Multilingual Lexical Network from the Archives of the Digital Silk Road

Mohammad Daoud
LIG, GETALP
Université Joseph Fourier
Mohammad.Daoud@imag.fr

Kyo Kageura
Graduate School of Education
The University of Tokyo
kyo@p.u-tokyo.ac.jp

Christian Boitet
LIG, GETALP
Université Joseph Fourier
Christian.Boitet@imag.fr

Asanobu Kitamoto
The National Institute of Informatics (Tokyo)
Kitamoto@nii.ac.jp

Mathieu Mangeot
LIG, GETALP
Université Joseph Fourier
Mathieu.Mangeot@imag.fr

Abstract

We are describing the construction process of a specialized multilingual lexical resource dedicated for the archive of the Digital Silk Road DSR. The DSR project creates digital archives of cultural heritage along the historical Silk Road; more than 116 of basic references on Silk Road have been digitized and made available online. These books are written in various languages and attract people from different linguistic background, therefore, we are trying to build a multilingual repository for the terminology of the DSR to help its users, and increase the accessibility of these books. The construction of a terminological database using a classical approach is difficult and expensive. Instead, we are introducing specialized lexical resources that can be constructed by the community and its resources; we call it Multilingual Preterminological Graphs MPGs. We build such graphs by analyzing the access log files of the website of the Digital Silk Road. We aim at making this graph as a seed repository so multilingual volunteers can contribute. We have used the access log files of the DSR since its beginning in 2003, and obtained an initial graph of around 116,000 terms. As an application, We have used this graph to obtain a preterminological multilingual database that has a number of applications.

1 Introduction

This paper describes the design and development of a specialized multilingual lexical resource for the archive constructed and maintained by the Digital Silk Road project. The Digital Silk Road project (NII 2003) is an initiative started by the National Institute of Informatics (Tokyo/Japan) in 2002, to archive cultural historical resources along the Silk Road, by digitizing them and making them available and accessible online.

One of the most important sub-projects is the Digital Archive of Toyo Bunko Rare Books (NII 2008) where 116 (30,091 pages) of old rare books available at Toyo Bunko library have been digitized using OCR (Optical Character Recognition) technology. The digitized collection contains books from nine languages including English. The website of the project attracts visitors from the domain of history, archeology, and people who are interested in cultural heritage. It provides services of reading and searching the books of Toyo Bunko, along with variety of services. Table 1 shows the countries from which DSR is being accessed. The table

Proceedings of the 6th Workshop on Ontologies and Lexical Resources (Ontolex 2010), pages 19–27,
Beijing, August 2010

shows that around 60% of visitors are coming from countries other than Japan. The diversity of the visitors' linguistic backgrounds suggests two things: 1) Monolingual translation service is not enough. 2) It shows that we can benefit from allowing them to contribute to a multilingual repository. So we design and build a collaborative multilingual terminological database and seed using the DSR project and its resources (Daoud, Kitamoto et al. 2008). However, Discovering and translating domain specific terminology is a very complicated and expensive task, because (1) traditionally, it depends on human terminologists (Cabre and Sager 1999) which increases the cost, (2) terminology is dynamic (Kageura 2002), thousands of terms are coined each year, and (3) it is difficult to involve domain experts in the construction process. That will not only increase the cost, but it will reduce the quality, and the coverage (number of languages and size). Databases like (UN-Geo 2002; IATE 2008; UN 2008) are built by huge organizations, and it is difficult for a smaller community to produce its own multilingual terminological database.

