Chapitre I. Corpus des Traditions Prophétiques (*Hadith*)

1. Introduction

Le Coran et la Sunna sont les deux principales sources de la théologie islamique. Le Coran est le recueil des paroles que le Prophète a reçues en état de Révélation (6235 versets). La Sunna (encore appelée Hadith ou Tradition Prophétique) est un immense corpus littéraire qui s'est cristallisé aux 3ème/9ème siècle après une longue période d'élaboration. Il est né de la nécessité historique de compléter le Coran ou de l'interpréter dans le cas où il était silencieux ou incomplet. Il existe actuellement six grands recueils de Hadiths qui font autorité dans le monde musulman. Etant donné que le domaine d'application de notre thèse est les Traditions Prophétiques, il est nécessaire de présenter ce domaine et de poser sa problématique [Sentürk, 2007].

La deuxième section de ce chapitre traite des concepts liés aux corpus textuels proprement dits. La troisième section est consacrée à la description des corpus arabes. Dans la quatrième section, nous exposerons en détail la structure des corpus prophétiques, nous présentons leurs historique, leurs méthodes, leurs finalité et nous passons en revue tous les travaux à caractère commercial et scientifique qui visent à informatiser ces corpus. La dernière section sera réservée pour la conclusion du chapitre.

2. Les corpus proprement dits

2.1. Définition

Un corpus en linguistique désigne l'aspect normatif de la langue : sa structure et son code en particulier. "Corpus" est généralement opposé à statut, qui correspond aux conditions d'utilisation de la langue. Cette opposition est commune dans l'étude des politiques linguistiques. En littérature, un corpus regroupe un ensemble de textes ayant une visée commune. Un corpus peut être constitué de documents différents (tableau, extrait de texte...) et ces documents divers ont un point en commun. En général c'est le thème qui fait figure de leur ressemblance. Il faut avoir une technique particulière pour le déchiffrer.

2.2. Le corpus dans la science

Les corpus sont des outils indispensables et précieux en traitement automatique du langage naturel. Ils permettent en effet d'extraire un ensemble d'information utile pour des traitements statistiques. D'un point de vue informatif, ils permettent d'extraire des tendances et notamment de construire des ensembles de n-grammes. D'un point de vue méthodologique, ils apportent une objectivité nécessaire à la validation scientifique en traitement automatique du langage naturel. L'information n'est plus empirique, elle est vérifiée par le corpus. Il est donc possible de s'appuyer sur des corpus (à condition qu'ils soient bien formés) pour formuler et vérifier des hypothèses scientifiques.

2.3. Méthodologie d'utilisation d'un corpus

Il serait maladroit d'un point de vue méthodologique d'appliquer des traitements statistiques sur le corpus qui a permis de faire ressortir un classement ou une modélisation du langage. Lorsque l'on travaille avec des corpus, il convient donc de séparer un corpus initial en deux sous-corpus:

- le corpus d'apprentissage, qui sert à retirer un modèle ou un classement à partir d'un nombre suffisant d'information ;
- le corpus de test, qui sert à vérifier la qualité de l'apprentissage à partir du corpus d'apprentissage.

Le calibrage des volumes des corpus se discute en fonction du problème, mais il est fréquent d'utiliser les 2/3 du corpus initial pour l'apprentissage et le tiers restant pour effectuer les tests.

Lorsque le volume du corpus initial n'est pas suffisant, il est possible de croiser les corpus de tests et d'apprentissage sur plusieurs expérimentations. La mesure de qualité des résultats (précision ou rappel) est alors plus précise, mais en aucun cas les corpus d'apprentissage et de tests n'ont été mélangés.

3. Les Corpus Arabes

Le corpus est la ressource la plus importante pour la recherche d'information. La langue arabe a un nombre limité de corpus disponible pour les recherches et les expérimentations, ces corpus sont en majorité construits autour des textes journalistiques. Dans la suite nous donnerons une brève description des corpus disponibles pour les expériences de la recherche d'information arabe.

3.1. Le Corpus LDC

Le corpus LDC ou « Arabic Newswire A Corpus » a été créé par David Graff et Kevin Walker au Consortium des Données linguistiques (LDC, 2001) de l'Université de Pennsylvanie. Le corpus est composé d'articles d'information en langue arabe de l'Agence France Presse (AFP) qui ont été publié entre le 13 mai 1994 et le 20 décembre, 2000. La matière source a été étiquetée en utilisant le TIPSTER-Style du Langage SGML, puis convertie au système d'encodage Unicode (UTF-8). Les données sont stockées dans 2337 fichiers compressés de données textuelles arabes. La taille de corpus est de 209 méga-octets dans le format compressé et de 869 Mo quand les données sont non compressées. Le corpus contient 383872 documents contenant 76 million de mots avec un ensemble approximatif de 666,094 mots uniques.

3.2. Le Corpus Textuel du Journal An-Nahar

Ce corpus1 comprend des articles publiés dans le journal libanais An-Nahar (La journée) de 1995 à 2000, stockés sous forme de fichiers HTML. Pour chaque année il y a 45,000 articles avec 24 million de mots. Chaque article inclut des informations telles que le titre, le nom de journal, la date, le pays, le type, et la page.

3.3. Le Corpus Al-Hayat

Le corpus Al-Hayat2 était développé au cours d'un projet de recherche de l'université d'Essex, en collaboration avec l'université Open. Il contient les articles du journal Al-Hayat enrichies par des informations additionnelles pour les rendre utiles pour les taches d'ingénierie de la langue et de développement des applications de recherche d'information. Les données ont été distribuées en sept bases de données à sujet spécifique conformément aux sujets traités par le journal Al-Hayat: Général, Voiture, Ordinateur, Nouvelles, Économie, Science, et Sport. Les balises, les nombres, les caractères spéciaux, et les ponctuations ont été supprimés. La taille totale du fichier est de 268 méga-octets. Cet ensemble de données contient 18639264 mots distincts dans 42,591 articles.

[1] Association européenne des ressources langagières, 2001.

[2] Association européenne des ressources langagières, 2001.

3.4. Le corpus «Gigaword Arabe»

Récemment, les corpus LDC, Al-Hayat et An-Nahar ont été agrégés dans une ressource commune intitulé Gigaword arabe [Maamouri & Cieri, 2002]. Par la suite, une quatrième source a été ajoutée, le contenu arabe Xinhua de l'agence de presse Xinhua (Consortium des Données Linguistiques, 2003). Gigaword arabe est un archive complet de données journalistique textuelles acquis par le LDC à partir de différentes sources d'informations arabes. Il consiste en 319 fichiers, avec une taille approximative de 1.1 giga-octets dans le format compressé; et de 4,34 giga-octets dans le format non-compressé, contenant un total de 391619000 mots. Tous les fichiers textuels de ce corpus ont été convertis dans le système d'encodage UTF-8.

3.5. Le corpus des textes parallèles arabe-anglais de l'O.N.U

Le corpus des textes parallèles arabe-anglais de l'O.N.U a été rassemblé, nettoyé, et aligné à partir du site Web de Nations Uni3 par [Xu et al., 2001] à la BBN. Le corpus contient 34575 paires de document et 3270200 paires de phrase. Les documents constituant ce corpus ont été publiés par l'O.N.U. entre janvier 1993 et décembre 1999. Un crawleur à objectif spécial réalise l'extraction des documents à partir du site Web de l'O.N.U. Les documents extraits, lesquels étaient dans le format Wordperfect ou Microsoft Word, sont convertis en fichiers plein texte UTF-8. Les auteurs ont alignés alors le corpus au niveau phrase avec un logiciel développé à la BBN. Le corpus aligné a été en outre nettoyé en supprimant les paires des document-phrase incorrectes.

3.6. Autres Corpus arabes

Bien que très petit quant à leurs taille, deux autres corpus sont moins mentionnés vue qu'ils ont été créés par des efforts individuels avec des systèmes à capacités limités dans le domaine de l'informatisation arabe. [Abu-Salem, 1992] et [Hmeidi, 1995] de l'Institut de Technologie d'Illinois (IIT) ont créé le corpus SACS. Il consiste en 242 résumés arabes rassemblés des débats de la Conférence Nationale en Informatique de l'Arabie Saoudite, contenant un total de 46968 mots. Chaque résumé comprend 36 champs y compris le titre, les auteurs, les sources, et le résumé. Chaque champ commence par trois caractères qui représentent le nom du champ suivis par un espace et un texte. [Hasnah, 1996] a créé le corpus Al-Raya, lequel comprend 187 articles sélectionnés du journal Al-Raya qui est publié à Qatar. Ce corpus contient un total de 219978 mots dont 30096 mots sont uniques.

