CRL Language Resources
Chinese and Arabic

James Cowie, Wanying Jin, Ahmed Abdelali, Hamid Mansouri Rad

MCCS-04-333

Computing Research Laboratory
MSC 3CRL
New Mexico State University
Las Cruces, NM 88001

The Computing Research Laboratory was established by the
New Mexico State Legislature
under the Science and Technology Commercialization Commission
as part of the Rio Grande Research Corridor

Table of Contents

Chinese Arabic Resources

This document summarizes the lexicon resources that CRL has developed or accumulated during its past projects, including Arabic and Chinese ontological lexica. It also includes the lexicon CRL has acquired from other sources, for example, the online parallel texts (in medical domain) acquired from Victoria, Australia; the slang and dialect resources in Arabic and Chinese (http://www.sa4sa.com/#7) as well as LDC's CALLHOME spoken corpora in Egyptian Arabic and Chinese.

CRL lexicon resources:

- Ontology with about 5000 concepts. We are using the last CRL version on **messene:9030**.

  (ABSORB DEFINITION VALUE "to suck up or take in, as a sponge" 0)
  (ABSORB IS-A VALUE NATURAL-EVENT 17)
  (ABSORB AGENT SEM *NOTHING* 10)

- English ontological lexicon with about 30,000 entries. Each entry contains information about word frequency from worldnet, (e.g. 1.149968E-5 in COMMENT line), definition, example, syntax, and semantics. The revision is needed to make sure each entry has the complete information. The most important work is to update the ontological concept mapping in semantics. We aware there are many errors in semantics.

  <RECORD><lexitem>**absorb**</lexitem><pos>V</pos><ws>3</ws><slot>**COMMENTS**</slot><lang></lang>
  <filler>1.149968E-5</filler><uid>0</uid></RECORD>
  <RECORD><lexitem>**absorb**</lexitem><pos>V</pos><ws>3</ws><slot>**DEFINITION**</slot><lang></lang>
  <filler>learn, larn, acquire knowledge, gain knowledge, acquire skills</filler><uid>0</uid></RECORD>
  <RECORD><lexitem>**absorb**</lexitem><pos>V</pos><ws>3</ws><slot>**EXAMPLE**</slot><lang></lang>
  <filler>(of knowledge or beliefs)</filler><uid>0</uid></RECORD>
  <RECORD><lexitem>**absorb**</lexitem><pos>V</pos><ws>3</ws><slot>**SYNTAX**</slot><lang></lang>
  <filler>(root ("V - Subject\(NP case:Nominative;\) - DirectObject\(NP case:Accusative;\)"
  (AGENT (Subject (NP (case:Nominative)))) (THEME (DirectObject
  (NP (case:Accusative))))))</filler><uid>0</uid></RECORD>
  <RECORD><lexitem>**absorb**</lexitem><pos>V</pos><ws>3</ws><slot>**SEMANTICS**</slot><lang></lang>
  <filler>(top (root:LEARN))</filler><uid>0</uid></RECORD>

- Arabic Ontological Lexicon
  About 3500 English entries with the Arabic translation

```
account-V1
(top ("root:ACCOUNTING FORMALITY:medium RESPECT:medium SIMPLICITY:low"))
DEFINITION  /((STYLE (ACCEPTABILITY:medium FIGURATIVE:no)))
Translations:حساب;حسب;

  (SYNTAX
   (1 (root
    (AGENT
     (Subject
     (NP
```

```
    (case Nominative)
   )
  ))
  (THEME
   (Adjunct1
   (PP
    (prep for)
   )
  ))
 ))
)
```

- Arabic-English bilingual dictionary:

The Arabic-English contains 67640 unique Arabic headwords with more than 120000 English translations

غادر:(aracode=r060;<leave>)
غادر:(aracode=r089;<depart>)
غادر:(aracode=r089;<leave>)
غادي:(aracode=r015;<coming>)
غادي:(aracode=r024;<coming>)
غادي:(aracode=r026;<coming>)
غار:(aracode=r013;<cave>)
غار:(aracode=r015;<attack>)
غار:(aracode=r015;<foray>)
غار:(aracode=r015;<raid>)

- Arabic-English proper names dictionary:

Around 1000 entries combines state capitals, contemporary personalities and head of states.