Country	Visitors	language	Books in the same language
Japan	117782	JA	2 books
China	30379	CH	5 books
USA	15626	EN	44 books
Germany	8595	GE	14 books
Spain	7076	SP	-
Australia	5239	EN	See USA
Italy	4136	IT	1 book
France	3875	FR	14 books
Poland	2236	PO	-
Russia	1895	RU	7 books
other	87573	Other	There are many books in different language
Total	284412		

Table 1. Countries of the DSR visitors (from jan/2007 to dec/2008)

In the next section we will give definitions for the basic concepts presented in this article, in particular, the preterminology and its lexical network (graph). Then, in the third section we will show the automatic approach to seed the multilingual preterminological graph based on the resources of the DSR. And then, we will discuss the human involvement in the development of such a resource by providing a study of the possible contributors through analyzing the multilinguality and loyalty of the DSR visitors. In the fifth section we will show the experimental results. And finally, we will draw some conclusions.

2 Multilingual Preterminological Graphs

2.1 Preterminology

Terminological sphere of a domain is the set of terms related to that domain. A smaller set of that sphere is well documented and available in dictionaries and terminological databases such as (FAO 2008; IEC 2008; IDRC 2009)... However, the majority of terms are not multilingualized, nor stored into a database, even though, they may be used and translated by the community and domain experts. This situation is shown in Figure 1, where the majority of terms are in area **B**. Preterminological sphere (area **B**) of a domain is a set of terms (*preterms*) related to the domain and used by the community but it might not be documented and included in traditional lexical databases.

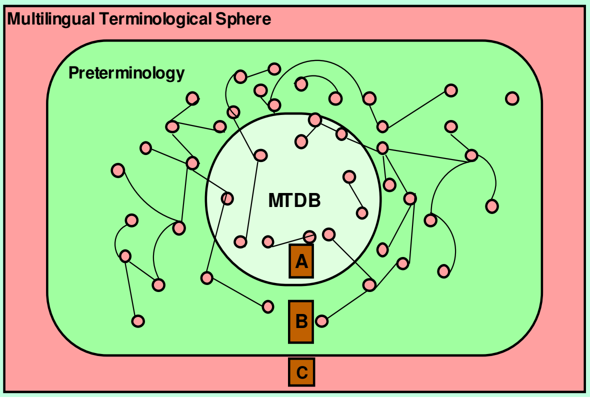


Figure 1. Preterminological sphere

Every year thousands of terms are coined and introduced in correspondence to new concepts, scientific discoveries or social needs. Most of these terms are produced in the top dominant languages, i.e. English. Interested people from different linguistic backgrounds would find suitable translations to new terms and use it amongst them. For example, the term 'status update' is used by people who visit social networking websites like facebook.com. Translation of this term to Arabic might not be available in area **A** of Figure 1. However the Arabic community found a translation that is acceptable which is تحديث الحالة. So this term is in the area **B**. We are trying to use what is in area **A**, and what can be contributed from **B** to build preterminology (Daoud, Boitet et al. 2009).

2.2 Structure of MPG

We are building preterminological resource as a lexical network (graph) to handle the diversity of the resources that we use. A multilingual preterminological graph *MPG(N,E)* is a finite nonempty set *N={n1,n2, ...}* of objects called Nodes together with a set *E={e1,e2, ...}* of unordered pairs of distinct nodes of *MPG* called edges. This definition is based on the general definition of a graph at the following references (Even 1979; Loerch 2000). *MPG* of domain *X*, contains possible multilingual terms related to that domain connected to each other with relations. A multilingual lexical unit and its translations in different languages are represented as connected nodes with labels.

In an *MPG* the set of nodes **N** consists of *p,l, s, occ,* where *p* is the string of the preterm, *l* is the language, *s* is the code of the first source of the preterm, and *occ* is the number of occurrences. Note that *l* could be undefined. For example: *N={[silk road, en, log],[Great Wall of China, en, ,wikipedia, 5], [الصين, ar, contributorx,6]}*, here we have three nodes, 2 of them are English and one in Arabic, each term came from a different source. Note that English and Arabic terms belong to the same *N* thus, the same MPG.

An *Edge e={n, v}* is a pair of nodes adjacent in an *MPG*. An edge represents a relation between two preterms represented by their nodes. The nature of the relation varies. However, edges are weighted with several weights (described below) to indicate the possible nature of this relation.

