Collaborative Construction of Arabic Lexical Resources

Mohammad Daoud (1), Daoud Daoud (2), Christian Boitet (1)

(1) Laboratoire LIG - Université Joseph Fourier
385, rue de la Bibliothèque, 38041 Grenoble, France
{Mohammad.Daoud, Christian.Boitet}@imag.fr
(2) Princess Sumaya University
P. O. Box 1438 Al-Jubaiha 11941 Jordan
Daoud@batelco.jo

Abstract

The absence of free usable lexical and syntactic resources and tools for Arabic makes it a “pi-language” (poorly informatized). This may be the main reason for making the process of transferring knowledge into it very difficult, especially in technical domains. To demonstrate the size of the problem: the translation of EOLSS (EOLSS 2008) (Encyclopedia Of Life Support Systems), which contains about 200,000 pages, probably requires to find or create translations for at least 250,000 terms, an effort which is estimated at about 25,000 working hours.

We are proposing a collaborative methodology that involves domain experts, linguists, terminologists and normal Internet users in the process of developing a domain-dedicated Arabic terminological database by facilitating their contribution and collaboration. The collaborative process would replace the expensive and infeasible (especially for Arabic) traditional approach. As an intermediate phase towards an Arabic terminological database, we aim at constructing a *preterminological* database (pTMDB) which *contributability* should be far higher than a true terminological database.

Introduction

Nowadays there is a massive production of terms in various domains (50 terms daily, 17,500 yearly), most of them being initially created in English. This massive production makes it harder to build a large enough term bank (Alfadhel 2007) for Arabic, because it is first needed to absorb these new findings at the content level (science, technology), before doing a comparable effort of term creation at the linguistic level. That is why there are many modern terms that do not have corresponding Arabic terms (Yassin 2003). The problem does not come from some linguistic resistance of Arabic: on the contrary, the Arabic language historically proved to be very apt at absorbing new cultures and scientific terms. Before the year 600, Arabic was the language of the people who lived in the Arabian Peninsula. After the advent of Islam, millions of people from different races and cultures became Muslims, and Arabic language successfully absorbed the terms and expressions of their cultures and scientific heritage.

Previous attempts made in this area were few and scattered. Moreover, they focused on solutions relying on high expertise on the linguistic side, not considering the possibility to use modern technology and exploit efforts of various groups, especially domain experts, to build computerized term banks accessible by different users, clients as well as volunteer or paid contributors.

The only notable exception is the work on *unified Arabic terminology* started and done for many years by Pr Lakhdar El-Ghazal with his group at IERA in Rabat (Ghazal 1977), until he retired around 1995. They built an impressive computerized lexical 3-lingual database (Arabic, French, and Latin) of 800,000 term pairs from more than 120 sources, and examples of unified terminology for some domains (police, insurance, banking…).

In this paper we propose a system for acquiring domain-specific Arabic lexical and terminological data (LexAR) using various approaches, most importantly, the collaborative approach, where people who are interested in the domain would contribute to LexAR through an online platform. The collaborative approaches would be used instead of the traditional approaches which have many limitations that we will explain. However, acquiring contribution from online users has its problems as well, such as the motivation, quality of the contributed data, coverage, etc. In an attempt to avoid these problems and reduce their negative effect, we propose to first build a kind of *intermediate contributable repository*, or more precisely a *preterminological multilingual database* (pTMDB). Volunteers will contribute to the pTMDB, which will then be used as “raw material” by terminologists to build LexAR.

This paper is organized as follows: in section 2, we will describe the current scene, section 3 will present the main concept of LexAR and its relation with the Arabic-pTMDB, section 4 will describe various approaches to acquire “Arabic preterminological data”, section 5 will present other useful applications for such an Arabic-pTMDB, and we will conclude by trying to estimate the cost, feasibility and efficiency of the approach.

