社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－CH－68 (4)
2005／10／28

エジプト口語アラビア語コーパスの構築に向けて

中道静香* 永崎研宣† 小田淳一‡
*大阪大学大学院言語文化研究科 †山口県立大学
‡東京外国語大学アジア・アフリカ言語文化研究所

本報告では言語研究および語学教育に利用可能な口語アラビア語コーパスのデザインを検討し、その上で報告者らが現在構築を進めているエジプト映画コーパスについて紹介する。正書法が定まっていない口語アラビア語を対象とするコーパス構築においては、まずエンコーディング・入力・出力の方法に関する基盤整備が重要な作業となるが、本報告においてはエジプトの口語アラビア語（エジプト方言）の言語的特徴と表記上の特徴をふまえ、有効と思われるコーパス作成の方策を提示する。ここで紹介するコーパスは、子音情報中心の入力用ラテン文字テキスト、アラビア文字テキスト、母音情報を加えた転写文字テキストの３種から構成され、多様な検索・出力に対応する柔軟性を備えている。

Towards the construction of a corpus of Egyptian colloquial Arabic

NAKAMICHI Shizuka*, NAGASAKI Kiyonori†, ODA Jun'ichi‡
*Graduate School of Language and Culture, Osaka University
†Yamaguchi Prefectural University
‡Research Institute for Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies

This paper presents a design for implementing a corpus of Egyptian colloquial Arabic. The corpus is intended for use in linguistic research and language education. Because it does not have well-defined orthography, creating a corpus for colloquial Arabic presents some difficulties. We examine grammatical features and the writing system of Egyptian colloquial Arabic, and suggest an improved method for inputting and outputting the data. We also introduce a corpus of Egyptian films as a sample, the text of which has been encoded using the following character codes: Unicode Latin-1, Unicode Arabic and an additional Latin-based code. This design will provide more flexible search and output options.

1．アラビア語コーパス

1．1 アラビア語コーパスの現状

コーパスの開発及びコーパスに基づく言語研究の分野において、アラビア語は後発にあたる言語である。その要因としては、まずアラビア語表記に使用されるアラビア文字の問題が挙げられよう。アラビア文字には、1）書記方向が右から左である、2）大半の文字は後続文字との連結を起こす、3）一つの文字が語頭・語中・語末で異なる形をもつ、という特徴がある。このようなアラビア文字の処理は、ラテン文字使用者のために開発されたコンピュータ上では明らかに困難を伴うものであった。PCが世界規模で普及して以降は、アラビア語処理の状況も大きく改善されたが、エンド・ユーザーが特殊なPC環境を整えることなしに、ある程度満足のいくアラビア語の入出力（日本語や他の言語との混在を含む）を行えるようになったのはここ数年のことである[1]。

一方、主に欧米の大学や研究機関では、1980

[1] Mac OS8.5 日本語版（98/10 発売）、Windows 2000 日本語版（00/4 発売）以降。

年代以降複数のアラビア語コーパス・プロジェクトが立ち上げられ、アラビア語処理に関わる技術開発だけではなく、コーパスに基づいた辞書編纂や言語学的研究などの成果が出されている。主なアラビア語コーパス・プロジェクトをデータの媒体（内容）別に整理すると以下のようになる（括弧内の年号はコーパス編纂事業の開始年）。

1) Written Data（新聞・雑誌記事、学術誌論文、宗教書、小説など）

a. Leuven Corpus (1990) by Catholic University Leuven,Belgium (http://www.kuleuven.ac.be/ilt/arabic/en_gesprokencorpus.htm)
b. Arabic Newswire Part 1 (1994) by LDC [Linguistic Data Consortium], University of Pensylvania,USA (http://www.ldc.upenn.edu/Catalog/index.jsp)
c. Nijmegen Corpus (1995) by Nijmegen University, the Netherlands (http://www.let.kun.nl/wba/Content2/1.4.5_Nijmegen_Corpus.htm)
d. CLARA [Corpus Lingiae Arabicae] (1997) by Institute of Ancient Near Eastern Studies, Charles University, Prague, the Czech Republic (http://www.elsnet.org/arabic2001/zemanek.pdf)
e. ARCOLEX [Arabic Raw Corpora for Lexical purposes] (DIINAR-MBC project) (1998) by Lumière-Lyon 2 University, France (http://sites.univ-lyon2.fr/langues_promodiinar/Accueil.htm)
f. Al-Warrāq (1998) by Al-Qarya al-Elekutoroniyya, Abu Dhabi, UAE (http://www.alwaraq.com/)
g. Arabic Gigaword (2002) by LDC
h. CCA [Corpus of Contemporary Arabic] (2004) by University of Leeds, UK (http://www.comp.leeds.ac.uk/eric/latifa/research.htm)