The following are the weights that label the edges on an MPG: *Relation Weights rw*: For an edge *e={[p1,l1,s1], [p2,l2,s2]}, rw* indicates that there is a relation between the preterm *p1* and *p2*. The nature of the relation could not be assumed by *rw*. *Translation Weights tw*: For an edge *e={[p1,l1,s1], [p2,l2,s2]}, tw* suggests that *p1* in language *l1* is a translation of *p2* in language *l2*. *Synonym Weights sw*: For an edge *e={[p1,l1,s1], [p2,l1,s2]}, sw* suggests that *p1* and *p2* are synonyms.

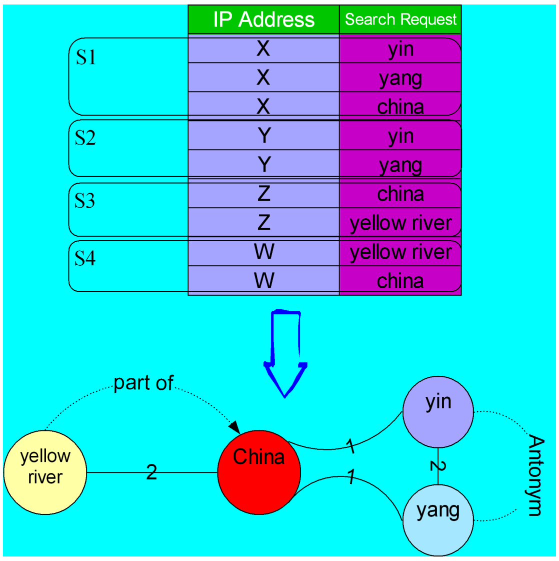
3 Automatic Initialization of DSR-MPG

Basically we seeded DSR-MPG, through two steps, the firs one is the automatic seeding, which consists of the following: 1) Initialization by finding interesting terms used to search the website of the DSR. 2) Multilingualization, using online resources. 3) Graph Expansion using the structure of the graph it self. The second step is the progressive enhancement, by receiving contributions from users, through set of useful applications. In this section we will discuss the first three steps. In section 4, we will discuss the human factor in the development of DSR-MPG.

3.1 Analyzing Access Log Files

We analyze two kinds of access requests that can provide us with information to enrich the MPG: (1) requests made to the local search engine of DSR (2) requests from web-based search engine (like Google, Yahoo!...). These requests provide the search terms that visitors used to access the website. Moreover, we can understand the way users interpret a concept into lexical units. For example, if we find that five different users send two search requests *t1* and *t2*, then there is a possibility that t1 and *t2* have a relation. The graph constructor analyzes the requests to make the initial graph by creating edges between terms in the same session. *rw(x,y)*, is set to the number of sessions containing *x* and *y* within the log file.

For example, *rw(x,y)* = 10 means that 10 people thought about *x* and y within the same search session. Figure 2 shows an example of a produced graph. The method did not discover the kind of relation between the terms. But it discovered that there is a relation, for example, three users requested results for "*yang*" followed by "*yin*" within the same session. Hence, edge with weight of 2 was constructed based on this.



	IP Address	Search Request
S1	X	yin
	X	yang
	X	china
S2	Y	yin
	Y	yang
S3	Z	china
	Z	yellow river
S4	W	yellow river
	W	china

Figure 2. Example of constructing an MPG from an access log file

3.2 Multilingualization Using Online Resources

Many researchers focused on the usage of dictionaries in digital format to translate lexical resources automatically (Gopestake, Briscoe et al. 1994) (Etzioni, Reiter et al. 2007). We are concerned with the automatic utilization of these resources to acquire multilingual preterminological resources through the following: 1) Wikipedia 2) online MT systems 3) online dictionaries.

