ABSTRACT: It could be said that specialized discourse literacy (academic and professional) has just begun to be explored in Chile. Advances along this line must start from a deep, empirical analysis of real data. Thus, one way to access the specialized written genres employed by academia is to begin from the tenet that all materials read by students during their university training reveal relevant data about disciplinary discourse and knowledge. This article presents information about a research project, in its early stages, at Pontificia Universidad Católica de Valparaíso, Chile, that focuses on the collection, construction, and description of an academic corpus based on texts collected in the academic and professional areas of four disciplinary domains of knowledge: Industrial Chemistry, Construction Engineering, Social Work, and Psychology. A review of the concepts of specialized, academic, and professional discourse is presented. This is followed by a description of the procedures of collecting and organizing the Academic Corpus PUCV-2006, which is comprised of almost 100 million words. In addition, a preliminary text typology of the 652-text corpus is provided.

KEYWORDS: Academic discourse, professional discourse, corpus linguistics.

1. Introduction

The past decade has witnessed a remarkable surge of interest in the study of linguistic variation through the discourse of scientific disciplines. This focus on diversity helps explain the divergent construction of specialized knowledge within the discourse communities. The well-founded perception that there is no such thing as a unified disciplinary discourse and that it is more useful to speak of disciplinary discourses in the plural form (Hyland, 2000), has become increasingly recognized by researchers (Bhatia, 2004). Empirical findings from diverse linguistic approaches have documented the relevance of analysis based on corpus as a way of advancing and describing linguistic and discourse variations in greater detail through the disciplines and through types of prototypical texts (Biber, 1988, 1994, 2005, 2006; Martin & Veel, 1998; Wignell, 1998; Williams, 1998; Swales, 1990, 2004; Flowerdew, 2002; Parodi 2005, 2006, 2007).

This article describes a project currently being carried out at the Pontificia Universidad Católica de Valparaíso, Chile. This project involves the collection, construction, and description of a corpus of written discourse beginning with collected texts in four disciplinary domains of knowledge in both the academic and professional fields. The first part of this article presents some theoretical referents that frame the research. The second part establishes the parameters of the constitution of the corpus, presents a projection of its eventual size, and undertakes a general description of the text types that have been found.

2. Background

2.1. Theoretical framework

Given the lack of a solid didactic base and the scarcity of knowledge about materials to improve discourse competencies of university students and professionals, the research has been aimed in this direction, that is, towards the study of written discourse on complementary
fields of university academics and the professional workplace. The initial steps consisted of collecting and studying the written texts that university students read (and which provide them with knowledge particular to their chosen discipline) in four academic degree programs and the written texts in that form the core of daily communication in the professional workplaces that correspond to these same four areas.

In this context, it is relevant to state certain fundamental assumptions for this research. On the one hand, the approach taken towards ‘discourse’ is decidedly interdisciplinary and of a psycho-sociolinguistic nature. Hence, the texts chosen are linguistic units immersed in a cognitive and social context, that is, whose function is determined cognitively and contextually. In this way, the texts are linguistic units with closed meanings in virtue of producers/speakers and readers/hearers in particular contexts and with defined purposes, with prior knowledge constructed from human cognition in specific social contexts and stored in their minds. In other words, the texts are conceived of as processes and products of cognition and context, and, at the same time, as forming supports that, in part, help people construct their world and their environment.

2.2. Specialized discourse: academic and professional

As a general frame of reference, the notion of specialized discourse (SD) is described as a one that embraces all of the objects of study of this research. This is followed by a critical discussion about the concepts of academic discourse (AD) and professional discourse (PD). The use of the term “specialized discourse” is currently widely accepted by the majority of language scholars. However, from its initial use, it has been used to express a variety of meanings. Hence, SD includes a varied set of text types, but with certain prototypical traits. It is precisely this idea of a heterogeneity of texts within a scale of gradation that Parodi is applying when he approaches the notion of SD. According to this notion, SD must necessarily be understood as a continuum in which texts are aligned along a diversified gradient that runs from a high degree to a low degree of specialization. Thus, SD could be conceived of as a hyperonime of AD and PD. In this sense, Parodi (2005) defines SD by using a series of characterizing linguistic traits. One aspect that does not directly apply to this option, but which does appear in the prototypical traits proposed by Parodi (2005), is the use of a specialized lexicon. It is highly important to apply this trait to SD, given that it constitutes one of the principal aspects that characterize this type of discourse (Cabré, 1993; Burdach, 2000; Cabré, Doménech, Morel & Rodríguez, 2001; Ciapuscio, 2003; Cabré & Gómez, 2006). Academic and professional discourses are made operational through a set of texts that can be organized along a continuum in which the texts are linked together, from general school discourse to university academic discourse to professional discourse in a workplace environment. This is presented graphically in Figure 1:
2.2.1. Academic discourse