2) Spoken Data（ラジオ・テレビ放送のニュース番組、演説、インタビュー、トークショー等）

a. Leuven Corpus (1990) by Catholic University Leuven
b. Broadcast News Speech (2001) by LDC

3) Conversational Data（電話での会話）

a. CALLFRIEND (1995) by LDC
b. CALLHOME Egyptian Arabic Speech (1997) by LDC
c. DARPA Babylon Levantine Arabic Speech and Transcripts (2005) by LDC

Written Dataのコーパスを構築する機関の数は、WWWが普及し始めた 1990 年代から飛躍的に増えた。辞書編纂を目的に作られたものが多いが (a, c, d, e) 、他にもアラビア語教育や自然言語処理に供するために作られたもの (b, g, h) などがある。コーパスの規模（単語数）は様々で、32 万語 (a) から 4 億語 (g) [2]までの開きがある。ただし、どのコーパスも多かれ少なかれWWW上で公開されている新聞・雑誌記事[3]を利用しているため、内容が重複している場合も多い。コーパスの形式は圧倒的に未加工のアラビア文字テキストが多く、XMLや独自のマークアップを行っているコーパスは一部で

[2] これが上限ではない。Buckwalter （[1]の著者の一人）などは 20 億語以上のコーパスを独自に作成しているようである。

[3] 例えば、An-Nahar, Al-Hayat, l'Agence France Presseなど。レバノン紙An-Naharは、新聞記事データベースの販売も行っている (http://www.annahar.com.lb/Narchives/archives.htm) 。

ある。Written Dataに関しては今後も早いペースで増えていくと予想されるが、次の段階としては生データをいかに加工し、コーパスとしての精度を上げるかが重要となるだろう[4]。

一方、Spoken, Conversational Data を扱うコーパスは極端に少ない。Written Data が既存の電子テキストを入手・加工すればよいのに対し、Spoken, Conversational Data のほうは音声情報の転写・入力という時間と労力のかかる作業が必要だからである。例えば、全体として 300 万語の規模をもつ Leuven Corpus でも Spoken 部分は 70 万語にとどまる。

ここで注意しなければならないのは、Spoken が意味する内容である。これは、主にテレビ・ラジオ放送の発話を収集したものだが、ニュースのように一方向に発信されるものから、インタビューやトークショーのような対話形式のものまで様々なデータを含んでいる。また、この Spoken Data が必ずしも口語アラビア語（アラビア語で「アンミーヤ」）ではないということにも留意する必要がある。上記の Written Data に用いられているのは「フスハー」と呼ばれる正則アラビア語である。これは時代・地域によって語彙や文体等に違いがあるものの、理念的には、規範化がなされた 8 世紀以降、アラブ世界共通の書き言葉として使用されてきたアラビア語である。現代においてもフスハーはアラビア語使用者の共通語とみなされており、書記目的としてだけではなく「公的な場で話す」際にも使われる。テレビやラジオでのニュースはフスハーが用いられており、インタビューなどの対話（例えばアナウンサーと評論家の間などやや形式ばった対話）ではフスハーに準ずる言葉が使われる。ただし、あらかじめ用意された原稿を読み上げるわけではないので口語的要素が含まれる場合もあり、くだけた対話であれば口語の出現率はさらに上がる。つまり、この Spoken コーパスに関してはアラビア語の複数の変種が混在している可能性があり、データの取り扱いには注意が必要となる。

最後の Conversational Data とは電話での会話を採取したものである。こちらは、ほぼ口語アラビア語で構成されていると考えてよく、a のエジプト方言、b のレバント方言（レバノン・シリア・パレスチナを含む）の二種類が完成している。いずれもペンシルバニア大学の LCD で作成されたもので、音声からテキストへ変換する音声認識技術の開発を目的とする。