Current Scene

History

One of the earliest people who worked on collecting Arabic lexical resources is Khalīl ibn Ahmad Al Farāhīdi (died 776 C.E.), in his book (“Kitab Al-‘ayn” means the book of the letter ‘ayn ع) the name came from the fact that the dictionary follows a phonetic order starting from the pharyngeal sound ع. Many dictionaries followed as there was an urgent demand on learning Arabic from new Muslims with different linguistic backgrounds, for example Al-Azhari’s (died 986) (tahthib al-lugha, " تهـــذيب اللغـــة "),

and Al-Jawaheri's (died 1004) (al sihah, الصحاح), which followed a phonetic or alphabetic order. Later people showed an interest in the conceptual organization rather than regarding the exclusive consideration of the lexical units. "ketab al ta'rifat" (the book of definitions) for Al-Jurjani (died 1417) was a clear example of this interest, amongst others in the fields of arts, Islamic studies, basic sciences, and more.

Directly after the Ottoman era, Arabs realized the lexical shortcomings of modern standard Arabic, especially when they tried to translate the western works into Arabic. Attempts to coin equivalences were (and are still) scattered and not sufficient. That is why many intellectuals use loanwords from European languages in their domains.

Available online systems

Notable examples

Recently, many free online systems offer the service of translating technical terms from one language to another, using a bilingual/multilingual terminological database, but few of them deal with Arabic terminology. Table 1 shows some examples of such systems.

Name	Number of terms (ML)	Number of Languages	Domains	Provider
IATE (IATE 2008)	1.400,000 terms	23 languages, without Arabic	General, 155 domains	EU
UNTerm (UN 2008)	80,000 terms	6 languages, with Arabic	100 subjects related to the UN	UN
FAOTerm (FAO 2008)	58,000 terms	7 languages, with Arabic	FAO related domains and bodies	FAO
Electropedia (IEC 2008)	20,000 terms	9 languages, with Arabic	Electrical terminology in 75 cats.	IEC
UMD (WHO-EMRO 2009)	N\A	11 languages, with Arabic	Medical Terms	WHO
WDMG (IDRC 2009)	750	3, with Arabic	Water demands domain	IDRC
The Great Terminological Dictionary (OQLF 2008)	3,000,000 terms	Fr, En, and Latin	200 categories	OQLF

Table 1: Current systems

From the table we can conclude the followings:

- It is clear that the providers have mature resources and experience to build such databases. In fact, those online systems are continuations of efforts started decades ago, and they have been compiled using material from older and existing databases. For example, IATE, provided by the EU, has been constructed by compiling previous databases, namely EURODICAUTOM, EUTERPE, and (TIS).
- The number of terms in the table is the number of multilingual records, but some proportion is not complete with regard to languages, especially so for Arabic, for example the English Wikitionary has more than one million entries while the Arabic one has less than 22,000, and that is not even a specialized dictionary.

Traditional Construction Method

The traditional process of constructing a terminological database, such as the databases shown in the previous section, is as follows (Cabre and Sager 1999). As illustrated in Figure 1, the process usually starts with a thematic analysis of the domain, to find its logical components. After that, a team of terminologists consults related documents to find the most important terms for each sub-domain. Then, for each targeted language, a team of terminologists proposes one or more terms in the target language as translations of each term extracted in the source languages, along with descriptive information. Finally, each entry is verified to reach agreement on its correctness.

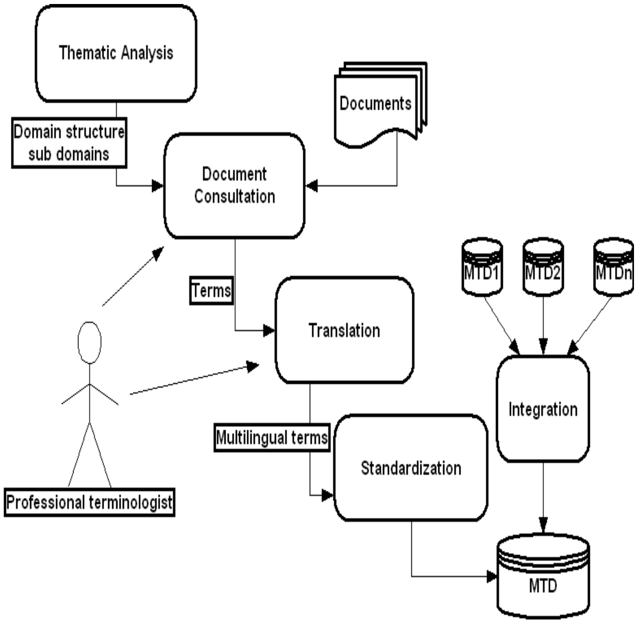


Figure 1: traditional method for constructing MTDs

The initial problem that we can figure out here is the exhausting human efforts that are put in extracting, translating, and verifying the terms. In the following subsection we will discuss the problems of this approach in details.