There is no doubt that any newcomer to the study of AD will find a diversity of approaches and perspectives that make an initial understanding of the field quite difficult. As Flowerdew (2002) suggests, there has been little systematic research into exactly what AD is. When one undertakes the study of this type of discourse the following questions need to be addressed: a) Are there any existing criteria that accurately define this type of discourse? If so; b) What type of criteria are they? Below three approaches to AD will be considered: 1) a functional communicative approach, 2) a contextual approach, and 3) a textual approach.

In functional communicative terms, AD is characterized by the predominance of description with persuasive and didactic purposes. Furthermore, it is a type of discourse that expresses credibility and prestige, that is to say, an authorized discourse about the subject it is dealing with. From the contextual criteria, AD is that which is used in academic contexts or for academic purposes (Kennedy, 2001; Flowerdew, 2002; Dudley-Evans & St. John, 2006). However, it is evident that academic purposes are varied and not always easy to determine, which makes the criteria rather complex. This is due to the fact that AD do not have a clear limit and may be confused or assimilated by other types of discourse in nearby fields, such as technical-scientific discourse, professional discourse, pedagogic discourse, or institutional discourse (López, 2002; Flowerdew, 2004). From a contextual viewpoint, Hyland (2000) argues that in order to characterize AD, identifying the interactions and the participants involved in the interactions is indispensable. In other words, analyzing the texts as social practices is critical. This approach includes an analysis of the mediums in which these texts circulate and are used (Gunnarsson, 1997). In this sense, AD is considered a manifestation of a specific community (Valle, 1997). In global terms, due to the fact that AD is oriented towards the transmission of knowledge, generally through definitions, classifications, and explanations (Wignell, 1998), its linguistic features attempt to produce an effect of clarity and even objectivity, avoiding ambiguity and erroneous interpretation. Moreover, in order to achieve conciseness, it tends towards an economy of words, an absence of empty adjectives and the elimination of redundancy and repetition. It maintains a controlled syntax in a set, established order, and it has a higher proportion of nominalizations expressed in words than in
non-academic discourse (Ciapuscio, 1992; Halliday, 1993; Lang, 1997; Gotti, 2003; Charaudeau, 2004; Parodi & Venegas, 2004; Cademártoni, Parodi & Venegas, 2006).

Non-linguistic aspects frequently found in this type of discourse, such as chemical formulas, physics equations, virtual recreations, mathematical representations, and symbols, must also be examined. In addition, items such as graphs, tables, figures, diagrams, and other graphic representations are found in this discourse. In view of the above aspects, Lemke (1998) suggest speaking about a text as a hybrid, semiotic system. This has also been studied as multimodal texts (Kress & van Leeuwen, 1996).

2.2.2. Professional discourse

Without a doubt, the problem that faces AD in its search for a strict characterization is similar to what occurs in the study of professional discourse (PD). This happens because in some cases these two terms tend to become confused, for example, when the term PD is used in a general sense that includes AD, and vice versa. In spite of the fact that these two objects of study can, to a certain extent, be theoretically confused, in the present investigation these two types of discourse will be differentiated based on the criteria of the environment of collection. That is to say, PD will be that which is collected in contexts of professional use and circulation, whereas AD will be that which is collected in contexts of academic use and circulation. Nevertheless, there is an inevitable area of overlap or intersection between these two types of discourse. Therefore, research objectives include identifying and describing those texts that are used in both domains, and form a nexus between the academic and professional domains.

It is worth noting the work of Bazerman y Paradis (1991a) and their perspective on the notion of PD. These authors suggest that the structure of PD is founded on a textual dynamic that gives form to a profession. Bazerman y Paradis (1991b) present a series of related articles describing the way in which professional communities organize themselves based on their own relevant texts. PD, in this sense, are those texts which bring together specific knowledge of the world, which, in turn, constitute the purposes of the professional community (Berkenkotter, Huckin & Ackerman, 1991; Doheny-Farina, 1991; Bathia, 1993; Christie & Martin, 1997).

