EMCOR: a medical corpus for terminological purposes[1]

Tamara Varela Vila and Elena Sánchez Trigo, University of Vigo

ABSTRACT

This article presents the bilingual (French-Spanish) comparable corpus EMCOR and how it has been used to create terminology resources to assist in the translation of medical texts. First of all, attention is paid to the features and interest of the chosen thematic sub-domain, that of the group of rare diseases (RDs) included under the name of inborn errors of metabolism (IEMs). This is followed by an explanation of the criteria used to design EMCOR, its compilation methodology and the processing tasks carried out. Finally, we show the various analytical techniques performed to extract the terms used in naming the pathologies chosen, as well as the concept hierarchy and the bilingual glossary created using the corpus. The study forms part of the research on text translation within the field of biomedicine, and more specifically RDs, on which we are currently working. This is a novel line of research in the medical sub-domain based on the creation of multilingual corpora to produce resources for translators, interpreters or scientific and technical writers that can also be put at the disposal of other users, such as health professionals or even patients themselves.

KEYWORDS

Translation resources, medical texts, corpus, terminology, inborn errors of metabolism (IEMs), rare diseases (RDs).

1. Introduction

The appearance of new technologies in the sphere of translation has brought a decisive change to the characteristics and volume of translation resources, enabling a series of tools to be developed that have transformed not only the task of translation itself, but also research in the field. One of the most useful of these tools are specialised multilingual electronic corpora.

More specifically, over the last decade the world of translation has witnessed the progressive standardisation of the work on electronic corpora that commenced in the 1990s in Translation Studies, a methodology that has produced extremely productive results. For this reason, corpus-based translation studies (CTS) are now one of the most fruitful lines of research and, according to some scholars, have even become the new paradigm in translation studies (Corpas Pastor 2008: 49).

However, the real usefulness of a corpus is closely linked to two key factors: the criteria used in the selection and compilation processes (Abaitua 2000: 2; Bowker and Pearson 2002: 9) and the methodology used to exploit them. Thus, a correctly compiled corpus can provide reliable and authentic examples of different linguistic phenomena as well as terminological and collocational information (Zanettin 2002: 240) of

great relevance in specialised texts. Work on corpora should thus form part of a properly focused research programme of substantial interest (Tymoczko 1998: 658).

In keeping with these premises, this article presents the steps taken to create the EMCOR corpus, a comparable corpus (French-Spanish) specialising in a sub-domain of the field of biomedicine (inborn errors of metabolism, IEMs), and the findings of the terminological analysis performed on it. It is divided into three sections, reflecting the different stages of the research undertaken. Firstly, given that the starting point for the selection of the corpus was a thematic area, we present the principal characteristics and focus of the chosen sub-domain: the group of rare diseases (RDs) collectively referred to as IEMs. This is followed by a description of the design criteria for EMCOR, the methodology used in its compilation and the processing to which it was subjected. Thirdly, reference is made to the results obtained from using the corpus, in particular, the terminological analysis carried out and the bilingual resources, in French and Spanish, that were created: the concept hierarchy of the IEM field in both languages and a bilingual glossary of names of the various diseases included in this domain. The paper ends with the main conclusions of our research.

2. Why a corpus specialising in inborn errors of metabolism?

IEMs are part of the wider group of pathologies generally referred to as RDs, also known as 'minority', 'orphan' or 'uncommon' diseases. They constitute a broad and varied set of diseases (6,000-8,000) sharing the common characteristic of reduced frequency (affecting fewer than 1 in 2,000 people). They also share their genetic origin (80% of cases) and they all are serious, chronic, degenerative, incapacitating and life-threatening diseases. Although there is currently no cure available for any of these diseases, they can be treated, thereby improving both the quality and expectancy of life of sufferers, also benefiting those close to them.

As has been pointed out above, each RD affects a small number of people, but the total number of sufferers is large. According to EURORDIS, 6-8% of the EU population (i.e. 24-36 million people) suffer from an RD, a significant figure in absolute terms. For this reason, there is now a demand that blanket action should be taken against RDs as a whole.

A series of initiatives is now under way to make RDs more visible and they are now beginning to be seen as a priority target for public health policy[2]. IEMs, the specific set of RDs to which the corpus we created relates, are a group of genetic pathologies caused by the alteration of a protein or enzyme that blocks a metabolic process.

They account for a significant number of RDs (30% of the total) and are the subject of much attention by the medical research community. These

two aspects make it possible to have access to a recently-produced, varied and quite extensive set of texts dealing with the diseases in question.

IEMs, as well as the RDs sub-domain, are of particular interest for corpus-based research, because it is a novel field of current interest in which there is a social demand for more widespread information and for multidisciplinarity. The texts produced in this respect will thus be produced in many languages, of high quality and in a state of permanent review and revision.

The diversity of IEMs also makes them particularly attractive from a terminological research standpoint and makes it possible to analyse an extensive but at the same time clearly defined range of terms.

The information given above allows us to delimit and justify the interest of the sub-domain being studied. Furthermore, it also provides a series of elements that will act as guidelines when determining the criteria to be followed when creating the corpus, which will now be explained in greater detail.

3. The EMCOR corpus: design and processing

The EMCOR corpus was compiled for the purpose of carrying out a study of the terminology associated with IEMs, bearing in mind the needs of translators of medical texts. We therefore designed a comparable bilingual corpus (French and Spanish) of complete written texts that are representative from the point of view of the latest scientific and social knowledge and approaches.

Given the existence of various points of view regarding what constitutes a comparable corpus, we should here point out that in this respect we adhere to the proposals put forward by Tognini-Bonelli (2001: 7) and Zanettin (1998: 617). We consider EMCOR to be a comparable corpus because it is composed of texts originally written in two languages (Spanish and French in this case). None of the texts is a translation of any of the others and the same selection criteria were used in all cases, thereby guaranteeing their comparability.

