ASp
la revue du GERAS

76 | 2019
Varia

Retrieving the specialised substance from a corpus of professional discourse in the field of Records and Information Management

Extraction de la matière spécialisée d'un corpus de discours professionnels dans le champ de la gestion de l'information et de la documentation

Philippe Millot

OpenEdition Journals

Electronic version
URL: http://journals.openedition.org/asp/6031
DOI: 10.4000/asp.6031
ISSN: 2108-6354

Publisher
Groupe d'étude et de recherche en anglais de spécialité

Printed version
Date of publication: 1 November 2019
Number of pages: 49-70
ISSN: 1246-8185

Electronic reference
Philippe Millot, « Retrieving the specialised substance from a corpus of professional discourse in the field of Records and Information Management », *ASp* [Online], 76 | 2019, Online since 01 November 2020, connection on 01 November 2020. URL : http://journals.openedition.org/asp/6031 ; DOI : https://doi.org/10.4000/asp.6031

Retrieving the specialised substance from a corpus of professional discourse in the field of Records and Information Management

Extraction de la matière spécialisée d'un corpus de discours professionnels dans le champ de la gestion de l'information et de la documentation

Philippe Millot

Introduction

1 English for Specific Purposes (ESP) in France and abroad has focused on the research and practice in broad academic domains such as law, medicine, economics, and business to name only a few. However, university specialisation and professionalisation, especially at master's degree level, has intensified over the last decades. The phenomenon has led to the growing need for teaching English in small specialised domains and an equally growing need for the description of the language and the discourses which may sometimes be unique to a domain. Although responding to those needs may be considered as a costly endeavour since much of the effort is concentrated on a very small proportion of learners, it also presents benefits such as offering a rewarding learner experience and the expansion of ESP frontiers. This article is a contribution to this sort of endeavour: we explore the "aboutness" (Scott & Tribble 2006) of a professional domain which is rather unknown to most ESP scholars in order to inform teaching and research through a corpus-driven methodology which may potentially be used for other domains.

2 Records and Information Management (RIM) may certainly be considered as a small, specialised domain whose discourses and culture have largely been left unattended by ESP scholars, thus resulting in a lack of knowledge and conception of what "RIM

English" actually covers. Like many specialised domains, however, RIM has a vivid culture[1] and a long history reaching as far back as the librarians' and archivists' early attempts at systematically organising knowledge in the late 19th century and early 20th through leading figures such as Melvil Dewey in America or Paul Otlet in Europe (Buckland & Lee 1995; Cox 2000; Millot 2018). This history and culture has generated domain-bound discourses which could potentially be transmitted by ESP practitioners to RIM students, either "for their own sake", or as general frameworks from which pedagogical activities could be designed.

3 But how can RIM discourse be observed and characterised? Four different approaches may be considered. Ethnomethodology, including conversation analysis, may certainly be considered as the historical approach to professional discourse analysis with seminal studies by Sacks and Garfinkel (1970) who created conceptual tools for understanding the structure of conversations in workplace contexts. The second approach may be termed the "business or organisational discourse approach" which developed widely in the late 1980s and early 1990s with the publication of pioneering work such as Bhatia's model of professional genres (Bhatia 1993) or Yates and Orlikowski's characterisation of communicative practices in large organisations (Yates & Orlikowski 1992). A third approach may be termed the "informational approach" adopted by terminologists and special or technical librarians who developed tools such as terminological databases and thesauri for organising professional knowledge within the corporations themselves (Mees 1920). Dupont de Nemours is a case in point where very large amounts of technical information (including technical reports, patents and thesauri) have systematically been organised ever since the 1920s (Kvalnes 1999). The final approach for analysing professional discourse developed in the 1990s and early 2000s may be termed the "corpus approach" in which data and metadata are collected in order to represent and account for professional discourse. Although the theories and the concepts differ sometimes markedly from one approach to another (e.g., the "qualitative" ethnomethodological approach versus the "quantitative" corpus studies), they are generally considered complementary and, in a way, they contribute to the making of a general theory of professional discourse.