4. Les Corpus de traditions prophétiques « Hadith »

4.1. Bref historique

Le hadîth rapporte les actes ou paroles du Prophète Muhammad (paix et salut sur lui), ainsi que des actes ou paroles d'autres personnes qui ont eu lieu en sa présence sans entraîner des réactions de rejet de sa part. L'ensemble des hadîths représente un corpus de traditions qui constitue la sunna (le chemin) du Prophète. Les sunnites se dénomment ainsi comme les « gens de la sunna et de la communauté». Les hadîths n'ont été mis par écrit que bien après le décès du Prophète (paix et salut sur lui). Comme les versets du Coran, ils étaient mémorisés et diffusés oralement. Certaines traditions rapportent que par crainte de confusion entre le Coran et les hadîths, le Prophète avait interdit la mise en écriture des hadîths. Les compagnons ont donc consigné le texte du Coran par écrit et ce n'est que plus tard, quand il fut devenu familier au plus grand nombre, que l'autorisation de mettre les hadîth par écrit fut donnée par le prophète.

Pourtant malgré cette autorisation les compagnons hésitèrent. Il faudra attendre le deuxième siècle de l'hégire pour que cette prévention disparaisse. Les premiers recueils de hadîths furent collectés à la fin du premier siècle par le calife Omeyyade 'Umar Ibn 'Abd al 'Azîz. Le choix fut rapidement fait de ne garder que les seuls hadîths munis de plus en plus systématiquement de la

[3] www.ods.un.org/ods

chaîne de transmission complète des rapporteurs afin de leur donner une crédibilité. Des recueils furent réalisés en les classant en fonction de celui qui les rapportait. Ces musnads sont donc intitulés en fonction du nom de tel ou tel compagnon : musnad d'Abû Bakr ou d'Abû Hurayra par exemple. Le nombre de hadîth est très important en fonction des sources. Chacun des sept compagnons en a rapporté plus de mille. Abû Hurayra dépasse à lui seul cinq mille hadîths. L'honneur d'avoir transmis le plus grand nombre de tradition parmi les compagnons revient à Abû Hurayra. Un de ses disciples Hammâm ibn Munabbih compila un recueil de traditions (Hadîths) apprises de son maître. Ce recueil a pour titre «Sahîfat Hammâm» et fut incorporé dans le second volume du «al-Musnad» d'Ibn Hanbal.

Al-Muwatta' « La Voie Aplanie », de l'Imâm Mâlik Ibn Anas (m 179 H) est considéré comme le premier recueil connu de hadiths en Islam après la SAHÎFA de Hammâm. Abû Zahra dans son livre sur l'Imâm Mâlik sa vie, et son époque, ses opinions et son Fiqh déclare : "l'histoire ne connaît pas de recueil de Hadîth et de Fiqh plus ancien qu'Al-Muwatta'...Aucun auteur avant Mâlik ne devait connaître la notoriété de ce dernier avec son Muwatta', qui nous est parvenu tel qu'il a été rédigé par son auteur. C'est pour cela que nous disons de lui qu'il est le premier recueil de Hadîth et de Fiqh à avoir été composé.[1]

Plus tard, le célèbre recueil "Al-musnad" de l'Imâm Ahmad Ibn Hanbal(m 241 H) rapportent les hadîths classés selon les compagnons qui les rapportent. Il regroupe tous les hadîths d'un compagnon dans un même chapitre quel que soit leur sujet. Il contient quarante mille hadîths sélectionnés sur une base de soixante seize mille.

La classification thématique des musannafs a été choisie au détriment de celle des musnad jugée trop confuse. Les grands recueils élaborés au troisième siècle ne firent pas l'unanimité dans la communauté, et il faudra attendre le siècle suivant pour que l'on reconnaisse leurs mérites respectifs et que la notoriété de certains s'impose définitivement comme des références en matière de compilation fiable.

Six recueils canoniques sont ainsi reconnus par les sunnites et deux sont particulièrement considérés : celui de Al Bukhârî (256/870) et de Muslim (261/875) qui ne recensent que des hadîths réputés sains ou authentiques c'est à dire Sahîh.

Al Bukhârî aurait entendu plus de 600 000 hadîth et mémorisé plus de 100 000 hadîth. Si l'on exclut les répétitions, son recueil contient 2762 hadîth. Cela signifie que les critères de validation d'un hadîth étaient particulièrement stricts en ce qui le concerne. Son ouvrage est organisé suivant les catégories du Fiqh (droit) mais inclus aussi d'autres thèmes comme les commentaires du Coran, des hadîth sur la création, les convenances, la société, le Paradis, l'Enfer, les sectes, le retour du Messie…Il est considéré par une majorité de sunnite comme l'ouvrage le plus important après le Coran.

Malgré des critères de sélection jugés moins stricts, celui de Muslim jouit aussi d'une grande notoriété. A côté de ces deux ouvrages (Sahîhayni), viennent d'autres recueils contenant des hadîths d'un degré de fiabilité inférieur. Ces recueils «sunan», sont par ordre de crédibilité ; Al-Tirmithî (279/893), Abû Dâwûd (275/889), Nasâ'î (303/915), et Ibn Mâjah (275/889). Ces six recueils Sahîhet sunan ont supplanté tous les autres. Ils ont fait l'objet de nombreux travaux de commentateurs et de multiples études.

Pourtant ils n'épuisèrent pas toute la matière du hadîth et d'autres recueils jugés tout aussi important furent réalisés par la suite. Ces ouvrages adoptèrent parfois une classification différente :

- Dictionnaire (en fonction de l'ordre alphabétique des maîtres, des pays, des tribus…)
- Suppléments rapportant des hadîth répondant aux critères de sélection d'Al- Bukhârî et non cités par lui.
- Des opuscules compilant des hadîths traitant d'un même thème.

Le corpus des hadîth s'est achevé à la fin du troisième siècle. Parmi ces ouvrages ceux de Abû Zakariyyâ' Yahyâ An-nawawî : « Le jardin des vertueux », anthologie thématique de hadîth illustrés de citations coraniques ou le recueil des « quarante hadîth ».

4.2. Problème de l'authenticité :

Entre le deuxième et le huitième siècle, la masse des hadîth en circulation a considérablement augmenté dont certains d'une origine forte douteuse. La source d'autorité que représente le hadîth a en effet rendu tentante sa falsification. Certains de ces hadîth ont pu être introduit dans leurs ouvrages par des auteurs réputés qui se souciaient moins de l'origine de ces hadîths que du fait que leur contenu leur paraissait conforme à la tradition et utile au croyant. Il était très important de préserver l'héritage en ne consignant que les hadîth sûrs. La méthode de ce discernement porte sur deux composantes :

- L'isnâd : la chaîne des transmetteurs successifs rapportant ce hadîth.
- Le matn : le texte proprement dit.