The decision to create a comparable corpus stems from their proven usefulness not only for terminologists and lexicologists, but also for translators, since they enable different languages or variants to be compared under similar communicative circumstances whilst avoiding the possible distortion arising from translations in a parallel corpus.

3.1. Text selection and compilation criteria

The process of selecting the texts that constitute EMCOR commenced after an initial documentation stage that gave us a deeper insight into the chosen sphere. Our starting point was the information on IEMs given in the handbook *Errores Congénitos del Metabolismo. Guía Divulgativa* (Hospital Sant Joan de Déu 2009). This is a reference work that includes a series of leaflets dealing with the 40 most common nosological entities within this sub-domain. This list of pathologies was used as a basis for searching for the texts that constitute the corpus.

In order to guarantee the quality of the EMCOR texts, our sources of information were chosen in accordance with the criteria indicated by leading specialists in the corpus field: Bowker (1996), Meyer and Mackintosh (1996), Sinclair (1996) Pearson (1998) and Aston (2002). The primary aspects governing our selection were therefore reliability, authority, accessibility, originality, specificity, exhaustivity and the target reader. These criteria allow us to specify the quality criterion proposed by Sinclair (1996).

The above-mentioned criteria meant that a variety of search strategies were used to select texts. One of these was to access sources of specific documentation on IEMs, such as: CIBERER, REDEMETH, SEIEM or AECOM. We also accessed more general sources of documentation on RDs, the sphere within which IEMs fall, such as: EURORDIS, Alliance Maladies Rares, National Organization for Rare Diseases, Canadian Organization for Rare Disorders, Instituto de Investigación de Enfermedades Raras or CRE Enfermedades Raras (Creer). Searches were also undertaken in specialised databases, such as ScienceDirect and Elsevier, and in general-purpose browsers using key words.

The criteria referred to above and used in the text selection process in order to guarantee their quality were complemented by a further set of criteria. As a first step we applied the four criteria proposed by Sinclair (1996: 6-8): quantity, quality, simplicity and documentation. Since our aim was to create a specialised corpus, which by their very nature tend to be more homogeneous and smaller than a reference corpus, quality (i.e. reliability, authority, accessibility, originality, specificity, exhaustivity) was given priority over quantity, although all efforts were made to establish a balance between these two aspects. It should also be noted that the texts included in the corpus have been clearly documented and tagged for the sake of transparency.

In addition to these general criteria, we adopted a series of more restrictive criteria linked to the intended outcome of our research, these being:

- Text mode: texts were limited to written texts in digital form that met the quality criterion referred to above.
- Full texts: one of our requirements was that the corpus should not only contain the terminology of interest, but also the contexts in which it is used.
- Text genre: after a prior analysis of the most commonly used text genres in this field and their availability, we selected the following: original research papers, review articles, case reports, article abstracts and patient information.
- Degree of specialisation: with a view to being representative of a discourse community (Swales 1990: 30), EMCOR includes not only specialised texts (original research papers, review articles, case reports and article abstracts), but also semi-specialised texts(patient information).
- Authority: all the texts were produced by specialists in the field or by prestigious bodies, such as hospitals, patient associations, health organisations and the like.
- Timescale: given that we wanted to discover the latest terminology currently used in this field, the texts selected were all produced during the period 2001-2010, thus ensuring that they are representative of the scientific state-of-the-art in the sub-domain of choice.
- Languages: the selected texts were written in Spanish or French, with no restrictions on the origin of their authors provided that the publications met the quality criteria in full. Nor were translations ruled out, since it was thought that whilst they may make less use of idiomatic constructions, the fact that they were drawn from reputed sources would mean that terminology was being used appropriately.

Given that our aim, as indicated above, has been to create a comparable corpus in French and Spanish, the same compilation criteria were used to select texts in both languages.

Having introduced the general criteria underlying the text selection process, let us now examine how these were applied specifically to the compilation of the sub-corpora that together make up EMCOR.

3.2. Corpus balance and comparability

We must first point out that each of the sub-corpora (French and Spanish) is divided into sections representing the different text genres included in EMCOR, on the basis of external criteria.

Structuring the two sub-corpora involved making decisions on various issues that arose in relation to the balance and comparability between them.

Thus, on the one hand, and putting comparability first, the internal balance of each sub-corpus was not strictly based on the number of texts or tokens per disorder. The number of texts written about a disorder is closely linked to its prevalence: the more people there are who suffer from a given disorder, the more research there is into the latter, resulting in a greater amount of related scientific output. For this reason it was decided to include the largest possible number of tokens for each disease, bearing in mind the restrictions imposed by the need to give priority to the comparability between sub-corpora.

On the other hand, and in the interests of representativeness, we also wanted to include approximately the same number of tokens per text genre and disorder in each sub-corpus, to ensure that they would be as similar as possible. Since the volume of texts on a given disease was greater in one language than in the other, a decision had to be taken as to which of the two conflicting criteria should prevail, comparability or quantity. In the former case we would have to eliminate texts from the language with the higher number of tokens in order to ensure an even number in both languages. In the latter case, there would be less comparability between the two sub-corpora, but the number of tokens would be higher, which would be an advantage when it came to extracting terminology. It was finally decided that the criterion of quantity should prevail, so texts were only eliminated when the difference in the number of tokens was very high[3].

The application of these criteria thus provided an initial selection of 121 texts in Spanish and 137 in French covering the various diseases included within the IEM domain. A further 6 texts in Spanish and 4 in French dealing with IEMs in general were also chosen, since this not only enabled us to extend the variety of terminological units in the corpus, but also to increase balance and comparability in terms of a number of tokens per text genre.