1．2　口語アラビア語コーパスの意義

以上のアラビア語コーパスの現状を見ると、Written Data、すなわち正則アラビア語のコーパスが圧倒的な数を占めており、口語アラビア語とのバランスの悪さが目に付く。口語アラビア語はアラブ人にとっての母語にあたり、単にレジスター（使用域）の問題として片付けられるものではない。外国語としてのアラビア語教育分野においては、伝統的に正則アラビア語が対象となってきたが、近年は正則アラビア語に加えて口語アラビア語も対象に入れるべきであるという考え方が定着している[10: 35-37]。また、アラビア語の言語学的研究においても、自然言語である口語アラビア語を扱うものが増えている。従って今後口語アラビア語コーパスの必要性が高まることは十分予測される。

本来口語アラビア語は、「書く」目的には使用されなかったため、その一次資料（記録）の少なさが口語アラビア語研究上の障害となってきた。しかし、19 世紀には口語アラビア語

[4] このようなコーパス加工技術に関する研究は近年増加しており、様々な提案や研究報告がなされている（例えば[4]を参照）。

で書かれたテキストが現れ、さらに20世紀初頭の映画メディアの誕生以降は、口語アラビア語の音声が映像とともに記録されるようになった。出版・映画大国エジプトにおいては、上記のような一次資料が豊富に存在する。本報告はこのようなエジプトの口語アラビア語(エジプト方言)資料を活用して、言語研究・教育ツールとしての口語アラビア語コーパスを作成し、これまでの一次資料不足を改善したいと考えている。またこれは、既存のアラビア語コーパスの手薄な部分を補うという点においても寄与できるものと思われる。

2. 口語アラビア語の表記上の問題

2. 1 二つの表記法

本節では、口語アラビア語コーパスを構築するにあたっての問題点について述べる。規範化された正則アラビア語には正書法が確立しているが、口語アラビア語では地域によって使用される語彙、音韻、形態が大幅に異なり、正書法も定められていない。つまり、既存の正則アラビア語コーパスの形式を、そのまま口語アラビア語コーパスに適用できるわけではない。まずは音価の転写規則から検討する必要がある。

最も大きな問題は、口語アラビア語を記述するのに二通りの表記法が存在することである。この背景には、口語アラビア語の体系的な記述がアラブ人の手によってではなく、ヨーロッパ人東洋学者によって始められたという事実がある。子音文字で構成されるアラビア文字が母音を正確に表記できないという理由から、東洋学者らは口語アラビア語の表記にラテン文字を用いた(アラビア語固有の音価に対しては、独自の付加記号をつけた文字もしくはIPA記号を追加)。そして、これが口語アラビア語記述のスタンダードな表記法となった。

一方、アラブ人(主にエジプト人)の間でも、19世紀以降意図的に口語アラビア語を用いた著述が試みられ、現代においては口語アラビア語による作品、特に戯曲が数多く創作されている。そこで採用されているのは、当然ながらラテン文字表記ではなく、正則アラビア語の正書法にある程度まで準拠したアラビア文字表記である。口語にしか存在しない形態や語彙についても、その多くは認知された慣習的表記法をもつが、書き手によっては異なる表記法をとることもある。なお、母音情報を欠くという点は、アラビア語母語話者である書き手・読み手の双方にとって全く問題にはならない。

この口語アラビア語に対する二種類の表記法の並存は現在まで続いており、言語研究にはラテン文字系の転写文字が、そしてアラブ人の著作物にはアラビア文字が使われるという傾向がある。コーパス作成を念頭に置いた場合、この二種類の表記法にはいずれも一長一短がある。転写文字表記は、子音と母音の両方が記述されるため、特に外国人学習者にとっては便利であり、語の解釈を誤ることも少なくなる。しかしエジプト方言では、前後の音環境によって長母音の短母音化、短母音の消失、子音の同化等が起こりうる。これらの現象を忠実にデータに反映させるとすれば、表記に揺れが生じることになり、検索には何らかの工夫が必要となる。一方、アラビア文字表記は、ある程度正則アラビア語の正書法に則るため、音環境による変化に関わらず表記には揺れが生じない。つまり一つの辞項は一種類の文字列に限定される。母音情報が含まれないため、検索時に母音による区別ができず検索結果が絞り込みにくい面はあるが、検索もれがないという多少の利点はある。またアラビア文字のほうが視覚的に把握しやすい場合もある。本コーパスでは、双方の