Limitation

As a result of depending on professional terminologists, building an Arabic term base is very expensive, especially when it comes to identifying and extracting a term, defining it, and translating it into another language. An entry may need hours of professional time. The following points are the rest of the most important limitations:

1. Coverage:

(a) Lexical coverage: in the traditional approach, extracting the terms is done by manually consulting related documents, which might not have a reasonable amount of technical terms of the domain, especially with the absence of Arabic technical documents in some domains in the first place. Beside, without computer help, human terminologists may miss some terms, although they are contained in the set of documents.

(b) Linguistic coverage: there might not be enough Arabic terminologists in specific domains who can handle such a job.

2. Involvement of domain experts: the traditional approach may create a gap between the terminologists and the subject matter experts, as they are not directly involved in the decision and the technicality of creating adequate entries in Arabic. The main reason why intellectuals in the Arabic world use English or French terms is perhaps that the translated term was coined by a terminologist without the consultation of a domain expert, who would not have sanctioned it.

3. Because terminology is alive, thousands of terms are produced yearly, and a rigid approach will not solve the problem, especially in the case of the Arabic language, where the pace of scientific production seems to be slower.

Collaborative methods

As a results of the limitations of the traditional methods we described previously, and the recent success of some collaborative applications, a possible and (we think) necessary solution is to depend on volunteers (not versed in terminology) knowing quite well the domain at hand in building an Arabic *preterminological* database.

For example, ITOLDU (Bellynck, Boitet et al. 2005) collected 17000 English-French terms in 20 technical domains from 250 French students (learners of English). Yakushite.net (Murata, Kitamura et al. 2003) is another example, where users contribute to bilingual dictionaries (organized following a domain hierarchy) that are used to enrich both the online Pensée machine translation system and the human translation aids. Also, Papillon (Sérasset 1994; Sérasset 2004; Serasset, Brunet-Manquat et al. 2006) is a Jibiki-based (Mangeot 2006) general purpose collaborative multilingual lexical database.

The problem with building terminological databases collaboratively is that it is difficult to attract domain experts to contribute: in the examples mentioned above, one can not expect massive contribution from normal people who are only visiting the database, and one can even less expect volunteers to replace professional terminologists. A volunteer could translate a term, but s/he may not be able to give full descriptive information about a term (its definition, usage, domain, context…), and, if s/he may, it will certainly not be in the way a professional terminologist would.

Beside, one should think of the general knowledge gathering problems described by (Richardson and Domingos 2003), in the domain of collecting multilingual terminology, which are:

(1) quality
(2) consistency
(3) relevance
(4) scalability
(5) and most importantly the motivation to contribute.

Another point to be considered is that such a database should be *seeded*, so that visitors can find initial data to start contributing. For that, using online resources seems to be a very promising option. Projects such as MultiMatch (Jones, Fantino et al. 2008), and PanImage (Etzioni, Reiter et al. 2007) use Wikitionaries (Wikitionary 2008), Wikipedia (Wikipedia 2008), and other online dictionaries to solve this problem.

Arabic-based pTMDB Towards "LexAR"

The Vision of LexAR

LexAR focuses on building a centralized web-based information system used to gather and disseminate multilingual terminological data containing Arabic terms. This system should provide the following functionalities:

- Creation and edition of a term entry.
- Advance search for particular terminological entries.
- Inclusion of some semantic attributes and reference to specialized ontologies of the tackled domains, and encyclopedic information if available.
- Exchange facilities with other systems.
- In-built possibility to account for Arabic dialectal variants.