EMCOR thus consists of a total of 268 texts and 458,718 tokens, the Spanish sub-corpus having 127 texts and 225,101 tokens and the French sub-corpus 141 texts and 233,617 tokens. The result is a specialised corpus of a size comparable to the standards of representativeness of a corpus of this kind and of sufficient length to enable us to carry out a terminological analysis representative of the IEM field.

The table below details the most relevant statistical information concerning the composition of the sub-corpora in Spanish (ES sub-corpus) and French (FR sub-corpus): their size in bytes, the number of tokens and types, the type/token ratio, mean token length, sentences and paragraphs.

	ES SUB-CORPUS	**FR SUB-CORPUS**
Size (bytes)	1,539,751	1,549,319
Tokens (running words) in text	225,101	233,617
Tokens used for word list	217,429	225,402
Types	15,372	15,173
Type/token ratio	7.07	6.73
Standardised type/token ratio	40.84	39.88
Mean token length	5.42	5.33
Sentences	8,697	9,667
Mean (in tokens)	25.00	23.32
Paragraphs	143	177
Mean (in tokens)	1,520.48	1,273.46

Table 1. Statistical information on the EMCOR corpus

Of the data provided, specific mention should be made of the type/token ratio, which is commonly used to indicate the lexical richness of a text, since the higher the ratio, the greater the lexical variety should be. However, in general terms, the more specific the corpus, as in our case, the lower its lexical wealth, since the vocabulary tends to be less varied than in ordinary language texts and the same terminology tends to be used throughout a given text in order to avoid any ambiguity (Sager 1990; Pearson 1998; Pérez Hernández 2002). In this regard, our analysis reveals that EMCOR is sufficiently representative of the thematic area of IEMs from the point of view of both size and lexical characteristics.

The breakdown of the content of EMCOR by text genre is given below:

TEXT GENRES	**ES SUB-CORPUS**	**FR SUB-CORPUS**
Original research paper	28,853	30,213
Case report	49,695	52,583
Patient information	30,989	32,269
Review article	112,951	115,507
Abstract	2,613	3,045
TOTAL	**225,101**	**233,617**

Table 2. Tokens per text genre in the EMCOR corpus

As can be observed, the number of tokens per text genre is quite similar in both sub-corpora, in our opinion an important aspect when establishing comparisons between the quantity and type of terminology contained in the corpus as a whole.

3.3. Text processing

The texts that were finally selected were transformed into plain text to allow them to be processed by computer. Each text was assigned an alpha-numeric code before being included in EMCOR, enabling it to be identified by means of two tables containing the most relevant information about each text in each of the two sub-corpora.

All the texts were manually cleaned up to remove author names, repeated headings and footnotes, references, abstracts in languages other than that of the text itself and tables limited to numerical content. Although this step might be a limitation on possible future uses of EMCOR, it was considered that it would reduce the number of possible tagging errors. Furthermore, it also means that only the tokens really belonging to the body of the text would be counted. This is a procedure that has been used by other authors when creating a specialised corpus, such as Vihla (1999: 38) in the case of the Medicor medical corpus.

3.3.1. Documenting and tagging the corpus

All the texts that make up EMCOR are structured according to the Text Encoding Initiative Consortium (TEI 2009) guidelines with regard to tags and the DTD (Document Type Definition), which should accompany (or be taken as given in) any SGML document to enable it to be correctly interpreted. According to this standard structure, all texts must have a header and a body. The information contained in the headers of the EMCOR documents can be classified into two groups:

- Electronic document data: title (and code number), name of the compiler, the date it was included in the corpus, its length, the institution or person publishing it and bibliographical data.
- Encoding data: code of the language in which it is written, text genre and nature (original or translation).

In the case of tagging, the most widely used kind is morphological tagging, which identifies the different parts of a sentence. Tags of this nature are particularly useful when more exact searches of the corpus are being carried out.

EMCOR was POS-tagged in order to allow the search for syntactic patterns specific to the field of medicine in the second stage of analysis (see section 4.2). For this purpose, we used TreeTagger, a free tool developed by Helmut Schmid at the Institute for Computational Linguistics at the University of Stuttgart that can annotate texts written in German, English, French, Italian, Dutch, Spanish, Bulgarian, Russian, Greek, Portuguese and Chinese. TreeTagger, which performs part-of-speech tagging and lemmatisation tasks, uses a binary decision tree to calculate the probabilities of a word being paired with a tag. This probability is the

result of the path followed down the tree until a leaf is reached, and has an estimated accuracy of almost 97% (Schmid 1994).

4. Exploiting the EMCOR corpus: analyses and results

We used WordSmith Tools 5.0 to study the most frequently occurring terms and patterns in EMCOR and then we produced the concept hierarchy for the IEM field, from which we were able to create a bilingual glossary of the names given to the various pathologies.

Lack of space prevents us from including all the data we obtained, such as concordances, frequencies and the like, as well as the above-mentioned concept hierarchy and glossary, so we will limit ourselves to presenting a summary of the most relevant data, with examples.

4.1. Analysis of the most frequently-occurring tokens

Two frequency lists, one for each language, were used to determine the most frequently-occurring tokens in the respective sub-corpora. The list of the 50 most common tokens in this specialised field allowed us to demonstrate that the corpus is appropriate for the domain we wish to study and to discover the most relevant conceptual fields in this sphere.
Since the most commonly-occurring tokens in any corpus tend to be grammatical words, we had to use an exclusion list to prevent WordSmith Tools from including such words in the analysis.

