Extraction automatique de fréquences lexicales en arabe et analyse d'un corpus journalistique avec le logiciel AraConc et la base de connaissances DIINAR.1

Ramzi Abbès, Joseph Dichy

Université Lumière Lyon 2, ICAR-CNRS
ramzi.abbes@univ-lyon2.fr – joseph.dichy@univ-lyon2.fr

Abstract

This paper presents the reader with an experiment in extracting lexical frequency lists in Arabic, from a "raw" newspaper corpus of around 2 million words. The software used was one of the analyzers founded on the **DIINAR.1** knowledge database (*DIctionnaire INformatisé de l'ARabe, version 1*), **AraConc**, which has been devised to extract concordances and frequency lists in Arabic by R. Abbès within the SILAT research team. (The SILAT group, "*S*ystèmes d'information, *I*ngénierie et *L*inguistique *A*rabe, *T*erminologie", is included in the ICAR Lab, CNRS/Université Lumière-Lyon 2 and ENS-LSH.) The frequency lists have been built to serve as a basis for experiments in cognitive psycholinguistics related to the "masked priming" paradigm that have been developed at LPC (Laboratoire de Psychologie Cognitive, CNRS and Université de Provence, directed by Jonathan Grainger). Word frequency is a crucial parameter of items selected for experiments in this paradigm. Authors account for the compilation of the requested frequency lists. Many figures related to ambiguity rates, linguistic phenomena and graphic realisations that need to be taken into account in the statistical analysis of a 2 million-word newspaper corpus, will be given. Results include a first general characterisation of Arabic newspaper corpora considered from a statistical standpoint.

Résumé

Ce travail présente une expérience d'extraction de listes de fréquences lexicales en langue arabe à partir d'un corpus journalistique brut de deux millions de mots (du quotidien *Al-Hayât*) au moyen du logiciel **AraConc** d'extraction de concordances et de listes de fréquences en arabe, réalisé par R. Abbès. C'est l'un des outils développés autour de la base de connaissances **DIINAR.1** (*DIctionnaire INformatisé de l'ARabe, version 1*) par l'équipe SILAT (« Systèmes d'information, Ingénierie et Linguistique Arabe, Terminologie ») au sein du laboratoire ICAR, CNRS/Université Lumière-Lyon 2 et ENS-LSH. Ces listes ont été extraites pour servir à des expérimentations en psycholinguistique cognitive menées au Laboratoire de Psychologie Cognitive (CNRS/Université de Provence, dir. Jonathan Grainger) dans le paradigme de l'amorçage rapide. Celles-ci ont recours au paramètre de la fréquence relative des mots. Nous rendons compte du travail de réalisation de ces listes et présentons plusieurs résultats chiffrés concernant les taux d'ambiguïté, certains phénomènes linguistiques et les réalisations graphiques à prendre en compte pour l'analyse statistique d'un corpus journalistique de 2 millions de mots. Ces résultats incluent une première caractérisation générale des corpus journalistiques arabes d'un point de vue statistique.

Mots-clés : TAL arabe, base de données DIINAR.1, AraConc, fréquences lexicales, corpus, concordances, presse arabe, pratiques d'écriture, corpus journalistiques en arabe, psycholinguistique.

1. Introduction

L'importance des listes de fréquence lexicales dans l'enseignement ou la lexicographie a été fortement soulignée dès le début du XX^{e} siècle. Aux lexiques de référence parus dans les années cinquante pour l'anglais [West 1953] ou le français [Gougenheim 1958] répondait en

arabe la liste de fréquences établie à partir d'un corpus de presse par Brill [1940], et reprise par Landau [1959], qui a ajouté un corpus de textes de littérature contemporaine.

Plusieurs langues disposent de corpus écrits de référence « nettoyés », harmonisés et étiquetés, qui sont des sources cruciales pour les fréquences lexicales. Au célèbre *Brown Corpus* ont succédé l'American National corpus (www.americannationalcorpus.org) et le British National Corpus (www.natcorp.ox.ac.uk). En français FRANTEXT (http://atilf.atilf.fr/frantext.htm) remplit la même fonction. De pareils outils manquent pour l'instant en arabe, mais différents projets de constitution de corpus annotés devraient aboutir dans les prochaines années à des résultats du premier intérêt[1], si certains problèmes (catégories, fonctions, relations lexico-grammaticales) trouvent des solutions qualitatives satisfaisantes. Ce travail s'est appuyé sur une partie du corpus brut de 10 millions de mots compilé au sein du projet DIINAR-MBC[2]. L'interrogation s'est faite au moyen du logiciel AraConc [Abbès 2004a et 2004b] qui fait appel à la base de connaissances lexicales DIINAR.1 (DIctionnaire INformatisé de l'ARabe, version 1 [Dichy, Braham, Ghazali, Hassoun 2002] – www.elda.org et http://diinar.univ-lyon2.fr).

2. Objectifs de l'expérimentation informatique menée sur un corpus arabe

La démarche expérimentale présentée ici relate l'analyse automatique d'un corpus journalistique contemporain, en réponse à des besoin provenant de l'un des champs de la psychologie cognitive, qui élabore des protocoles expérimentaux pour mettre en évidence l'influence de certains paramètres sur la reconnaissance des mots. Les tests effectués auprès de sujets humains se basent sur des échantillons de mots dans lesquels la fréquence joue un rôle crucial [Grainger 2003]. Le Laboratoire de Psychologie cognitive (LPC, CNRS/Université de Provence, dirigé par Jonathan Grainger), a conçu en collaboration avec nous, puis mené, à l'École normale de Fès, une série d'expériences inscrites dans le paradigme de l'amorçage rapide (*masked priming*), qui visaient à identifier l'influence de la racine dans le processus mental de reconnaissance des mots arabes écrits. La réalisation de ces expériences nécessitait la constitution de listes de mots et de racines appartenant à l'arabe écrit d'aujourd'hui, et dont les items étaient soit de haute, soit de basse fréquence. Les résultats sont favorables à l'hypothèse d'un rôle de la racine dans la reconnaissance des mots arabes : des mots de basse fréquence, mais dont la racine a une fréquence élevée, sont reconnus plus rapidement par les lecteurs dans une proportion très significative [Grainger, Dichy, El-Halfaoui, Bamhamed 2003]. Des travaux similaires ont été menés sur l'hébreu [Frost, Forster, Deutsch 1997, 2000].