利点を生かすために二つの表記方法で出力しつつ、検索の問題点を克服する手法を検討する。

2．2　コード化と転写文字

本節では，アラビア文字のコード及び転写文字に関する一般的問題について触れる。アラビア文字はかつて、プラットフォームごとに独自の文字コード（Windows Arabic (CP1256), IBM Arabic (CP420), Mac Arabic, ISO Latin / Arabic (ISO 8859-6)など）が用いられていた。当時はWindows と Mac のアラビア文字コードの間には互換性がなくデータ共有が困難であったし、WWW が普及し始めた頃も、HTML ファイルの作成に使用された文字コードが原因で、アラビア語テキストが文字化けを起こす不具合が生じた。しかし、Unicode が普及した現在では、上記の問題はほぼ解決されている。プラットフォームに依存しない文字コードとしてのUnicode は、少なくともアラビア語に関しては機能しているため、本コーパスでも Unicode のアラビア文字を採用する。

一方、ラテン文字系の転写文字については、未だに多くの問題が残っている。一つには、転写文字に対する研究者間のコンセンサスがないことが挙げられよう。出版物においてさえも、転写文字の選択は著者に依存しているといってよい。Macユーザーには、Jaghbubというアラビア語転写用文字フォントがフリーで提供されており、これを用いれば正則アラビア語や一部の口語アラビア語を表記することができる。しかしWindowsにはこのようなフリーフォントが少なく、市販の転写文字フォントを用いるしかないのが現状である。もちろんこれらの異なる転写文字の間に互換性はなく、したがって入力の際には、できる限りOSの違いに依存しない文字コードによって行う必要がある。転写文字にUnicodeを用いる検討も行っているが、これについては今後の課題としたい[5]。

3．本コーパスの概要

3．1　データ内容

口語アラビア語のデータ源としてはエジプト映画を採用した。その利点として、1）収集面：エジプトは映画産業が盛んで 20 世紀前半から現在にかけて製作された多数の映画ビデオ・DVD が入手できる、2）言語データとしての価値：音声だけでなく映像という付加情報が含まれており再現性もある、3）語学教育ツールとしての可能性：多言語においては映画を用いた語学教育が盛んであるがこの手法はアラビア語にも応用できる、などが挙げられる。

すでに文字起こしが終わっているのは、以下の 5 作品である（丸括弧内は、監督名と制作年）。

Bayna l-Qaṣrayn (Ḥasan al-ʾImām, 1964)「バイナル・カスライン」

Al-Ḥarām (Henrī Barakāt, 1965)「ハラーム」

Al-ʾArḍ (Yūsuf Šāhīn, 1970)「大地」

Duʿāʾ al-Karawān (Henrī Barakāt, 1959)「チドリの祈り」

Bāb al-ḥadīd (Yūsuf Šāhīn, 1958)「鉄の門」

映画の選択基準としては、無意味な台詞が多く含まれるコメディ映画は避け、主にドラマを取り入れた。はじめの 4 作品はエジプトの有名作家による小説を映画化したものである（このうち 3 作品には邦訳もある）。映画の台詞の転写と入力はすべて手作業によって行われた。手順としては、上記映画作品の登場人物が発した台詞をその役名とともにアラビア文字に起こし、

[5] Unicodeに準拠しているフリーのIPAフォントとして、現在はDoulos SIL fontの利用を検討中である。

その手書きデータをラテン文字で入力するというものであるが、今回はエジプト方言母語話者の協力を得ることができた。

3．2　コーパスの構成

コーパス全体は3つのユニット（入力用・検索用ラテン文字テキスト、出力用アラビア文字テキスト、出力用転写文字テキスト）で構成される。検索に関して主軸となるのは、入力用テキストと、そこから変換スクリプトによって変換されたアラビア文字テキストである。各ユニットの詳細は以下の通りである。

1）入力用ラテン文字テキスト：コーパスの中核にあたる部分（検索用）

本コーパス構築のために新たに規定した、正則アラビア語正書法をベースにしたラテン文字表記で入力されたもの。ラテン文字とアラビア文字とは一対一で対応、つまり音韻情報は原則として子音のみである。アラビア文字の直接入力を避けた理由は、アラビア文字表記では一部の接続詞と前置詞、冠詞、接尾代名詞が他の実詞と連結してしまい、語の単位が明示できないからである。後のテキスト加工の効率のために、品詞ごとにハイフン等で区切りを入れる必要があり、これにはラテン文字が適している。また、加工の際にもラテン文字のほうが書記方向に煩わされることがなく柔軟性がある[6]。