4 With the development of online information and corpus analysis tools, accessing and dealing with specialised knowledge is no longer a quantitative problem since many resources (specialised journals and magazines, job advertisements, etc.) are now accessible, convertible to all kinds of digital formats, and may be processed by the ESP specialist. However, qualitative aspects of corpus design in terms of text choice, text categories, and analytical objectives are certainly more challenging in that they require knowledge of the targeted domain to a certain extent. Other factors may also influence corpus design such as the general rationale (will the corpus be ultimately used for particular teaching purposes or is corpus development based on purely scientific ones?), the institutional context in which the corpus is developed (a university or a private publishing house) and the financial resources made available for maintaining the corpus over the years. These factors are critical: most small, specialised corpora are more or less isolated objects (i.e., they are no longer used or made available), which considerably reduces the chances of developing standards.

2. Professional discourse analysis and the corpus

2.1. The specialised nature of professional discourse

5 One of the most obvious reasons why professional discourse may be considered as a deeply specialised one is the well-documented fact that the legitimacy and the credibility of professionals is largely based on the acquisition and the certification of domain-bound knowledge. The specialised character of professional discourse has also received ample developments in fairly recent publications in the field of discourse analysis (Gillaerts & Gotti 2005; Flowerdew & Gotti 2006; Gunnarsson 2009; Gillaerts & *et al.* 2012; Bhatia & Bremner 2017) especially with Bhatia (2002) whose model of professional expertise is based on disciplinary knowledge. Since the early 2000s, the centrality of specialised domains in professional discourse has become a regular characteristic of what counts as professional discourse, especially in the many cases in which English is used as a professional lingua franca. In these cases, specialised domains do not provide mere contextual data but form the data themselves. These data, which are sometimes referred to as "technical vocabulary" (Nickerson 2000), "specific vocabulary", "concepts and terminology" (Kankaanranta & Planken 2010), contribute to the making of a shared professional culture which facilitates international communication, especially when some participants "do not know English so well" (*op. cit.* 391). The fact that professional discourse is intrinsically specialised implies that data should, at least to some extent, comprise the discourses that allow the transmission of that form of knowledge whether formally (e.g., scholarly literature, professional press, and so on) or informally (e.g., interactions within the accomplishment of professional duties).

6 A very large body of the literature on professional discourse actually pays little attention to the specialised dimension of professional discourse. Most studies in this area stem either from interactional studies or from genre analysis where the data are essentially defined according to their structure within a general conversational analysis framework on the one hand, and, on the other, by the social function of the text in specialised communicative situations. Although these approaches have undeniably contributed to our understanding of professional discourse, they tend to relegate the specialised domains involved in the professions to the realms of secondary characteristics, the primary foci being the interactional characteristics (i.e. how the participants use discourse to interact with each other) and the intertextual ones (i.e. how the text interacts with other texts or text types).

2.2. Datasets and corpora in professional discourse analysis

7 The data collected for the study of professional discourse cover an extremely broad category of text types ranging from the most formal ones conveying official information about the organisation (e.g., annual reports or corporate websites) to the most informal ones conveying information about what happens backstage (e.g., meetings, conversations or email communication). In the early studies of professional discourse (Sacks 1972; Drew & Heritage 1992), the data were collected as ethnographic records which were generally described qualitatively, the data being systematically considered within the context in which they were produced thus allowing fine-grained interpretations. This qualitative approach is still very much in use today but it is

typically complemented with a quantitative approach. In spite of a strong tradition for the qualitative study of professional discourse with sometimes very small datasets (Maier 1992), the latter approach has gained in popularity and has led to an increase in corpus sizes. Since the early and mid-1990s, corpus sizes, once only expressed by numbers of texts – Swales and Rogers studied 100 missions statements from 100 large US organisations (Swales & Rogers 1995); Gimenez's early study of email communication consisted of 61 messages (Gimenez 2000) – have gradually become expressed in numbers of words.