Wikipedia (Wikipedia-A 2008) is a rich source of preterminology, it has good linguistic and lexical coverage. As of December, 2009, there are 279 Wikipedias in different languages, and 14,675,872 articles. There are 29 Wikipedias with more that 100000 articles and 91 languages have more than 10,000 articles. Beside, Wikipedia is built by domain experts. We exploit the structure of Wikipedia to seed an MPG, by selecting a root set of terms, for each one of them we fetch its wikipedia article, and then we use the language roll of the article. For example, we fetch the article (Cuneiform script) En: http://en.wikipedia.org/wiki/Cuneiform_script, to reach its translation in Arabic from this url: http://ar.wikipedia.org/wiki/كتابة_مسمارية

We use also online machine translation systems as general purpose MRDs. One of the main advantages of MT systems is the good coverage even for multiword terms. The agreement of some MT systems with other resources on the translation of one term enhanced the confidence of the translation. Another positive point is that the results of MT provide a first draft to be post edited later. We used 3 MT systems:

- Google Translate (Google 2008) (50 languages)
- Systran (Systran 2009) (14 languages)
- Babylon (Babylon 2009) (26 languages)

Here is an example of translating the term “great wall of China” into Arabic.

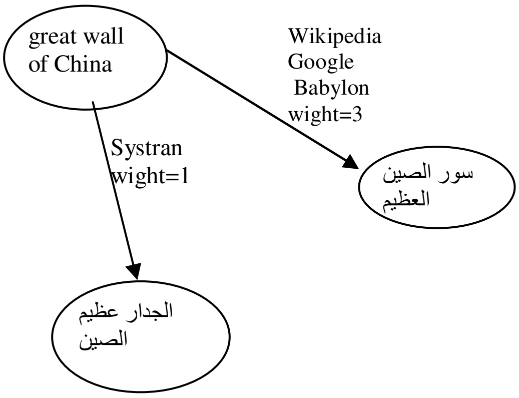


Figure 3. MPG sample nodes

In a similar way, we used several online repositories; to make good use of what is available and standardized, to initializing the MPG with various resources, and to construct a meta-system to call online dictionaries automatically. We used IATE (IATE 2008) as an example of a terminological db, and Google dictionary (Google 2008). The concept is similar to the concept of using online translations, where we construct an http request, to receive the result as html page.

3.3 Graph Expansion

And then, the Graph is expanded by finding the synonyms according to formula (1) described at (Daoud, Boitet et al. 2009). After finding synonyms we assume that synonyms share the same translations. As Figure 4 shows, *X1* and *X2* have translations overlaps, and relatively high *rw*, so that suggest they are synonyms. Therefore we constructed heuristic edges between the translations of *X1* and *X2*.

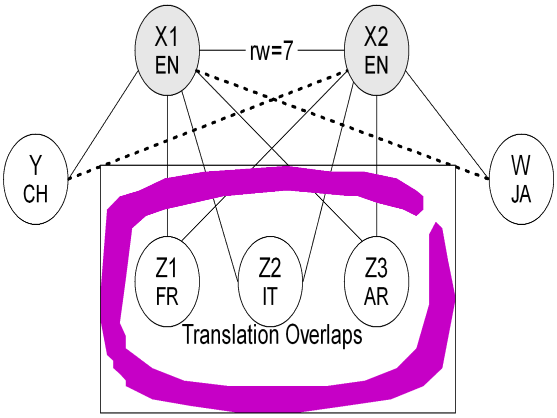


Figure 4. Graph expansion

4 Human Involvement in the Development of DSR-MPG

After initializing the graph, we target contributions from the visitors to the DSR website. In this section we will start by analyzing the possibility of receiving contributions from the visitors, and then we will introduce some useful applications on the DSR-MPG that can help the visitors and attract them to get involved.