La section 3 présente les outils AraConc et DIINAR.1 et rappelle la représentation du mot graphique en arabe qui en a sous-tendu la conception. La section 4 porte sur les caractéristiques du corpus journalistique arabe de deux millions de mots analysé pour les besoins de l'expérimentation, et livre des informations statistiques dont plusieurs peuvent être

[1] Signalons, parmi d'autres, le travail en cours à l'université de Louvain [Van Mol & Paulussen 2004], et celui de la Charles University de Prague : le corpus CLARA [Zemanek 2001], la constitution d'un corpus associé à des arbres syntaxiques (tree-bank) [Smrz, Snaidauf, Zemanek 2002] et sa mise en commun avec le Linguistic Data Consortium [Maamouri et al. 2004].

[2] DIINAR-MBC : « DIctionnaire INformatisé de l'ARabe, Multilingue et Basé sur Corpus » - projet n° 961791 du programme de Coopération avec les Pays Tiers et les Organisations Internationales (INCO-DC) de la Commission européenne. Dates : février 1998 à décembre 2000 ; coordination : J. Dichy, université Lumière-Lyon 2.

considérées comme ayant une portée générale en ce qui concerne les corpus arabes de presse. La section 5, enfin, sera consacrée aux calculs de fréquences effectués et à leurs résultats en ce qui concerne les racines et les mots, ainsi que la démarche d'établissement des listes de fréquences nécessaires au protocole expérimental en psycholinguistique cognitive que l'on vient de décrire.

3. Présentation sommaire des outils informatiques (DIINAR.1 et AraConc) et de la modélisation linguistique sous-jacente

DIINAR.1 et AraConc sont basés sur une modélisation du mot graphique en arabe [Dichy et Hassoun, éds. 1989 ; Dichy 1990]. Celui-ci est constitué d'une séquence de morphèmes soumis à une relation d'ordre : les proclitiques, le préfixe, la base, les suffixes et l'enclitique. Ces morphèmes sont en inventaire fini [Dichy 1997], et se combinent en pré-bases et post-bases. Ils renferment des informations morphosyntaxiques : personne, genre, nombre, paradigme de conjugaison et mode pour les verbes ; cas, nombre, genre, etc. pour les noms et les adjectifs. Les proclitiques incluent des conjonctions ou des prépositions mono-consonantiques, la marque du futur, l'article, etc. ; les enclitiques sont des pronoms complément. Les entrées lexicales nominales peuvent se trouver liées à un suffixe de manière figée.

3.1. La relation lexique-grammaire et la base de connaissances lexicales DIINAR.1

L'un des principaux problèmes de l'analyse morphologique de l'arabe provient de la nécessité de prendre en compte le noyau lexical du mot graphique dans les grammaires « gérant » les relations entre les éléments constitutifs de cette unité [Dichy 1997], d'où : (a) la conception d'analyseurs et de générateurs morphologiques construits avec une grammaire mettant en œuvre la conception des relations entre les formants du mot graphique et (b) l'association à chacune des entrées de la base de connaissances lexicales DIINAR.1 de *spécificateurs morphosyntaxiques* permettant à la grammaire des formants du mot de traiter les relations entre le noyau lexical et les morphèmes situés à droite et à gauche de lui. DIINAR.1 a été construite pendant les années 1990 par des chercheurs lyonnais (J. Dichy, M. Hassoun avec la collaboration de N. Gader) et tunisiens (S. Ghazali, A. Braham avec la collaboration de M. Ghénima).

3.2. Le logiciel AraConc et l'identification des racines

AraConc est un logiciel de concordance et de calcul de fréquences des mots arabes construit par R. Abbès [2004a, 2004b] ; il fait appel aux informations de DIINAR.1. En entrée AraConc traite un ensemble de textes écrits en arabe, qu'il analyse mot par mot. Les analyses morphologiques et les emplacements du mot sont stockés dans des fichiers spécifiques, en vue de différents regroupements et de l'établissement de concordances. Le logiciel propose des étiquettes morphosyntaxiques associées aux entrées de DIINAR.1 comme la racine, le patron (ou schème), le cas, la transitivité, le paradigme de conjugaison, etc. Les étiquettes sont indépendantes, pour permettre à l'utilisateur de procéder à des regroupements selon les critères choisis par lui. Dans ce travail nous utilisons uniquement la racine (au sens sémitique ; en arabe, la racine est un triplet ou un quadruplet de consonnes [Dichy 1990]).

Dans la grande majorité des mots graphiques, la racine ne peut être identifiée de manière immédiate : il faut procéder pour l'extraire à une segmentation des mots, qui comportent souvent plusieurs possibilités de segmentation et plusieurs réalisations entièrement pourvues

de signes de vocalisation. Dans le meilleur des cas, les différents découpages donnent la même racine, faute de quoi une intervention manuelle est indispensable.

4. Caractérisation générale d'un corpus journalistique de 2 millions de mots

Le corpus provient d'un quotidien généraliste, édité en Europe et destiné aux lecteurs du monde arabe, *Al-Hayât*. Les événements traités ne sont pas spécifiques à une seule contrée et les journalistes sont natifs de divers pays arabes. L'échantillon sélectionné contient 4.338 articles, cumulant un total de 2 006 631 *mots* regroupés sous 149 990 *items*[3] hors ponctuation, chiffres et mots en alphabet latin. Les articles couvrent plusieurs rubriques du journal de l'année 1995. Une première catégorie de textes (1 075 347 mots) concerne le développent des titres de la une. Ils traitent pour la plupart de politique, d'économie, etc. La deuxième rubrique (866 764 mots), comprend les éditos, pages culturelles, débats, courrier des lecteurs… Enfin, une minorité de textes (64 520 mots) relève d'un domaine spécialisé (rubrique « automobile »). Cette classification permet d'observer les éventuelles influences de chaque type de texte sur les fréquences ou la distribution des items. Le recours à plusieurs types d'écrit a pour objet de n'écarter *a priori* aucun type de vocable ou de terme. Le large éventail couvert par les textes et la taille du corpus (2 millions de mots) occultent statistiquement les mots relatifs aux domaines de spécialité et laisse émerger le vocabulaire commun. En outre, la dimension du corpus suffit pour creuser les écarts entre les mots de haute fréquence ou de basse fréquence. L'analyse met en évidence un classement invariable, que nous avons observé à partir d'un seuil donné : on constate qu'à partir de 300 000 mots, les indices de fréquence des items et des mots ne sont plus altérés par les nouveaux textes [Abbès 2004b].