The analysis carried out reveals that the most frequently occurring words in the EMCOR corpus come from to a range of lexical fields. Firstly, we find vocabulary in general use belonging to the sphere of biomedical science such as the Spanish words *enfermedad/es, paciente/s, diagnóstico, tratamiento, síndrome, síntomas, trastornos*, whilst in French we have *maladie, patient, diagnostic, traitement, syndrome, symptôme, troubles*, etc. Within this group, together with the above-mentioned forms *enfermedad/es* or *maladie/s*, it is worth noting the presence of a large number of words indicating the existence of a problem affecting the organism, such as the Spanish examples *déficit, deficiencia, defecto, síndrome, alteraciones, trastornos* and *retraso*. We also encounter other words from the sphere of biology, and more specifically that of genetics, an extremely important aspect in the diagnosis of pathologies of this nature, such as *mutación, metabolismo, gen, mitocondrial, cadena, aminoácido*, etc. in Spanish, or *mutation*, *métabolisme*, *gène* or *ADNmt* in French.

Finally, it is interesting to note the occurrence of the token *I* in both sub-corpora and that of *C*, this time only in the French sub-corpus in French. Searches carried out in both sub-corpora revealed that *I* is used as a numeral to refer to certain diseases, such as *glucogenosis tipo I*, *mucopolisacaridosis tipo I*, *tirosinemia tipo I*, *aciduria glutárica tipo I*, etc.

In the case of *C*, we discovered that in the French sub-corpus it is used in the terms *cytochrome c oxydase* and *ubiquinone-cytochrome c réductase*, as well as in others such as *protéine C*, *maladie de Niemann-Pick de type C*, etc.

The frequency analysis also revealed that 70% of the 50 most frequently occurring tokens in the Spanish sub-corpus have their equivalent amongst the 50 most frequently occurring tokens in the French sub-corpus, from which it can be deduced that although the content of the two sub-corpora is by no means identical, the most frequently occurring vocabulary is relatively common to both.

Furthermore, this analysis highlighted a significant presence of words used with a general meaning in common parlance which acquire a specialised value in the texts forming the corpus. This kind of word (such as *deficiencia*, *síndrome* or *enfermedad*) were then used as a basis for carrying out more precise searches and discovering terminological units belonging to this field. The presence in both sub-corpora of terms from other areas, such as genetics, indicates that EMCOR is representative of the multidisciplinary nature of the IEM sphere, referred to in the first section of this paper.

4.2. Analysis of the most commonly used patterns

The second of the analyses performed enabled us to extract most of the terminology contained in the corpus relating to the names of the pathologies included in the IEM group of disorders. The process has to be carried out systematically in order to obtain the largest possible number of candidate terms. The first step was to perform a search (using the POS-tags described in section 3.3.1) for all the syntactic patterns that our experience has shown us to occur with the greatest frequency in this field. These included the following, which are common to both Spanish and French:

- Common noun + preposition + common noun
- Common noun + preposition + proper noun
- Common noun + adjective + adjective
- Common noun + preposition + noun + adjective
- Common noun + adjective + preposition + noun.

This list of patterns enabled us to study the various sets of candidate terms obtained from the corpus and decide which were in fact terminological units. We will now comment on some of these results by way of example.

The first search we carried out in the sub-corpus in Spanish was for the pattern 'common noun + preposition + proper noun'. As was logically to be expected, a large percentage of the results returned corresponded to

names of diseases in the IEM group: *enfermedad de Pompe, síndrome de Leigh, síndrome de Hurler, enfermedad de Hurler, enfermedad de NPC, deficiencia de MHBD, enfermedad de Segawa, síndrome de Zellweger*, etc. The search also produced several sets of initials that correspond to the names of certain pathologies, such as *SLN* (*síndrome de Lesch-Nyhan*), *MPS* (*mucopolisacaridosis*) or *NPC* (*enfermedad de Niemann-Pick C*). Others, by contrast, refer to deficits, such as *deficiencia de OTC* (*ornitín carbamil transferasa*) or *déficit de HPRT* (*hipoxantina guanina fosforribosil transferasa*). It should be pointed out that the fact that initials appeared amongst the patterns we were looking for is due to errors deriving from the tagging procedure. In Spanish sets of initials are assigned the tag <ACRNM>, but in French, given that no specific tag for acronyms exists, these sets of initials were tagged in accordance with the function they perform in the sentence.

On the other hand, the analysis enabled us to exclude some of the candidate terms, such as *ciclo de Krebs*, which, although belonging to this area of biomedicine, is not a term that refers to entities within the IEM sphere, and we therefore decided not to include it in our classification. Others, such as *aparato de Golgi*, were also excluded because they fall outside the IEM sub-domain.

As mentioned above, we also made a search for the pattern 'common noun + preposition + common noun + adjective'. We were particularly interested in nouns with a high frequency of occurrence, such as *defecto*, *déficit* and *deficiencia*. A large number of diseases in this domain are in fact referred to by the name of the deficiency that causes them, such as *defecto de creatina cerebral* or *deficiencia de adenilosuccinato liasa*.

Taking this premise as our starting point, we decided to run a search for the sequence 'déficit de' in our corpus, since it is a frequently recurring pattern. We thus obtained a list of 145 concordance lines, which we studied in detail in order to determine whether the deficits in question belonged to our field of interest.

By frequency, the most significant entries on the list were *déficit de fosfoglicerato mutasa* (also referred to as *glucogenosis tipo X*), an inborn error of the carbohydrate metabolism, and *déficit de fosfoglicerato quinasa*, an inborn error of the energy metabolism. Other disorders found by the search included *déficit de fosforilasa muscular*, *déficit de arginasa*, *déficit de lactato deshidrogenasa* and *déficit de adenilsuccionato liasa*. However, we also detected the presence of certain deficits that bear no relation to the IEM sphere, such as *trastorno por déficit de atención* and a variety of vitamin deficits.

