ACTAS DEL III CONGRESO INTERNACIONAL DE LINGÜÍSTICA DE CORPUS

LAS TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES: PRESENTE Y FUTURO EN EL ANÁLISIS DE CORPUS

Editores:
María Luisa Carrió Pastor
Miguel Ángel Candel Mora

Editores
María Luisa Carrió Pastor
Miguel Ángel Candel Mora

ACTAS DEL III CONGRESO INTERNACIONAL DE LINGÜÍSTICA DE CORPUS.

LAS TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES: PRESENTE Y FUTURO EN EL ANÁLISIS DE CORPUS

EDITORIAL

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Esta editorial es miembro de la UNE, lo que garantiza la difusión y comercialización de sus publicaciones a nivel nacional e internacional.

Primera edición, 2011

ISBN: 978-84-694-6225-6

Ref. editorial: 6032

ÍNDICE

Turigal: compilation of a parallel corpus for bilingual terminology extraction

Adonay Custódia Dos Santos Moreira

School of Technology and Management

Polytechnic Institute of Leiria

Turigal, a parallel corpus of tourism advertising material, has been devised to support the creation of a bilingual term bank on tourism. The corpus consists of texts – printed brochures, guidebooks 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. For the moment, it contains 1,285,764 words and is included in the Linguistic Corpus of the University of Vigo (CLUVI). This paper describes the methodology used in the compilation of Turigal. First, we examine the process of text collection and storage. Then, we discuss Pearson's (1998) set of criteria for corpus design and text selection which has been considered when compiling our corpus. Finally, we present the alignment and tagging of Turigal.

Keywords: parallel corpus, corpus design, alignment, tagging.

1. Introduction

These last few years have witnessed an increase in research involving the compilation of large quantities of texts and their respective translations, as well as the development of techniques for processing those bilingual term banks (Bowker & Pearson, 2002; Biber, Conrad & Reppen, 2004; McEnery & Wilson, 2004). The present study is an example of such research as it uses a Portuguese-English unidirectional parallel corpus as a starting point for the retrieval of terminology. The main goal of our research is to exploit one of the possibilities offered by parallel corpora: the compilation of bilingual term banks. *Turigal*, a parallel corpus of tourism advertising material, has been devised to support the creation of a bilingual term bank on tourism. Our comprehensive term bank is comprised of pragmatic (context of use, relative frequency of terms), linguistic (gender, number, grammatical category, lemmas, synonyms) and conceptual information (thematic tree, semantic relations). It can eventually be useful to assist translators working in this industry, 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 through the use of printed or electronic multilingual texts.

First, we look at the collection and storage of texts. Then, we discuss Pearson's (1998) set of criteria for corpus design and text selection – namely size, constitution, publication, author, factuality, technicality, audience, intended outcome, setting and topic – which has been considered when compiling our corpus. And lastly, we present the alignment and tagging of *Turigal*.

2. Compilation of a parallel corpus for bilingual terminology extraction

2.1. Text collection and storage

The corpus on which the term bank is based consists of texts (printed brochures, guidebooks 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. For the moment, it contains 1,285,764 words (469,873 words in the leaflets and 815,891 words in the webpages; 632,193 words in Portuguese and 653,571 in English) and it is included in the *Linguistic Corpus of the University of Vigo* (Gómez Guinonart, 2003) and available for free consultation at http://sli.uvigo.es/CLUVI. Since our terminological approach is based on corpus – specifically a parallel corpus, where the meaning of terms arises from their context of use – and this corpus is determined by the purpose for which it will be used, we have named it "special purpose parallel corpus". This expression is adapted from Pearson (1998: 48)'s term – "special purpose corpus" – and it designates a parallel corpus built for specific purposes. Thus, within the present research, a "special purpose parallel corpus" consists of a corpus of original texts and their translations, which is used for terminological purposes.[1]

1 Within this research, a corpus is a set of texts of a given field, which have been written and used by specific groups of people and selected according to a specific purpose. In this case, the purpose is the extraction of bilingual terminology.

Due to the texts' format – brochures/guidebooks and hypertexts – the apparently simple task of storing them as plain text turned out to be time-consuming. On the one hand, the printed brochures and guidebooks had different formatting types (size, font, text layout and page configuration), different colours and quite often texts in Portuguese, English and other languages were kept side by side on the same page. All brochures and guidebooks had to be scanned. On the other hand, working with webpages, though more productive in terms of the quantity of texts obtained, also required substantial post-processing, since many webpages had formatting codes which prevented easy access. Webpages had multiple input formats and in the process of text conversion some chunks of text would sometimes disappear and hence had to be manually typed. Moreover, texts which were not translated or were only in English had to be disregarded. Finally, some newly implemented sites were extremely slow and one could only open a page at a time, which slowed down the storage process.