The size of LexAR depends on the number of domains; a too optimistic vision is to have Arabic terms for all the available English or French at a large repository such as IATE (1.4 millions), but that is quite unrealistic, knowing the history of IATE, however starting by a reasonable set and focusing on the relevance of the included entries can bring eventual collaborative enlargement to database, and that what we are going to discus in the following subsection.

The concept of pTMDB

We are aiming at a massive contribution in term translation (in any direction) rather than at fully complete terminological records:

$Cx = Ts1, Ts2…Tsn \rightarrow Tt1, Tt2…Ttm$,

Where C is the contributed entry, Ts is the source term, and Tt is the term in the target language.

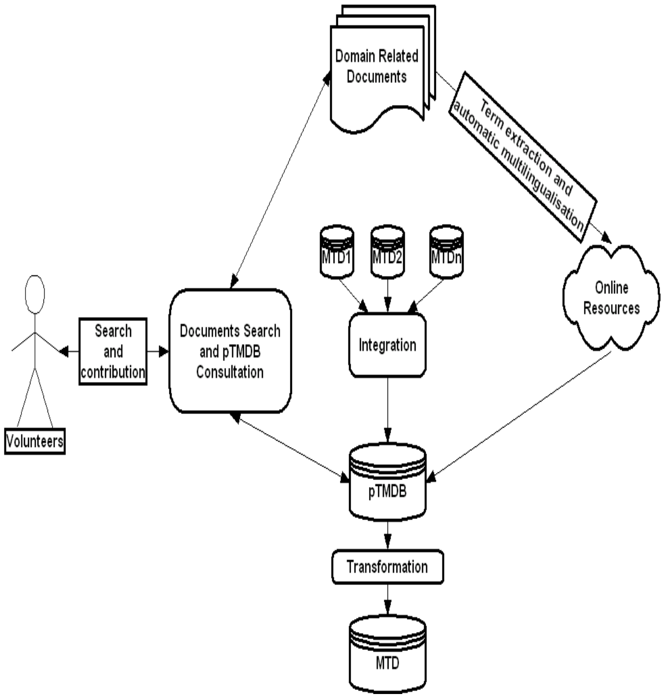


Figure 2: a proposed method for constructing pTMDBs

Such a pTMDB will be built by normal internauts interested in the domain and contribute spontaneously, with a sense of community, that is, to make it easier in the future for people of their language to access and use important information.
Figure 2 shows a possible process for building a pTMDB , using three main resources:
(1) preexisting Arabic databases to be integrated;
(2) volunteer contributors of the Arabic community (with the help of a domain related documents);
(3) online resources (with the use of domain-related documents).

Proposed acquisition methods

Semi-automatic method

In this approach, preterminological data is collected from multilingual online resources. As presented in the second section, the traditional methods depend on specialists in finding the terms from the related documents.
For a pTMDB, we assume there is a set of available related documents. We send them to an automatic term extraction engine, and produce candidate translations of the extracted terms automatically, using multilingual online resources such as wikipedias, online dictionaries...

Conscious contribution method

It is a participative approach, where volunteers contribute while they are visiting the web community not specifically for contribution, but for some other interesting activities, such as browsing archived data, searching for images, etc.

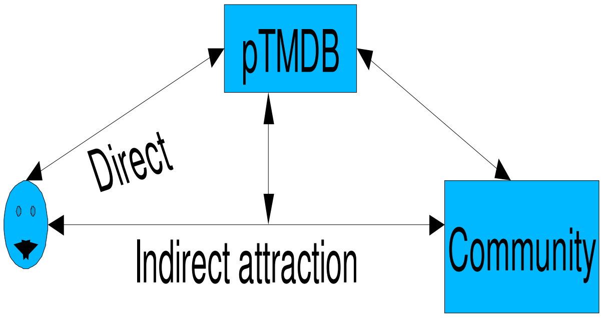


Figure 3: attractiveness of an online community

As Figure 3 shows, the volunteers will not be attracted directly to the pTMDB itself, but rather to the online community, which will act as a proxy to the pTMDB.