2）アラビア文字テキスト：出力・検索用

入力用ラテン文字テキストから変換スクリプトにより自動生成されるアラビア文字テキスト。これにより、アラビア文字による生テキストが完成し、アラビア文字列での検索が可能となる。アラビア語用コンコーダンサーなど既存ツールも利用できると思われる。また1）に比べて視覚的把握が容易なため、入力ミスチェックにも役立つ。

3）転写文字テキスト：出力用・（検索（予定））

ラテン文字基盤の転写文字によって表記された、母音情報を含むテキスト。これは本コーパスの中では付随的な位置づけのユニットである。作業手順としては、1）のテキストから変換スクリプトにより転写文字へ自動変換した上で、映画の音声情報に従って確定した母音をそのテキストに挿入する。同じ語彙であっても、地域方言や前後の音環境によって異音が生じるため、これを機械的に処理することは難しいので当面は手作業に頼らざるを得ないだろう。このテキストは、読み方の確認あるいは提示用として利用できる。また、これを用いて適切な検索を行なうために、音環境による揺れを吸収するシソーラス的な辞書の構築も併せて検討したい。

3．3　入力の際の規則

1）入力文字

入力文字とアラビア文字は、以下のように対応させた（数字はUnicodeにおける各アラビア文字のコード）。各アラビア文字に対応する入力用ラテン文字は原則として一文字とし、大文字と小文字は区別する。ただし、一部ラテン文字二文字に、アラビア文字一文字を対応させたものもある。

---

6 アラビア文字とラテン文字(ASCII)との対応づけやハイフネーションについては、[1], [2], [3], [5]などでも採用されているが、目的（コーパスか組版）や扱うアラビア語の種類によって、その内容に多少のばらつきがある。ここでは、エジプト方言コーパスに適すると思われる規則を提案したが、今後さらなる標準化が期待される。

$	ء	0621	H	ح	062D	`	ع	0639
'a	آ	0622	X	خ	062E	G	غ	063A
'u	أ	0623	d	د	062F	f	ف	0641
'U	ؤ	0624	_d	ذ	0630	q	ق	0642
'i	إ	0625	r	ر	0631	k	ك	0643
'I	ئ	0626	z	ز	0632	l	ل	0644
@	ة	0629	s	س	0633	m	م	0645
A	ا	0627	c	ش	0634	n	ن	0646
b	ب	0628	S	ص	0635	h	ه	0647
t	ت	062A	D	ض	0636	w	و	0648
_t	ث	062B	T	ط	0637	Y	ى	0649
j	ج	062C	Z	ظ	0638	y	ي	064A

2）その他の規則

a. ハイフネーション

二つのスペースに挟まれた文字列の中に複数の語が含まれる場合、各語はハイフン "-" で区切られる。また動詞（未完了形）の語幹に接頭される接頭辞（現在の進行・習慣と未来を表すものの二種類）は、 "=" でつなぐ。

前後に連結する語の例（冠詞、接続詞、前置詞等）：Al-ktAb (the-book), w-qlt (and-I said) l-hA (to-her)

動詞の接頭辞の例：b=yktb (Present=he write), H=yktb (Future=he write)

b. 重子音の表記

二つの子音間に母音が含まれない場合、アラビア語の正書法では一つの子音字のみで表記される。この表記では、例えば動詞の基本形 (CvCvC) と派生形第２形 (CvCCvC) が区別されない点に問題がある。そこでラテン文字による入力においては、二つの連続子音を /ss/ のようにスラッシュで括る。間に母音を含む場合（つまり正書法で二つの子音字が書かれる場合）は、単に二つの子音を並べる（例：ss）。アラビア文字への変換プログラムは、スラッシュで囲まれた子音字を一文字だけ出力する。

以下に各テキストのサンプルを示す。

Ex. 1 入力用ラテン文字テキスト

```
'int `Arf yA fhmy  yA bny 'ubw-k S`b qwy .
w-AnA xAyf@ y`ny yqwl l-k l/ss/@ bdry .
H=Aqwl l-k klm@ b/ss/ Asm`y .
```

Ex. 2 自動変換されたアラビア文字テキスト

إنت عارف يا فهمي يا بني أبوك صعب قوي
وانا خايفة يعني يقول لك لسة بدري
حاقول لك كلمة بس اسمعي

Ex. 3 自動変換された転写文字テキスト

ʔnt ʕrf y fhmy y bny ʔbw-k ṣʕb qwy.
w-n xyf yʕny yqwl l-k lss bdry.
ḥ=qwl l-k klm bss smʕy.