4.1 Analyzing Possible Contributors of the DSR

We are trying to analyze access log files to find out the possible contributors to a preterminological multilingual graph dedicated to an online community. This kind of information is necessary for the following reasons: 1) it provide feasibility analysis predicting the possibility of receiving contribution to a multilingual preterminological repository. 2) it gives information that can be used by the collaborative environment to personalize the contribution process for those who prove to be able to contribute.

In the analysis process we are using the following information that can be easily extracted the access records:

- Key terms to access the historical resources of the Digital Silk Road, whether it is the local search engine, or any external search engine.
- Access frequency: number of access requests by a visitor over a period of time.
- Language preferences
- Period of visits

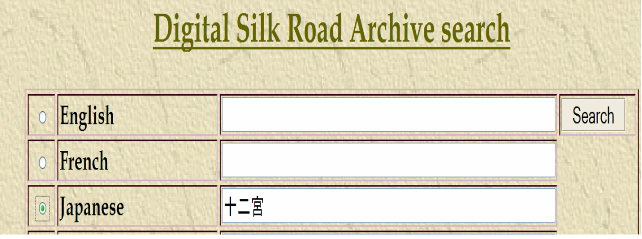
Knowing these points helps determining the possible users who might be willing to contribute. A contributor should satisfy the following characteristics: 1) *Loyalty* 2) *Multilinguality*. A multilingual user is a visitor who uses multilingual search terms to access the online resources. We rank users based on their linguistic competence, we measure that by tracking users' search requests, and matching them with the multilingual preterminological graph, users with higher matches in certain pair of languages are ranked higher. A *loyal user* is a user who visits the web site frequently and stays longer than other users. Users based on how many months they accessed the website more that k times.

4.2 DSR-MPG Applications

For a historical archive like the DSR, we find that reading and searching where the most important for users. Log files since 2003 shows that 80% of the project visitors were interested in reading the historical records. Moreover, around 140000 search requests have been sent to the internal search engine. So we implemented two applications (1) "*contribute-while-reading*" and (2) "*contribute-while-searching*".

4.2.1 Contribute While Searching

Physical books have been digitized and indexed into a search engine. We expect users to send monolingual search requests in any language supported by our system to get multilingual answers. Having a term base of multilingual equivalences could achieve this (Chen 2002). A bilingual user who could send a bilingual search request could be a valid candidate to contribute. We plan that users who use our search engine will use the DSR-pTMDB to translate their requests and will contribute to the graph spontaneously. As Figure 5 shows, a user would translate the search request, during the searching process; the user can ask to add new translation if s/he was not happy with the suggested translation, by clicking on "Add Suggestions" to view a contribution page.

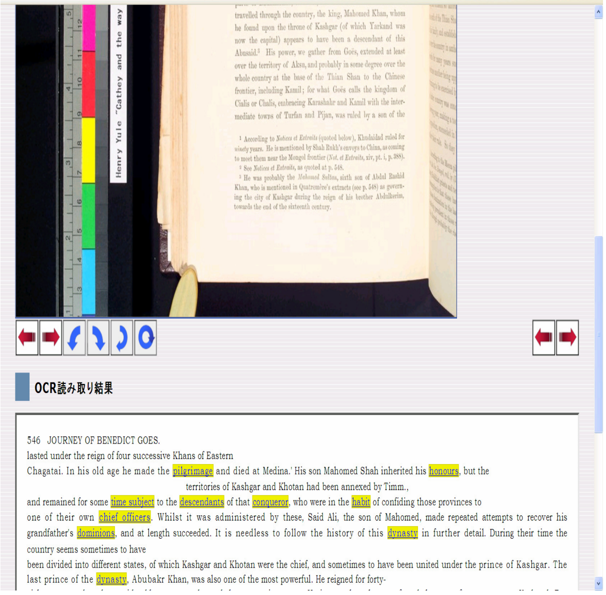

Digital Silk Road Archive search

○ English
○ French
● Japanese 十二宮

Figure 5. A Japanese user translating his request

4.2.2 Contribute While Reading

The other application is trying to help users from different linguistic backgrounds to translate some of the difficult terms into their languages while they are reading, simply by selecting a term from the screen. As shown in Figure 6, readers will see a page from a book as an image, with its OCR text. Important terms will be presented with yellow background. Once a term is clicked, a small child contribution/lookup window will be open, similar. Also user can lookup/translate any term from the screen by selecting it. This application helps covering all the important terms of each book.