الثـــاني ربيـــع:(aracode=r058;proper=time_period;<rabi thani >)
اردوغان طيـب رجب:(aracode=r058;proper=name;<Recep Tayyip Erdogan>)
رجب:(aracode=r058;proper=time_period;<rajab >)
رضـا:(aracode=r058;proper=name;<reza>)
رفســـــنجاني:(aracode=r058;proper=famname;<rafsanjani>)
رمضـان:(aracode=r058;proper=name;<ramadhan >)
رمضـان:(aracode=r058;proper=time_period;<ramadhan >)
رودريغـــيز:(aracode=r058;proper=name;<rodriges>)
روزنبـــورغ:(aracode=r058;proper=city;<rosenberg>)
روســـيا:(aracode=r058;proper=country;<russia>)
روما:(aracode=r058;proper=city;<rome>)
ئـــيرزا:(aracode=r058;proper=country;<zaire>)

- Arabic-English onomasticon dictionary:

About 186210 translation and transliteration for entities names, places …etc.

بـــلاف كوفـــي:(aracode=r058;<coffee bluff>)
لنـــــدينغ بـوت كوفـــي:(aracode=r058;<coffee pot landing>)
بوينـــــت يكوف:(aracode=r058;<coffee point>)
بـــي كوفـــي:(aracode=r058;<coffee bay>)
تشـــــيز كوفـــي:(aracode=r058;<covey chase>)
ران كوفـــي:(aracode=r058;<coffee run>)

ريـج كوفـــي:(aracode=r058;<coffee ridge>)
ســـــبرينغز كوفـــي:(aracode=r058;<coffee springs>)
ســـــيتي كوفـــي:(aracode=r058;<coffee city>)
كريـك كوفـــي:(aracode=r058;<coffee creek>)
مراكش:(aracode=r058;<marrakech>)
اينـك لابوراتـــــوريس الكـــــترونيكس مـديكال .:(aracode=r058;<medical electronics laboratories inc .>)
اينـك لابوراتـــــوريس الكـــــترونيكس مـديكال:(aracode=r058;<medical electronics laboratories inc>)
ســـــبرينغز مـديكال:(aracode=r058;<medical springs>)

- Arabic dialects and slang from the web
  The Internet forums provides a significant amount of data in local dialects and using local slang, alsmot every region in the Arab world has it own dialects. The dialects could be grouped in 4 major groups

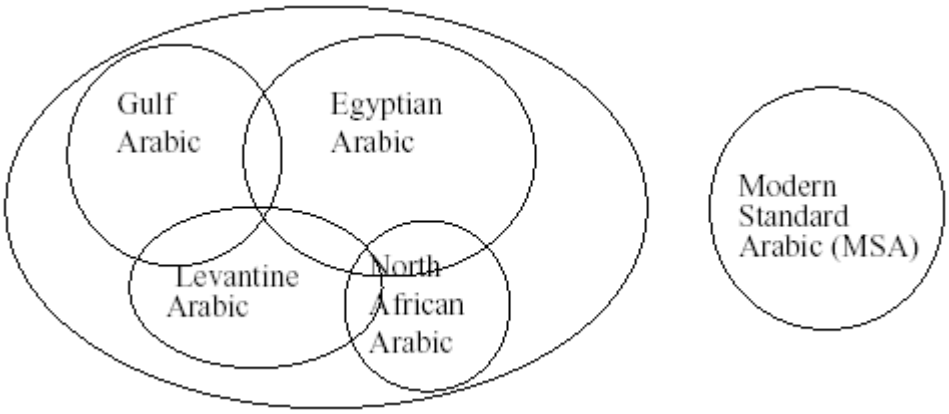

Figure 1. Arabic language variants

Resources

- http://www.sa4sa.com/#7 lists more than 100 Arabic internet forum

علينـــا رأسـها يكـــبر لا اسمها شـــيلو ~~~>> ؛؛ عـــــــذوب ؛؛
يمكـن كثـــير للنـــت تجـي تقـدر ما ظروف بـس موجوده الدبـــه
بالمنتـــــدى تشـــــرفنا ما بـس موجوده عاد ذي رومـانس معكـم اضـــيفو
للمنتـــدى يجـي ما بالشـــات بـس موجود بعـد حـظ
عليـك كلهـم كذا فيهـم وش
الأيميـــل علـى إليـك رسـالة بإرســال إما IP عـرفو
وقـاموا جهازك فـي Port بفتـــح البـــاتش وقـام جهازك إلـى مخفـي بـــاتش بإرســال قـاموا او بإختراقـــه
... سـامي أخ
..... آخر شـيء أي من أكـثر الارم الـزون مشـكلة هذه أعتقـد
... مصدرها وتلحـق ألارمات تعـد لحـتى بـس البرنـــامج ركبـت وكانـك