4.1. Statistiques générales

Le tableau 1 récapitule le comptage général établi à partir du corpus.

[3] **Conventions** : nous entendons par ***MOT*** toute séquence de caractères arabes délimitée par deux séparateurs (blanc ou autre marqueur de séparation, tel que la ponctuation) ; si une séquence de caractères de ce type se répète 2, 3, n fois, elle correspond à 2, 3 ou n *mots,* mais constitue un seul et même ***ITEM***.

1	2	3	4	5	6	7	8	9	10	11
	Mots		**Items**		**Nouveaux mots**		**Nouveaux items (N.I.)**			
N° de fich.	**Cumul**	**Nombre par fichier**	**Nombre par fichier**	**% par fichier**	**Nombre par fichier**	**% par fichier**	**Nombre par fichier**	**% N.I. / Nouveaux mots**	**% N.I. / Items**	**% N.I. / Mots**
1	83 254	83 254	18 570	22%	83 254	100%	18 570	22%	100%	22%
2	154 948	71 694	16 794	23%	11 588	16%	8 573	74%	51%	12%
3	231 027	76 079	17 350	23%	8 558	11%	6 932	81%	40%	9%
4	303 338	72 311	16 992	23%	6 674	9%	5 558	83%	33%	8%
5	384 447	81 109	18 420	23%	6 925	9%	5 783	84%	31%	7%
6	467 352	82 905	19 137	23%	6 474	8%	5 433	84%	28%	7%
7	543 883	76 531	18 418	24%	5 487	7%	4 795	87%	26%	6%
8	627 916	84 033	18 662	22%	5 052	6%	4 286	85%	23%	5%
9	710 664	82 748	18 281	22%	4 419	5%	3 905	88%	21%	5%
10	790 643	79 979	17 611	22%	3 672	5%	3 309	90%	19%	4%
11	853 380	62 737	15 414	25%	2 957	5%	2 654	90%	17%	4%
12	923 497	70 117	16 367	23%	3 094	4%	2 765	89%	17%	4%
13	996 332	72 835	16 820	23%	3 299	5%	2 909	88%	17%	4%
14	1 075 347	79 015	17 895	23%	3 417	4%	3 105	91%	17%	4%
15	1 153 223	77 876	23 354	30%	9 979	13%	8 211	82%	35%	11%
16	1 233 933	80 710	24 416	30%	8 399	10%	7 370	88%	30%	9%
17	1 316 066	82 133	24 820	30%	8 328	10%	7 220	87%	29%	9%
18	1 384 551	68 485	21 657	32%	6 505	9%	5 602	86%	26%	8%
19	1 467 949	83 398	24 923	30%	7 532	9%	6 578	87%	26%	8%
20	1 545 689	77 740	23 069	30%	6 053	8%	5 285	87%	23%	7%
21	1 623 027	77 338	23 037	30%	5 661	7%	5 012	89%	22%	6%
22	1 714 564	91 537	26 839	29%	7 586	8%	6 589	87%	25%	7%
23	1 789 979	75 415	22 962	30%	5 401	7%	4 893	91%	21%	6%
24	1 869 785	79 806	23 752	30%	5 368	7%	4 776	89%	20%	6%
25	1 942 111	72 326	22 942	32%	5 209	7%	4 742	91%	21%	7%
26	2 006 631	64 520	14 598	23%	9 771	15%	5 135	53%	35%	8%

Tableau 1: Statistiques générales

Commentaire des colonnes. Colonne (1) : numéro d'ordre attribué à chaque fichier analysé. Colonne (2) : cumul des mots (hors ponctuation, chiffres ou mots en caractères latins). Colonne (3) : nombre de mots par ficher. Colonne (4) : nombre d'items par fichier. Colonne (5) : pourcentage d'items par rapport au nombre de mots dans un fichier, soit le rapport colonne(4)/colonne(3). Colonne (6) : apport de chaque fichier en nouveaux *mots*. Colonne (7) : proportion de nouveaux mots dans le nombre total de mots du fichier. Colonne (8) : apport de chaque fichier en nouveaux *items*. Colonne (9) : proportion de nouveaux items dans les nouveaux mots par fichier. Colonne (10) : pourcentage de nouveaux items dans le nombre d'items par fichier. Colonne (11) : proportion de nouveaux items dans le nombre total de mots.

Commentaire des lignes. Lignes 1-14 : articles, développement de la une. Lignes 15-25 : éditos, pages culturelles, débats, courrier des lecteurs… Ligne 26 : rubrique « automobile ».

4.2. Évolution des items

Les colonnes numéro quatre et cinq montrent un taux relativement constant d'items dans les textes d'une même rubrique. Il varie entre 22 et 25% pour le développement des titres de la une, et avoisine les 30% sous la rubrique littéraire. Il est de l'ordre de 23% dans les pages « automobile ». Cette diversité est probablement due à la richesse de l'expression dans certaines rubriques. Les sujets traités dans les rubriques littéraires, culturelles, débats… offrent une plus grande liberté d'expression aux auteurs et leur permettent de puiser dans un large éventail lexical, ce qui n'est pas le cas des rubriques politiques, où le langage est plus codé et où les formules sont souvent pré-formatées.