546 JOURNEY OF BENEDICT GOES.
lasted under the reign of four successive Khans of Eastern
Chagatai. In his old age he made the pilgrimage and died at Medina.' His son Mahomed Shah inherited his honours, but the
territories of Kashgar and Khotan had been annexed by Timm.,
and remained for some time subject to the descendants of that conqueror, who were in the habit of confiding those provinces to
one of their own chief officers. Whilst it was administered by these, Said Ali, the son of Mahomed, made repeated attempts to recover his grandfather's dominions, and at length succeeded. It is needless to follow the history of this dynasty in further detail. During their time the country seems sometimes to have
been divided into different states, of which Kashgar and Khotan were the chief, and sometimes to have been united under the prince of Kashgar. The last prince of the dynasty, Abubakr Khan, was also one of the most powerful. He reigned for forty-

Figure 6. Translate while reading

5 Experimental Results

In this section present we will present the experiment of seeding DSR-MPG, and the results of discovering possible contributors from the visitors of the DSR.

5.1 DSR-MPG Initialization

To build the initial DSR-MPG, we used the access log files of the DSR website (dsr.nii.ac.jp) from December 2003 to January 2009. The initial graph after normalization contained 89,076 nodes. Also we extracted 81,204 terms using Yahoo terms. 27,500 of them were not discovered from the access files. So, the total number of nodes in the initial graph was 116,576 nodes, see Figure 7 for sample nodes.

After multilingualization, the graph has 210,781 nodes containing terms from the most important languages. The graph has now 779,765 edges with $tw > 0$. The important languages are the languages of the majority of the visitors, the languages of the archived books, and representative languages a long the Silk Road. DSR-MPG achieved high linguistic coverage as 20 languages have more than 1000 nodes on the graph. To evaluate the produced graph, we extracted 350 English terms manually from the index pages of the following books:

Ancient Khotan, vol.1: http://dsr.nii.ac.jp/toyobunko/VIII-5-B2-7/V-1/

On Ancient Central-Asian Tracks, vol.1:http://dsr.nii.ac.jp/toyobunko/VIII-5-B2-19/V-1

Memoir on Maps of Chinese Turkistan and Kansu, vol.1: http://dsr.nii.ac.jp/toyobunko/VIII-5-B2-11/V-1

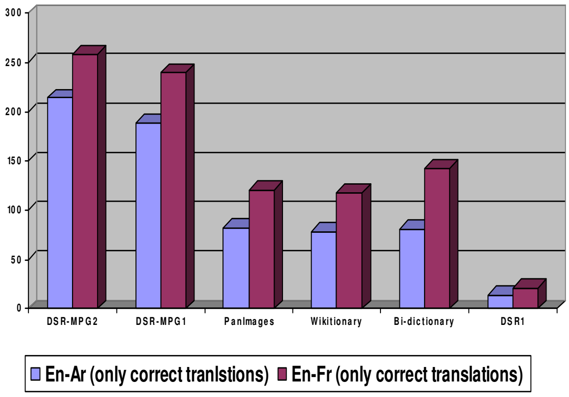


Figure 7. A comparison between DSR-MPG, and other dictionaries. The En-Ar bi-dictionary is Babylon (Babylon 2009), and the En-Fr bi-dictionary was IATE.