Macrostructure and the superstructure have not been extensively examined by any analysis of PD. However, studies can be found that apply these categories to administrative language (MAP, 1995). López (2002) applies some rhetorical microstructures to the analysis of a text about economic policy, but there are no systematic studies with respect to how these rhetorical microstructures are distributed in each professional field. With regard to linguistic traits, the lexical level is the area that has received the greatest attention of academic studies, especially the specific terminological analysis associated with each particular profession (Ciapuscio, 2003). At the morphological level, an issue that has been extensively investigated is the role that nominalizations perform (Chafe, 1982, 1985; Biber, 1986; Ciapuscio, 1992; Halliday, 1993; Lang, 1997; Parodi & Venegas, 2004; García, Hall & Marín, 2005; Cademártoni, Parodi & Venegas, 2006).

3. The research

3.1. Project in progress

Given the above framework, this research proposes to carry out a descriptive-comparative study of linguistic-textual order, beginning with the texts that are read in the academic and professional fields in the areas of Basic Sciences and Engineering, as well as
the Social and Human Sciences. This will be accomplished by examining an academic corpus and a professional corpus using a methodology based on the linguistics of corpus (Sinclair, 1991; Leech, 1991; Stubbs, 1996, 2006; Simpson & Swales, 2001; Tognini-Bonelli, 2001; Teubert, 2005; Parodi, 2006, 2007), with multi-dimensional and multi-trait analysis. The academic corpus will be taken from a university setting, whereas the professional corpus will be comprised of texts found in the work settings that correspond to the four university degree programs from which these same professionals graduated. The texts that comprise the corpora are chosen following criteria that are highly representative and accurately reflect the workplace environment.

The academic field will be defined by four degree programs offered by Pontificia Universidad Católica de Valparaíso, Chile: Industrial Chemistry, Construction Engineering, Social Work and Psychology. The professional level is comprised of a set of texts that are used in the professional circles in which graduates of these degree programs work.

3.2. Constitution of Corpus PUCV-2006

As stated above, the aim is to collect 100%, or as close as is possible, of the written material that is required reading or reference material for all requisite courses of the respective university degree programs and for the professionals working directly in the respective fields of work that correspond to the graduates who studied and obtained their professional titles from those four degree programs. The specific methodology of this research is divided into different stages according to the status and focus of each one:

Procedures: Academic Corpus

- Construction of a database with the complete curricula of the four degree programs (including the syllabi of all required courses).
- Construction of a database with obligatory bibliographic references of all required courses.
- Collection of complementary materials that the professors provide through prepared note files and photocopied materials.
- Preparing a survey for all the professors of each of the four degree programs, which included a request for the complementary materials mentioned above.
- Searching the Internet to find those titles already available in digital format, thus minimizing time spent on digitalization.
- Collection of the texts from the corresponding libraries.
- Photocopying of each text in order to maintain a database in paper format.
- Forming and training a team of ten people to do the work, such as scanning and compiling the digital texts.

4. Preliminary results

Two types of results will be presented below. Firstly, we provide the definitive constitution of the Academic Corpus PUCV-2006 in quantitative terms, as distributed among the academic disciplines and the four course degree programs. As part of this quantitative analysis the current status of the project is presented, including the work in progress, in terms of the percentage of text already digitalized and currently operating in the online system, El Grial. In addition, the first typological approximation of the texts that comprise the total
Academic Corpus of PUCV-2006 is described. This description defines nine types of text identified and quantifies their overall occurrence in the total corpus. In Table 1, below, the figure 652 represents the total number of texts of the Academic Corpus PUCV-2006.

<table>
<thead>
<tr>
<th>Scientific Field</th>
<th>University Degree Program</th>
<th>Number of texts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic Sciences and Engineering</td>
<td>Industrial Chemistry</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>Construction Engineering</td>
<td>92</td>
</tr>
<tr>
<td>Social and Human Sciences</td>
<td>Social Work</td>
<td>188</td>
</tr>
<tr>
<td></td>
<td>Psychology</td>
<td>296</td>
</tr>
<tr>
<td><strong>Total texts</strong></td>
<td></td>
<td><strong>652</strong></td>
</tr>
</tbody>
</table>

Table 1. Constitution of the Academic Corpus PUCV-2006: number of texts

The first thing that catches one’s attention is the high degree of heterogeneity. Moreover, there is both a progressive increase in and a substantial difference between the quantity of texts in the fields of Engineering and the Basic Sciences and the Social and Human Sciences, as well as a considerable difference between the specific degree programs themselves. A preliminary interpretation might lead one to believe that students in the Social and Human Sciences have to read much more than students in Engineering and the Basic Sciences. One might even risk saying that Psychology students read up to four times more than Industrial Chemistry students. However, before making such claims, one must take into account a very relevant factor – the length of each text in question, which may completely alter one’s first impression of the data.