Unconscious contribution

In this approach, users are assigned to translate certain related terms. Here, users could be volunteers or paid terminologists, with various levels of proficiency.
ITOLDU is an example of this category, as students are asked to contribute to a bilingual (EN-FR) lexical database.

Useful applications of the Arabic-oriented pTMDB

pTMDB-Based Arabic IR system

As explained in (DAOUD, KITAMOTO et al. 2008), a DSR-pTMDB is used at the Digital Silk Road project to serve a CLIR system, where the exact same archive search engine is used to attract contributions. An Arabic pTMDB could also be used in an Arabic search engine, to translate search terms. Having a term base of multilingual equivalences could achieve this (Chen 2002) (Oard 1999). A similar repository is used for the Panimages project as well.

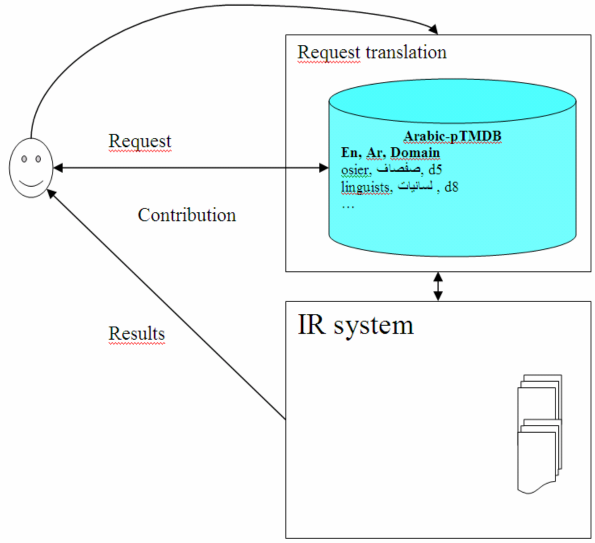


Figure 4: Arabic pTMDB used in a CLIR system

Figure 4 shows an outline of such a companionship between an Arabic-pTMDB and a search engine.

Translation Aid Tools

An Arabic-pTMDB could provide an online consultation service, for translators, general public; such a service would even attract more contributors.

pTMDB → LexAR

As we described earlier, the ultimate objective of our proposed system is to build a domain dedicated repository of Arabic specialized terminology and lexical units (LexAR) which would include the languages mainly used in these domains (English, French…).

We are proposing to use an initial Arabic pTMDB. Which can be used as is for number of useful applications, including an online CLIR system.

But such a pTMDB would go through a process to be transformed into a standardized LexAR.

This process will rely on the following factors:

(1) The redundancies and the measures of collaborative translations agreements within the pTMDB, in another word optimizing the quality of the contributed data and rating each entry.

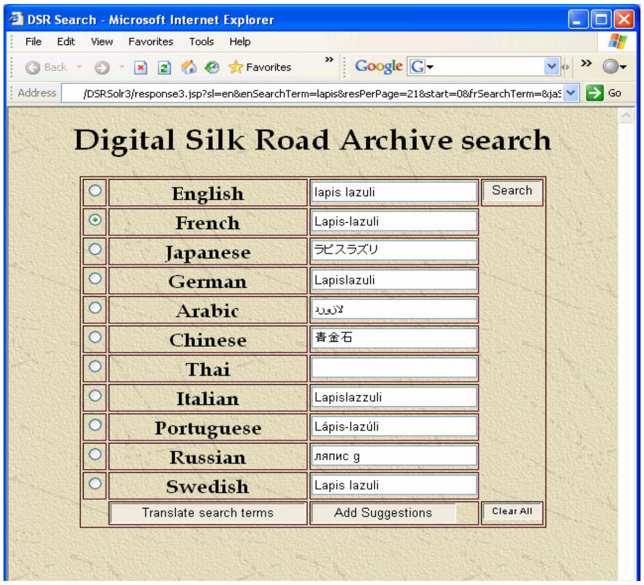
(2) The help of professional terminologists, who would verify, improve and certify the entries. That is like the traditional way, and it is necessary for LexAR to reach a professional quality level. The difference is that starting from a pTMDB built from various resources and with quality ratings should considerably ease the task of the terminologists.