We assume that the terms available in these books are strongly related to the DSR. Hence, we tried to translate them into Arabic and French. Figure 7 compares between DSR-MPG, and various general purpose dictionaries. Out of the 350 terms, we found 189 correct direct translations into Arabic. However, the number reached 214 using indirect translations. On the other hand, the closest to our result was PanImages, which uses Wikitionaries and various dictionaries, with only 83 correct translations. DSR-MPG1 is the translations obtained from formula 1, DSR-MPG2 represents the translations obtained from indirect translations, which increased the amount of correct translation by

25 terms in the case of En-Ar. The result can be progressively enhanced by accepting contributions from volunteers through the applications we described in the section three and the generic nature of MPG makes it easy to accept contributions from any dictionary or terminological database.

Around 55200 root English terms were used as a seed set of terms; these terms were selected from the initial DSR-MPG. Around 35000 terms have been translated from Wikipedia into at least 1 language, mostly in French, German. Wikipedia increased the density of the graph by introducing around 113,000 edges (with *tw*).

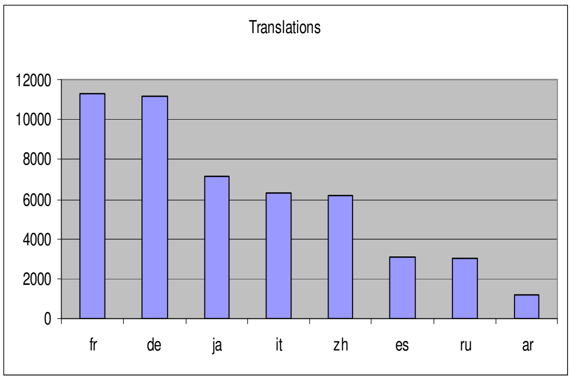


Figure 8. Number of translated terms in sample languages using Wikipedia

Naturally MT would achieve better coverage; we checked the results for Arabic, we selected 60 terms randomly from the root set, around 25 terms were translated correctly. 13 terms needed slight modification to be correct.

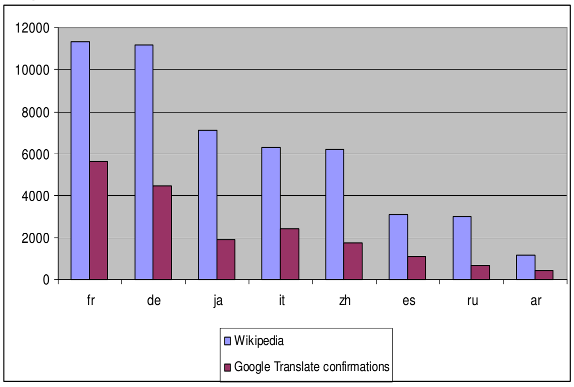


Figure 9. Terms translated by Google MT and matched the translation of Wikipedia

5.2 DSR Possible Contributors

With K=2, meaning that a *multilinguality competence* is counted only if the two terms sent by a user has to have more than 2 points of translation weight on the MPG.

The highest score was 33, achieved by this IP: p27250-adsao05douji-acca.osaka.ocn.ne.jp. That means that this user sent 33 multilingual search requests. We have another 115 users with score higher than 5.

For example, the following two request, sent by one user:

p27250-adsao05douji-acca.osaka.ocn.ne.jp
&input=peshawar
p27250-adsao05douji-acca.osaka.ocn.ne.jp
&input=ペシャワール

On the DSR-MPG the translation weight between peshawer and ペシャワール = 5, thus this IP earned a point. With k=10, means that a user should send 10 requests to earn a *loyalty point*, only 309 users earned 12 point (for 12 months), 43 of them has more than 3 points.

6 Conclusions

We presented our work in constructing a new lexical resource that can handle multilingual terms based on the historical archive of the Digital Silk Road. Multilingual Preterminological Graphs (MPGs) are constructed based on domain dedicated resources, and based on volunteer contributions.

