Textual corpora for the analysis of terminological dependency: challenges and premises

Miguel Sánchez Ibáñez*

Departamento de idiomas, Universidad Católica de Murcia, Avenida Jerónimos, 135, 30107 Murcia

Abstract

The present paper aims to define and the main criteria which must be taken into account in order to build a textual corpus for the study of terminological dependency. After having completed a detailed analysis on the way Spanish depends on English to create neologisms in the field of Alzheimer Disease, (Sánchez Ibáñez, 2013), we propose a set of basic principles and strategies to keep both uniformity and asymmetry as opposite but yet complementary notions along the whole corpus building process, as a first step to keep track of this kind of linguistic subordination between languages. The thematic specificity, the chronological proximity or the genre of texts compiled have turned out to be very useful factors for our purposes. The conclusions we have drawn confirm that an accurate, bilingual corpus, built according to a set of well-defined parameters, is a valid starting point for the study of terminological dependency.

Keywords: Terminological dependency; textual corpus; textual genres; linguistic asymmetry; language for special purposes.

1. Introduction

Terminological dependency is quite a common phenomenon which is currently taking place in many communities whose languages are not the most privileged ones, in a global context, to convey specialized concepts. If we think of, for example, a group of scientists working at a cutting-edge level on a very specific domain, we tend to imagine them speaking English, as well as publishing their discoveries in this language. However, many of those

* Corresponding author.
E-mail address: msanchez@ucam.edu
scientists may not have English as their native language, and thus, they may have to put their own language aside and start using the privileged, dominant one, in order to legitimize the notions they talk about. In case they dare to use their own language, they may still import many structures and vocabulary from English. The resultant linguistic phenomenon from this unequal linguistic exchange is what we call terminological dependency (Portelance, 1987; Carbajal, 2004; Muñoz & V aldivieso, 2006; Loubier, 2008; Garrido, 2010; García Palacios, 2009; García Palacios & Humblye, 2012; Sánchez Ibáñez, 2013), which could be defined as follows:

Terminological dependency [is] a linguistic phenomenon arising from a unidirectional transfer of specialized denominations between two languages. Terminological neology in potentially dependent codes such as Spanish reflects the way the importation of units from English involves a set of linguistic asymmetries that affects the conceptual configuration of specialized domains by the importing community of scientists. (Sánchez Ibáñez & García Palacios, 2013)

In the following paragraphs, we will try to suggest a strategy to build a textual corpus for the study of this phenomenon.

2. Corpus for terminological dependency: concept and basic criteria

The study of different areas of Linguistics through the building of corpora has experienced a constant evolution along the past decades, due to two main factors: first, the technical progress which has speeded up the way big amounts of data can be organized in order to retrieve information from them; secondly, thanks to the main role played by the in vivo approaches in many studies. This has been the case of terminology. As a consequence of its own disciplinary evolution towards the study of terms in real contexts and not as unmovable, static elements, this discipline has taken corpora as a very useful resource to detect and study terms in context. In words of Cabré (2007), “la denominada Lingüística de Corpus permite explorar exhaustivamente las producciones lingüísticas y, con ello, ofrece al lingüista muestras de datos que mediante un análisis manual no llegan a la misma profundidad”.

Corpora have been defined in many different ways, as their uses, purposes and features evolved along the past years. Sinclair (1995) provided us with some of their key features when he explained that a corpus was nothing but “a collection of pieces of language that are selected according to explicit linguistic criteria in order to be used as a sample of the language”, implying that explicit purposes were needed in order to build meaningful corpora. Regarding the role of this collections of language in terminology, L'Homme (2004) points out that: “Au moment d'entreprendre une recherche, le terminographe réunit un ensemble de textes représentatifs du domaine dont il compte décrire la terminologie. L'ensemble constitué par ces textes est appelé corpus”. Taking into account these two approaches, we set four main funding criteria which helped us compile the kind of material we thought that might be profitable for our final aim:

- Chosen languages: we chose English and Spanish as the languages studied for several reasons: first, we chose English because it is the referential language for science in a global context, and the most used for the publication of specialized texts1. Plus, we focused on the Spanish language because we observed that the study and the systematization of its terminology is needed in order to enhance the transfer of science and knowledge expressed in this language.
- Thematic specificity: the domain chosen for the study was very specific, something which conditioned and limited the searching process for the both subcorpora we finally built. The aim of delimiting the selection of texts

---

1 English keeps on being the preferred langue in the Internet: 45% of the webpages are written in this language. Other European languages which are also used for this purpose are German (5,90%), French (4,41%), Spanish (3,80%), Italian (2,66%) and Portuguese (1,39%) (Source: Unión Latina: http://dtll.unilat.org/LI/2007/es/resultados_es.htm).
depending on the topic they talked about, and as systematically and objectively as possible, was the incentive
which led us to design two different protocols to select our documentary sources.