Both types of texts – printed texts and hypertexts – were then submitted to an Optical Character Recognition (OCR) programme and then to a spelling correction programme, in order to check the text generated by the OCR. The texts which remained practically illegible after the OCR was applied to them were manually typed.

All graphs, addresses and pictures have been removed as well as some information considered irrelevant for our terminographical purposes, such as proper names (people and companies) and addresses. As far as hypertexts are concerned, these have been saved sequentially, according to the "site map", whenever this was available. To allow for the alignment of texts – the process of matching each phrase in Portuguese to its English translation – all texts which have originally been saved individually were joined together in a single text, creating a larger text or "super text".

A total of 3,484 Portuguese and English webpages from 17 websites were stored as plain text and were subsequently aligned. It should be noted, however, that the number of stored webpages was substantially higher. Despite having similar titles, many webpages in Portuguese could not be matched to their English version, due to a lack of correspondence in content. Those webpages just in one language had to be disregarded as well.

With regard to printed promotional texts, a total of 110 brochures/guidebooks were stored. Table 1 shows the Internet addresses (URL) and the number of brochures/guidebooks gathered from Tourism Regions.

Table 1: List of websites and brochures/guidebooks from Tourism Regions.

PORTUGUESE TOURISM REGIONS	**URL USED IN *TURIGAL***	**NUMBER OF BROCHURES/ GUIDEBOOKS USED IN *TURIGAL***
Algarve	—	3
Alto Minho	http://www.rtam.pt	13
Alto Tâmega e Barroso	http://www.rt-atb.pt	2
Centro	http://www.turismo-centro.pt	22
Dão Lafões	http://www.rtdaolafoes.com	4
Douro Sul	—	—
Évora	http://www.rtevora.pt	—
Leiria / Fátima	http://www.rt-leiriafatima.pt	2
Nordeste Transmontano	—	2
Oeste	http://www.rt-oeste.pt	—
Planície Dourada	http://www.rt-planiciedourada.pt	2
Ribatejo	—	7
Rota da Luz	http://www.rotadaluz.pt	11
S. Mamede	http://www.rtsm.pt	1
Serra da Estrela	—	2
Serra do Marão	http://www.rtsmarao.pt	—
Setúbal / Costa Azul	—	15
Templários	http://www.rttemplarios.pt	6
Verde Minho	—	3

Table 1 only lists bilingual websites whose texts have been collected. At the time this research was undertaken there were 19 Tourism Regions, some of which – *Douro Sul*, *Ribatejo*, *Serra da Estrela*, *Setúbal/Costa Azul* e *Verde Minho* – did not yet have English websites. Sometimes, only the titles of the webpages were in English.

Turigal also comprises texts sourced from the Azores and Madeira Regional Tourism Boards, and the following Regional Tourism Promotion Agencies: ADETURN, ARTA, ATA Azores, ATA Algarve and ATL. Table 2 indicates the URL and the number of brochures/guidebooks collected from these organizations.

Table 2 – List of websites and brochures/guidebooks from Regional Tourism Boards and Regional Tourism Promotion Agencies.

REGIONAL TOURISM BOARDS AND REGIONAL TOURISM PROMOTION AGENCIES	**URL USED IN *TURIGAL***	**NUMBER OF BROCHURES / GUIDEBOOKS USED IN *TURIGAL***
ADETURN	http://www.visitportoenorte.com	2
ARTA	—	1
ATA Azores / Azores Regional Tourism Board	http://www.visitazores.org	7
ATA Algarve	http://www.visitalgarve.pt/	—
ATL	http://www.visitlisboa.com	—
Madeira Regional Tourism Board	http://www.madeiraislands.travel/pls/madeira/wsmwhom0.home	5

Most bilingual brochures/guidebooks displayed in Tables 1 and 2 were obtained in the aforementioned organizations or received by mail, after being requested by telephone or e-mail. Others were fetched from a Tourism Fair held in Lisbon in January 2007.

2.2. Corpus design and text selection

We have used Pearson's (1998: 58-62) set of criteria for corpus design and text selection – namely size, constitution, publication, author, factuality, technicality, audience, intended outcome, setting and topic – to outline our special purpose corpus.