The DSR-pTMDB prototype

The Digital Silk Road project (Ono, Kitamoto et al. 2008) is an initiative started by the National Institute of Informatics (Tokyo) in 2002, to archive cultural historical resources along the Silk Road, by digitizing them and making them available and accessible online. We are trying to build a pTMDB dedicated to this project and its resources.

The DSR-pTMDB (http://dsr.nii.ac.jp/pTMDB/) now serves a CLIR system dedicated to the digitized archive of DSR. The books in this archive where used to extract monolingual lists of important terms, more than 20000 of which was translated into up to 12 languages (including Arabic) using Wikipedia as a multilingual encyclopedia which is a clear example of the semi-automatic acquisition approach we described earlier.

Table 2 shows some examples of the extracted terms and its translations into some of the languages included at the DSR-pTMDB.


DSR Search - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Favorites
Google
Address
/DSRSolr3/response3.jsp?sl=en&enSearchTerm=lapis&resPerPage=21&start=0&frSearchTerm=&ja
Go

Digital Silk Road Archive search

○	English	lapis lazuli	Search
●	French	Lapis-lazuli	
○	Japanese	ラピスラズリ	
○	German	Lapislazuli	
○	Arabic	لازورد	
○	Chinese	青金石	
○	Thai		
○	Italian	Lapislazzuli	
○	Portuguese	Lápis-lazúli	
○	Russian	ляпис g	
○	Swedish	Lapis lazuli	
	Translate search terms	Add Suggestions	Clear All

Figure 5: DSR CLIR system interface

English	French	Arabic	German	Chinese	Italian	Swedish
majapahit	Majapahit	إمبراطورية ماجاباهيت	Majapahit	满者伯夷	Majapahit	Majapahitriket
brahma	Brahmâ	براهما	Brahma	梵天	Brahma	Brahma
pilgrimage to mecca	Hajj	حج	Haddsch	朝覲	Hajj	Hajj
ancient egypt	Égypte antique	مصر القديمة	Altes Ägypten	古埃及	Antico Egitto	Forntida Egypten
cuneiform writing	Cunéiforme	كتابة مسمارية	Keilschrift	楔形文字	Scrittura cuneiforme	Kilskrift
kharosthi script	Alphabet kharoṣṭhī	*****	Kharoshthi-Schrift	佉卢文	Kharoshthi	Kharosti

Table 2: sample multilingual entries of DSR-pTMDB

DSR-pTMDB’s entries are classified based on the quality (Q) of their source. Q for an entry translated by Wikipedia is *** out of five stars, while a terms entered by terminologist would have five stars out of five. The next phase will be a process of transforming DSR-pTMDB into a MTDB, which involve verifying and post editing all of the entries contributed by the volunteers or by Wikipedia, and enhancing their quality.

Conclusion

In this paper we proposed and analyzed LexAR, an Arabic-based multilingual repository of technical terms and specialized lexical units. With the absence of truly up to date Arabic equivalences of the very dynamic terminology of the major language, the need for LexAR is becoming an unavoidable necessity.

LexAR can help bridging the scientific gap between the Arabic world and the First World countries by easing the knowledge transfer into and from the Arabic language. Previous attempt are still scattered, does not benefit from new technologies and can not meet the dynamicity of terminology at this era.

We showed that traditional ways in building such a database could be infeasible, expensive, and it could not achieve satisfactory results. On the other hand collaborative methods may not be able to meet the requirements of a standardized LexAR, thus, we proposed a more contributable Arabic-pTMDB that will act as an intermediate repository of "raw material" for LexAR.

Arabic-pTMDB could be used as is in number of useful applications, even without transforming it into LexAR, such as a building a pTMDB-based CLIR-system similar to the Digital Silk Road's.

Acknowledgments

This research is supported by two international internships to the first author from the National Institute of Informatics (NII Tokyo/Japan). And we would like to thank Pr. A. Kitamoto the director of the Digital Silk Road Project.

References

Alfadhel, A. (2007). "Saudi Terminology Data Bank." Retrieved 11/2007, 2007, from http://gdis.kacst.edu.sa/resources.html.

