GLODOC: A MULTILINGUAL SPECIALIZED GLOSSARY

Lola García-Santiago* and Ana Belén Lozano-Carvajal**
*CSIC, Unidad Asociada Grupo SCImago, Madrid, Spain
Department of Library and Information Science, University of Granada, Granada, Spain.
** University of Granada, Granada, Spain.

1. Introduction

In modern society, knowledge has an international and multilingual nature. The internationalization process is also reflected in the university environment. In these centres, the development of strategies for multilingual management is promoted with the aim of providing academics and professionals a multilingual training, which is needed throughout life in the immediate environment, in Europe, and worldwide.

This knowledge is embodied in scientific documents that are generated daily. This phenomenon was significantly accelerated, on one hand, by the emergence and development of the Internet and, on the other, by the proliferation of specialized and common information sources as well as by the ease of access for researchers, professionals, and users. The constant updating of information requires good selection and translation to enable the understanding of new ideas.

A major issue in terms of understanding, at the stage of information selection and retrieval, is that a word can have many possible meanings, but its real meaning is determined by context, and sometimes even more so when a specialized vocabulary is used.

GloDoc has been developed in response to this demand by society, with this being a multilingual glossary specializing in the area of communication and documentation, and it is available online. It works with three languages, i.e. pairing Spanish with French or German. In this paper, we describe the objectives of GloDoc and its characteristics.

2. Objectives

Frequently while reading texts in another language, although we may know the words, we are unable to make sense of the sentence, and a general dictionary hardly helps us. GloDoc seeks to clarify the meaning of an expression in an academic context and in a particular speciality.

From an academic viewpoint, translation and interpretation students are intended to acquire specialized vocabulary from the most recent sources, particularly

when these terms are still sometimes not registered in dictionaries and encyclopaedias. The tool presented here is used to apply personal skills for self-learning through the creation of the student's own documents. From an instrumental standpoint, practitioners, researchers, and students specializing in the field of Communications and Information Science are meant to have a terminology tool that facilitates the translation of specialized vocabulary in the area of documentation, information, and communication technologies.
The building of this database reflects two objectives inherent to all terminology resources:

1) to allow the representation of the current evolution of the language from new words and meanings which arise continuously in language;
2) to allow users to explore the wealth of knowledge terminology (e.g. grammatical information, thematic context, etc.).

The project is also based on the following specific objectives:

- To create a corpus of texts specializing in communication and documentation in Spanish, French, and German;
- To avoid using a bridge language to translate expressions;
- To take the simple terms and specific compound expressions of the area;
- To develop a terminology database;
- To access public online GloDoc.

3. Methodology

For the development of this terminology tool, the following were established: a) selection of the corpus of scientific papers from the field, b) the selection of terminology, c) construction of the entries, d) the development of the database terminological data, and e) online implementation.

3.1. Selection of the corpus of scientific papers from the field

For these aims, a varied collection was built from scientific articles available on the web. These documents are specialized and deal with any subject in information science, communication or documentation area. The collection of original documents was divided by language and analysed individually. Consequently, three independent groups of items were formed, based on the languages of French, German, and Spanish. Then, words and compound expressions were identified to supply the database with terms according to the language.

These papers come from international institutions such as the IFLA or from scientific journals, which are usually monolingual. These included electronic documents in doc or pdf format for easier text editing. Each was analysed and each sentence was identified with a number.

3.2. The selection of terminology

For the terminological retrieval of the articles from the corpus, we chose the manual process. On the one hand, better results are attained than with an automatic retrieval program, due to the limited number of texts to be analysed. On the other hand, stop words could be included within the compound expressions that an automatic terminological retriever would overlook unless a more sophisticated list of exceptions was compiled. That is, the loss of possible terminological expressions is avoided by achieving greater exhaustiveness of the entries.

In principle, only specialized expressions were included in our glossary. This selection was easier in French than in German, where we found more compound words in general and in specific contexts. At the same time, we discovered a wide range of examples of paradigmatic usage for some expressions.

Specialized secondary documents such as specialized dictionaries, glossaries and encyclopaedias are used for:

- Undertaking lexical analysis
- Producing semantic markers and notes
- Validating and complementing the results, especially in case of acronyms
- Disambiguating terms
- Finding terminological equivalents to Spanish (only with German terms)
- Finding synonyms which can be found on the list of terms

However, it should be mentioned that all synonyms were accepted because we have not built a documental language but rather a specialized terminological tool with expressions that are used frequently in the area selected for this project.

3.3. Construction of the entries

After selection of the terms to be included in the database, they were standardized in order to achieve effective indexing and search functions. The most common criteria followed in dictionaries and terminological tools to index the different expressions were used. Moreover, this information is explained in the section “help” of GloDoc to guide to the user in order to carry out successful searches. Thus, the terms identified were stored as entries in the form of masculine singular for nouns and adjectives, the infinitive form for verbs, while these

rules were combined for composite expressions. Then, each standardized entry was tagged with a single identification number.
A limitation arises from not being able to search in all of the variants of a word. The indexing rules permit construct queries with roots or fixed expressions. However, we should not forget that the sentence does include the word or the expression in the original form.
Despite not being our main aim, more abbreviations have been included in German for practical reasons, as in this language there are more occurrences of abbreviations in any type of text, whether a general or scientific paper.
In the field of context phrases, edition and review work was performed to avoid the loss of semantic meaning and/or adapt to the technical limitations of the database. Finally, other information of the lexical and thematic types were added to the Spanish language.
The abbreviations which appear in the texts have a differentiated treatment due to their importance and complexity. Both the denomination developed as well as its equivalent in Spanish have been identified as to whether they exist, or in English if not.