At first glance, a 1,285,764 word-corpus is small; however, one should take into consideration its time-consuming conversion to electronic form and the shortage of bilingual promotional texts from official organizations. Like Pearson (1998: 57), we believe a special purpose corpus does not have to be as big as a general purpose corpus. *Turigal* is considered to be sufficiently representative of all bilingual (Portuguese-English) promotional materials published and distributed by the official organizations responsible for the internal and external tourism promotion of Portugal in 2007, the year the texts were collected. The fact that it is a corpus difficult to compile can make it even more interesting to be studied, since there will certainly be fewer people interested in spending time in its collection.

The corpus contains complete written informative/promotional texts, of different size, in Portuguese and their translations into English. These freely available texts 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. They come in multiple formats – brochures, guidebooks and websites – and they are all full texts (not extracts) from written sources. Most brochures and guidebooks have no publishing date, but the ones which do are mostly from 2005 and 2006.

All texts have been published by official tourism bodies. This validates texts as a potential source of terminology in the area of tourism.

The case of authorship is particularly complex in our corpus, since most websites, guidebooks and brochures do not mention the authors and translators of texts. Websites frequently give the name of the company responsible for creating the websites, but do not indicate the name of people responsible for creating their texts. However, since texts are published by official tourism bodies, one assumes their authors are experts in tourism or someone with technical qualifications in the area.

Pearson also considers the criterion of "factuality" (1998: 61). According to the author, texts must be factual or should represent what is known or believed to exist. In our study, this criterion is particularly ambiguous with respect to promotional texts: on the one hand, we can consider them factual, since they display a specific tourism product that can be purchased by consumers; on the other hand, the language that is used is not in any way factual. Dann (1996) and Buck (1977) identified some features of this language of tourism. According to Dann, the language of tourism only speaks in a positive manner of the services and attractions that it is promoting (1996: 65). Thus, it is a hyperbolic language, full of clichés, which is used to capture tourists' attention at all costs. Referring specifically to the language of tourism brochures, Buck remarks that these are naturally fraudulent, as they send preconceived messages that affect tourists' expectations and perceptions. Hence, we have considered that the criterion of "factuality" indicated by Pearson has no relevance to our work.

As for "technicality", Pearson makes a distinction between technical (written by specialists for specialists) and semi-technical texts (written by specialists for a specific target audience). Our texts fall into this second category. However, the audience of our texts is not the student or professional working in the discipline, but the average citizen with a lower level of expertise in the area.

The intended outcome of our texts is informative, but mostly promotional, and the setting corresponds to communication between relative experts and the uninitiated. In her work, Person rejected texts which fitted this communicative setting, on the grounds that they were not likely to contain a high density of terms.

Regarding the last criterion – topic – all our promotional texts belong to the area of tourism.

It should be noted that the criteria selected by Pearson to classify her corpus were not considered in the same way for our project. Such criteria should suit the objectives of each research project and our objectives were distinct from hers. Pearson wanted to select specialized texts likely to contain metalinguistic statements which could be used to formulate definitions (1998: 62). Her aim was to provide specific groups of researchers

with useful definitions of terms. Ours is to create a descriptive terminological resource designed to meet the needs of a specific group of users – translators.

2.3. Text alignment and tagging

Texts were aligned with the program *TRANS Suite 2000 Align* (Cypresoft, 2000) and 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, 2005). Each text has a header with information about text type (brochure, guidebook or website), its title in Portuguese and English, author, translator, publisher, year, URL, and date of access to the website and to the brochure, whenever the latter had no indication of publishing date.

As for the alignment itself, although a source sentence usually corresponds to a sentence in the translation, on some occasions one source sentence corresponds to two or more translated sentences or vice-versa, i.e., two or more source sentences correspond to one sentence in the translation. 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 tagging, 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. Here are some examples of the tagged *Turigal* parallel corpus.

This is an example of an omission in the translation, which means that a source segment or sentence has no correspondence in the translated text. The omitted segment is placed in bold, between the tags "<hi type="supr">" and "</hi>".

Table 3: Example of an omission in corpus Turigal.

<tu>

<tuv lang="PT-PT">

<seg>Se gosta de desportos radicais, nada como fazer uma descida no Rio Minho (Rafting), tendo já em Melgaço Associações que preparam tudo (profissionalmente) [[hi type="supr"]] **para que a descida seja um êxito** [[/hi]]. </seg>

</tuv>

<tuv lang="EN-GB">

<seg>If you are a radical Sports lover, try to the descending of the river Minho (rafting) in Melgaço, there you can also contact a Professional Association to organise all . </seg>

</tuv>

</tu>

This second example is an addition, in which the translator decides to add segments that do not exist in the source text. The added segments are placed in bold, between the tags "<hi type="incl">" and "</hi>".