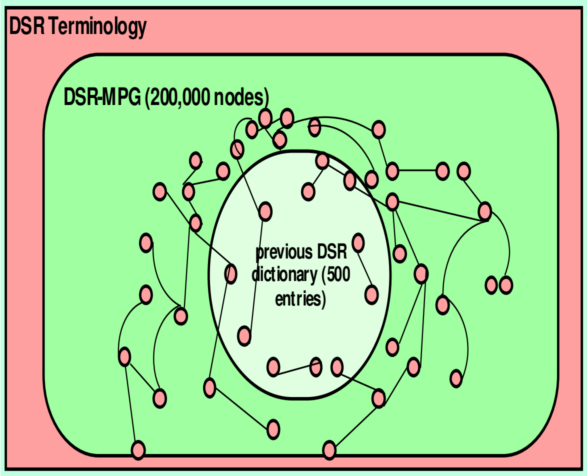


Figure 10. DSR preterminology

It compiles terms available in the preterminological sphere of a domain. In this article we defined the framework of the construction of preterminology, and we described the approach for using access log files to initialize such preterminological resource by finding the trends in the search requests used to access the resources of an online community. Aiming at a standardized multilingual repository is very expensive and difficult. Instead of that, MPGs tries to use all available contributions. This way will enhance the linguistic and informational coverage, and tuning the weights (*tw*, *rw*, and *sw*) will give indications for the confidence of the translation equivalences, as the *tedges* accumulate the agreements of the contributors and MDRs (online resources).

We used the resources of the Digital Silk Road Project to construct a DSR-MPG and some applications that attract further contribution to the MPG. DSR-MPG achieved high linguistic and informational coverage compared to other general purpose dictionaries, Figure 10. Furthermore, the generic structure of the MPG makes it possible to accept volunteer contributions, and it facilitates further study of computing more lexical functions and ontological relations between the terms. We made a study on the possibility of receiving contributions from users, by analyzing the access log file to find multilinguality and loyalty of the DSR visitors; we found 115 users with the needed linguistic capacity 43 of them scored high loyalty points. This gives an indication of the future of the contributions. These measures are just estimations and expected to go high with the help of the MPG-DSR applications.

References

Babylon. (2009). "Babylon Dictionary." Retrieved 5/5/2009, 2009, from http://www.babylon.com/define/98/English-Arabic-Dictionary.html.

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., C. Boitet, et al. (2009). Constructing multilingual preterminological graphs using various online-community resources. the Eighth International Symposium on Natural Language Processing (SNLP2009), Thailand.

Daoud, M., C. Boitet, et al. (2009). Building a Community-Dedicated Preterminological Multilingual Graphs from Implicit and Explicit User Interactions. Second International Workshop on REsource Discovery (RED 2009), co-located with VLDB 2009, Lyon, France.

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

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

Even, S. (1979). Graph Algorithms, Computer Science Press.

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

Google. (2008). "Google Dictionary." Retrieved 1/9/2008, 2008, from http://www.google.com/dictionary.

Google. (2008). "Google Translate." Retrieved 1 June 2008, 2008, from http://translate.google.com.

Gopestake, A., T. Briscoe, et al. (1994). "Acquisition of lexical translation relations from MRDS." Machine Translation Volume 9, Numbers 3-4 / September, 1994: 183-219.

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.

Kageura, K. (2002). The Dynamics of Terminology: A descriptive theory of term formation and terminological growth.

Loerch, U. (2000). An Introduction to Graph Algorithms Auckland, New Zealand, University of Auckland.

NII. (2003). "Digital Silk Road." Retrieved 1/9/2008, 2008, from http://dsr.nii.ac.jp/index.html.en.

NII. (2008). "Digital Archive of Toyo Bunko Rare Books." Retrieved 1 June 2008, 2008, from http://dsr.nii.ac.jp/toyobunko/.

Systran. (2009). "Systran Web Tranlstor." Retrieved 20/12/2009, 2009, from www.systransoft.com/.

UN-Geo (2002). Glossary of Terms for the Standardization of Geographical Names, UN, New York.

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

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