4.3. Les nouveaux mots et les nouveaux items

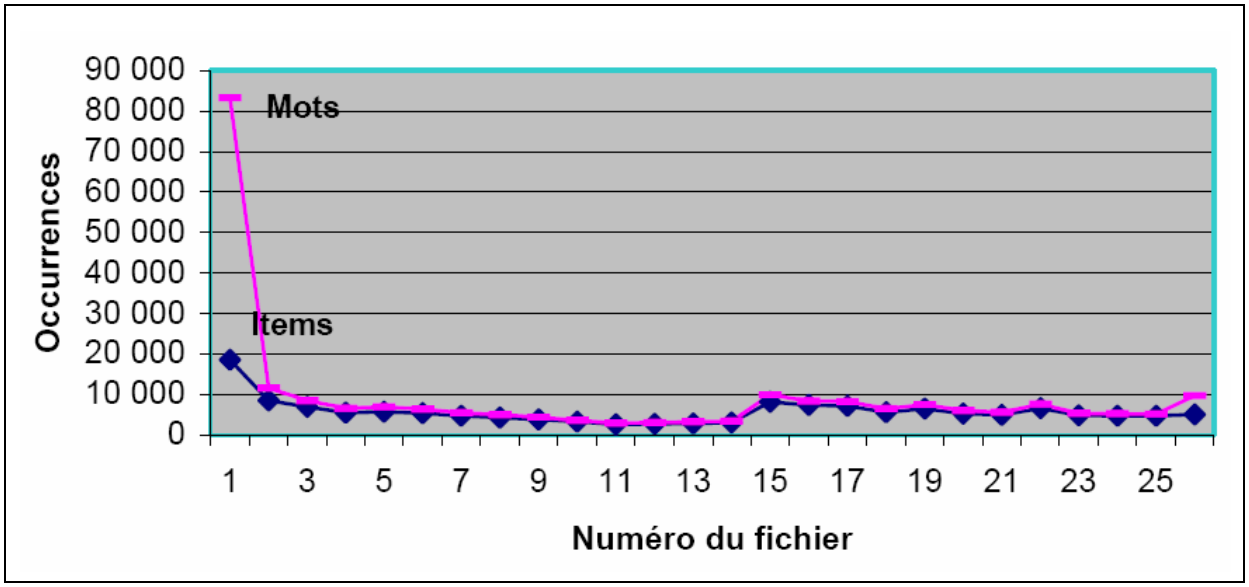

Figure 1 : Évolution des mots et des items nouveaux au fur et à mesure du traitement

Les colonnes du tableau 1 sont illustrées sous forme de graphe dans la figure ci-dessus. On peut ainsi visualiser la nature décroissante des apports en nouveaux mots, et par conséquent en nouveaux items, de chaque fichier.

4.3.1. Une courbe décroissante

Les colonnes du tableau 1 sont illustrées dans la figure ci-dessus sous forme de graphe. La courbe commence par une décroissance spectaculaire de l'apport en mots nouveaux (tous les mots du premier fichier étant évidemment nouveaux). Les contributions en nouveaux éléments décroissent considérablement jusqu'à une relative stabilité vers de faibles pourcentages. De légers pics sont enregistrés à chaque fois qu'est introduit un nouveau type de textes (par ex., fichier 15), mais le comportement global de la courbe reste similaire. L'essentiel du vocabulaire est recensé dès les premiers 100 000 mots. Le reste du traitement consiste à confirmer les tendances de fréquence. Nous n'avons plus observé de variation importante dans le classement par ordre de fréquence à partir du seuil de 300 000 mots. Chaque nouveau fichier creuse d'avantage l'écart entre les items de haute fréquence et les items de basse fréquence. Vers un certain seuil optimal, l'ajout des nouveaux textes n'a plus d'influence sur le classement. Ceci peut s'expliquer par deux raisons :

- L'apport en nouveaux items des nouveaux textes devient négligeable par rapport à la totalité de la liste. La colonne 9 du tableau 1 montre qu'au fil du traitement le nombre de nouveaux mots est sensiblement égal au nombre de nouveaux items : il tend vers 90% (sauf pour la rubrique « automobile », ligne 26). Cela signifie que les nouveaux items sont introduits dans de très faibles proportions, pratiquement une occurrence pour chaque item.
- L'écart des occurrences entre les items situés en haut de la liste et les items du bas de la liste devient très important. En conséquence, le classement n'est plus vraiment altérable par de nouveau fichiers.

4.3.2. Les pics

On observe toutefois une légère croissance lorsque l'on change de type de texte. Ce phénomène s'explique par l'introduction de nouveaux vocabulaires spécifiques aux rubriques du journal. Ces items, à l'image du début du traitement sont introduit en masse. Par la suite ces sommets sont atténués et la courbe reprend sa tendance générale.

On observe à la fin de celle-ci un pic plus important d'occurrences de nouveaux mots et une faible quantité d'items. Cela coïncide avec l'insertion de textes du troisième type ci-dessus : les articles sur l'automobile comportent beaucoup de vocabulaire spécialisé et de mots étrangers translittérés en caractères arabes. La majeure partie des noms de pièces, de marques ou encore d'options n'est pas traduite dans un équivalent arabe. Pour illustrer cela, voici deux extraits de phrase : غولف كابريوليه تي دي آي..., « GOLF CABRIOLET T.D.I. » ; مئة ألف هوندا في سويندون..., *Mi'at 'alf HONDA fî SWINDON...*, « cent mille Hondas à Swindon » (les mots translittérés sont en lettres capitales). La présence de telles occurrences explique le nombre élevé de nouveaux mots et, proportionnellement, le faible pourcentage de nouveaux items (53%). Les noms translittérés, qui ne se déclinent pas, sont très rarement à l'origine de formes dérivées : d'où le nombre réduit des items correspondants. S'agissant d'items empruntés, l'analyseur morphologique a bien évidemment détecté dans ce fichier un faible taux de racines arabes.

Ces observations sur le comportement des items lors de l'évolution du corpus, nous a conduits à arrêter le traitement au nombre seuil de deux millions de mots. Nous arrivons en effet à ce niveau à un dénombrement satisfaisant des items avec des écarts significatifs entre les hautes et les basses fréquences. Cette limite paraît significative pour ce type de corpus, bien qu'elle reste à vérifier pour d'autres cas. Comme pour le seuil de 300 000 mots, les propriétés de corpus arabes de cette dimension seront l'objet d'une publication ultérieure.