Table 4: Example of an addition in corpus Turigal.

<tu>

<tuv lang="PT-PT">

<seg>São as cumeadas da serra do Gerês, as Terras de Bouro, as praias de riba Minho, as Terras Soajeiras, os contrafortes da Senhora da Peneda e da Senhora do Sameiro, Barcelos e as margens ridentes do Cávado.</seg>

</tuv>

<tuv lang="EN-GB">

<seg>[[hi type="incl"]] **Minho is there for you to discover it:** [[/hi]] the peaks of the Serra do Gerês [[hi type="incl"]] **(mountains)** [[/hi]], the municipality of Terras de Bouro, the beaches of Riba Minho, the territory around the Serra do Soajo, the spurs of Senhora [[hi type="incl"]] **(Lady)** [[/hi]] da Peneda and Senhora do Sameiro, Barcelos and the luxuriant banks of the Cávado river. </seg>

</tuv>

</tu>

Finally, a reordering, i.e., changing the position of segments in the translation, compared to their position in the source text. Since the alignment is always from source text to translation, the reordered segment always has to match the source sentence. The reordered segment is placed in bold, between the tags "<hi type="reord" x="1" >" and "</hi>". The tag "<ph x="1"/ >" indicates the original position of the reordered segment.

Table 5: Example of a reordering in corpus Turigal.

<tu>

<tuv lang="PT-PT">

<seg>- Azulejos da nave, historiados, Barrocos e monocromáticos, de fabrico Lisboeta, alusivos a Santa Cruz e à vida de Santo Agostinho - Púlpito da autoria de Nicolau de Chanterenne, sendo considerado uma obra prima do Renascimento. </seg>

</tuv>

<tuv lang="EN-GB">

<seg>- The nave with historiated and monochromatic baroque tiles made in Lisbon and representing Santa Cruz and Saint Augustin's life. - A Pulpit, made by Nicolau de Chanterenne, [[hi type="reord" x="1"]] **and considered a masterpiece of the Renaissance;** [[/hi]] </seg>

</tuv>

</tu>

<tu>

<tuv lang="PT-PT">

<seg>Data de 1521.</seg>

</tuv>

<tuv lang="EN-GB">

<seg>dating back to 1521 [[ph x="1"/]] </seg>

</tuv>

</tu>

Encoding omissions, additions and reorderings in the corpus allows the automatic search of these translation strategies and facilitates their analysis. However, the purpose of the present research is not the study of translation strategies, but the use of an aligned parallel corpus for term extraction.

3. CONCLUSIONS

The main objective of this paper was to provide insight into the compilation of parallel corpus *Turigal*. This corpus was built for terminological purposes: to enable us to retrieve terms, get examples of use and translation equivalents, and distinguish multiple meanings

of terms. Our ultimate goal was the creation of a descriptive terminological resource which is available for free consultation at <http://sli.uvigo.es/termoteca/> (Gómez Clemente & Gómez Guinovart, 2006).

REFERENCES

BIBER, D., CONRAD, S., & REPPEN, R. (2004). *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

BOWKER, L., & PEARSON, J. (2002). *Working with Specialized Language: A Practical Guide to Using Corpora*. London and New York: Routledge.

BUCK, R. (1977). The ubiquitous tourist brochure: explorations in its intended and unintended use. *Annals of Tourism Research*, 4, 195-207.

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

DANN, G. (1996). *The Language of Tourism: a Sociolinguistic Perspective*. Oxon: Cab International.

GÓMEZ CLEMENTE, X. & GÓMEZ GUINOVART, X. (dir.) (2006-). *Termoteca - Banco de Datos Terminolóxico da Universidade de Vigo*. Vigo: Universidade de Vigo. <http://sli.uvigo.es/termoteca/>

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

MCENERY, T. & WILSON, A. (2004). *Corpus Linguistics: An Introduction*. 2nd ed.. Edinburgh: Edinburgh University Press.

PEARSON, J. (1998). *Terms in Context*. Amsterdam – Philadelphia: John Benjamins Publishing Company.

SAVOUREL, Y. (2005). *TMX 1.4b Specification*. Localisation Industry Standards Association. Retrieved from <http://www.lisa.org/standards/tmx/specification.html>.