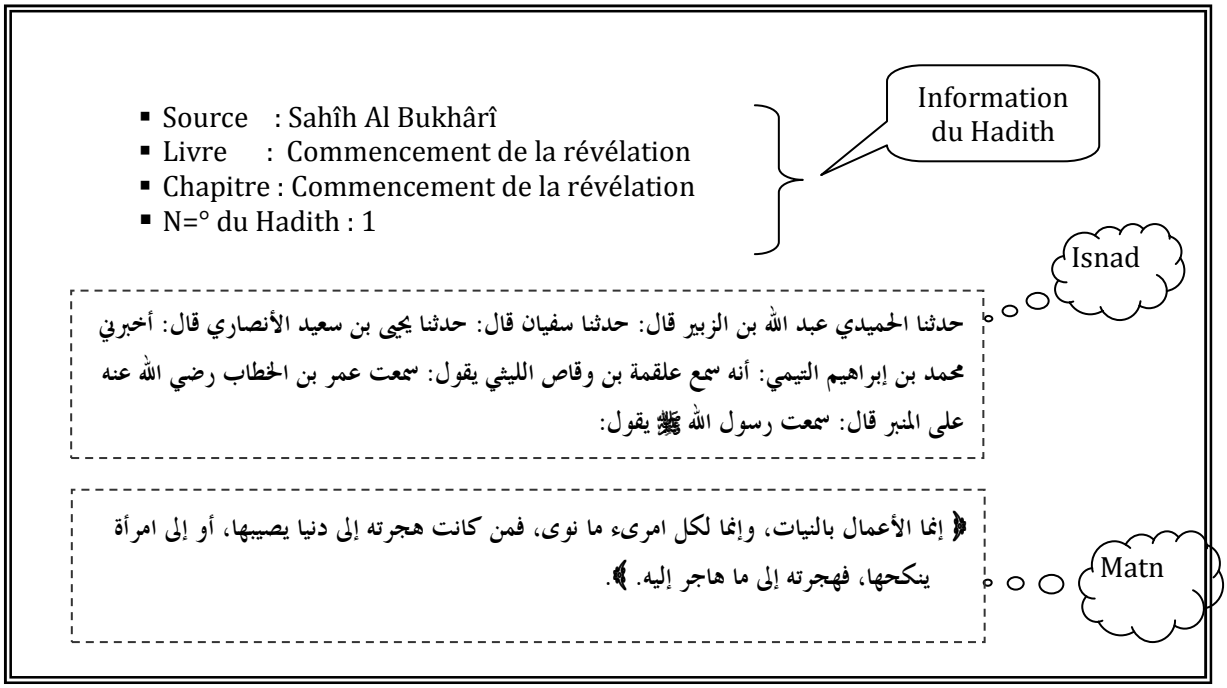

Fig. 1.1. Composants du Hadith.

4.2.1. L'isnâd

Concernant l'isnâd l'investigation va porter sur l'identité exacte des autorités figurant dans la chaîne de transmission, leur biographie et qualités morales, leur probité, leur bonne mémoire et leur exactitude. Ceci permet de repérer des homonymies ou de débusquer des incohérences, par exemple si deux transmetteurs placés côte à côte dans une même chaîne ont vécu à des époques différentes. En fonction des qualités citées plus haut, les transmetteurs sont classés en huit catégories selon un mérite décroissant depuis le «digne de foi» : thiqa jusqu'à celui dont la transmission est rejetée : «Matrûku-l-hadîth». Mais les traditionalistes ne sont pas toujours en accord sur la classification d'un même transmetteur. Une fois les différents transmetteurs repérés, reste à établir une classification des isnâds considérés dans leur ensemble c'est à dire en fonction de la validité globale de la chaîne :

- Isnâds ininterrompus.
- Isnâds interrompus quand il manque un ou plusieurs transmetteurs.

4.2.2. Le Matn

Cette science a pour objet le locuteur du hadîth, on distingue :

- Hadîth qudsi : Ces hadîths sont transmis par le Prophète mais inspirés directement par Dieu. Ils différent du Coran dans la mesure où dans le Coran le sens et la lettre proviennent de Dieu, tandis que dans le hadîth qudsi seul le sens vient de Dieu, la formulation appartenant au Prophète, il commence par : « le prophète a dit : Dieu dit ».
- Hadîth marfû‘ ou sharîf ou nabawî : rapporte les paroles, actes, approbations…du Prophète. C'est de lui dont on parle quand on dit hadîth sans autre précision ou qualificatif.
- Hadîth mawqûf est attribué à un compagnon et concerne ses actes ou paroles.
- Hadîth maqtû‘ est attribué à un suivant (génération suivant celle des compagnons).

4.2.3. Degré de validité

Le Coran déclare {vous avez dans l'envoyé de Dieu un excellent exemple} «sourate: 33, verset: 21». La transmission de la sunna nécessitait donc un tri entre ce qui était acceptable et le reste. Cette classification terminale se fait en trois grandes catégories :

- Sahîh : sain, authentique et exempts de déficiences. Son isnâd est ininterrompu et les autorités fiables et exactes.
- Hassan : bon, l'exactitude des transmetteurs est insuffisante.
- Da‘îf : faible, c'est dans cette catégorie que se placent les différents types de hadîths rejetés, le fictif (mawdû‘) étant le pire de tous.

Seuls les hadîth sahîh et hassan font autorité et sont susceptibles d'argumenter un point de doctrine ou de jurisprudence. Quant aux hadîths faibles, ils ne servent que s'ils sont appuyés par d'autres références plus solides ou s'ils concernent les bonnes choses (communément admise et qui ne contredisent pas les textes authentiques) (fadâilu al-a‘mâli).

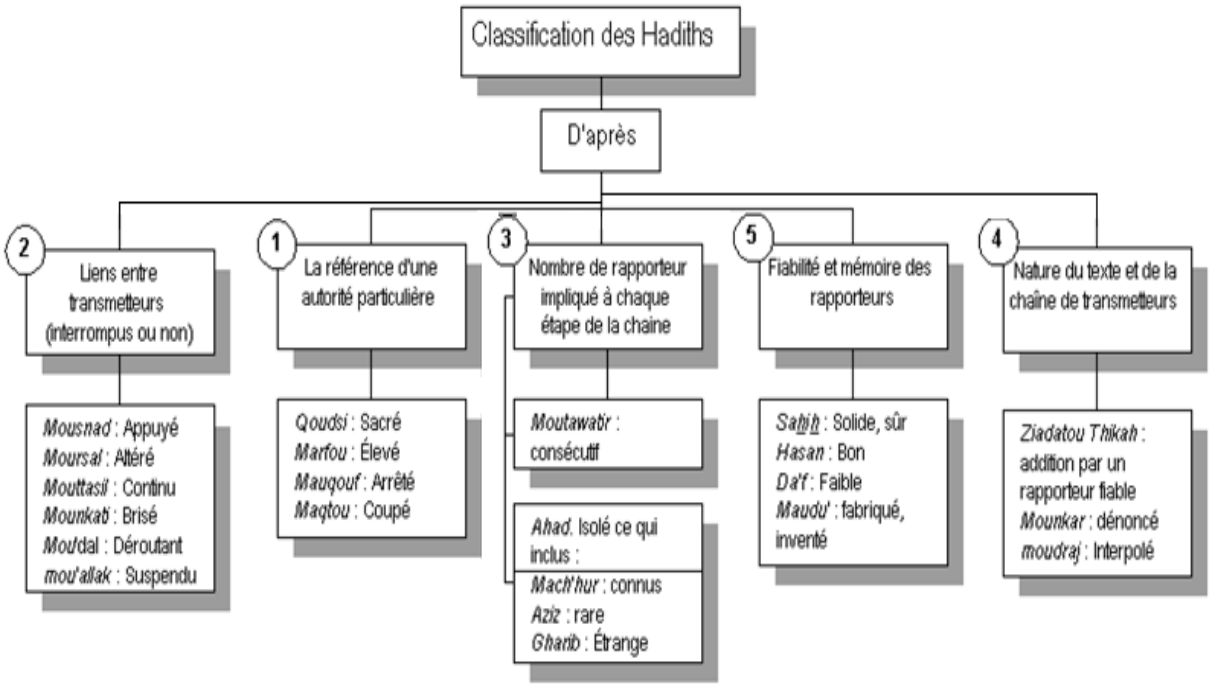

Fig. 1.2. Classification du Hadith.

4.2.4. Authentification du "Hadith"

Si on veut réaliser l'authentification du hadith numéro : 35/145/4553 du « Sunan Abû Dâwûd ». On doit suivre les étapes suivantes:

- Trouver le hadith dans sa source originale : « Sunan Abû Dâwûd ».
- Trouver le titre du Chapitre N° 35 : « les morales ».
- Trouver le titre de la Section N° 145 : « être debout pour quelqu'un ».
- Extraire les mots-clés du hadith.
- Déduire les thèmes partiels du hadith.
- Trouver l'ensemble des recueils de Sunnah où se trouvent des hadiths contenant ces mots-clés (recherche par mots-clés).
- Trouver l'ensemble des recueils de Sunnah où se trouvent des hadiths proche sémantiquement du hadith en question (recherche thématique).

Cette première étape connue sous le nom de "Âazou el hadith" représente la base du processus d'authentification, elle peut être réalisée par plusieurs méthodes:

- Le taraf du hadith: cette méthode utilise la première lettre du Hadith pour faire une recherche par ordre alphabétique dans les différents recueils prophétiques qui indexent les hadiths selon l'ordre alphabétique des lettres.
- Le matn du hadith: cette méthode utilise un mot du Matn pour faire une recherche par mots-clés dans les différents recueils prophétiques qui indexent les hadiths selon leurs mots. Ces recueils utilisent les noms ou les verbes pour représenter les hadiths.
- Le premier narrateur de la chaîne de transmission: cette méthode utilise le nom de premier narrateur du hadith pour faire une recherche dans les différents recueils prophétiques qui indexent les hadiths selon les noms des premiers narrateurs.
- Le thème du hadith : cette méthode se base sur la connaissance du thème de hadith pour faire une recherche dans les différents recueils prophétiques qui indexent les hadiths selon leurs thèmes.