8 Corpus size is a clear indicator of the sort of professional data that the corpus contains. Corpora representing front stage activities (i.e. activities that are easily accessible such as CEO's letters to shareholders or annual reports) may reach several tens or hundreds of thousands of words as is the case with Corpus of Environmental Impact Assessment (Flowerdew 2008) or Cho and Yoon's earning calls corpus (Cho & Yoon 2013), both containing 250,000 words. Much larger corpora of frontstage discourse have been developed but they tend to be the exception rather than the norm. The Wolverhampton Corpus of Written Business English (Fuertes-Olivera 2007; Cheng 2017) comprises more than ten million words and contains a fairly wide variety of genres including product descriptions, company press releases, annual financial reports, business journals, academic research papers, political speeches and government reports, all dealing with business or professional topics. On the other hand, corpora representing backstage activities (i.e. activities that are not easily accessible such as business conversations or professional emails) are typically small. Handford and Matous' corpus of oral interactions on construction sites contains 12,000 words (Handford & Matous 2011). Gillaerts' corpus of organisational emails hardly exceeds 14,000 words (Gillaerts 2012). Koester's Corpus of American and British Office Talk (ABOT) contains 34,000 words (Koester 2010) which is very similar in size to Millot's corpus of professional emails which reaches 33,000 words (Millot 2017).

9 Many studies in professional discourse use corpus linguistics and its frameworks in order to investigate patterns of language in large collections of natural texts in professional contexts (Cheng 2017). Apart from a few in-depth studies exploring communication from various angles (Nickerson 2000) thus giving a global view of practices within a particular setting, corpora in professional communication studies are generally designed for exploring particular aspects of professional communication such as Someya's business letter corpus (1999). Although corpus-based or corpus-driven studies of professional genres represent the greatest number of them by far, other aspects of professional communication such as specialised, professional cultures (Millot 2017), politeness strategies, gender issues (Fuertes-Olivera 2007), business English as a lingua franca, multimodality, intertextuality, or the impact of new communication technologies within the organisation (Kessler 2010) are also considered. A growing number of studies also make use of corpus linguistics' analytical frameworks and tools. For example, in his study of the Business English Corpus, Nelson identified semantic patterns which are unique to business situations, some words having collocational patterns or semantic prosodies which systematically differ from "general English" (Nelson 2006). In his comparative study of accounting narratives in seven corporate settings based on word frequency analyses, Rutherford also identified special semantic patterns such as a tendency to opt for positive words even in settings where the business situation was technically rather negative (Rutherford 2005). Similar

results were found in financial disclosure genres by Camiciottoli who shows that connectives play a fundamental role in the achievement of rhetorical strategies emphasising success (Camiciottoli 2010). Corpus analytical tools such as concordance and collocation tools have also been used to explore intertextuality and interdiscursivity in legal cases through the identification of lexico-grammatical patterns spanning across legal genres and whose mastery by novice writers may be considered as the evidence of the acquisition of a professional competence (Candlin & Maley 1997).

2.3. Professional discourse and the concept of register

10 Tools used for corpus analysis have clearly revived register analysis whose foundations were laid by Halliday and Hasan in the 1970s (Halliday & Hasan 1976). Since then, particularly with Biber and Conrad's contributions in the field, the concept of register has been defined as "a variety associated with a particular situation of use (including particular communicative purposes)" (Biber & Conrad 2009: 6). As Anthony (2018) claims, it has naturally become popular in ESP circle not only because most register studies have focused on lexico-grammatical patterns but also because most language teachers wish to transmit appropriate lexico-grammatical knowledge to their students. However, as Halliday (1994) also claims, registers reach beyond lexico-grammar into the world of semantics:

> A register can be defined as the configuration of semantic resources that the member of a culture typically associates with a situation type. It is the meaning potential that is accessible in a given social context. Both the situation and register associated with it can be described to varying degrees of specificity; but the existence of register is a fact of everyday experience – speakers have no difficulty in recognizing the semantic options and combination of options that are 'at risk' under particular environmental conditions. (op.cit. 26)

11 It has been evidenced, more particularly in research on English as a lingua franca (Cogo & Dewey 2012), that professionals have no trouble recognizing those options. The specialists' ability to find their ways in the world of specialised meanings may be explained by the fact that they have received "the set of knowledge and/or know-how directly involved in implementation of the purpose of the domain" (Van der Yeught 2016: 59). As a consequence, the professionals who work in English tend to easily come to terms with their own register because, even if English is a foreign language, they are already familiar with what their professional world means. In other words, professionals are familiar with Halliday's "environmental conditions" which may be defined in this study as workplaces (e.g., companies or administrations) where professionals conduct activities – sometimes realised conventionally through genres – that call upon specialised domains as they were defined by Van der Yeught, that is, as "complex intentional universes" (*op. cit.* 51). However, those conditions are particularly complex in that they may be quite antagonistic even within the same organisation. In RIM for example, when an organisation complies with new data security standards, the information professional becomes in charge of implementing procedures that may conflict with how other staff conduct their own activities. This fact of life means that the specialised substance within a workplace is highly diverse, each workplace generating as many specialised registers as professional domains. Professional discourse in general may therefore be considered as a meta-register comprising a constellation of professional registers which are semantic configurations experienced

by professionals who draw upon specialised domains or specialised domain sets in order to carry out their day-to-day activities. In this framework, a professional register consists of the discourses which are experienced by specific types of professionals and corpora can be used to observe this discursive experience.