Expert from Arabic forum
http://www.al-atheer.com/vb/showthread.php?s=b0d594944a6d8bbf40a267678ef55f27&threadid=21415

caDub ~~~>> $ylu ismaha la yakbar rasha calyna
ymkn kTyr llnet tigy tqdr ma Zruf bs mwgwdh Aldbah
Difw macakom rowmAns Dy cAd mawguwdah bs mA t$arafnA bAlmuntadaN
HaD bacd mawguwd bAl$at mA ygy llmuntadaN

wa$ calyk fyhum kaDA kolhum calyk
crfow "IP" ImA bIrsal risalaT Ilyk calN AlIymayl Aw Qamow BirsaAl bAt$ maxfy lIN gihazek wQaAm AlbAt$ bfatH "Port" fi gihazek wqaAmow bKtiraqh
Ax saAmy….
Actaqid haDh mo$kilat Alzown Alarm
WkaAnk rakabt Albarnamag bas lHataN tcd AlarmAt wtalHaq maSdarha

English Transcription of the Arabic text from the forum.

The following table highlights new phenomena of foreign word borrowing and usage in Arabic dialects, words extracted from the above transcription.

Word	English	Comment
llnet	to the net	**ll** prep+det "to the"
rowmAns	romance	
bAl$at	in the chat	**bAl** prep+det "in the"
AlIymayl	the email	**Al** determinant marker
bAt$	Batch	
AlbAt$	the batch	**Al** determinant marker
Alzown Alarm	zone alarm	
AlarmAt	the alarms	**At** plural marker suffix

- LDC resouirces includes spoken and transcribed Arabic dialect (`CALLFRIEND Egyptian Arabic, CALLHOME Egyptian Arabic Speech`).

198.09 200.86 A: تـــانى أرجع و كده مثـلا ؛لشـــبرا عنـدنا من زى أمشـى {laugh}
201.38 202.42 B: برضـــه كـلام ده و طب
202.67 210.48 A: لا هى ميزة البلـــد انها هنا am- يعـــنى من آمن البـــلاد البـــيرة الكبـــيرة
لوحده يمشـــى حد إن جدا خطر <English Chicago> يعـــنى جدا خطر الكبـــيرة
211.19 212.11 B: احنـا اللـــى ده هو_ما طيب bin-
211.16 213.19 A: بالليـــل ولا الصـــبح وقـــتى اى فـــى
212.72 214.15 B: اه! طب
213.47 215.88 A: هنا لكـن am- يعـــنى امان هنا البلـــد امنـة
216.07 217.26 B: من بالـــك خد برضـــه لا
216.84 220.21 A: يعـــنى امان فيعتـــــبر هنا بتـــاعهم الـــريف يعـــنى تعتـــبر تعتـــبر
220.82 223.56 B: معروف إعمل نفســـنا من بالنـــا ناخـد احنـا مبـرض ده هو_ما لا
222.31 227.17 A: م! تمـام لله_الحمد بالنـــا واخـدين لا ilH- لا م!
225.48 231.93 B: يعـــنى منمشـــيش برضـــه لوحـدنا منمشـــيش نفســـنا من بالنـــا ناخـد لا هو اللـــى