L'application de ces méthodes va permettre de construire l'arbre des hadiths, l'authentification consiste donc à comparer les différentes chaînes de transmission des hadiths et de connaitre leurs degrés de véracité. Le résultat de cette étude est la connaissance du degré de véracité du hadith recherché. Un exemple d'authentification est donné ci-dessous pour le hadith numéro 4553 de «Sunan Abû Dâwûd»:

- سنن أبي داود، حديث رقم 4553 :

حدثنا أبو بكر بن أبي شيبة حدثنا عبد الله بن نمير عن مسعر عن أبي العنبس عن أبي العدبس عن أبي مرزوق عن أبي غالب عن أبي أمامة قال: ﴿خرج علينا رسول الله ﷺ متوكئا على عصا فقمنا إليه فقال لا تقوموا كما تقوم الأعاجم يعظم بعضها بعضا﴾

1- Recherche par mots-clés:

Les mots "مُتَوَكِّئًا", "عَصَا", "الأَعَاجِمُ" et "يُعَظِّمُ" sont lemmatisés vers leurs formes canoniques et utilisés pour effectuer une recherche par mots-clés. Les hadiths trouvés sont {21158, 21176} du « Mussnad Ahmed », et {3826} du « Sunan Ibn Mâjah ».

2- Recherche par thème :

Les sous-thèmes "القِيَامُ تَحِيَّةٌ لِلْقَادِمِ - être debout pour la salutation du venant " et " الأَمْرُ بِمُخَالَفَةِ المُشْرِكِينَ - l'ordre d'opposer les infidèles" sont utilisés pour effectuer une recherche par thème. Le résultat de cette recherche est le même que la recherche par mots-clés.

- مسند أحمد، حديث رقم 21158:

حدثنا بن نمير حدثنا مسعر عن أبي العنبس عن أبي العدبس عن أبي مرزوق عن أبي غالب عن أبي أمامة قال: ﴿خرج علينا رسول الله ﷺ وهو متوكئ على عصا فقمنا إليه فقال لا تقوموا كما تقوم الأعاجم يعظم بعضها بعضا فكأنا اشتهينا أن يدعو الله لنا فقال اللهم اغفر لنا وارحمنا وارض عنا وتقبل منا وأدخلنا الجنة ونجنا من النار وأصلح لنا شأننا كله فكأنا اشتهينا أن يزيدنا فقال قد جمعت لكم الأمر﴾.
حدثنا محمد بن عباد حدثنا سفيان حدثنا مسعر عن أبي عن أبي عن أبي منهم أبو غالب عن أبي أمامة عن النبي ﷺ مثله أو نحوه،

- مسند أحمد، حديث رقم 21176 :

حدثنا يحيى بن سعيد عن مسعر حدثنا أبو العدبس عن رجل أظنه أبا خلف حدثنا أبو مرزوق قال قال أبو أمامة: ﴿خرج علينا رسول الله ﷺ فلما رأيناه قمنا قال فإذا رأيتموني فلا تقوموا كما يفعل العجم يعظم بعضها بعضا قال كأنا اشتهينا أن يدعو لنا فقال اللهم اغفر لنا وارحمنا وارض عنا وتقبل منا وأدخلنا الجنة ونجنا من النار وأصلح لنا شأننا كله﴾

- سنن ابن ماجه، حديث رقم 3826 :

حدثنا علي بن محمد حدثنا وكيع عن مسعر عن أبي مرزوق عن أبي وائل عن أبي أمامة الباهلي قال: ﴿خرج علينا رسول الله ﷺ وهو متكئ على عصا فلما رأيناه قمنا فقال لا تفعلوا كما تفعل أهل فارس بعظمائها قلنا يا رسول الله لو دعوت الله لنا قال اللهم اغفر لنا وارحمنا وارض عنا وتقبل منا وأدخلنا الجنة ونجنا من النار وأصلح لنا شأننا كله قال فكأنما أحببنا أن يزيدنا فقال أوليس قد جمعت لكم الأمر﴾

Le résumé de ce processus est donné dans la figure ci-dessous:

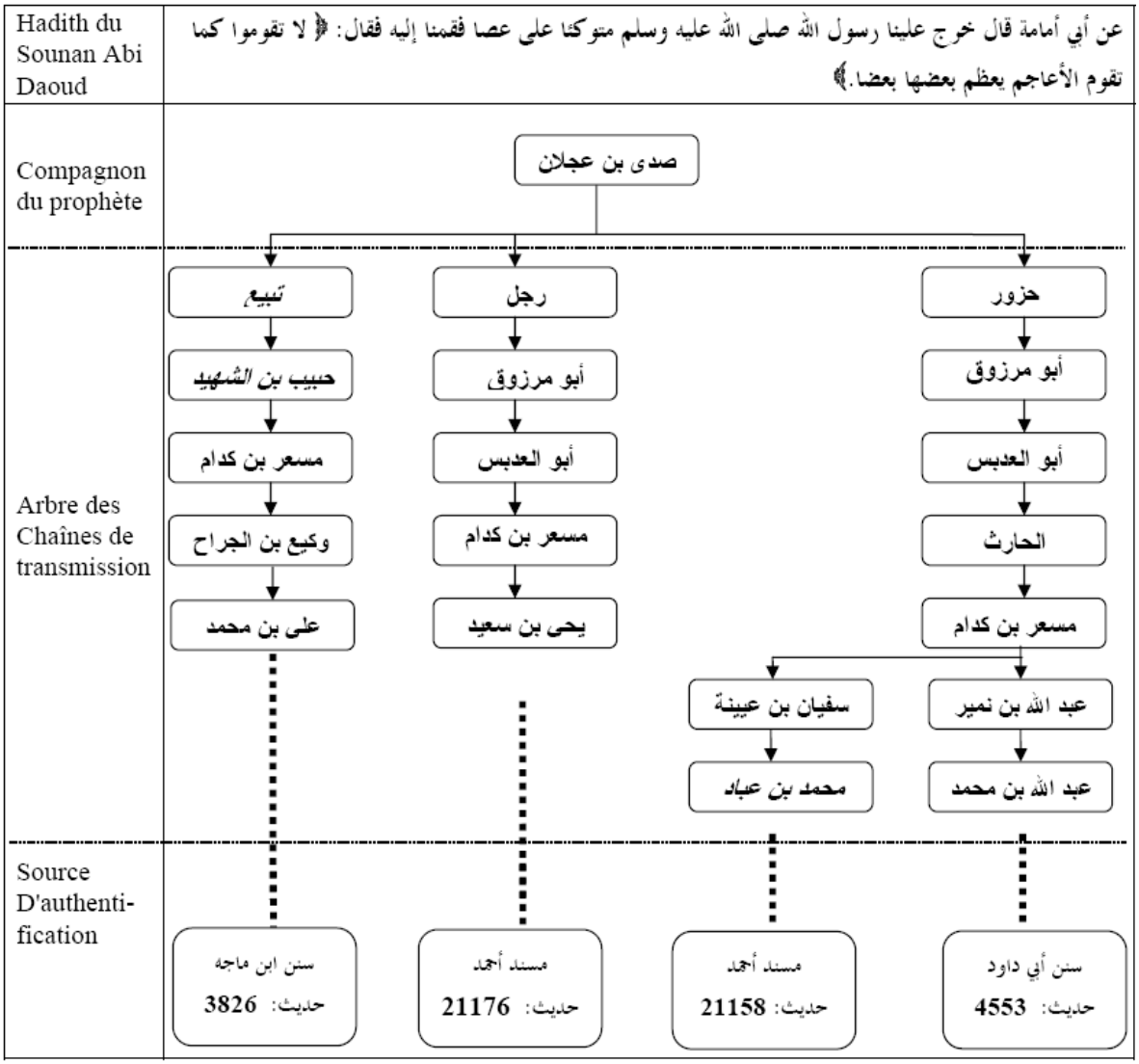

Fig. 1.3. Résultat du processus d'authentification.