3. Method: Building a specialised corpus in RIM

3.1. Corpus structure

12 The corpus is designed so as to represent three planes of professional discourse, namely Theory, Mediation, and Practice (see Table 1). These planes represent fundamental functions in the profession, from the most theoretical ones to the most practical ones.

Table 1: The RIM Corpus structure (2019)

Categories	Text types	Nb of texts	Contents	Nb of words
Theory	Research articles	133	Articles from the *International Journal of Records Management* (2008-2016)	829,867
Mediation	Professional magazine articles	629	*Information Management* (2013-2016)	530,293
Practice	Job advertisements	40		25,798
	Professional norms	1	ISO 15489	8,901
		1	RIM Core Competencies	31,574
Total				1,426,433

13 Our corpus seeks to identify the "aboutness" (Scott & Tribble 2006) of Records and Information Management (RIM) whose general intention may be defined as the organisation of institutional knowledge (Greer *et al.* 2013). More precisely, the domain is defined by the Association of Records Managers and Administrators as "the field of management responsible for establishing and implementing policies, systems, and procedures to capture, create, access, distribute, use, store, secure, retrieve, and ensure disposition of an organisation's records and information" (ARMA 2016: 43). RIM professionals typically work in large organisations such as multinationals or public institutions. They are often considered as corporate or organisational librarians in that many records and information managers have received education and training in library science, archive science, documentation and, more generally, information science. The profession has therefore inherited from a great deal of scientific concepts, the most recent ones pertaining to the very broad field of digital information. Although the domain clearly benefits from a French-speaking school of thought with scholars such as Suzanne Briet or Paul Otlet, a very large part of that knowledge has been conceived in English.

14 RIM theory is disseminated by some journals, each journal having its own scientific or regional scope. Among them, the *Records Management Journal* appeared as a particularly interesting one for our research, since most articles deal with RIM in business or institutional contexts. The articles chosen here were published in the *Records Management Journal* between 2013 and 2016. As such, they bring insights into the contemporary, theoretical issues and concepts in records management. As is the case with many professions, theoretical knowledge is mediated through "professional magazines" where experts in the field present the readers with the latest trends and issues in the profession such as the economic and legal impacts of digital information in the corporate world, the need for compliance with international regulations (e.g. General Data Protection Regulation) or the development of new tools or concepts such as Cloud services. This sort of "mediated" knowledge corresponds to the "mediation" part of the corpus which is based on a set of 24 issues of *Information Management* (2013-2016),[2] a professional magazine published by ARMA. Finally, our corpus includes a set of "practical genres" which, as we defined above, directly contribute to the achievement of actions. Although the profession is codified, job titles tend to vary from one organisation to another, "information managers" being sometimes referred to as "records and information management specialists", "coding and records managers", or "electronic records archivists". The practical part of our corpus also includes the ISO norm on records management (15489) which is used for the practical implementation of information policies. We also included the description of records management core competencies. The document is published by ARMA and used by RIM professionals and human resources managers within the recruitment process. Although the "practical genres" are quite heterogeneous in terms of communicative purposes, the category enables us to accumulate texts that are used by professionals within the course of their practical experience of the domain.

3.2. Comparing RIM with general English

15 Retrieving the specialised substance from a corpus is a challenging task because the retrieval first depends on our conception of the "specialised" items and then on how the items are retrieved technically. Regarding the latter, concordance tools such as Antconc (Anthony 2014) can be used to identify lexical items such as words and phrases which are specifically frequent in a corpus. However, the conceptual universe of even small, specialised domains goes far beyond the lexical, surface-level domain and reaches the realms of genre, register, and semantics. Following our definition of specialised domains (see section above), some meanings may be specialised because they describe and organise the domain, as is the case for terms, or because they result from a specific organisation in the text (i.e. some words in a specialised genre may be particularly frequent in a specialised corpus because the genre in question has no equivalent in the general language). In addition to this, some items may have a low frequency in a specialised corpus although they are part of a statistically salient semantic field.