3.4. The development of the database terminological data

An Access database was designed, allowing us more flexibility to relate tables while the terminologies were collected. Thus, a group of tables was generated by language as was another group of tables of common characteristics.
The field context was limited to 255 characters, and consequently some of the original sentences had to be edited.
Also, the terminology was analysed to identify the semantic and linguistic equivalence relationships of synonymy. Synonyms in the same language were searched and, once the entries had been established, equivalents were sought between the different languages pairs (Spanish-French-Spanish, and Spanish-German-Spanish).
As mentioned above, GloDoc does not use a bridge language to translate. Language pairs were chosen to build the terminological database from the articles directly. In other words, due to the limited number of documents analysed, we cannot find the same number of expressions for the three languages. Consequently, we do not have all terms with their equivalence in other languages and their contexts.
To offer this glossary online and to avoid security problems, we imported the final tables to a SQLite Database which is hosted by a University of Granada Server. All these tables have the fields which are shown online at the moment.

They include terms, equivalences, sentences and, in Spanish terms, grammatical characteristics.

3.5. Online implementation

We have developed an online glossary where the query interface allows us to recover information easily from French or German to Spanish and vice versa from the terminological database without using a intermediate or bridge language such as English.

GloDoc is presented in the three languages which can be used to search. The interface permits the user to interrogate our database, which contains the terms together with sentences, categories, and the relationships between them. The information requested can be recovered with a given term in this language. The language is selected directly by placing the requested work in the search field, as each one corresponds to a given pair of selected languages.

This means that, at the moment of the search, GloDoc works as though it were a bilingual tool when it shows the information recovered from the database (Figure 1). In fact, the interface provides all the registers for which the entry begins by the term placed in the search field, terms as well as composite expressions.

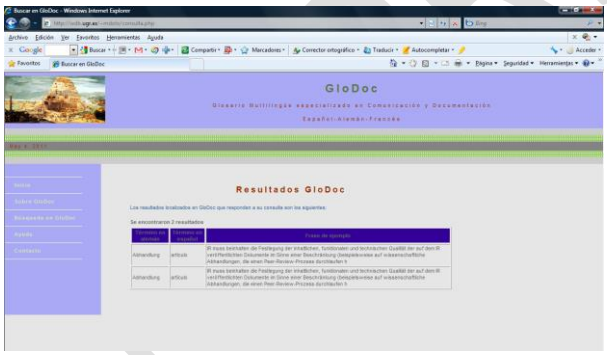

Figure 1: Example of GloDoc results

The terminological information in Spanish is broader (Figure 2). Grammatical information has been included together with a descriptor that frames the term within the large thematic groups of the speciality. Also, a field has been included to identify the location of the term or expression within the article. Consequently, it shows the importance given to this entry by the author of the document or by the magazine.

Figure 2: Example of Spanish terms in GloDoc

Concerning theme, a relationship of sub-areas was developed such as documental processes, the Internet, information sources and systems, information industry and commerce, information technologies and communication, documental languages, publicity, and marketing, among others. This list is a mere orientation for the most novice users in the field of communication and documentation, but it is not held to be an exact, precise, or rigorous classification.

4. Results

The results of this project can be found on the Internet (http://wdb.ugr.es/~mdolo) and are useful to professionals in the fields of translation and documentation who work with specialized texts. It is also a tool that helps make work easier and more effective, due to the online access in a very short time. The GloDoc interface allows the user to enter a word or expression from the field for one of four pairs of languages, and provides a table with the required information.

This tool is useful in finding context and collocations in scientific documents within information science, especially taking into account that we are managing real/actual expressions.

5. Conclusions

GloDoc is an innovative project in the field of contextual translation and terminology. This work has been shown to be useful for the translator in retrieving and representing knowledge from corpus techniques, as it enables a better understanding of the specialty field and its main concepts, as well as a better knowledge of its phraseology. In addition, the classification of the terms is important to integrate specialized and multilingual knowledge within a general

framework for non-specialized translators. In any case, it is important for the analysis of corpus and also the use of languages other than English without using intermediate languages. Another strong point of this project is that GloDoc is made available to the public on a free basis.

Although the project is linked directly to the university sphere, an effort was made to make it freely available as another resource to break down language barriers.

In any case, it is important to develop corpus analysis with the use of languages other than English.

Without doubt, this is a resource at the crossroads of terminology and documentation. Its functionality is designed to improve the work of both professional translators and specialists or researchers in information science, communication, and documentation.

References

Dorr, B. J. (1997): Large-Scale Dictionary Construction for Foreign Language Tutoring and Interlingual Machine Translation. In: *Machine Translation*, 12/4, 271-322

Fantinuoli, C. (2006): Specialised Corpora from the Web and Term Extraction for Simultaneous Interpreters. In: Baroni, M. / Bernardini, S. (eds.): *Wacky! Working Papers on the Web as Corpus*. Bologna: GEDIT, 173-190

Muresan, S. / Popper, S.D. / Davis, P.T. / Klavans, J.L. (2003): Building a Terminological Database from Heterogeneous Definitional Sources. In: *Proceedings of the National Conference on Digital Government Research*. Boston, 1-4.

Velardi, P. / Navigli, R. / D'Amadio, P. (2008): Mining the Web to Create Specialized Glossaries. In: *IEEE*, 23/5, http://www.dsi.uniroma1.it/~velardi/IEEE-IS-accepted.pdf

Peñas, A. / Verdejo, F. / Gonzalo, J. (2001): Corpus-based terminology extraction applied to information access. In: *Proceedings of the Corpus Linguistics 2001*. Lancaster: UCREL, 458-465

Zanettin, F. (1998): Bilingual Comparable Corpora and the Training of Translators. In: *META* 4, 616-630