Another frequently occurring sequence in our corpus is *enfermedad de*, due to the large number of eponyms used, particularly in those pathologies with a number of sub-types, such as glycogenosis or

mucopolysaccharidosis. The results of the search carried out for this expression included the names of pathologies such as *enfermedad de Alpers* (also referred to as *síndrome de Alpers, síndrome de Alpers-Huttenlocher* and *polidistrofia de Alpers*), *enfermedad de Schilder* (*adrenoleucodistofia cerebral infantil*), *enfermedad de Sly* (*mucopolisacaridosis tipo VII* o *MPS VII*) or *enfermedad de Andersen* (*glucogenosis tipo IV*).

As has already been pointed out, not all the terminological units obtained from searches of this kind belong to the IEM field. This is the case, for example, of *enfermedad de Alzheimer* and *enfermedad de Huntington*, which, although belonging to completely different spheres, appear more than once in EMCOR.

Once we had finished extracting candidate terms from the sub-corpus in Spanish, we discovered that a large number of these diseases have similar names in French. For example, an examination of concordances revealed that *enfermedad de Canavan* is assigned a very similar denomination in French, namely *maladie de Canavan*, as well as the more technical name of *acidurie N-acetyl aspartique*. Furthermore, by examining concordances we were able to find the French equivalent of the Spanish terminological unit *lipofuscinosis ceroidea neuronal*, which is *céroïde-lipofuscinose neuronale.* Similarly, by means of other searches we were able to find the various sub-types of this pathology: *CLN1* (*lipofuscinose infantile* o *maladie de Santavuori-Haltia*), *CLN2* (*lipofuscinose infantile tardive* or *maladie de Jansky-Bielschowsky*), *CLN3* (*lipofuscinose juvenile, maladie de Spielmeyer-Vogt-Sjögren* or *maladie de Batten*) and *CLN4 (lipofuscinose adulte* or *maladie de Kufs*).

In this section we have given a small sample of the various searches we carried out on the corpus using syntactic patterns as our search parameter. The early results we obtained enabled us to define more specific searches that produced accurate results. Furthermore, the combination of searches in both sub-corpora allowed us to use the results obtained in one language to enhance our extraction of term candidates in the other.

5. Construction of a concept hierarchy for inborn errors of metabolism

The construction of a concept hierarchy for terms used to designate IEMs was by no means an easy task. In this regard, the absence of reliable sources of information on all the diseases that had to be classified and the abundance of synonyms for one and the same nosological entity were the two major problems[4].

In order to build this hierarchy, in addition to the sources already mentioned in section 2 above, we took as our starting point the

classification of rare diseases produced by the specialised Internet portal Orphanet.

We also consulted other sources of information, such as the International Statistical Classification of Diseases and Related Health Problems, the Online Mendelian Inheritance in Man (OMIM), CISMEF and PubMed, amongst others.

Similarly, we made use of a wide range of specialised sources, amongst which we can mention articles by authors such as Faid (2008) or Ribate Molina (2009). All of these sources were useful in helping us to discover the different perspectives from which pathologies of this kind can be classified.

Initially, we followed the general classification of IEMs proposed by Orphanet, enabling us to organise all the terminology in four major areas:

- Inborn errors of the complex molecule metabolism
- Inborn errors of the carbohydrate metabolism
- Inborn errors of the energy metabolism
- Inborn errors of the metabolism caused by intoxication.

Once this initial sub-division had been established, we used the information contained in EMCOR to locate all the terms extracted in their corresponding group and to exclude those which were at first considered to belong to this sphere, but were later revealed not to form part of it after a more detailed study of their characteristics.

The complexity of this classification process was due to a variety of reasons, the first of these being that in some cases the information about a given disease represented in the corpus did not always allow us to determine its main characteristics, and thus which group it belonged to. In cases such as these it was necessary to use documentation from an external source, such as those referred to above.

A second aspect that caused a certain amount of difficulty in the process was the large number of synonyms and near-synonyms used in this field. In some texts two designations were used as synonyms, although a closer analysis revealed a generic-specific relationship between them. For example, *glycogénose par déficit en phosphorylase kinase* seemed to be a *glycogénose type VI* but after some research we concluded that there are two subtypes (*glycogénose type VIA* and *glycogénose type VIB)* and *glycogénose par déficit en phosphorylase kinase* corresponds to *glycogénose type VIA*.

Another issue that cannot be overlooked is the presence of variations in spelling and syntax. In this regard, it must be pointed out that the synonyms themselves, variations in spelling and syntax and even initials

were included in the classification, with no distinctions being drawn between them.

Once we obtained the classification, it then became necessary to choose the preferred term for each pathology, i.e. the term we considered to be the most appropriate one to use when referring to that specific concept. These preferred terms were placed first in the concept hierarchy, followed by their synonyms, separated by a slash. Choosing the preferred terms was by no means easy, since although some are used with a much higher frequency than their synonyms and can therefore be taken as unique identifiers, there were other cases in which there was no predominant form or the predominant form was considered inadequate. For example, in the corpus, the number of occurrences of *enfermedad de Canavan* is much higher compared to *aciduria N-acetilaspártica*. However, we prefer to avoid using an eponym as unique identifier if another denomination exists. In cases like these the responsibility for the final decision lies with the terminographer.

Dubuc (1992: 107) proposes four criteria for evaluating synonyms and deciding which is the most appropriate for use as a unique identifier: frequency, manageability, suitability and motivation.