16 In order to identify the specificity of our corpus regarding general language, we conducted a double, data-driven analysis. First, we retrieved a list of broad semantic categories contained in the RIM corpus. The identification was conducted by using Wmatrix (Rayson 2008) which attributes semantic tags to all words in the corpus and then calculates which semantic tags are particularly salient in the target corpus (our

corpus) compared with a reference corpus. The latter corresponded to the written part of the British National Corpus Sampler which seeks to represent written British English. We selected the written register in order to avoid the bias resulting from the presence of oral components in the analysis.

17 As the introduction to the annotating system indicates,[3] the tags are based on the UCREL Semantic Analysis System (USAS) consisting of broad, general English categories which then develop into fine shades of semantic subcategories including synonyms, hypernyms and hyponyms. The system includes fields such as "money", "commerce in industry", "science and technology", "government and public" or "education" as well as broader ones such as "general and abstract terms" or "names and grammar". More importantly, the USAS system develops into a number of sub-fields some of which prove potentially useful for the study of specialised languages. For example, the "linguistic actions, states and processes" category includes the subcategory "communication" which then develops into even finer meanings which were fundamental starting points for our study (e.g. "paper documents and writing", "information technology and computing" or "general ethics"). We used these very broad categories as first filters for the identification of semantic fields which could potentially cover the RIM domain.

3.3. Beyond the comparison: Defining specialised semantic dimensions

18 Professional domains encapsulate many kinds of social relations and equally many genres and registers. For example, the medical domain includes different kinds of communicative practices such as scientific articles, conference talks, doctor-patient interactions or medical reports. Although these practices may be shared to some extent with other domains, what makes a specialised domain specific is probably its semantic configuration or, more philosophically, its "intentional universe". In the case of medicine, most stakeholders including medical staff, patients and anyone involved in medical activities (e.g., companies developing vaccines) have more or less similar interpretations of concepts such as "disease" and "diagnosis" because they are driven by a similar intention which is to cure or to relieve patients. Similarly, information managers and information scientists tend to have similar interpretation of "information", "document" or "record" because these notions have been conceptualised within the domain and only the insiders – those who have been trained or have experience – in this specialised domain can interpret these notions from a professional perspective.

19 In order to understand how RIM professionals interpret their own world, we started with the study of two types of published works written by the expert themselves. The first type may be considered as the "historical literature", that is, the works published by specialised historians. The second type may be considered as the professional norms and white papers which are published by professional associations such as the Association of Records Managers and Administrators in America or the Information and Records Management Society in Britain. More particularly, the "Records and Information Core Competencies" handbook published by the ARMA provides a fine-grained description of the profession divided into six domains, namely "business functions", "RIM/ Information governance practices", "risk management",

"communication and marketing", "information technology", and "leadership" (ARMA 2017: 2). Given the size of the corpus and the scope of this study, these categories were used to create three broad semantic dimensions, namely "organisational", "technical" and "promotional", which we consider as areas of meanings covering a potentially unlimited number of linguistic realisations.

20 Dimension 1 ("Organisational") covers parts of speech which describe business functions (e.g., supervising staff, budgeting, or providing customer service), business environments (companies, departments, etc.), and leadership (i.e., the words referring the motivation of people and groups in the achievement of RIM goals). Dimension 2 ("Technical") covers the field of both the systematic management of information (including sharing, using, accessing as well as those involved in compliance with information regulations), and information technology (i.e. the words used in the maintenance and development of information systems). The dimension also includes RIM job titles. Lastly, dimension 3 ("Promotional") deals with two kinds of semantic subcategories. The first one corresponds to the parts of speech referring to education and training in RIM issues (e.g. degrees, qualifications, and certification) as well as the discourse showing that RIM practitioners seek to reach higher levels in the profession. The latter aspect includes parts of speech where expert information is reported to other professionals (e.g. survey results, summary of findings). The second subcategory is based on the idea that RIM professionals are not mere technicians doing technical jobs in specific business contexts but they also promote their domain by adopting attitudes that will make the domain indispensable to stakeholders (e.g. positive or negative attitudinal markers, intensifiers or attenuators). The promotion of the domain may lead to "promotional" actions such as "compliance" (with particular information laws), "adoption" (of particular information schemes) or "regulation" (of information access). It may also lead to emphasise risks which may in turn justify the implementation of risk management procedures. The promotional dimension also includes institutions and organisations such as the Office of the Information and Privacy Commissioner for British Columbia which plays an active role in the promotion of information standards and norms.