- **Chronological proximity**: Our aim was not only to detect terms which were relevant in the chosen domain, but
  also neological, so the chronological factor turned out to be other parameter we had to establish, as a crucial
  factor to delimit which texts we could choose and which ones we could not. In fact, Pearson (1998: 51)
  recommends that “a corpus which is being used for terminological studies may require that the material be less
  than ten years old”. It was necessary, thus, that the texts, besides being relevant for the domain, were also recent.
- **Bilingual study**: the multilingualism was an indispensable condition to delve into one of the issues we wanted to
  study: the assumption that the translator and the scientist made part actively and decisively in the term creation
  process. In consequence, it was necessary to establish common criteria for the search of terms in both languages,
  in order to ease the contrastive analysis: issues as basic as the level of specialization of the chosen texts or the
genre soon became very important factors which cannot be ignored when interpreting the conclusions drawn
from this study.

3. **English subcorpus**

We decided to take the configuration of the English subcorpus as our starting point for several reasons: the access
to acceptable documentary sources was extremely unbalanced, and we found a huge amount of specialized texts
written in English, with a great variety of registers and thematic specificity. Also, we proved that English was the
international language of reference used by scientists to talk about Alzheimer Disease.

In order to have a group of both thematically and linguistically appropriate pieces of language for our study, we
decided to extract all the texts to compose the English subcorpus from Medline², a bibliographical database where
many papers regarding Medical and Health Sciences are indexed. All these works have been collected by the
National Library of Medicine of the United States since 1947. Currently, Medline hosts 21.6 millions of documents
coming from 5,582 specialized journals and books. Its webpage registered almost two billions of visits only in 2011.

3.1. Information retrieval in Medline: the Medical Subject Headings.

Even if the figures we have just mentioned are impressive, they were not the only factor we took into account to
choose Medline as our main documental source. We finally decided to use it because of the information retrieval
system Medline uses, organized through what is called MeSH (Medical Subject Headings). Those headings are a
group of descriptors designed by the National Library of Medicine, forming a thesaurus of more than 25,000 items,
used to index each one of the papers added to Medline. On average, each indexed article has between ten and fifteen
descriptors. The combination of different descriptors with other parameters such as the period of time or the kind of
journal allows the users to delimit and ease their queries.

In our case, the Medical Subject Heading we wanted to work with was Alzheimer Disease, indexed in the
database since 1963, so we decided to design a retrieval strategy in order to collect papers regarding this medical
condition where some kind of disciplinary innovation could be found. We thought that this disciplinary innovation
could imply a lexical updating to convey all those new concepts and discoveries, something which might be useful
for our purposes. In consequence, we delimited three main aspects:

- **Chronological**: we decided to focus on the papers published from 2000 to 2010.

²Previous works, such as Racine’s (2004), leaned us towards the use this resource for our analysis. This author (ibid.: 12) enumerates several
features which define Medline as a very valuable source of text to build a specialized corpus: references are chronologically, all the texts are
digitally adapted and easily exportable, and the thematic organization is extremely thorough.
• Textual: likewise, we selected just the titles and the abstracts of each article, since we considered them to be the most representative parts of the text, supposed to contain a higher terminological (and neological) density.

• Thematic: our aim was to find the most productive subdomain within Alzheimer Disease along the past decade, because we thought that this progress would imply an increase in the creation of relevant terminology.

After having defined this, we made a two-step retrieval process:

a) In the first place, we proved that the use “Alzheimer disease”, either as Subject Heading or as keyword, was in clear progress along the chosen period. The retrieval was done in three different ways: using just “Alzheimer Disease” as keyword (focus), including in the search all the descriptors considered by Medline as equivalents in meaning to “Alzheimer Disease” (explode), and searching for the papers with “Alzheimer Disease” as keywords. In the three cases, the increase is clear, as shown in the Figure 1.

b) Secondly, and due to the huge amount of papers selected in the previous step, we searched for papers indexed between 2000 and 2010 under the “Alzheimer Disease” heading and the three Subject Headings appearing most often together with it, which, according to Medline, were Amyloid beta-peptides, Tau proteins and Apolipoproteins E.

This way, we aimed to refine the search and to increase the possibilities to find texts with a big neological density. The following figures show the evolution of the amount of papers indexed by Medline under the descriptor “Alzheimer Disease” and each one of the other three.
We deduced from these searches that the most profitable combination was “Alzheimer Disease” and “Amyloid Beta – Peptides”. In other words, the amount of papers in which these two Subject Headings appeared together from 2000 to 2010 increased much more clearly than in the other two possible combinations. This rising proved a growing amount of relevant texts, something which was likely to evidence the presence of neologisms thematically linked to our field of interest. That is why we decided to build our English subcorpus with the titles and abstracts of all the papers indexed by Medline from 2000 to 2010 under the Subject Headings “Alzheimer Disease” and “Amyloid Beta – Peptides”.

4. Spanish subcorpus

After having focused on the English part of our corpus, the next step was to collect a representative group of texts written in Spanish. In consequence, we considered carefully the most pertinent criteria to follow and soon we realized that the online-available materials in Spanish were very different to those found for the English subcorpus. Our searches proved that there were no databases as vast as Medline, maybe due to several reasons: first, Medline itself, as a linguistic and disciplinary referent for publishers and scientists, limits and even blocks the proliferation of resources in other languages. Second, specialists prefer to adapt themselves to the international canons, in order to obtain more prestige and recognition, rather than setting their own ones. In consequence, publications regarding Alzheimer Disease written in Spanish are mainly devoted to educational purposes, or are not as specialized as those published in English.