All these criteria were taken into account when deciding which were to be the preferred terms in the concept hierarchy we were constructing. A further factor that we considered to be important in this regard was coherence, at both intra- and inter-linguistic level. In the event of there being several different terms for a group of similar diseases, we decided that the preferred term should be the one common to all of them. For example: *enfermedad de Pompe* is a usual designation for *glucogenosis tipo II* (also called *GSD II* or *déficit del ácido alfa-1,4-glucosidasa).* However, in order to provide a coherent classification of glycogen storage diseases (*glucogenosis type I, Ia, Ib, Ic, Id, III, IIIa, IIIb*...), we chose *glucogenosis tipo II* as a preferred term. As far as inter-linguistic coherence is concerned, our choice of preferred term was determined by the similarity between them, provided that the terms in question were commonly used in both French and Spanish. This means that the names are as close to each other as possible in the two languages, thereby producing less confusion in the user's mind.

The figure below gives an example of the concept hierarchy we produced in Spanish.

1. Error innato del metabolismo / Error congénito del metabolismo / Enfermedad metabólica hereditaria
    1.1. Error innato del metabolismo de las moléculas complejas
        1.1.1. Condrodisplasia punctata ligada al X
        1.1.2. Defecto congénito de la glicosilación / Síndrome de Glucoproteínas Deficientes en Carbohidratos / CDG
            1.1.2.1. *Defecto de la N-glicosilación*

1.1.2.1.1. *CDG-Ia / Déficit de fosfomanomutasa*
1.1.2.1.2. *CDG-Ib / Déficit de fosfomanoisomerasa*
1.1.2.1.3. *CDG-Ic / Déficit de glucosiltransferasa 1*
1.1.2.1.4. *CDG-Id / Déficit de manosiltransferasa 6*
1.1.2.1.5. *CDG-Ig / Déficit de manosiltransferasa 8*
1.1.2.1.6. *CDG-Ih / Déficit de glucosiltransferasa 2*
1.1.2.1.7. *CDG-Ii / Déficit de manosiltransferasa 2*
1.1.2.1.8. *CDG-Ij / Déficit de UDP-GlcNAC*
1.1.2.1.9. *CDG-Ik / Déficit de manosiltransferasa 1*
1.1.2.1.10. *CDG-Il / Déficit de manosiltransferasa 7-9*
1.1.2.1.11. *CDG-IIa / Déficit de N-acetil-glucosaminiltransferasa 2*
1.1.2.1.12. *CDG-IIb / Déficit de glucosidasa 1*
1.1.2.2. Defecto de la O-glicosilación / Trastorno de la O-glicosilación
1.1.2.2.1. Defecto de la O-galactosilación
1.1.2.2.1.1. Síndrome de Ehlers-Danlos tipo Via
1.1.2.2.2. Defecto de la O-xilosilación
1.1.2.2.2.1. Exostosis múltiple hereditaria / HME
1.1.2.2.2.1.1. HME tipo I
1.1.2.2.2.1.2. HME tipo II
1.1.2.2.2.1.3. HME tipo III
1.1.2.2.2.2. Síndrome de Ehler-Danlos progeroide
1.1.2.2.3. Defecto de la síntesis de O-manosil glicanos / Distrofia muscular congénita

Figure 1. Example of the concept hierarchy in Spanish

As can be seen, the concept nodes, i.e. the non-terminological sequences used to organise the concepts within the classification with greater precision, appear in **bold** type. *Italics* are used to highlight terms that, in spite of not occurring in the corpus, are introduced in the glossary following the recommendation of the experts in order to provide a clearer vision of the field. This is the case of terms that only appear in one of the sub-corpora, but not in the other.

Once the concept hierarchy had finally been constructed, it was checked by several experts in the RD field[5]. This review process provided us with an endorsement of the reliability of the classification of terms used in the IEM field.

A further result of this research project was obtained by combining the French and Spanish concept hierarchies to produce a bilingual glossary that we hope will go some way towards alleviating the current scarcity of resources available in the IEM sphere.

See below for examples of the Spanish>French / French>Spanish glossary (Figure 2 and Figure 3, respectively).

A	
acidemia argininosuccínica	Cf. *déficit de argininosuccinato liasa (ASL)*
acidemia glutárica	acidémie glutarique Cf. *aciduria glutárica*
acidemia glutárica tipo I	acidémie glutarique de type I Cf. *aciduria glutárica tipo I* Cf. *deficiencia congénita de glutaryl-CoA deshidrogenasa*
acidemia glutárica tipo II	acidémie glutarique type 2 Cf. *aciduria glutárica tipo II*
acidemia isovalérica	acidurie isovalérique

Figure 2. Example of the Spanish>French section of the glossary

M	
MADD	Cf. *acidurie glutarique type 2*
maladie d'Alpers	enfermedad de Alpers
maladie de Batten	Cf. *CLN3*
maladie de Canavan	enfermedad de Canavan Cf. *acidurie N-acétyl aspartique*
maladie de Cavanagh	Cf. *CLN5*
maladie de Danon	enfermedad de Danon Cf. *glycogénose de type IIb*
maladie de Fabry	enfermedad de Fabry
maladie de Farber	enfermedad de Farber

Figure 3. Example of the French>Spanish section of the glossary

The glossary (Varela Vila *et al*. 2011) follows an alphabetical order and contains all the terms from the two concept hierarchies, including acronyms, on which it is based. It is our considered opinion that acronyms should neither be treated separately nor excluded from the glossary, since they play an extremely important role in this field, where they are used to refer to a large number of pathologies. Furthermore, there is a link from each term to its equivalent term in the other language, as well as a link forward to the corresponding preferred term.

Conclusions

This paper describes the criteria and methodology used to compile the EMCOR corpus and the way in which the latter has been exploited for terminological purposes.

In the creation of this comparable corpus in French and Spanish on the group of pathologies that together constitute the IEM domain, priority was given to three fundamental aspects: representativeness, quantity

(458,718 tokens) and a set of clearly defined quality criteria that would enable us to guarantee the reliability of the results obtained. As we have seen, there is an adequate balance between the two sub-corpora that together constitute EMCOR, both in terms of the total number of tokens and the number of tokens per text genre, each sub-corpus containing terminology used to refer to the diseases classified as belonging to this domain.

