entre deux politiciens, deux partis politiques, ou une réunion entre un parti politique et un politicien, doit nécessairement prendre les descripteurs «réunions» avec les noms des deux partis.
- Un discours ou une annonce est décrit par «annonces, discours, déclarations», selon la situation.
- Concernant les accrochages entre les miliciens, les miliciens et l'armée, etc., il faut adopter une méthodologie stricte et surtout objective. Le descripteur «accrochages» est utilisé avec les noms des partis et mentionnés par ordre alphabétique.
- Toute visite d'une personnalité politique en fonction, hors de son pays, doit nécessairement être accompagnée du descripteur «visites» et du descripteur «relations extérieures» avec les noms des deux pays concernés. La même règle s'applique si un assassinat ou une explosion se passe contre des ressortissants d'un pays dans un autre.
- Pour enrichir le thesaurus et le rendre plus malléable avec plus d'une centaine d'entrées par jour, il faut ajouter presque tous les «noms propres» qui passent dans les titres des articles en premier lieu, puis ceux qui étaient dans l'article. Ce qui nous ramène à la clarté des titres mentionnée auparavant. Et par «noms propres», on veut dire tous les noms des personnes, des syndicats, des partis politiques, des sociétés, des conférences, etc.

La liste des descripteurs

Ils sont classés par ordre alphabétique en deux listes séparées : la liste des «auteurs» et la liste des sujets.

Un index papier fut publié pour la période 1992-1994. C'est un travail pénible et difficile, car il s'agissait d'indexer les articles avec des descripteurs tels que les vedettes-matières utilisées dans un index-papier. Il s'agissait d'imprimer les descripteurs comme entrée reliées chacune au nom du pays concerné.

Par ex :

Étudiants-France

France-relations extérieures-Liban

Liban-parlement

Les titres qui paraissent sous les descripteurs sont publiés par ordre chronologique afin de conserver le déroulement des faits.

UNE PREMIÈRE ESQUISSE

Le premier numéro indexé était celui du 1er janvier 1992. Année de publication : 59, numéro de série : 18121. C'est un numéro de 12 pages et de 95 entrées.

Dans cette première page du 1er janvier 1992 (dossier attaché no 1), le nombre d'entrées s'élève à 7. Deux auteurs sont relevés : *Gibran Hayek*, dans le second article de la page, première colonne de droite, et *Ahmad Ayache*, dans le cinquième article, colonne cinq, de cette même page.

Dans ce numéro, on note :

■ **Table 4 - « descripteurs spécifiques » au 1^{er} janvier 1992**

Auteurs	27
Desc. 1 = réunions	7 fois
Desc. 2 = déclarations	10 fois

Le nombre total d'entrée pour l'année 1992 est de 41 501. Et pour donner une idée de l'importance des « descripteurs spécifiques » ajoutés à la version du thesaurus de l'OCDE et utilisés pendant l'année 1992, un bilan donne les résultats suivants :

■ **Table 5 - « descripteurs spécifiques » à l'année 1992**

Réunions	3 978 entrées	Discours	762 entrées
Déclarations	5 197 entrées	Visites	1 832 entrées
Annonces	671 entrées	Cadeaux	2 entrées

CONCLUSION

L'indexation est et restera toujours un moyen de traitement d'information très importante même en la présence de logiciel de reconnaissance optique de caractère (OCR) ou par la recherche par mots du texte. La gestion des descripteurs pour indexer un journal quotidien est une entreprise difficile car elle demande une grande connaissance des sujets actuels et historiques. Elle rassemble plusieurs domaines, tels l'histoire, la politique, l'art, l'éducation, l'économie, le sport, etc.

Et relier ces sujets ensemble dans le système structuré qu'offre un thesaurus est intéressante pour le chercheur comme pour l'indexeur. C'est la recherche qui nous dévoile l'intersection des sujets surtout si la période couverte par la banque de donnée s'étend sur plusieurs décennies. Et au chercheur de découvrir, par l'accumulation des articles, des réponses à ces questions :

Quels sont les facettes qui éclairent ce problème ?

Est-ce que l'on doit considérer d'autres sujets en relation avec ce problème ?

Comment furent résolus les problèmes similaires ?

Des questions que seul un index bien agencé parviendra à résoudre.

Références

- Abou Assi, J. (2009, février), «Les médias libanais: entre confessionnalisme et recherche de crédibilité», *Confluences Méditerranée*, 69, p. 49-59.
- Khalifa, C. A. (2001), dans *La liste des en-têtes arabes standardisée pour les bibliothèques, centres de recherche et bases de données*, Caire: Dar El Chourouq.
- El-Zarkali, K. (1980), *Qamouss el Iilam : Tarajem li Achhar el Rijal wal Nissa' min el Arab wal Moustaaribin wal Moustachrigine*, Beyrouth: Dar El Ilm Lylmalayeen.
- El Hajj, L. (1956, 7 octobre), Mahalyat «Al-Nahar» kayfa taktibo hadthan: Louis El Hajj Yachrah kayfa yousna'o min al khabar. *An-Nahar* (6388), p. 1.
- Gibran Tueni baad 25 sana* (1973) [Beyrouth].