Bellynck, V., C. Boitet, et al. (2005). ITOLDU, a Web Service to Pool Technical Lexical Terms in a Learning Environment and Contribute to Multilingual Lexical Databases Springer Berlin / Heidelberg.

Cabre, M. T. and J. C. Sager (1999). Terminology: Theory, methods, and applications J. Benjamins Pub. Co.

Chen, A. (2002). "Cross-Language Retrieval Experiments at CLEF 2002." in CLEF-2002 working notes,.

Daoud, M., A. Kitamoto, et al. (2008). CLIR-Based Collaborative Construction of a Multilingual Terminological Dictionary for Cultural Resources. Translating and the Computer 30, London/UK.

EOLSS. (2008). "EOLSS." Retrieved 12 January 2009, from http://www.eolss.net/.

Etzioni, O., K. Reiter, et al. (2007). Lexical translation with application to image searching on the web. MT Summit XI, Copenhagen, Denmark.

FAO. (2008). "FAO TERMINOLOGY." Retrieved 1/9/2008, 2008, from http://www.fao.org/faoterm.

Ghazal, L. (1977). The New Methodology for Assigning Arabic Terminology. IERA, Rabat.

IATE. (2008). "Inter-Active Terminology for Europe." Retrieved 10/10/2008, 2008, from http://iate.europa.eu.

IDRC. (2009, 10 January 2009). "The Water Demand Management Glossary (Second Edition)." from http://www.idrc.ca/WaterDemand/IDRC_Glossary_Second_Edition/index.html.

IEC. (2008). "Electropedia." Retrieved 10/10/2008, 2008, from http://dom2.iec.ch/iev/iev.nsf/welcome?openform.

Jones, G. J. F., F. Fantino, et al. (2008). Domain-Specific Query Translation for Multilingual Information Access Using Machine Translation Augmented With Dictionaries Mined From Wikipedia. Proceedings (CLIA-2008), Hydrabad, India.

Mangeot, M. (2006). Dictionary Building with the Jibiki Platform. Software Demonstration. Proc. of EURALEX 2006, Torino, Italy.

Murata, T., M. Kitamura, et al. (2003). Implementation of collaborative translation environment 'Yakushite Net' MT Summit IX. New Orleans, USA.

Oard, D. (1999). Global Access to Multilingual Information. Fourth International Workshop on Information Retrieval with Asian Languages. Taipei-Taiwan.

Ono, K., A. KITAMOTO, et al. (2008). Memory of the Silk Road -The Digital Silk Road Project-. Proceedings of (VSMM08), Project Papers, Limassol, Cyprus.

OQLF. (2008). "Le grand dictionnaire terminologique." Retrieved 1/9/2008, 2008, from granddictionnaire.com/.

Richardson, M. and P. Domingos (2003). Building large knowledge bases by mass collaboration. International Conference On Knowledge Capture, Sanibel Island, FL, USA

Sérasset, G. (1994). "Interlingual lexical organisation for multilingual lexical databases in nadia." COLING-94 volume 1: pages 278-282.

Sérasset, G. (2004). A Generic Collaborative Platform for Multilingual Lexical Database Development. COLING 2004, p. 73-79, Geneva, Switzerland, Aug. 2004. .

Serasset, G., F. Brunet-Manquat, et al. (2006). Multilingual legal terminology on the Jibiki platform: the LexALP project. Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Sydney, Australia, Association for Computational Linguistics.

UN. (2008). "United Nations Multilingual Terminology Database." Retrieved 10/10/2008, 2008, from http://unterm.un.org/.

WHO-EMRO. (2009). "The unified health lexicon." Retrieved 10 January 2009, from http://www.emro.who.int/umd/.

Wikipedia. (2008). "Wikipedia." Retrieved 1 June 2008, 2008, from http://www.wikipedia.org/.

Wikitionary. (2008). "Wikitionary." Retrieved 1/9/2008, 2008, from http://en.wikipedia.org/wiki/Wiktionary.

Yassin, Y. A. (2003). "Why Arabic Is the Most Difficult Language for Localization." Globalization Insider Volume XII(Issue 3.6).