Ex. 4 転写文字テキストに母音を追加したテキスト

ʔinta ʕaarif ya fahmi ya bni ʔabuu-k ṣaʕb qawi.
w-ana xayfa yaʕni yiqul l-ak lissa badri.
ḥa=qul l-ak kilma bass ismaʕi.

4．今後の展望

今後は、ラテン文字テキストに対するマークアップ等の加工による、より精度の高いコーパスの作成及び、転写文字テキストの有効な検索手法の検討による，効率的なコーパス検索を目指すと共に，アラビア語対応のコンコーダンスツール（XAIRA[7]やaConcorde[8]）も併用して、主に動詞の用例を調べるための具体的処理を試みる予定である。

付記

本報告は、平成17年度科学研究費補助金（若手研究(B))「口語アラビア語研究のためのコーパス開発とその応用」（研究代表者：大阪大学大学院言語文化研究科助手中道静香）及び平成17年度科学研究費補助金（特別推進研究(COE)）「アジア書字コーパスに基づく文字情報学の創成」（研究代表者：東京外国語大学アジア・アフリカ言語文化研究所教授ペーリ・バースカララーオ）による研究成果の一部である。

＜参考文献＞

[1] Buckwalter, T. and Maamouri, M. 2004. *Guidelines for Transcribing Levantine Arabic: AOST Transcription.* (http://www.ldc.upenn.edu/Projects/EARS/Arabic/Guidelines_Levantine_TRA.htm)

[2] Buckwalter, T. and Maamouri, M. 2004. *Guidelines for Transcribing levantive Arabic: MSA-based Transcription.* (http://www.ldc.upenn.edu/Projects/EARS/Arabic/Guidelines_Levantine_MSA.htm)

[3] Haralambous, Y. and Plaice, J. 1997. Multilingual Typesetting with Ω, a Case Study: Arabic. *International Symposium on Multilingual Information Processing '97,* Tsukuba.

[4] Khoja, S., Garside, R. and Knowles, G. 2003. A tagset for the morphosyntactic tagging of Arabic. In A. Wilson et al. (eds.) *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World,* Lincom-Europe, Munich.

[5] Legally, Klaus. 2004. *ArabTex Typesetting Arabic and Hebrew: User Manual Version 4.00.* (http://129.69.218.213/arabtex/html/arabdoc.pdf)

[6] Maanouri, M., Buckwalter, T. and Cieri, C. 2004. *Dialectal Arabic Telephone Speech Corpus: Principles, Tool design, and Transcription Convention.* Paper presented at the NEMLAR International Conference on Arabic Language Resources and Tools, Cairo, Sept. 22-23, 2004.

[7] Maanouri, M., Graff, D., Jin, H., Cieri, C. and Buckwalter, T. 2004. *Dialectal Arabic Orthography-based Transcription & CTS Levantine Arabic Collection.* Paper presented at the Parallel STT-NA Tracks Session of the EARS RT-04 Workshop, Palisades IBM Executive Center, New York, Nov. 10, 2004.

[8] Roberts, A., al-Sulaiti, L. and Atwell, E. 2005. aConcorde: towards a proper concordance for Arabic. In *Proceedings of Corpus Linguistics conference 2005,* University of Birmingham, UK.

[9] Van Mol, M. and Paulussen, H. 2001. AraLat: A relational database for the development of bilingual Arabic dictionaries. In *Proceedings of Asialex 2001, Asian Bilingualism and the Dictionary,* Seoul, August 2001, 206-211.

[10] al-Sulaiti, L. 2004. *Designing and Development a Corpus of Contemporary Arabic,* Master thesis, School of Computing, the University of Leeds.

---

[7] http://www.oucs.ox.ac.uk/rts/xaira/

[8] http://www.comp.leeds.ac.uk/andyr/software/aConCorde/