4.3. Finalités des Corpus prophétique

La législation islamique consiste en l'ensemble des lois et des statuts promulgués par Allah en vue de régir ou d'organiser toutes les relations de l'humanité, à savoir les relations de l'homme avec Allah, avec lui-même ou avec les autres.

La pensée islamique est l'ensemble des connaissances et des sciences tirées du Message, telles que la pensée économique et politique, la philosophie, les Fondements de la Jurisprudence (Uçûl al-Fiqh), l'Éthique, l'exégèse etc. Le Saint Coran et la Sunnah du Prophète sont la source de la pensée, de la connaissance et de la législation islamique, comme l'affirme clairement le Livre d'Allah en différents endroits:

- «Il (le Prophète) ne parle pas sous l'empire de la passion. C'est seulement une Révélation qui lui a été faite.» (Sourate al-Najm, 53 : 3 - 4)

- «Prenez ce que le Prophète vous donne, et abstenez-vous de ce qu'il vous interdit...»

(Sourate al-Hachr, 59 : 7)

- «(le Prophète dit): Il ne m'appartient pas de le (le Coran) changer de mon propre chef: Je ne fais que me conformer à ce qui m'a été révélé.» (Sourate Yûnis, 10 : 15)

- «O vous qui croyez! Obéissez à Allah! Obéissez au Prophète et à ceux d'entre vous qui détiennent l'autorité. Portez vos différends devant Allah et devant le Prophète; si vous croyez en Allah et au Jour du Jugement, c'est mieux ainsi; et c'est le meilleur arrangement.» (Sourate al-Nisâ', 4 : 59)

- «Vous avez, dans le Prophète d'Allah, un bel exemple...» (Sourate al-Ahzâb, 33 : 21)

- «Dis: "Obéissez à Allah! Obéissez au Prophète!". S'ils se détournent, le Prophète n'est alors responsable que de ce dont il est chargé et vous n'êtes responsables que de ce dont vous êtes chargés. Si vous lui obéissez, vous serez bien dirigés, il incombe seulement au Prophète de transmettre en toute clarté ses messages.» (Sourate al-Nour, 24 : 54)

- «Chaque Prophète envoyé par Nous ne s'exprimait, pour l'éclairer, que dans la langue du peuple...» (Sourate Ibrâhîm, 14 : 4)

Ainsi, le Coran définit la Sunnah du Prophète et toutes ses paroles comme étant une partie intégrante du Message divin éternel. En se fondant sur ces données coraniques les sommités de la science et du savoir ont considéré la Sunnah comme la seconde source de la législation, de la pensée et de la culture islamiques.

5. Le Hadith et le traitement de l'information

5.1. 60 ans de travail sur le lexique des Orientalistes [Wensinck et al., 1936-1988]

Le travail monumental de publication entrepris dès 1936 par le professeur Arent Jan Wensinck, instigateur du projet de la Concordance et Indices de la Tradition Musulmane, (décédé en 1939) et qui devait durer jusqu'en 1969, était le résultat de très nombreuses années de recherches et de coopération internationale portant sur ce qu'on entend par la tradition musulmane. Ces «traditions», recherchées, recueillies, collationnées par multiples «traditionnistes» musulmans de différentes époques et de différents lieux, suggèrent, par leur masse et leur diversité mêmes, une confrontation et une mise en concordance.

S'étant cet objectif, A.J.W s'est appuyé sur un certains nombre de compilations de traditions, à propos desquelles un consensus s'était établi parmi les savants musulmans. Dès 1922 Wensinck avait publié dans les Communications de l'Académie Royale d'Amsterdam son intention de compiler une Concordance de la Tradition Musulmane. Ce projet, il avait déjà voulu le réaliser depuis longtemps. En effet, sa conception même datait d'avant 1916. Wensinck avait développé le projet sous l'influence, entre autres, des conseils de Christiaan Snouck Hurgronje. En 1922, le

travail allait déjà en bon train. Il fut alors décidé que l'on préparerait des concordances des Traditions des Six Livres canoniques, auxquels furent ajoutés trois autres collections: le Muwatta' de Mālik, les Sunan d'al-Dārimī, et le Musnad d'Ahmad ibn Hanbal, qui était particulièrement incommode à lire.

En 1927 Wensinck publia « *A Handbook of Early Muhammadan Tradition* », qui toutefois ne saurait être considéré comme un remplaçant de la Concordance. Le Handbook ne donnant pas les contextes et isnāds des Traditions relatives, mais seulement, en ordre alphabétique, les sujets paraphrasés en anglais, suivis des références aux places où l'on peut les trouver dans les collections de Traditions dépouillées. Comme il va de soi, les mêmes fiches servirent de base tant pour le Handbook que pour la Concordance. La grande différence entre les deux livres était que la Concordance a la prétention de donner, en arabe, non seulement les renvois mais aussi les mots importants et substantiels dans leur contexte. Tout cela ne devait rester d'ailleurs qu'un idéal, la Concordance n'étant pas conçue comme un instrument d'études linguistiques, mais plutôt comme une aide pour les étudiants qui désiraient trouver facilement le contenu des collections canoniques de Traditions Islamiques.

Il devenait alors possible, grâce à un classement des mots et idées par ordre alphabétique; de retrouver l'existence, l'emploi et la localisation bibliographique des traditions comportant ce terme ou cette idée. Comme ces traditions (*hadîth*), appelées dans leur ensemble, Tradition ou Voie (*sunna*), représentent, à coté du Coran, la source textuelle sacrée de l'islam, on devine l'intérêt à la fois scientifique que religieux du travail accompli par Wensinck., ses collaborateurs et successeurs.

Les collaborateurs, invités par Wensinck à dépouiller les livres canoniques, avaient-ils la permission d'omettre les mots les plus communs, ce qu'ils firent nécessairement d'une manière subjective. Ce même élément subjectif joua un rôle important dans la phase de rédaction. Seuls les mots sans aucun but pour l'ouvrage furent omis tout à fait, les mots assez communs étant donnés sans leur contexte. Ceci avait un but non seulement pratique mais également économique, car si la Concordance avait été achevée sans omettre un mot ou une phrase, l'ouvrage aurait eu plusieurs fois l'ampleur actuelle. Personne n'en aurait pu tirer profit et les frais d'impression auraient été insurmontables.

Parmi les successeurs de Wensinck, Witkam et Ravan on produit en 1988 un volume d'index (« indices ») qui complète et enrichit cet incomparable outil de travail, lequel, dans l'univers spécialisé de l'islamologie, trouve une place de choix à côté de l'*Encyclopédie de l'islam* et de la *Geschichte der Arabische Litreratur* (G.A.L.) de Brokelman.

5.2. Etat de l'art des recherches à caractère commercial

Avec les progrès technologiques et l'émergence de ce que l'on appelle l'ère de l'information, et à l'aide des ordinateurs, des logiciels, et du réseau Internet, de nombreux efforts ont été consacrés au service de la tradition prophétique pour faciliter l'accès aux actes et paroles du Prophète Muhammad (paix et salut sur lui) pour l'ensemble de la nation musulmane dans tous les coins de la terre. Dans cette section, nous présentons les logiciels les plus connues dans le domaine de traitement informatique de la tradition prophétique.

5.2.1. La compagnie « *Harf* » pour les technologies de l'information[4]

La société «*Sakhr computer program*» a été fondé en 1982, comme l'une des branche de la société «*Universal electronics*». En 1985, la société *Sakhr* a crée un département au nom du «*Centre du Patrimoine Islamique*» concerné par la production de programmes islamique, et s'est connue par la suite sous le nom de «*Société Harf pour la Technologie de l'Information*». Elle est l'une des premières sociétés travaillant dans ce domaine, car elle était la première à mettre les principales sources islamiques sur des supports électroniques.

[4]Site de la société *Harf* : http://www.harf.com.

L'*Encyclopédie des Traditions Prophétiques*[5] «Les Neufs Livres ou Al-kotôb Al-Tissâa» est l'un des programmes les plus importants produits par cette société. C'est une bibliothèque complète des traditions prophétiques existants dans les neuf livres, à savoir: le Sahîh de Al Bukhârî, le Sahîh de Muslim, les Sunan de Darimî, les Sunan de Tirmithî, les Sunan de Abû Dâwûd, les Sunan de Nasâ'î, les Sunan d'Ibn Mâjah, le Muwatta' de Mâlik et le "Musnad" de Ahmad avec leurs commentaires respectives. Cette bibliothèque contient plus de 62000 tradition prophétique, sur un nombre de 25 mille pages.