We were finally conditioned by the need to choose texts from the same textual genre than those compiled for the English subcorpus, since we realized that the hypothetical terminological equivalencies we were supposed to find would be more obvious if both subcorpora were made up of the same kind of texts. Thus, we considered the textual genre as a determining factor for the creation of new terms. Keeping in mind the essential evidences, we observed that we could base our search in Spanish on that criterion, even if the Spanish publications were not as specialized as the English ones. That way, we managed to adopt one of the few constants of our analysis: all the texts compiled, either for the English or for the Spanish subcorpus were scientific papers or, at least, well-defined parts of them. We found a journal which fulfilled our needs of thematic and chronological delimitation and which was easily accessible on the Internet: the Revista de Neurología, which aims to publish articles on the main discoveries and progresses taking place in the field of Neurosciences. It is also the official bulletin of the Spanish Society of Neurology and of more than fifteen similar regional associations. 56 biannual volumes have been published so far, but the amount of numbers included in each one of them varies significantly. Its webpage offers to the users all the papers published since the volume 26 (January-February 1998), separately and in easily downloadable formats, something which was very helpful.

The information retrieval system was much less complex than in Medline even if it is bases on similar principles: every paper is indexed according to a group of thematic areas, which subdivide the field of neurosciences into several subdomains. Similarly to what happened with the English subcorpus, the main parameter followed to select
the texts was the thematic specificity. In this case there were some limitations that did not exist in Medline: Firstly, the absence of a big list of Subject Headings and of lists of equivalents for each one of them; secondly, the impossibility to make combined searches, something which was extremely useful in the case of the English subcorpus. Finally, the fact that “Alzheimer Disease” was not a Subject Heading itself turned out to be another significant impediment.

Due to all these differences, we thought of a particular strategy for the selection of texts written in Spanish: we decided to select all the papers indexed under the tag “Neurodegeneración”. After this first filter, we observed that there was a significant group of texts which did not concern the Alzheimer Disease, so we applied a second selection: among all the articles indexed under “Neurodegeneración”, we chose just those having “Alzheimer” as a keyword. Once again, the existence of recurrent textual elements became a decisive factor for the compilation, something which proves that the textual genres and the way information is organized in them ease the detection of neologisms. The following table summarizes and puts in contrast the main features of the two subcorpora:

<table>
<thead>
<tr>
<th>Feature</th>
<th>Subcorpus 1</th>
<th>Subcorpus 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>English</td>
<td>Spanish</td>
</tr>
<tr>
<td>Number of words</td>
<td>620, 513</td>
<td>1, 114, 678</td>
</tr>
<tr>
<td>Textual markers used during the search process</td>
<td>Alzheimer disease, amyloid beta-peptide</td>
<td>Neurodegeneración, Enfermedad de Alzheimer</td>
</tr>
<tr>
<td>Chronological period</td>
<td>2000-2010</td>
<td>2000-2011</td>
</tr>
<tr>
<td>Level of specialisation</td>
<td>High - Very high</td>
<td>Medium - High</td>
</tr>
<tr>
<td>Information retrieval system</td>
<td>Combination of Medical Subject Headings (MESH)</td>
<td>Combination of thematic areas and keywords</td>
</tr>
<tr>
<td>Parts of the texts selected to build the subcorpus</td>
<td>Titles and abstracts</td>
<td>Full papers</td>
</tr>
<tr>
<td>Symmetry between specialists</td>
<td>Total. In both cases, texts had been conceived for and by scientists</td>
<td></td>
</tr>
<tr>
<td>Selection criteria</td>
<td>Thematic specificity, Chronological proximity, chosen, cognitive adequation, sources’ reliability</td>
<td></td>
</tr>
<tr>
<td>Accessibility of the material</td>
<td>Open source in .pdf format convertible into .txt.</td>
<td></td>
</tr>
</tbody>
</table>

5. Conclusions

After having built both subcorpora, we could draw some conclusions: first, we proved that, as we had ventured before starting the whole process, it was necessary to build a couple of subcorpora which were somehow comparable but never parallel, since the asymmetry we wanted to prove thanks to the texts compiled should find its origin in the asymmetry of the documentary sources chosen.

Likewise, we observed that it was necessary to combine the notions of imbalance and uniformity in order to find evidences of dependency. The first element was reflected on the difficulties we had to select texts equally specific for both subcorpora. The unequal number of words needed to find comparable data was also a clear sign of asymmetry. On the other hand, uniformity was firmly based on the chronological period chosen in both cases, as well as on the textual genre of all the documents selected, both in English and in Spanish. It was also proved that in the case of the depending code, in this case, Spanish, more flexible and encompassing criteria were needed. In short, we could say that a corpus built to study terminological dependency clearly proves till what extent Corpus Linguistics can be a flexible discipline, able to adapt itself to the study of various and changing phenomena, such as the linguistic subordination.
References


