A TRANSLATION-ORIENTED TOURISM TERM BANK

Adonay Custódia dos Santos Moreira
PhD, Associate Professor, Polytechnic Institute of Leiria
adonay@ipleiria.pt

ABSTRACT

All institutions and businesses that export their products and services around the world need appropriate bilingual or multilingual term banks to fulfil their communicative and commercial needs.

This paper examines the methodology used in the creation of a translation-oriented tourism term bank based on Portuguese info-promotional texts and their English versions.

It addresses the following topics: corpus constitution; creation of a subject-field classification system for the area of tourism; extraction of candidate terms with term-extraction tools; extraction of semantic relations, and terminology record completion with conceptual, linguistic and pragmatic information.

KEYWORDS

Tourism Terminology, Term Banks, Corpus, Brochures, Websites.

1. INTRODUCTION

The main goal of this paper is to provide insight into our parallel corpus-based approach to the creation of a term bank in the subject area of tourism. The term bank under construction is based on a unidirectional parallel corpus of Portuguese-English tourism info-promotional material and is conceived as a tool for all those interested in finding linguistic, conceptual and pragmatic information on the terminology of tourism. Thus, it can become particularly useful for translators, who need to master the specialized lexical items and find their appropriate foreign language equivalents; tourism professionals who work in an increasingly multilingual society and would gain from access to a 'ready-made' bilingual list of terms; and tourism trade businesses that market products and services internationally with printed or electronic multilingual texts.

2. TERMINOLOGY BASED ON CORPUS

Theoretically, this project is firmly grounded on Teresa Cabré's Communicative Theory of Terminology (1999), according to which terms or terminological units are simply lexical units that activate a specialized value in a certain pragmatic-discursive context. It is context that creates the specialized value, hence our emphasis on a linguistic-textual theoretical and methodological approach. It is fundamentally a descriptive model that records language in use and therefore acknowledges the principle of conceptual variation. The methodology which has been developed can support the creation of term banks in other specialized areas.

Cabré's approach recognizes the role of discourse and text (we perceive text as an embodiment of discourse) in the study of descriptive terminology, and acknowledges the multidimensional nature of terms. Cabré describes terms as many-sided: they are simultaneously units of knowledge, language and communication. Thus, these linguistic, cognitive and communicative units only acquire a meaning and function in a given discursive context. Our creation of a descriptive bilingual terminological product is based on the use of a parallel corpus where terms can be analysed in their context.

While this textual context is important, within a specific application terms are selected on the basis of this application as well. We extract text units, properly contextualized in the subject area, with the aim of creating a specific terminological application – a term bank on tourism – to bridge an economical and social need which has been previously identified. Thus, we follow the path of both a user-oriented and corpus-based bilingual terminology. Since our terminological approach is based on a parallel corpus and this corpus is determined by the purpose for which it will be used, we have named it "special purpose parallel corpus". Within this research, a special purpose parallel corpus consists of a corpus of original texts and their translations, which is used for terminological purposes.

3. COMPILATION OF TOURISM TERM BANK

The *Turigal* corpus on which the term bank is based consists of texts (printed brochures and websites) in Portuguese and their translations into English, all of which were sourced from Portuguese Tourism Regions, Regional Tourism Boards and Regional Tourism Promotion Agencies, and stored as plain text. This corpus, which for the moment contains 1,285,764 words (632,193 words in Portuguese and 653,571 in English; 469.873 words in the brochures and 815.891 words in the web pages), is already included in the *Linguistic Corpus of the University of Vigo* – CLUVI (Guinonart, 2003) and freely available for consultation at http://sli.uvigo.es/CLUVI.

The bilingual info-promotional texts – either printed or hypertexts – are clearly consumer-oriented, since their purpose is to transmit information to potential buyers in order to persuade them to buy or consume products or services. *Turigal* is considered to be sufficiently representative of all bilingual info-promotional materials published and distributed by the official entities responsible for the internal and external tourism promotion of Portugal in 2007, the year the texts were collected. The texts were aligned with the program *TRANS Suite 2000 Align* (Cypresoft, 2000) and three translation strategies – omission, addition and reordering – were encoded. The alignment always starts with the source sentence, which means that the translation sentences were split or joined together to match the source sentence. Thus, aligning a parallel corpus also entails its manual annotation, since translating is not a linear task. Translators can omit words, phrases or sentences from the source text, insert new ones as well as reorder segments or whole sentences in the translation. The format chosen for storing the aligned parallel texts is an adaptation of the TMX format (Translation Memory eXchange), as this is the XML encoding standard for translation memories and parallel corpora (Savourel, 2004).

The extraction of Portuguese term candidates from the *Turigal* corpus was done with the help of *kfNgram* (Fletcher, 2007), a computer program which produces a list of the most frequent words in the corpus, as well as the most frequent sequences of compound nouns, up to five words. English term candidates were identified with the aid of *NATools* (Simões and Almeida, 2007), a statistical word aligner workbench which shows the most likely translations of each Portuguese term. Reference works on tourism were also used to assist in the identification of key terms that were not found with the previous methods.

A subject-field classification system for the subject area of tourism was also devised. The hierarchical ascription of terms to a specific branch of the classification system was done with a view to systematizing the subject field as well as clarifying the sense of each term in relation to each other.

In the tourism term bank, terminology records group together equivalent terms in Portuguese and English. It is semantic content, defined in a single record, which establishes a bridge between both languages. Each record also groups all inter and intra linguistic synonyms.

Our research also involved the extraction of semantic relations. It is possible to identify semantic relations with the help of a pre-defined set of patterns, namely those of hyponymy-hypernymy and meronymy-holonymy.

Therefore, the terminological treatment of terms drawn from the *Turigal* parallel corpus includes three types of information: conceptual (thematic tree and semantic relations); linguistic (lemma, grammatical category and type of variation – lexical, morphological, syntactic, orthographic and length) and pragmatic (context of use, relative frequency of terms and regional terms).

4. CONCLUSIONS

The aim of this project was to compile an updated bilingual term bank comprised of pragmatic, linguistic and conceptual information in the specific subject area of tourism. All institutions and businesses that export their products and services around the world need appropriate bilingual or multilingual term banks to fulfil their communicative and commercial needs.

BIBLIOGRAPHY

CABRÉ, T. (1999), *La Terminología: Representación y Comunicación. Elementos para una Teoría de Base Comunicativa y Otros Artículos*, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona.

CYPRESOFT (2000), *TRANS Suite 2000 Align*, Belgium.

FLETCHER, W. (2007), *kfNgram*, MD, Annapolis.

GUINOVART, X. (dir.) (2003-), *Corpus CLUVI – Corpus Lingüístico da Universidade de Vigo*, Universidade de Vigo, Vigo, http://sli.uvigo.es/CLUVI/.

SAVOUREL, Y. (2004), "TMX 1.4b Specification", *Localisation Industry Standards Association*, http://www.lisa.org/standards/tmx/specification.html, accessed 01.03.2011.

SIMÕES, A., and ALMEIDA, J. (2007), *NATOOLS Query Interface (NAT-QI)*, http://linguateca.di.uminho.pt/nat/nat.pl.