4.4. Spécificités des corpus bruts de la presse écrite

La transformation d'un corpus brut en un annuaire de mots et d'analyses nous offre par ailleurs de nouvelles vues sur les textes. Elle nous permet de distinguer et d'évaluer des phénomènes spécifiques. L'analyse automatique du corpus s'est en effet heurtée à plusieurs obstacles relevant pour la plupart des spécificités de l'écrit journalistique arabe contemporain. Dans cette sous-section, nous allons exposer certaines données chiffrées relatives aux éléments rencontrés dans le corpus. Il s'agit pour l'essentiel des signes de ponctuation, des chiffres, des mots non arabes, des abréviations et de lettres isolées.

• ***Les signes de ponctuation.*** Cette catégorie inclut toute séquence de caractères de ponctuation délimitée par des lettres ou des espaces. Nous partons du principe qu'un point est différent de trois points (même si toute succession de signes de ponctuation ne constitue pas

nécessairement une ponctuation unique). Dans notre corpus nous avons énuméré 201 séquences de ponctuations différentes, cumulant 330 489 occurrences.

• ***Les nombres.*** Nous regroupons toutes les séquences de caractères situées entre deux espaces et contenant des chiffres sous une seule occurrence. Cette méthode comporte l'inconvénient de conserver les signes de ponctuations collés à des chiffres. Mais elle a l'indispensable avantage de regrouper les dates, les nombres réels, les pourcentages. Nous avons dégagé 4 655 séquences de chiffres différentes, pour un ensemble de 34 463 occurrences.

• ***Les mots en caractères latins.*** Les mots en caractères non arabes, essentiellement en caractères latins, sont tout simplement regroupés selon leur forme graphique. Les items latins ne sont pas nombreux, et ne dépassent pas 1 495 items. Ils sont surtout très peu fréquents : 2 177 occurrences.

• ***Les abréviations et les lettres isolées.*** La liste des fréquences révèle la présence d'un nombre assez important de mots à une seule lettre dans les textes journalistiques. Ces lettres sont souvent utilisées dans les abréviations. Une lettre isolée est rarement la première et unique lettre du mot arabe. Elle peut désigner une variable, par exemple الفئة ب, *al-fi'a bâ',* « la catégorie B ». Mais elle correspond le plus souvent à la translittération – à une traduction phonétique en quelque sorte – de sigles étrangers, comme أ.ف.ب (ou encore ا.ف.ب) abréviation de AFP, « Agence France Presse ». Elle peut enfin, de manière en partie analogue à ce que l'on trouve dans des langues comme le français ou l'anglais, correspondre à l'abréviation d'un mot arabe, ainsi : ت pour تاريخ, *târîx,* « date », م pour ميلادي, *mîlâdî*, « du calendrier grégorien », ص pour صفحة, *ṣafḥa*, « page ».

Certains sons appartenant à des sigles étrangers ne sont pas translittérés au moyen d'une seule lettre : c'est le cas de la marque de voiture ب.ام « BMW », sigle qui se trouve lui-même abrégé en « BM » (la translittération arabe imite sans doute l'usage oral, mais elle évite en même temps la difficile notation de « W », dans sa prononciation anglaise ou française). Ce dernier usage pose d'autant plus de problèmes de notre point de vue que أم est par ailleurs un mot arabe fréquent et de surcroît ambigu.

Toutes les lettres isolées ne sont pas des abréviations. Les proclitiques prépositionnels ou de coordination comme بِ, وَ, فَ, لَ, كَ (bi-, wa-, fa-, li-, ka-…) sont notés sous forme isolée à coté des chiffres (ex : 54 362 ب), ou encore quand ils sont suivis de guillemets dans le cas de citations, de mots étrangers, d'expressions familières, de noms propres, ainsi : " صرّح ل ..."الحياة, *ṣarraḥa li-« l-ḥayât »...*, « il a déclaré à "Al-Hayât"... », ou de noms de lieu ou d'organismes, exemple : ..."حزب الله"ل, *li-« ḥizb al-lâh. »,* « au "Hezbollah" ».

4.5. Pratiques d'écriture – problèmes généraux

Le traitement automatique des textes de la presse écrite présente quelques difficultés dues, d'une part, aux structures propres du mot graphique en arabe, d'autre part, à certaines habitudes orthographiques dont l'impact sur la reconnaissance automatique n'est pas négligeable (comparer avec l'approche différente de [Buckwalter 2004]).

• **Hamza *et* 'alif.** Les scripteurs confondent souvent la *hamza* (أ – إ) et le *'alif* en début de mot. On trouve par exemple, 26 923 fois الى *'ilâ,* « à », et 2 089 fois إلى. On trouve également 33 901 ان indifférencié, contre 50 569 أن (conjonctions *'an* ou *'anna*) et 759 إن (conjonctions *'in* ou *'inna*).

Pour les items n'admettant qu'une seule réalisation, avec أ ('a ou *'u*) ou avec إ (*'a*) le système peut procéder à une correction automatique. Mais dans le cas de lectures multiples, il est

impossible de deviner la bonne orthographe en analyse hors contexte. Cela pose un problème supplémentaire pour la reconnaissance et nécessite souvent une intervention manuelle. Les estimations auxquelles nous parvenons dans ce corpus indiquent que le taux des items non reconnus par l'analyseur en raison de cette seule confusion s'élève à 5,79% de l'ensemble des items ou encore à 6,76% des mots.

• **Yâ'** ***et*** **'alif maqṣûra.** A la *hamza-'alif* initiale s'ajoute une autre confusion, située cette fois en fin de mot, entre ي (lettre *yâ'* finale) et ى (ou *'alif maqṣûra*). Le mot نادي, *nâdî,* « club », par exemple, peut être noté نادى, ce qui correspond à l'usage des typographes égyptiens (mais peut aussi être lu comme *nâdâ*, « convier, convoquer »). Cette confusion est une source d'erreur importante mais elle ne peut être comptabilisée, parce que la plupart des mots de ce type admettent des analyses avec l'une ou l'autre lettre. Avec des moyens automatiques, on ne peut identifier ou corriger que les cas où le mot n'admet aucune analyse avec la lettre ى. C'est ainsi que يبقى n'a aucune chance d'être reconnu en tant que يبقي, *yubqî,* « il laisse » (mot-à-mot « il fait rester »), parce que يبقى, *yabqâ,* « il reste » existe par ailleurs.