In any case, even without calculating length of text into the equation, one can state as a fact the there is a greater quantity of texts in use in one disciplinary domain than in another disciplinary domain. Indeed, it is quite accurate to point out that the number of texts found in Industrial Chemistry is far fewer than the number of text found in Psychology. However, a series of linguistic and academic factors may have a bearing on this, such as less exposure to a broad typological variety, a concentration of one very prototypical type of discourse, a higher degree of alignment with one particular type of discursive reasoning, etc.

In order to clear up this question as to the length of the texts and to gauge the overall size of the corpus, Table 2, below, provides figures that permit a more complete interpretation. It is worth pointing out that this table, to maintain the progressive order from lowest to highest, has changed the positions of Industrial Chemistry and Construction Engineering. Whereas Industrial Chemistry had a smaller number of texts, Construction Engineering now has considerably fewer number of total words.

<table>
<thead>
<tr>
<th>Scientific Field</th>
<th>University degree program</th>
<th>Number of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic Sciences and Engineering</td>
<td>Construction Engineering</td>
<td>14,653,760</td>
</tr>
<tr>
<td></td>
<td>Industrial Chemistry</td>
<td>22,103,620</td>
</tr>
<tr>
<td>Social and Human Sciences</td>
<td>Social Work</td>
<td>29,394,904</td>
</tr>
<tr>
<td></td>
<td>Psychology</td>
<td>30,136,590</td>
</tr>
<tr>
<td><strong>Total texts</strong></td>
<td></td>
<td><strong>96,288,874</strong></td>
</tr>
</tbody>
</table>

Table 2. Academic Corpus of PUCV-2006: number of words.
These figures constitute a true milestone. There is no other register of a written academic corpus available in the Spanish language of such dimension that is so representative and so thematically confined and focused. A corpus of such size, close to a 100 million words, in digital format, organized by subject matter and typology, becomes a fundamental tool for cutting edge research in corpus linguistics in Spanish. Adding the Academic Corpus PUCV-2006 to those already existing at our www.elgrial.cl website greatly enlarges and enhances the amount of available diversified material for enquiries and investigations.

Returning to the data, specifically, the number of texts relative to the number of words contained in the texts, the figures presented in Table 1 show a considerable disparity in the total number of texts deriving from the respective areas. For example, there are nearly four times as many Psychology texts compared to Industrial Chemistry texts. However, this apparent disparity is significantly reduced in terms of the number of words, where the total words in Industrial Chemistry texts, as shown in Table 2, are nearly 75% of those in Psychology. The greatest difference in terms of total words is between Construction Engineering and Psychology, where Psychology contains a bit over twice the number of words of Construction Engineering. In any case, whether in terms of the number of texts or the number of words, there is clearly a greater representation of reading materials from the Social and Human Sciences. These figures should undergo a more thorough analysis as more data and information becomes available. The remainder of this article will attempt, in part, to begin this process.

Precisely with this objective of carrying out a more in-depth analysis of the written material collected, a study and classification as to the types of texts will be examined. To do this, communicative-functional and textual-discursive linguistic criteria will be presented. By and large the principles of a complex taxonomy will be applied, as proposed by Parodi and Gramajo (2003). Below, in Table 3, nine types of texts are identified along with figures of their frequency of occurrence.