Experts from CALLHOME Egyptian Arabic Transcripts ar_4482.scr

198.09 200.86 A: am$i zayy min candina li+&$ubra masalan kida wi argac tAni {laugh}
201.38 202.42 B: Tab wi da kalAm barDu
202.67 210.48 A: la hiyya mIzaB il+balad hina innaha am- yacni aCman min il+bilAd il+kibIraB il+bilAd il+kibIraB xaTar giddan yacni <English Chicago> xaTar giddan inn Hadd yim$i li+waHdu
211.19 212.11 B: Tayyib mahu da illi iHna bin-
211.16 213.19 A: fi ayy waqti il+SubH walla bi+il+lEl
212.72 214.15 B: Tab %ah
213.47 215.88 A: lAkin hina am- yacni amAn hina il+balad hina amnaB
216.07 217.26 B: la barDu xud bAlak min
216.84 220.21 A: tuctabar tuctabar yacni il+rIf bitachum hina fa+yuctabar amAn yacni
220.82 223.56 B: la mahu da barDu iHna nAxud balna min nafsina icmil macrUf
222.31 227.17 A: %M la ilH- la waxdIn balna ilHamdulillA tamAm %M
225.48 231.93 B: la nAxud balna min nafsina manim$I$ li+waHdIna barDu manim$I$ yacni illi huwwa <Upper biygullak> biyiTlacu=a musallaHIn wi illi biyiTlacu=a mi$ cArif Eh

Experts from CALLHOME Egyptian Arabic Transcripts ar_4482.txt

- Chinese ontological lexicon has two versions:
  The small version at size of 2,250 has complete information for each entry. The Chinese word 吸收 corresponds to the English word **absorb** with the concept **ABSORB.**

<RECORD><lexitem>吸收<lexitem><pos>V</pos><ws>1</ws><slot>COMMENTS</slot><lang></lang>
<filler>**absorb**</filler><uid>0</uid></RECORD>
<RECORD><lexitem>吸收>/lexitem><pos>V</pos><ws>1</ws><slot>DEFINITION</slot><lang></lang>
<filler>a process in which one substance permeates another; a fluid permeates or is dissolved by a liquid or solid.</filler><uid>0</uid></RECORD>
<RECORD><lexitem>吸收</lexitem><pos>V</pos><ws>1</ws><slot>SEMANTICS</slot><lang></lang>
<filler>(top (root:**ABSORB**))</filler><uid>0</uid></RECORD>
<RECORD><lexitem>吸收</lexitem><pos>V</pos><ws>1</ws><slot>SYNTAX</slot><lang></lang>
<filler>(root ("V - Subject\(NP\) - DirectObject\(NP\) - Adjunct1\(PP prep:‘´Ó;\)"
(AGENT (Subject (NP))) (THEME (DirectObject (NP)))
(SOURCE (Adjunct1 (PP (prep:‘´Ó))))))</filler><uid>0</uid></RECORD>

Large version with size of 22,000 is automatically translated from English lexicon above using on-line LDC English-Chinese dictionary. All the Chinese translations have different meanings. Not every Chinese word has the meaning **ABSORB.** Disambiguation needs to be done manually.

<RECORD><lexitem>/吸收/吸进/缓和/合并/吸引/使全神贯注/吸取/摄/<lexitem><pos>V</pos><ws>1</ws>
<slot>COMMENTS</slot><lang></lang> <filler>**absorb**</filler><uid>0</uid></RECORD>
<RECORD><lexitem>/吸收/吸进/缓和/合并/吸引/使全神贯注/吸取/摄<</lexitem><pos>V</pos><ws>1</ws>
<slot>DEFINITION</slot><lang></lang> <filler>a process in which one substance permeates another; a fluid permeates or is dissolved by a liquid or solid.</filler><uid>0</uid></RECORD>

<RECORD><lexitem>/吸收/吸进/缓和/合并/吸引/使全神贯注/吸取/摄/</lexitem><pos>V</pos><ws>1</ws>
<slot>SEMANTICS</slot><lang></lang><filler>(top (root:**ABSORB**))</filler><uid>0</uid></RECORD>
<RECORD><lexitem>/吸收/吸进/缓和/合并/吸引/使全神贯注/吸取/摄</lexitem><pos>V</pos><ws>1</ws>
<slot>SYNTAX</slot><lang></lang><filler>(root ("V - Subject\(NP\) - DirectObject\(NP\) – Adjunct1\(PP prep:从;\)" (AGENT (Subject (NP))) (THEME (DirectObject (NP))) (SOURCE (Adjunct1 (PP (prep:从))))))</filler><uid>0</uid></RECORD>

- Ceta Chinese –English dictionary contains 218,000 entries. It includes regular words, proper name<PN>, transliteration <TL>,colloquial <COLLOQ>, figurative<FIG>, idiom<ID>, new Chinese usage<NCU>, etc. Examples are:

吸收 TO ABSORB, TO ASSIMILATE; TO RECRUIT, TO ENROLL
美国 〈PL〉 AMERICA, THE UNITED STATES OF AMERICA
惠灵顿 〈PL〉〈TL〉 WELLINGTON (CAPITAL, NEW ZEALAND)
傻里傻气 〈COLLOQ〉 STUPID, SILLY, FOOLISH
一分为二 〈NCU〉 ONE DIVIDES INTO TWO (MAO ZEDONG'S THEORY OF DIALECTICS IN WHICH EVERY PHENOMENON ENCOMPASSES TWO MUTUALLY OPPOSING, AND AT THE SAME TIME MUTUALLY UNITED ANTITHESES WHICH ARE SIMULTANEOUSLY IN THE STATE OF UNITY AND STRUGGLE AND CAPABLE, UNDER CERTAIN CONDITIONS, TO BE TRANSFORMED INTO THE OPPOSITE. ACCORDING TO MAO THIS PRINCIPLE MUST PROVIDE THE BASIS FOR ANALYSIS AND SOLUTION OF ALL CONTRADICTIONS, ESP. CONTRADICTIONS AMONG CLASSES)

- CRL proper name list
  - Foreign person's name list with 34,179 names. (See attachment)

阿巴约米	Abayomi	s
奥克塔维厄斯	Octavius	s, m
奥克塔维亚	Octavia	f

  - Foreign governmental figure's name

总理：库尔曼别克·巴基耶夫（２０００．１２．２１－ ）
Prime Minister: Kurmanbek Bakiyev
第一副总理：尼古拉·塔纳耶夫（２０００．１２．３０－ ）
First Deputy Prime Minister: Nikolai Tanayev
副总理：埃先古尔·奥穆拉利耶夫（１９９９．４．１２－ ）
Deputy Prime Minister: Esengul Omuraliyev

  - Foreign place names include country name, city name, state name, region names. (See attachment)
    - Country and capital names(about 240 countries):

安提瓜和巴布达		Antigua and Barbuda	安提瓜和巴布达	Antigua and Barbuda	
		圣约翰		St. John's	
澳大利亚		Australia	澳大利亚	the Commonwealth of Australia	
	堪培拉	Canberra			
奥地利	Austria	奥地利共和国	the Republic of Austria		
	维也纳	Vienna			

- State names:(US, Canada and China only)

怀俄明	Wyoming	夏延	Cheyenne
加利福尼亚	California	萨克拉门托	Sacramento
堪萨斯	Kansas	托皮卡	Topeka
萨斯客彻温	Saskatchewan	里贾纳	Regina
曼尼托巴	Manitoba	温尼伯	Winnipeg
安徽	Anhui	合肥	Hefei
福建	Fujian	福州	Fuzhou
甘肃	Gansu	兰州	Lanzhou

- 570 city names

德岛	Tokushima	Prefecture and city in Japan
德尔柏郡	Derbyshire	City of England
德黑兰	Teheran	Capital of the Iran
德累斯顿	Dresden	City of Germany
德里	Delhi	City of India

- 300 region names

安第斯山	Andes Mountains
巴波亚岛	Balboa Peninsula
包干维尔海峡	Bougainville Strait

Parallel texts on web:

Information of 206 countries at

http://www.bjfao.gov.cn/newsite/world/detail.asp?countryID=1

ID = 1 to 206

Victoria medical phrases of four diseases (Campylobater, Gastroenteritis, Giardiasis and Salmonellosis) in Chinese, Arabic and English parallel texts are at

http://www.dhs.vic.gov.au/phd/hprot/idci/camp.html
http://www.dhs.vic.gov.au/phd/hprot/idci/campdf/arabic.pdf
http://www.dhs.vic.gov.au/phd/hprot/idci/campdf/chinese.pdf
http://www.dhs.vic.gov.au/phd/hprot/idci/gastr.html
http://www.dhs.vic.gov.au/phd/hprot/idci/gastrpdf/arabic.pdf
http://www.dhs.vic.gov.au/phd/hprot/idci/gastrpdf/chinese.pdf
http://www.dhs.vic.gov.au/phd/hprot/idci/giar.html
http://www.dhs.vic.gov.au/phd/hprot/idci/giarpdf/arabic.pdf
http://www.dhs.vic.gov.au/phd/hprot/idci/giarpdf/chinese.pdf
http://www.dhs.vic.gov.au/phd/hprot/idci/sal.html
http://www.dhs.vic.gov.au/phd/hprot/idci/salpdf/arabic.pdf
http://www.dhs.vic.gov.au/phd/hprot/idci/salpdf/chinese.pdf