The concept hierarchies constructed in French and Spanish during the exploitation stage provide a clear overview of the chosen field of study and the characteristics of each disease, since each concept is defined by the position it occupies in the concept tree. Thus, each of the trees contains the terms used to refer to a total of 335 pathologies, covering the synonymy that is a characteristic feature of this sphere of knowledge in both languages.

We also combined the two concept hierarchies to create a bilingual glossary of the names of the various pathologies belonging to this sub-domain.

The focus of our research has been a thematic field of particular interest, involving a variety of disciplines and in which there is a social demand for the dissemination of information. Furthermore, the choice of languages, French and Spanish, gives the corpus an extra feature of interest: in a sphere in which English appears to be the dominant language, we have identified high-quality literature in both French and Spanish.

This study is part of the research line for the creation of resources for translators of medical texts on which we are currently working. It is thus envisaged not as the end point of a journey, but rather as one more contribution to the goal of obtaining a set of resources that satisfactorily describe the field of rare diseases and enable those concerned to use the specialised terminology of this domain correctly.

Having successfully produced this first version and analysis of EMCOR, as described above, our aim is to continue developing the corpus with a view to extending, improving and further exploiting it.

Bibliography

- **Abaitua, Joseba** (2000). "Tratamiento de corpora bilingües." Paper presented at the seminar *La ingeniería lingüística* (Soria, Spain, 17-21 July 2000). http://paginaspersonales.deusto.es/abaitua/konzeptu/ta/soria00.htm (consulted 17.01.2012).

- **Aston, Guy** (2002). "The learner as corpus designer." Bernard Kettemann and Georg Marko (eds) (2002). *Teaching and learning by doing corpus analysis*. Amsterdam: Rodopi, 9-25.

- **Bowker, Lynne** (1996). "Towards a corpus-based approach to terminography." *Terminology* 3(1), 27-52.

- **Bowker, Lynne and Jennifer Pearson** (2002). *Working with specialised language: a practical guide to using corpora*. London: Routledge.

- **Corpas Pastor, Gloria** (2008). *Investigar con corpus en traducción: los retos de un nuevo paradigma*. Frankfurt am Main: Peter Lang.

- **Dubuc, Robert** (1992). *Manuel pratique de terminologie*. Quebec: Linguatech Éditeur.

- **Marta, Valentina, Elena Sánchez Trigo and Tamara Varela** (2011). "Terminological Analysis in the Field of Medicine: The Translation of the Names of Assistive Products in the Book Occupational Therapy and Duchenne Muscular Dystrophy and a Proposal for their Classification." Sergio Maruenda-Bataller and Begoña Clavel-Arroitia (eds) (2011). *Multiple Voices in Academic and Professional Discourse: Current Issues in Specialised Language Research, Teaching and New Technologies.* Newcastle: Cambridge Scholar Publishing, 288-297.

- **Meyer, Ingrid and Kristen Mackintosh** (1996). "The Corpus from a Terminographer's Viewpoint." *International Journal of Corpus Linguistics* 1(2), 257-268.

- **Miquel Vergés, Joan and Elena Sánchez Trigo** (2010). "The social model of translation and its application to health-specialised search engines on the Internet. An example: the ASEM neuromuscular disease search engine." *Meta: Translators' Journal* 55(2), 374-386.

- **Pearson, Jennifer** (1998). *Terms in Context*. Amsterdam: John Benjamins.

- **Pérez Hernández, Chantal** (2002). *Explotación de los corpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento*. Madrid: CSIC/Elies. http://elies.rediris.es/elies18/ (consulted 17.01.2012).

- **Sager, Juan Carlos** (1990). *A Practical Course in Terminology Processing*. Amsterdam/Philadelphia: John Benjamins.

- **Schmid, Helmut** (1994). "Probabilistic Part-of-Speech Tagging Using Decision Trees." *Proceedings of International Conference on New Methods in Language Processing*. Manchester: UMIST. ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger1.pdf (consulted 17.01.2012).

- **Sinclair, John** (1996). *Preliminary Recommendations on Corpus Typology*. EAGLES Document, EAG-TCWG-FR-2, 1–13. http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html (consulted 17.06.2012).

- **Swales, John** (1990). *Genre Analysis.* Cambridge: Cambridge University Press.

- **Tognini-Bonelli, Elena** (2001). *Corpus linguistics at work*. Amsterdam/Philadelphia: John Benjamins.

- **Tymoczko, Maria** (1998). "Computerized Corpora and the Future of Translation Studies." Special issue of *Meta: Translators' Journal, L'approche basée sur le corpus/The Corpus-based Approach* 43(4), 652–660. http://www.erudit.org/revue/meta/1998/v43/n4/004515ar.pdf (consulted

18.06.2012). **Varela Vila, Tamara *et al*.** (2011). "Vocabulario francés-español de Enfermedades Raras: Errores Innatos del Metabolismo." *Panace@* 12(33), 35-78. http://medtrad.org/panacea/IndiceGeneral/n33-Tradyterm-VilaTrigoFerreiraHueso.pdf (consulted 17.01.2012).

- **Vihla, Minna** (1999). *Medical writing: modality in focus*. Amsterdam: Rodopi.

- **Zanettin, Federico** (1998). "Bilingual Comparable Corpora and the Training of Translators." *Meta: Translators' Journal* 43(4), 616-630. http://www.erudit.org/revue/meta/1998/v43/n4/004638ar.pdf (consulted 18.06.2012).