Certains mots cumulent les deux difficultés. Par exemple nous trouvons 678 fois le mot الاولى et 921 fois le mot الأولى (« première », dans les deux réalisations).

• ***Le caractère '-'.*** Les typographes font un usage fréquent du caractère '-' (appelé *kashida*), qui permet l'allongement du trait au milieu des mots, pour une meilleure lisibilité, pour limiter les espaces blancs sur une ligne justifiée, voire pour des raisons purement esthétiques. Or cet usage peut nuire aux analyses automatiques : ce caractère ne faisant pas partie de l'alphabet arabe, il est considéré comme un intrus par le système d'analyse automatique. Il faut donc recourir à un sous-programme particulier afin de l'éliminer.

• ***L'absence des signes de vocalisation.*** L'absence de signes de vocalisation dans les textes – à laquelle les lecteurs arabes sont accoutumés –, constitue pour l'analyse automatique une source de difficultés considérable. Certains signes diacritiques relatifs à la base (ou noyau lexical) sont indispensables pour la détection du sens du mot. Ils sont par conséquent indispensables pour le choix de la bonne analyse, particulièrement hors contexte. Les analyses peuvent en effet reconnaître dans un même item plusieurs patrons (وزن), voire plusieurs combinaisons de patrons et de racines.

• ***Mots étrangers translittérés en arabe.*** Les translittérations en arabe de mots étrangers ne concernent pas cette expérimentation, puisqu'ils n'ont pas de racine en arabe. Nous avons toutefois remarqué une influence de la deuxième langue des auteurs. On trouve par exemple pour « Milan », ميلانو, *mîlânû*, mais également ميلان, *mîlân*, selon que l'auteur est influencé par l'anglais (il imite de ce fait en partie la prononciation italienne) ou le français.

Quelques items étrangers méritent une attention particulière en raison de leurs fréquences élevées, comme دولار, *dûlâr,* « Dollar » et ses réalisations avec des clitiques ou des suffixes (الدولارات, دولاراً, الدولار, دولارات, بالدولار, للدولار, دولارا, والدولار, ودولارهم, البترودولار, دولاران, كالدولار, ودولارات, دولارين, بدولار – bi-dûlâr ; dulârât, wa-dulârât...) qui totalisent 814 occurrences.

• ***Mots non reconnus par l'analyseur.*** A ces difficultés d'écriture s'ajoutent quelques cas non reconnus par le système automatique parce qu'ils n'existent pas dans son lexique. Les mots non reconnus par l'analyseur ayant recours à la base de données DIINAR.1 restent relativement faibles. Mais dans l'écrit journalistique, le taux de mots non analysés augmente légèrement, en raison notamment de la présence de noms propres, et notamment de noms de lieux ou de pays. Ces items peuvent atteindre assez rapidement des fréquences importantes en fonction des événements relatés par le journal. On peut bien évidemment anticiper

l'importance de mots comme أمريكا, *'amrîka,* « Amérique » et les intégrer dans les noms de pays de la base. En revanche, il n'est pas évident de deviner à l'avance l'importance de noms comme *Clinton* كلينتن ou celui du/de la prochain(e) président(e) des États-Unis.

5. Résultats de l'extraction automatique des fréquences d'items et de racines

Passons maintenant aux résultats de l'interrogation automatique du corpus, en vue de l'établissement de listes de fréquence. Les mots outils, relativement peu nombreux (ils sont environ 450 dans DIINAR.1), sont, comme il est attendu, très utilisés. Ils occupent les premières places avec des fréquences très élevées par rapport aux verbes et aux noms. Nous avons par exemple dans notre corpus 76 727 التي, *'allatî,* « qui » (relatif), 14 077 على, *'alâ,* « sur » et 30 199 في, *fî,* « dans », alors que nous trouvons le premier nom رئيس, *ra'îs,* « chef » au seuil de 8 422 occurrences.

5.1. Taux de reconnaissance des racines

Dans environ 85% des items du corpus, la ou les racine(s) sont identifiables automatiquement. Parmi les 15% restants, nous trouvons les mots outils identifiés comme tels, les mots d'origine étrangère, les mots non inclus dans les lexiques associés à AraConc… Le listing obtenu à la sortie du traitement automatique et avant toute intervention manuelle nous livre 4 466 racines différentes. Par comparaison, la base de données lexicale DIINAR.1 contient près de 5 751 racines [Abbes, Dichy, Hassoun, 2004]. On notera que 79 057% des items admettent, à l'issue de l'analyse, une seule racine, 11.09% deux racines, 4.84% trois racines. Les 4.48% restant admettent quatre racines ou plus. Le nombre de racines dépend des analyses automatiques possibles du mot. (Rappelons que, dans DIINAR.1, il s'agit de racines existantes et attestées, et non de racines virtuelles.)

5.2. Racines et items

Sous chaque racine nous trouvons un ensemble d'items. Nous appelons « famille d'une racine » l'ensemble des items d'une racine présents dans le corpus. Le nombre moyen d'items par famille de racine est de 38,86, avec une valeur maximale de 599.

Les racines les plus fréquentes comportent – comme on pouvait s'y attendre – un nombre d'items important. C'est le cas la racine جمع /j-m-'/ (idée de « réunir »), qui a une fréquence de 13 146 occurrences et admet 549 items (voir ci-dessous, tableau 2).

Il faut cependant noter que le rapport entre la fréquence d'une racine et le nombre de ses items n'est pas fixe. Il arrive en effet que la fréquence d'occurrence des racines n'entraîne pas celle de leurs items. Ainsi, la racine دول /d-w-l/ (idée « d'État ») est très fréquente, avec 8 842 occurrences, mais n'admet que 190 items. De même pour حزب /ḥ-z-b/ (idée de « parti ») qui totalise 4 707 occurrences avec uniquement 91 items, ou encore pour كثر /k-ṯ-r/ (idée « d'être nombreux »), qui, avec 141 items seulement, comporte 4 755 occurrences. Ces valeurs sont faibles relativement aux racines ayant des fréquences d'occurrences avoisinantes.