Fig. 1.4. Interface du logiciel « Encyclopédie des Neufs Livres ; Al-Kotôb Al-Tissâa ».

5.2.2. Le center *Turath* des recherches en informatique[6]

C'est une société jordanienne créée en 1993 dans un objectif de servir les livres du patrimoine islamique de et d'en faciliter l'accès en utilisant les nouvelles technologies. Le Centre a un grand nombre de logiciels dans les différentes branches de la culture islamique. Le premier et le plus célèbre programme du centre a été annoncé en 1997, l'Encyclopédie Dorée de la Tradition Prophétique et ses Sciences - la deuxième version -, comprenne plus de 600 volumes et livres islamiques, elle intègre le service d'authentification automatique des traditions prophétiques. Cette encyclopédie couvre l'authentification automatique de plus de 200000 textes prophétiques avec leurs chaînes de transmission respectives, avec une base de données de biographies des narrateurs qui contient plus de 150000 biographies.

[5] http://www.harf.com/products/arb/hadeth.htm.
[6] Site de la société *Turath*: www.turath.com/arabic/index.php

Fig. 1.5. Interface du logiciel « l'Encyclopédie Dorée des Traditions Prophétiques».

5.2.3. La Société Al-Âariss[7]

Sous le slogan "Vers une bibliothèque complète de la science islamique", la société Al- Âariss a construit sa démarche de développement des programmes islamiques. La société dispose d'une série de programmes dans les différents domaines : bureautique, éducatif, familiale, scientifique et islamique. Le programme le plus important de cette compagnie est nommé *Bibliothèque Moderne des Traditions Prophétiques*. Ce programme, dans sa cinquième version, comprend plus de 2100 volumes et livres dans le domaine de la tradition prophétique et ses sciences. En plus de plusieurs livres de biographies, de langue et une dizaine de dictionnaires. Le nombre de titres figurant dans cette encyclopédie est de 359 titres.

[7] Site de la société Al- Âariss: www.elariss.com.

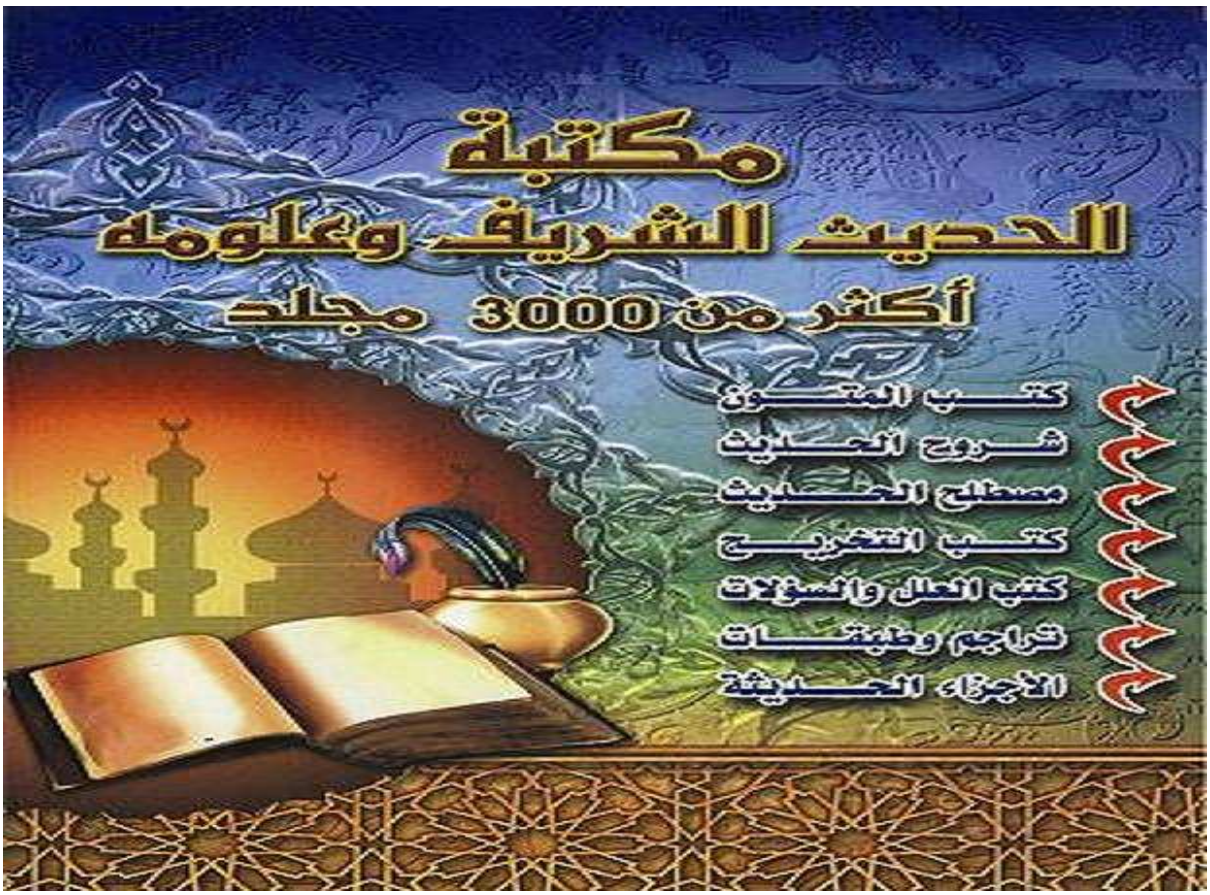

Fig. 1.6. Interface du logiciel « Bibliothèque Moderne des Traditions Prophétiques».

5.2.4. La société Abdelatif informatique

Cette société a été crée en 1990 comme une société commerciale, en 1996, elle s'est dirigé vers le domaine informatique par la mise en place de quelques programmes administratifs, en 1997, elle a commencé le travail dans le domaine des programmes islamique. En 1998, la société a produit son programme appelé "*Encyclopédie de l'Etudient Musulman* " pour la première fois dans le monde islamique. Par la suite, la société a continué de créer des programmes islamiques comme la *Biographie du Prophète,* l'*histoire islamique,* le *Coran,* les *Traditions Prophétiques*, la *Jurisprudence*, l'*Encyclopédie de Ibn Al-Kayim* et d'autres.

En 2000, la société a réussi à programmer l'authentification automatique des traditions prophétiques, et malgré l'existence de quelques erreurs, le programme est fonctionnel d'un pourcentage de 95%. Le plus important des programmes de cette société est l'*Encyclopédie Aisée de la Tradition Prophétique et ses Sciences*, elle comprenne environ 300 volumes et livres, ce programme offre la possibilité d'authentifier automatiquement tous les versets et les traditions existantes dans les bases de données des autres programmes de la compagnie. Le programme offre aussi la possibilité de dessiner automatiquement l'arbre des chaînes de transmissions de chaque tradition de l'encyclopédie.

Fig. 1.7. Interface du logiciel « Encyclopédie Aisée de la Tradition Prophétique et ses Sciences».

5.2.5. Le projet du logiciel *Muhaddith*[8]

L'idée de ce projet a commencé depuis une quinzaine d'années par un groupe d'étudiants de l'ancienne *Dar Al-Hadith*, de la société «madrasa », situé à Washington, États-Unis. Les responsables de ce projet sont des membres de *Dar Al-Hadith Al-Ashrafiya*, situé dans la ville de Damas. Le projet s'intéresse au développement et à la mise à jour du programme *Al-Muhaddith*, c'est un programme libre qui peut être téléchargé à partir d'Internet avec tous les fichiers de livres. L'objectif de ce projet, comme il est mentionné dans le site du programme, est le transfert du plus grand nombre possible du patrimoine islamique sous la forme électronique.

[8] Site du projet: www.muhaddith.org.