Slang and dialect:

- References:

1. Chinese-English Dictionary of Modern Chinese Slang, Fai Feng Publishing Co. in Hong Kong, 1998.

2. New Slang of China (最新中国俚语), New World Press (新世界出版社) in China, 2000.

3. A Modern Comprehensive English-Chinese Dictionary (现代综合大辞典), Shanghai Science and Technology publishing Co. (上海科技出版社), Third edition, 1995.

4. A New English-Chinese Dictionary (新英汉词典), Shanghai Translation Publishing (上海译文出版社),. Second Edition, 2001.

5. A Dictionary of Chinese with English Translations (汉英语林), Shanghai Jiao-Tong University Press (上海交通大学出版社), 1992,

6. An English-Chinese Dictionary of Science and Technology (英汉技术词典), Defense Industry Press (国防工业出版社), 1992

7. A Comprehensive Chinese-English Dictionary of Traditional Chinese Medicine(中医药大词典), World Library Publishing Co. (世界图书出版社), 1997.

- Examples: (See attachment)

Slang:

逗悶子: Joking around
没事儿别在这儿**逗悶子**.
(If you have) nothing to do here, don't be **joking around**.

侃: Boast; brag; talk big
你去跟他聊聊,他可能**侃**呢.
You have a chat with him. He is good at **bragging**.

Dialect:

今朝夜道吃啥？没地方**混**，我请侬吃饭好伐?
What do you eat tonight? (If) no place to get (free meal), what about I invite you for dinner?

侬讲！侬到底爱俄伐？侬今朝勿讲清爽,俄死百侬看!
You tell (me)! Do you love me after all? (If) you do not make it clear today, I let you see I will die.

Dialect from the CALLHOME spoken corpora:

The texts in call-home phone corpora look pretty standard. Not many slang or dialect.

463.02 467.74 A: 对 她 已经 在 &香港& 待 了 两 年 , 因为 她 男朋友 好象 亲戚 在 &香港& , ((什么)) 这 乱七八糟

388.13 391.44 A: 挺 咿 , 乱七八糟 一 大 堆 事情 我 前一阵子 他 前一阵 子 给 我 写 信 我 一直 没 给 (我)) [[distortion]]

294.97 299.23 A: 哎 , 对 , 晚上 再 吧 , 搞搞 了 半夜三更 的 , 肚子 饿 的 呱呱叫

300.01 303.45 A: 你 实在 不行 , 你- 你- 给 他 塞 五十 块钱 那个, 那个 , 那个 美金 给 他

258.62 262.98 A: 既然 上面 这么 说 了, 谁, 谁 肯 到 让 他 不 愉快 , 那 就 这么 ((着))了

226.74 227.67 B: <Shanghai_((咯么))> 谁 教 他 啦 ?

532.63 536.33 B: <Shanghai_呶> , 旧 的 我们 自己 在 利用 么 , 万一 <Shanghai_呶> 不好 的话 , 还好 看看 别的

655.67 666.76 A: 开幕式 好象 在 +芝加哥+ , 反正 我 当时 想 跟 你 买 一件 那个 , +芝加+ 足球赛 的 那种 T 恤 衫 哦 , 后来 想想 也 算了 , {laugh} 你 大概 也 不会 , 上班 也 不会 穿 {laugh} 的 {/laugh} .

Victoria medical phrases:

Chinese	English	Arabic
贾弟鞭毛虫病	Giardiasis	داء الجيارديات
贾弟鞭毛虫病**是什么病**?	What is Giardiasis?	الجيارديـــات داء ماهو
贾弟鞭毛虫病**有哪些症状**?	What are symptoms of Giardiasis?	الجيارديـــات داء أعراض ماهي
哪里有贾弟鞭毛虫?	Where are Giardia found?	الجيارديـــات توجـد أيـن
贾弟鞭毛虫病**如何**传**染**?	How does Giardiasis spread?	الجيارديـــات داء ينتشـــــر كيـف