Nous avons remarqué également que l'échelle de fréquence des items ne coïncide pas avec celles de leurs racines. Le deuxième item le plus fréquent dans le corpus (après *ra'îs,* « chef », § 3.1 ci-dessus) est العام, *al-'âmm,* « général » (adj., déterminé par l'article) avec 3 274 occurrences, mais la racine عمم /'-m-m/ (idée de ce qui est « général » ou « commun ») est classée 22^{e} avec une fréquence de 6 402. Le 10^{e} item de la liste, خلال, *xilâla,* « à travers », qui

totalise 2 557 occurrences est issu de la 83e racine خلل /*x-l-l*/ (idée de « percer », « faire une faille ») avec 3 904 occurrences.

Occ.	Item	Occ.	Item	Occ.	Item	Occ.	Item	Occ.	Item
700	المجتمع	303	جماعة	151	جمعية	87	للجامعة	62	لجميع
676	جميع	294	الاجتماعي	136	جمع	83	والاجتماعي	60	المجموعات
613	مجموعة	283	جامعة	121	اجتماعية	81	الجامعات	59	الجامع
508	الجامعة	271	المجموعة	120	تجمع	80	مجموعات	58	اجتماعه
503	الجماعة	238	التجمع	118	المجتمعات	80	يجمع	56	الجماعي
488	اجتماع	234	الجمعية	115	اجتماعاً	76	مجموع	54	للمجتمع
473	الجميع	226	الجماعات	110	مجتمع	70	المجمع	53	الجماعية
444	الاجتماع	224	اجتماعات	108	الاجتماعات	63	اجتمع	52	أجمع
438	الاجتماعية	204	جميعاً	102	جماعات	63	جمعة	48	والمجتمع
357	الجمعة	168	والاجتماعية	97	للجميع	62	جامع	,,,	,,,

Tableau 2 : Distribution des items de la racine جمع */j-m-ʿ/*

On constate en outre de fortes variations de la fréquence en fonction de l'association, dans l'item, des lemmes avec certains mots-outils clitiques. Considérons à titre d'exemple la racine قول /q-w-l/ (idée de « dire »). Ses occurrences sont au nombre de 10 585 ; elle regroupe dans sa famille, dans l'ordre décroissant, les items وقال, *wa-qâla,* « et il dit », 3 195 occ., قال, *wa-qâla,* « il a dit », 1 341 occ. Ce verbe est donc 2 à 3 fois plus fréquent avec le coordonnant que sans lui (ce qui peut s'expliquer en partie par le fait qu'il s'agit d'un verbe de parole, souvent situé en début de phrase ou de paragraphe). La racine نخب /n-x-b/ (idée de « choisir » – 4 663 occurrences) comporte, en tête de sa famille الانتخابات, *al-intixâbât,* « les élections », 2 256 occurrences suivi de l'item انتخابات, « élections » (sans l'article), avec 684 occurrences seulement.

La fréquence des racines n'est donc pas proportionnelle à celle des items. La dernière remarque ci-dessus indique en outre que la fréquence des items et celle de l'entrée lexicale correspondante peuvent elles-mêmes être soumises à d'importantes variations.

5.3. Liste de fréquences pour l'expérimentation en psycholinguistique cognitive

L'élément fréquence peut concerner différents critères linguistiques. Lors de cette expérience nous avons manipulé à la fois la fréquence des mots minimaux (sans proclitiques ni enclitiques) et celle des racines. Sachant qu'un item peut correspondre à plusieurs racines, nous avons associé à chacun de ceux que nous avons identifié l'ensemble des racines existantes pouvant être supportées par lui. Nous avons ensuite classé les items selon leur ambiguïté, c'est à dire selon le nombre de racines qu'ils peuvent admettre. Prenons quelques exemples.

● La racine سلم /s-l-m/ (idée de « salut », ou de « paix ») comporte 12 636 occurrences, correspondant à 426 items. Chacun des items de cette racine ne peut prendre que سلم en tant que radical. Ces 426 items sont donc analysés comme incluant la même racine. C'est donc un cas d'absence d'ambiguïté de ce point de vue.

● La racine قول /q-w-l/ (idée de « dire ») comprend 10 585 occurrences et 336 items, auxquels peuvent être associées 779 occurrences de racines, soit une moyenne de deux racines par item. Les radicaux de la racine قول sont donc hautement ambigus.

● La racine حكم (9 527 occurrences) regroupe 392 items, auxquels correspondent 446 occurrences de racines. Les radicaux de cette racine sont donc faiblement ambigus.

Pour connaître le degré d'ambiguïté de chaque racine nous avons calculé la proportion des items ambigus par racine. Ce coefficient est nul pour les racines non ambiguës :

(([Nb. Analyse] - [Nb. Items])/[Nb. Analyse]) * 100

Ainsi le coefficient est de 0% pour سلم, il est de 56.68% pour قال et il ne dépasse pas 12.1% pour حكم.

N°	Nombre d'occurrences	Racine	Nombre d'items	Nombre d'analyses	Degré d'ambiguïté
1	13146	جمع	549	551	0,36%
2	12636	سلم	426	426	0,00%
3	11286	عمل	505	506	0,20%
4	10998	رأس	221	233	5,15%
5	10585	قول	336	779	56,87%
6	9527	حكم	392	446	12,11%
7	9055	وحد	310	606	48,84%
8	8842	دول	190	281	32,38%
9	8815	قبل	416	416	0,00%
10	8803	أخر	215	270	20,37%
…	…	…	…	…	…

Tableau 3 : Coefficient d'ambiguïté pour dix racines

6. Conclusion

Cette expérience a permis de procéder à une évaluation réelle de l'analyseur morphologique inclus dans AraConc et de la base de connaissances lexicales DIINAR.1. Elle a également permis de mette en évidence certaines pratiques orthographiques et scripturales observables dans la presse arabe contemporaine, d'une manière quantifiée. Leur impact sur le traitement automatique de la langue et l'extraction des données d'un corpus textuel a pu, au passage, être évalué. Les ajustements du post-traitement et les heuristiques de départ n'ont pas évité l'intervention manuelle, indispensable en arabe, mais ils ont facilité l'extraction de la liste finale en conduisant rapidement vers les mots candidats à l'expérience de psycholinguistique à la conception de laquelle nous avons participé.