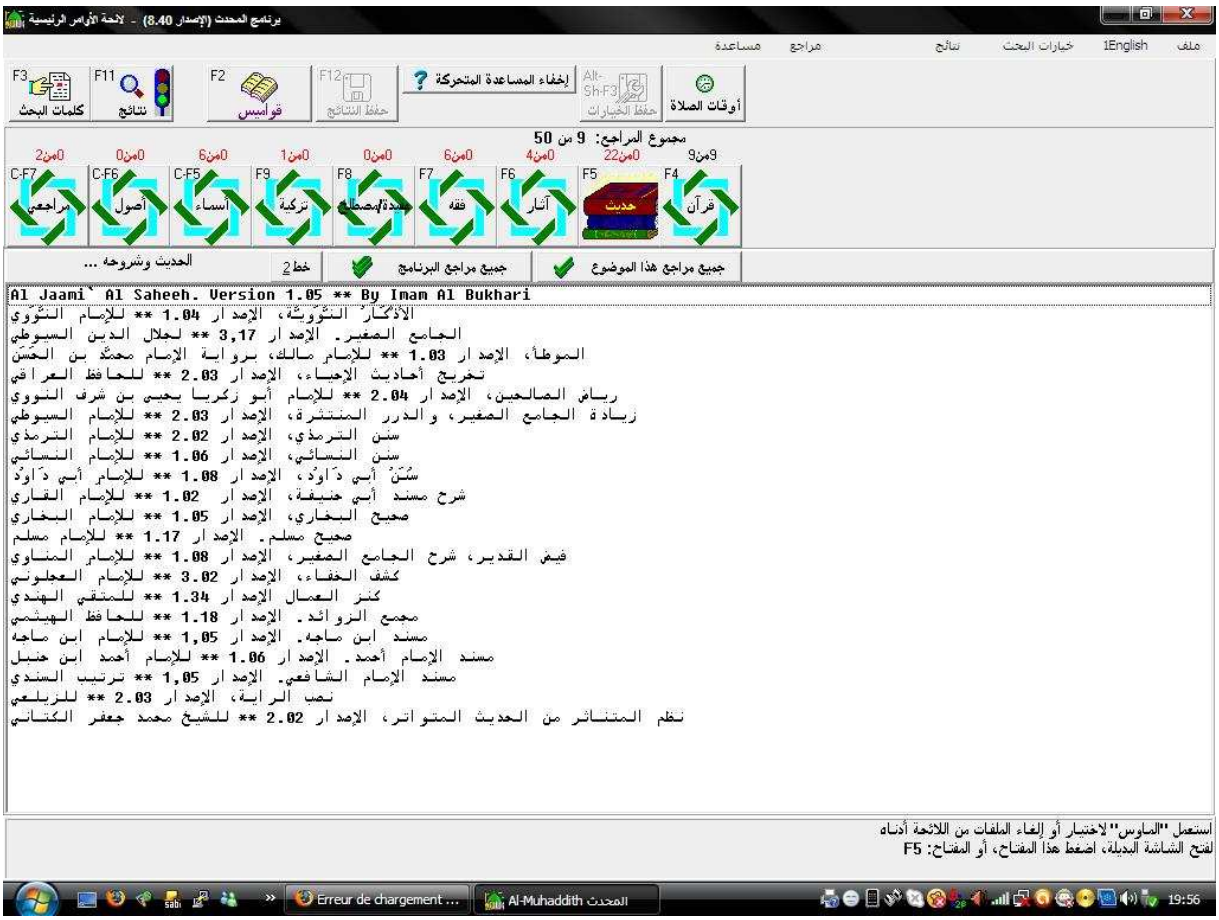

Fig. 1.8. Interface du logiciel « Al-Muhaddith».

5.2.6. Le Projet *Sunna* de la Fondation du Thesaurus Islamique[9]

Bien que l'usage des technologies de l'information ouvre de nouvelles perspectives dans le domaine de l'étude du hadith, le champ continue à souffrir une pénurie de projets sérieux dirigés plutôt par des universitaires que par des commerciaux. Pour remédier à ceci, la Fondation du Thesaurus Islamique, une fondation charitable à but non lucratif, enregistré dans le Duché de Liechtenstein, a été fondé en 1989 à l'instigation du Shaykh *al-Hussayni Hashim*, vice-recteur d'Al-Azhar au Caire. La fondation a un centre en Egypte et un autre en Royaume-Uni, et un ensemble de savants et de spécialistes en informatique affiliés à des organisations dans d'autres pays. L'objectif de la fondation est de servir la littérature dans le domaine des traditions prophétiques. Elle cherche à augmenter le nombre de publications prophétiques dans les formes imprimé et électronique, avec un niveau plus haut d'excellence, permettant à cette littérature de faire office à un modèle pour traitement et l'édition des autres corps substantiels de la littérature islamique et mondiale.

La fondation a publié son produit *Encyclopédie de Hadit* en 2001, Cette première collection de traditions prophétiques, consiste en 19 volumes, incluant les Sept Grandes collections de Hadith en arabe (Sahîh de Al Bukhârî, Sahîh de Muslim, Sunan de Tirmithî, Sunan de Abû Dâwûd, Sunan de Nasâ'î, Sunan d'Ibn Mâjah et le Muwatta' de Mâlik). En plus, la collection inclut une réimpression de cent ans de l'édition *Sultaniyya* du Sahîh de Al Bukhârî, dans trois volumes publiés par la maision Bulaq avec leurs commentaires marginaux précieux. La collection inclus aussi deux volume d'indices analytiques de tous les textes imprimés, intitulé *Maknaz al-*

[9] Site du projet: www.fonsvitae.com/thesaurus.html.

Mustarshidin qui inclut parmi ses références tous les traditions prophétiques pertinentes comptées dans le livre *Tuhfat al-Ashraf* d'al-Mizzi.

Accompagné du volume imprimé, l'encyclopédie inclut une base de données révolutionnaire sous forme d'un CD-ROM qui contient toutes les collections des traditions prophétiques fournies dans forme imprimée et auxquelles on peut y accéder par 21 différents genres de recherches de degrés variables de complexité.

Fig. 1.9. Interface du Site web du projet « Encyclopédie de Hadit».

5.3. Etat de l'art des recherches à caractère scientifique

5.3.1. Les travaux de H. Hammed and H. Tolba [Hammed & Tolba, 2000]

Cette recherche a utilisée des techniques de raisonnement adapté de l'intelligence artificielle dans la compréhension de la langue arabe. Le but de cette recherche était de développer un utilitaire basé sur la logique du raisonnement temporelle simple pour comprendre la langue arabe. Les chercheurs ont vérifié leur méthode par la compréhension du Sanad (Chaîne de transmission) du *Hadith* dans la langue arabe.

5.3.2. Les travaux de Hassan Mudhafar Rizzo [Rizzo, 2004]

Cette recherche vise à l'instauration d'une logique prophétique qui profite des capacités immenses des ordinateurs pour entamer l'informatisation de la tradition prophétique dans les domaines de l'étude des chaînes de transmission (*Assaneed*) et la critique des textes prophétiques (*Moutoun*) ainsi que l'énonciation des sources de lésion et des raisons qui touchent à l'authentification de textes prophétiques. Cette étude a été basée sur des systèmes experts utilisant les principes de base de l'ingénierie de connaissance pour la formulation d'un modèle informatique simulant les méthodes de travail des savants et des experts de Hadith. L'idée était de transformer les données existantes dans les livres du Hadith en informations classées selon des catégories thématiques, les informations sont aussi transformées en connaissances utilisables dans le processus d'authentification des textes prophétiques et de leurs chaînes de transmission. L'approche des bases de connaissance basées sur le mécanisme des règles d'inférence a été utilisée dans l'opération de transformation des informations en connaissances

précises. L'auteur a inspiré de l'ensemble de règles prophétiques citées par les *imams* experts de hadith dans leurs livres pour formuler la liste des règles du modèle de connaissance. Les résultats préliminaires de l'application du modèle de connaissance ont montré une grande nécessité à un volume important d'efforts d'archivage et d'indexation pour la création des bases de données utilisés comme une plateforme pour le système de connaissance.

5.3.3. Les travaux de Mohammed Naji Al-Kabi, Ghassan Kanaan, Riad Al-Shalabi, Saja I. Al- Sinjilawi et Ronza S. Al-Mustafa [Al-Kabi et al., 2005]

Cette étude explore l'implémentation d'une méthode de classification de texte pour classer les traditions prophétique en se basant sur l'hiérarchie de Sahîh de Al Bukhârî. Cette méthode adopte la technique TFIDF pour calculer les poids des termes dans les vecteurs des documents (Hadiths). Après la transformation de ses termes vers leurs formes canoniques correspondantes (racine), le texte prophétique est classé dans l'un des huit chapitres (classe) du Sahîh de Al Bukhârî. Un terme avec un poids élevé est considéré comme un bon descripteur pour un chapitre particulier, ceci veut dire que l'apparition de ce terme est très fréquente dans le chapitre, mais elle est rare dans le corpus entier de Sahîh de Al Bukhârî. Initialement, un ensemble d'entrainement a été utilisé pour la phase d'apprentissage, ensuite un ensemble de test a été utilisé pour évaluer l'exactitude du classifieur. L'exactitude moyenne pour cet échantillon était approximative de 83.2%.