<table>
<thead>
<tr>
<th>Corpus Typology PUCV-2006</th>
<th>Nº of texts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conferences</td>
<td>1</td>
</tr>
<tr>
<td>Didactic Guideline</td>
<td>58</td>
</tr>
<tr>
<td>Dictionaries</td>
<td>2</td>
</tr>
<tr>
<td>Disciplinary Texts</td>
<td>266</td>
</tr>
<tr>
<td>Manuals</td>
<td>259</td>
</tr>
<tr>
<td>Norms</td>
<td>16</td>
</tr>
<tr>
<td>Reports</td>
<td>19</td>
</tr>
<tr>
<td>Research Articles</td>
<td>29</td>
</tr>
<tr>
<td>Tests</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>652</strong></td>
</tr>
</tbody>
</table>

Table 3. Distribution by type of text

The organization of the information in Table 3 follows the alphabetical order for the names applied to each type of text. Simple, everyday names were selected, names that are easily accessible and clear in their usage. As one can observe, a quite heterogeneous panorama with clear concentrations emerges from this table. Tentatively, one can say that two types of texts are by far the most frequent: Manuals (M), with 259, and Disciplinary Texts (DT), with 266. This provides and overall initial situation that combines, on the one hand, disciplinary knowledge proper as such, DT, which provide specialized knowledge, sometimes
with a high degree of complexity, and Manuals, on the other hand, which, although oriented towards disciplinary knowledge, have a didactic character that generally uses more educational resources, such as graphs, tables, diagrams, etc., in a more systematic manner and that incorporate exercises and other practical applications in order for readers to develop and test their knowledge. When observing the remaining figures, one’s attention is drawn to the fact that there is sparse occurrence of the other types of texts. For example, Didactic Guideline (DG) which, although they are the third largest type numerically, seem to be under-represented, given that this material was collected from four degree programs, each with duration of four to five years. Perhaps even more surprising is the low representation of Research Article (RA). One would think that this means of transmitting specialized knowledge would occupy a more prominent position, particularly in Basic Sciences and Engineering. So the occurrence of only 29 RA out of a total of 652 texts in the corpus would seem to indicate that they also are under-represented, especially when one considers that this is the most typical medium for disseminating the latest findings of academic and professional research. Thus, the types of texts of this corpus appear to be clearly concentrated in M and DG. These reveal the two points of the continuum of discourses, but not two extreme points. Both the M and the DT are oriented towards greater specialization, but with a clear tendency towards mainstream dissemination of educational information. The DG, which represent the extreme point of the gradient on the end concerned with education per se, do not appear in sufficient numbers to be extremely significant to the overall corpus.

Lastly, in Table 4, the figures for the corpus that is currently digitalized are provided, in other words, the total number of words of the corpus presented in Table 2, nearly 100 million words, represented the total corpus in paper format. Below is the figure for the corpus that has already been processed digitally up to this time.

<table>
<thead>
<tr>
<th>Corpus collected on paper</th>
<th>Percent of the Total Corpus</th>
<th>Number of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus collected on paper</td>
<td>100%</td>
<td>96,288,874</td>
</tr>
<tr>
<td>Digital Corpus</td>
<td>70%</td>
<td>67,402,212</td>
</tr>
</tbody>
</table>

Table 4. Digital Academic Corpus PUCV-2006: 70% of the total.

According to these figures, 70% of the corpus, over 67 million words, has been digitalized. This corresponds to a total of 457 texts out of the total of 652. This corpus can already be found in digital format (with morphosyntactic marking) and is available through website www.elgrial.cl. To identify it, use the abbreviation Corpus PUCV-2006, which will also allow you to access the different typological classifications.

5. Final comments

As pointed out in the Introduction of this paper and as observed in the preliminary results of this project in progress, both the scientific-investigative implications per se and the pedagogical implications seem to be multiple and varied. Having accurate data at hand as to the types of written discourse, both within the halls of academia and the professional workplace, as well as an accurate characterization of their types and traits will be an significant contribution to the field.

With regard to the preliminary analysis of the constitution of the academic corpus PUCV-2006, the data presented in the areas of the Social and Human Sciences relative to the Basic Sciences and Engineering reveal differences both in the number of written texts in use as a means of transmission of disciplinary knowledge and in the variety of texts that are disseminated as specialized knowledge. Thus, it is evident that Psychology and Social Work
tend to employ a greater quantity of texts with a relatively more extensive range (at least in the number of words) during the course of their degree programs, when compared to Industrial Chemistry and Construction Engineering, which use a smaller number of texts and more limited range of texts in terms of the number of words.

As regards the types of texts that were identified, one observes an interesting heterogeneity, which would support the validity of the idea of a continuum of types of texts between poles of higher disciplinary complexity and greater concern for general educational teaching. In this sense, two textual types are clearly identifiable: Manuals and Disciplinary Texts. Both show special attention to disciplinary knowledge in a relatively specialized format.

References