Dans les listes extraites par nous, il reste encore une mine d'informations non encore explorées, et qui devraient permettre dans un avenir proche, de mettre en relation les fréquences des mots avec leurs significations, en diminuant de façon déterminante la part inévitable de travail manuel associée à toute analyse sémantique de qualité, à l'aide de procédures automatisées.

Bibliographie

Abbès R. (2004a). AraConc : un outils informatique pour le traitement des corpus de textes arabes : quantification et concordances. In Brahim BEN MRAD (éd.), *Actes du Colloque international de lexicographie : Dictionnairique et corpus*, 19-21 juin 2004. Tunis, Association des Lexicographes tunisiens.

Abbès R. (2004b). *La conception et la réalisation d'un concordancier pour l'arabe*. Thèse de doctorat en Sciences de l'Information. Lyon, INSA, décembre 2004.

Abbès R., Dichy J. and Hassoun M. (2004). The Architecture of a Standard Arabic Lexical database : some figures, ratios and categories from the DIINAR.1 source program. In *COLING'04, 20th International Conference on Computational Linguistics. Proceedings of the Workshop Computational Approches to Arabic Script-bases Languages*, 28.08.2004, Genève : 15-22.

Brill M. (1940). In coll. with Neustadt D. and Schusser P. *The basic word list of the Arabic daily newspaper, Qâmûs al-ṣaḥâfa al-ᶜarabiyya al-yawmiyya.* Jerusalem, The Hebrew University Press Association.

Buckwalter T. (2004). Issues in Arabic Orthography and Morphological Analysis. In *COLING'04, 20th International Conference on Computational Linguistics. Proceedings of the Workshop Computational Approaches to Arabic Script-bases Languages*, 28.08.2004, Genève : 31-41.

Dichy J. (1990). *L'Écriture dans la représentation de la langue : la lettre et le mot en arabe.* Thèse d'État, Université Lumière Lyon 2.

Dichy J. (1997). Pour une lexicomatique de l'arabe : l'unité lexicale simple et l'inventaire fini des spécificateurs du domaine du mot. *Meta 42*, printemps 1997. Québec, Presses de l'Université de Montréal : 291-306. www.erudit.org/revue/meta/1997/v42/n2/002564ar.pdf.

Dichy J., Braham A., Ghazali S. et Hassoun M. (2002). La base de connaissances linguistiques DIINAR.1. In Abdelfattah BRAHAM (éd.) *Actes du Colloque international sur le Traitement automatique de l'arabe*, 18-20 avril 2002. Manouba – Tunisie, Université la Manouba, pp. 45-56. www.elsnet.org/acl2001-arabic.html.

Dichy J. et Hassoun M. (éd.). (1989). *Simulation de modèles linguistiques et Enseignement Assisté par Ordinateur de l'arabe - Travaux SAMIA I*. Paris, Conseil International de la Langue Française.

Dichy J. et Hassoun M. (2005). The DIINAR.1-« معالي » Arabic Lexical Resource, an outline of contents and methodology. In *The ELRA Newsletter,* Vol. 10, n°2, April-June 2005 : 5-10.

Frost R., Forster K. & Deutsch A. (1997). What can we learn from the morphology of Hebrew ? A masked priming investigation of morphological representation. *Journal of Experimental Psychology : Learning, Memory and Cognition,* 23 : 829-856.

Frost R., Forster K. & Deutsch A. (2000). Decomposing morphologically complex words in a non linear morphology. *Journal of Experimental Psychology : Learning, Memory and Cognition,* 26 : 751-65.

Gougenheim G. (1958). *Dictionnaire fondamental de la langue française*, 2e édit. revue et augmentée. Paris, Didier.

Grainger J. (2003). Visual word recognition. In *The Encyclopedia of Cognitive Science*. Macmillan Publishers : 565-568.

Grainger J., Dichy J., El-Halfaoui M. et Bamhamed M. (2003). Approche expérimentale de la reconnaissance du mot écrit en arabe. Revue *Faits de langues (FDL)*, n° 22, 2003, Jean-Pierre Jaffré (éd.), *Dynamiques de l'écriture : approches pluridisciplinaires* : 77-86.

Landau Jacob M. (1959). *A word count of Modern Arabic prose.* New-York, American Council of learned Societies.

Maamouri M., Bies A., Buckwalter T., Mekki W. (2004). The Penn Arabic Treebank : building a large-scale annotated Arabic corpus. In Mahtab Nikkhou (ed.) *NEMLAR International Conference on Arabic Language Resources and Tools*, 22-23 September 2004, Cairo : 102-109.

Smrz O., Snaidauf J., Zemanek P. (2002). Prague Dependency Treebank : Arabic version. Annotation of Arabic-English Parallel Corpus. In Abdelfattah Braham (éd.) *Actes du Colloque international sur le Traitement automatique de l'arabe*, 18-20 avril 2002. Manouba – Tunisie, Université la Manouba : 147-160.

Van Mol M. and Paulussen H. (2004). Natural Language processing and Arabic : the Leuven tandem approach. In *JEP-TALN 2004, Arabic Language Processing.* Fez, 19-22 April 2004.

West M. (ed.) (1953). *A General Service List of English Words, with semantic frequencies and a supplementary word-list for the writing of popular science and technology.* London, Longman.

Zemanek P. (2001). Clara (Corpus Linguae Arabicae): An Overview. In *Proceedings of ACL 39th Annual Meeting - Workshop on Arabic Language Processing; Status and Prospect*, 9-11 juillet 2001. Toulouse : 111-112. www.elsnet.org/acl2001-arabic.html.