5.3.4. Les travaux de Mohammed Naji Al-Kabi et Saja I. Al- Sinjilawi [Al-Kabi et Al-Sinjilawi, 2007]

Cette recherche consiste à décrire la conception et l'implémentation d'une nouvelle méthode convenable pour la classification des traditions prophétiques en langue arabe. Cette méthode a été mise en œuvre sous Microsoft Visual Basic 6.0. Le but de cette étude est de trouver la méthode optimale qui peut être utilisée pour classer des textes prophétiques parmi les six méthodes (produit intérieur, cosinus, Jaccard, Dice, Bayésien Naïfs, et Euclidien). Un vecteur du document a été utilisé pour calculer et comparer quatre coefficients associatifs différents du modèle de l'espace vectoriel (VSM) basé sur le Produit Intérieur, Cosinus, Jaccard, et Dice. Les auteurs ont trouvés que la mesure Cosinus a dépassée les trois autres coefficients associatifs du modèle vectoriel. Les auteurs ont aussi comparés l'efficacité des mesures Cosinus, Bayésien Naïf, et Euclidien pour classer des textes prophétiques. Les résultats expérimentaux montrent que le classifieur Bayésien Naïf dépasse légèrement les autres méthodes. L'étude a montrée que des méthodes de calcul de similarité peuvent être combinées avec des classifieurs pour améliorer l'exactitude générale du système de classification. L'algorithme Bayésien Naïf a été utilisé comme classifieur combiné avec la méthode Cosinus pour déterminer la ressemblance entre un Hadiths classé et son chapitre (classe) correspondant dans le livre de Sahîh de Al Bukhârî.

5.3.5. Les travaux de Syed Irfan Hyder et Syed Ghazanfar Ahmad [Syed Ifran et Syed Ghazanfar, 2008]

Les travaux de recherche de ces deux auteurs visent à définir un graphe de représentation théorique des chaînes de narrateurs de traditions prophétiques « Hadiths » et une structure d'une base de données alignée convenable pour le stockage des données biographiques des narrateurs ainsi que d'autres événements historiques. Leur hypothèse était que l'usage des concepts informatiques comme la recherche algorithmique, l'interrogation des bases de données, les entrepôts de données les techniques avancées de fouille de données permettait d'assister d'une manière fiable les recherches faites dans le domaine du Hadith et les recherches liées à l'histoire et à la littérature islamique. Le résultat est de mettre la vaste littérature prophétique sous un format convenable pour des vérifications et des analyses en utilisant des techniques computationnelles faisables, et donc certaines classes d'analyse qui ont été considérés auparavant comme intraitable, peuvent devenir faisables en utilisant des technologies courantes.

5.3.6. Les travaux de M.Ghazizadeh, M.H.Zahedi, M.Kahani et B.Minaei Bidgoli [Ghazizadah et al., 2008]

Cette étude a essayé de montrer que le modèle théorique existant dans la science islamique peut aider à distinguer les Hadiths valides de ceux qui sont invalides. Les auteurs ont confirmés que la "*Science de Hadith*" avec la "*Science des Rejals*", qui est concentré sur l'examen du caractère de ceux qui ont rapportés le Hadith, peuvent contribuer ensemble pour prouver la validité de Hadith. L'objectif principal de cette étude était de déterminer le taux de validité d'un Hadith à travers un système flou en ce qui concerne quelques uns des paramètres. D'après le point de vue d'un expert du domaine, la base de connaissance a été conçue et le les règles essentielles ont été extraites. Le système a été implémenté en utilisant des logiciels pour les systèmes experts. Les échantillons de données prises à partir du volume1 du livre "*Al-Kafi*" ont été insérés dans la base de connaissances pour une estimation au moyen des informations documentaires. Les résultats déduits par le système expert conçu ont été comparé avec les points de vue de l'expert et la comparaison a montré que le système a été correct dans 94% des cas.

5.3.7. Les travaux de Manar Alkhatib [Alkhatib, 2010]

Dans cette étude, l'auteur a comparé l'efficacité de quatre algorithmes différents d'apprentissage automatique pour classer le Hadith dans 8 livres sélectionnés du Sahîh de Al Bukhârî. Les algorithmes d'apprentissage automatique utilisés sont : l'algorithme Rocchio, l'algorithme du K-PPV (K - Plus Proche Voisins), l'algorithme de Bayes Naïf et l'algorithme MVS (Machines à Vecteurs Supports). La technique TF_IDF a été utilisée pour le calcul des fréquences relatives des termes. Les documents prophétique ont été divisés en deux parties, 75% des documents (1350 Hadiths) ont été utilisés comme données d'apprentissage (construction du classifieur) et 25% des documents (150 Hadiths) ont été utilisés pour tester l'exactitude des modèles résultants.

5.4. Vers un lexique général et paramétrable, vers une bibliothèque électronique des traditions prophétiques

L'importance du Hadith aux yeux des musulmans lui permet d'être l'un des plus importants domaines de développement des logiciels en tous genres qui existent autour de lui, notamment pour l'étude des chaînes de transmission (*îlm Al-Assaneed*), l'étude des textes prophétiques (*îlm Al-Motoun*) et l'étude des narrateurs (*îlm Al-Rijal*). Après une étude approfondie des travaux et des recherches faites dans le domaine de traitement informatique des traditions prophétique, nous apercevons l'absence d'un modèle standard pour l'extraction de connaissances dans les bases de données prophétiques, nous constatons aussi, que les systèmes de recherche d'informations prophétiques existants sont incapables de fournir les informations suffisantes pour l'accomplissent de la tâche de prise de décision par les experts du domaine.

En effet, l'application des techniques de la fouille de textes sur les corpus prophétiques apporte de nombreux avantages aux applications qui concernent le traitement informatique des documents Prophétique. Elle permet en particulier d'automatiser une partie des traitements effectués sur les documents Prophétiques. Elle contribue, également, à améliorer la portabilité en fournissant un formalisme commun pour la représentation de ces documents sous forme d'une bibliothèque électronique. La manipulation du Hadith à l'ère de l'informatique nécessité aussi d'établir des liens entre les documents Prophétique, Le Coran, la science de Jurisprudence et les sciences de la langue arabe. Ce qui ne peut être rendu possible que par l'élaboration d'un lexique générale paramétrable sous forme d'un réseau lexical ou conceptuel ou sous forme d'une ontologie pour les documents prophétique.

A présent, notre recherche a pour but d'étudier les nouvelles méthodes de classification et de segmentation des connaissances dans les bases de données textuelles telles que celles des Traditions Prophétiques (Hadith). Ces méthodes représentent un processus de structuration de l'ensemble des informations et des réponses fournies par un système de recherche d'information. La classification thématique permet à l'utilisateur d'orienter son exploration en fonction des thématiques générales des classes et d'accéder ainsi plus rapidement au but de sa recherche. La segmentation permet de présenter à l'utilisateur les segments textuels jugés pertinents et de mieux positionner certains documents longs dans lesquels l'information recherchée ne constitue que l'une des thématiques abordée. Cela s'applique notamment à des données en quantité trop importante pour qu'une étude visuelle soit possible comme dans le cas du corpus du Hadith, par exemple.

6. Conclusion

Dans ce chapitre, nous avons étudié les Corpus Prophétiques, leur historique et leurs caractéristiques. Nous avons exploré les recherches faites dans le domaine de traitement d'information dans les corpus prophétiques et nous avons constaté le besoin d'un système d'extraction de connaissance à partir des bases de données prophétiques, qui facilitera l'indexation et l'interrogation des documents prophétique; nous avons aussi constaté la nécessité d'intégrer au système, des modules pour la classification et la segmentation thématiques des documents recherchés. Dans le reste de cette première partie, nous allons exposer les résultats expérimentaux de nos recherches pour la mise en œuvre d'un système d'extraction de connaissances prophétiques basé sur une approche de fouille de données textuelles. Dans la deuxième et la troisième partie de cette thèse, des méthodes de classification et de segmentation des documents sont proposées et évaluées dans le contexte de structuration des réponses d'un système de recherche d'information.