- — (2002). "DIY Corpora: The WWW and the Translator." Belinda Maia, Johann Haller and Margherita Ulyrich (eds) (2002). *Training the Language Services Provider for the New Millennium*. Porto: Facultade de Letras, Universidade do Porto, 239-248.

Sources of documentation

- "AECOM." http://www.ae3com.eu (consulted 17.01.2012).

- "Alliance Maladies Rares." http://www.alliance-maladies-rares.org (consulted 17.01.2012).

- "ASEM." http://www.asem-esp.org (consulted 22.06.2012).

- "ASEM-Galicia." http://www.asemgalicia.com (consulted 22.06.2012).

- "Canadian Organization for Rare Disorders." http://www.raredisorders.ca (consulted 17.01.2012).

- "CIBERER." http://www.ciberer.es (consulted 17.01.2012).

- "CISMEF." http://www.cismef.org (consulted 17.01.2012).

- "CRE Enfermedades Raras (Creer)." http://www.creenfermedadesraras.es (consulted 17.01.2012).

- "Elsevier." http://www.elsevier.es (consulted 17.01.2012).

- "EURORDIS." http://www.eurordis.org (consulted 17.01.2012).

- Faid, Valegh (2008). *Approches de glycomique apliques à l'étude des pathologies métaboliques des glycoprotéines*. PhD Thesis. Université Lille 1. http://ori-nuxeo.univ-lille1.fr/nuxeo/site/esupversions/e42ccf34-32b5-45be-86b4-3d6c99600838 (consulted 17.01.2012).

- Hospital Sant Joan de Déu (2009). *Errores Congénitos del Metabolismo. Guía divulgativa*. Barcelona: Unidad de Enfermedades Metabólicas Hereditarias – Hospital Sant Joan de Déu. http://pkuatm.org/guia-divulgativa (consulted 17.01.2012).

- "Instituto de Investigación de Enfermedades Raras." http://www.isciii.es/ISCIII/es/contenidos/fd-el-instituto/fd-organizacion/fd-estructura-directiva/fd-subdireccion-general-servicios-aplicados-formacion-investigacion/fd-centros-unidades/instituto-investigacion-enfermedades-raras.shtml (consulted 17.01.2012).

- "International Statistical Classification of Diseases and Related Health Problems." http://apps.who.int/classifications/apps/icd/icd10online (consulted 17.01.2012).
- "National Organization for Rare Diseases." http://www.rarediseases.org (consulted 17.01.2012).
- "Online Mendelian Inheritance in Man (OMIM)." http://www.ncbi.nlm.nih.gov/omim (consulted 17.01.2012).
- "Orphanet." http://www.orpha.net (consulted 17.01.2012).
- "PubMed." http://www.ncbi.nlm.nih.gov/pubmed (consulted 17.01.2012).
- "REDEMETH." http://www.cbm.uam.es/redemeth/informaciongeneral/dirnodos/directorio01.asp (consulted 17.01.2012).
- Ribate Molina, María Pilar and Feliciano Jesús Ramos Fuentes (2009). "Defectos congénitos de la glicosilación: no tan raros, pero grandes desconocidos." *Libro de ponencias del 58 Congreso de la Asociación Española de Pediatría* (Zaragoza, Spain, 4-6 June 2009), 96-102.
- "ScienceDirect." http://www.sciencedirect.com (consulted 17.01.2012).
- "SEIEM." http://www.eimaep.com (consulted 17.01.2012).

Biography

Tamara Varela Vila completed a degree in Translation and Interpreting at the University of Vigo (Spain) and a Masters degree in Multilingual Lexicology and Terminology – Translation at the University of Lyon 2 (France). She is currently doing her PhD at the University of Vigo, related to medical terminology. Contact: tvarela@uvigo.es.

Elena Sánchez Trigo is a Translation and Interpreting Full Professor in the Department of Translation and Linguistics of the University of Vigo (Spain), where she teaches subjects related to French-Spanish Translation. She has translated texts for the Spanish and Galician Federations for Neuromuscular Diseases (ASEM and ASEM-Galicia). Her

current research deals with medical translation, terminology, and the application of corpus linguistics to translation. Contact: etrigo@uvigo.es.

Notes

[1] This study forms part of the R&D&I project *Construcción eficiente de recursos lingüísticos multilingües* (INCITE08PXIB302179PR), funded by the Programa de Promoción Xeral da Investigación do Plan Galego de Investigación, Desenvolvemento e Innovación Tecnolóxica (INCITE) of the Regional Government of Galicia (Spain).

[2] This has led us to develop a line of research linked to the translations and translation revisions we carry out in this field. Regarding research activity, see for example Miquel Vergés and Sánchez Trigo (2010), concerning the MYOCOR corpus. The publications resulting from translations and translation revisions are available for consultation on the web pages of ASEM and ASEM-Galicia.

[3] For example, EMCOR includes six textual samples on *congenital deficiency of glycosylation*: four texts in Spanish (5,480 tokens) and two texts in French (5,865 tokens). We had initially selected two more texts in Spanish, but in the absence of other French texts on the same subject, their inclusion in the corpus would have resulted in a greater number of tokens in Spanish (12,500 tokens) than in French (5,865 tokens), and we therefore decided not to include them in the corpus. We thus succeeded in establishing a balance both in the number of tokens per disorder and per textual genre.

[4] The same problems were found when we created a classification on assistive products terminology (see Marta *et al*. 2011).

[5] Dr. Manuel J. Hens Pérez, Dr. Verónica Alonso Ferreira and Ms Ana Villaverde Hueso, of the Rare Disease Research Institute at the Carlos III Health Institute (Madrid). Dr. Bernardo Sopeña Pérez-Argüelles, researcher with the Internal Medicine Service at Meixoeiro Hospital, Vigo (Spain), also collaborated in the review.