21 We hypothesise that the "specialisedness" of RIM discourse lies in the combination of the three dimensions within short stretches of text. In order to test this hypothesis, we created a small corpus sample of 9,000 words (3,000 words taken from each part of the corpus) and tagged it manually for each dimension.

4. Results

4.1. Effects of specialisation on general semantic categories

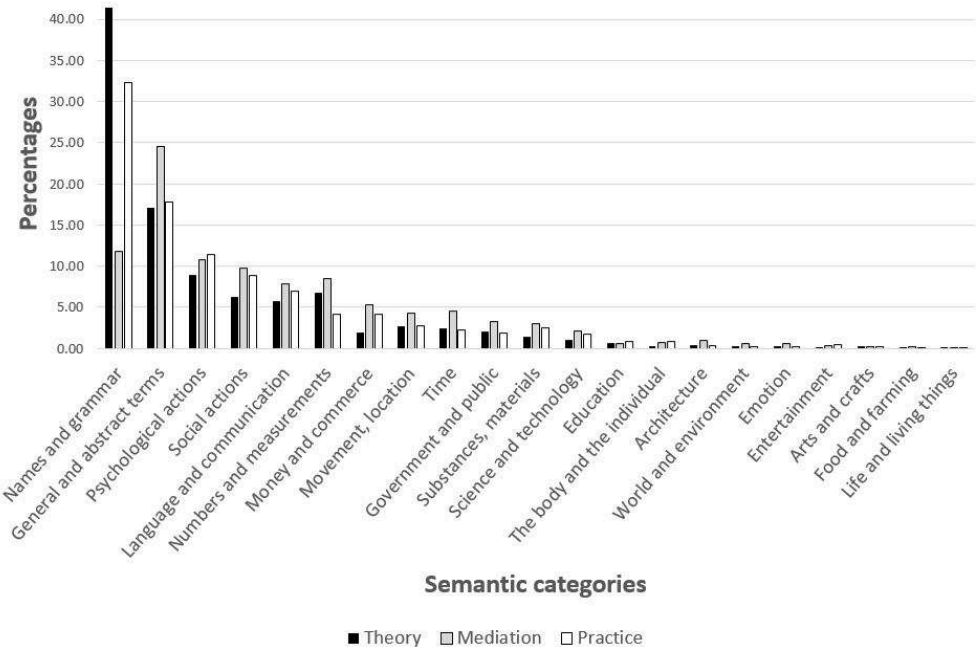
4.1.1. General distribution of semantic categories across the corpus

22 When the corpus was semantically tagged by Wmatrix, it appeared that all semantic categories were represented (see figure 1). Although "RIM English" is a specialised variety, it covers the entire semantic spectrum of general English with expected categories such as "science and technology" or "numbers and measurements" and less expected ones such as "food and farming" or "life and living things". Such a wide range of meanings finds its origin in the very nature of the domain. First, RIM is used in many kinds of contexts and therefore mobilises their semantic universes. Secondly, RIM

discourse is partly based on the use of metaphors (e.g., the cloud), which mechanically brings counterintuitive categories to the surface (e.g., the world and the environment).

23 However, some areas of meaning are clearly more salient than others. As figure 1 shows, the most frequent category is the "names and grammar" set which includes a very wide range of grammatical words as one might expect given the variety of genres – from scientific articles to job advertisements – contained in the corpus. The category is closely followed by the "general and abstract terms" tag set which yet again covers a broad category of meanings including those related to "making", "evaluating", "modifying" and "expressing emotions". The two most frequent categories are then followed by a heterogeneous set of further semantic categories including "psychological" and "social actions", "money and commerce", "science and technology" or "life and living things". It is interesting to note that these statistically lower categories were found in all corpus parts (theory, mediation, and practice).

Figure 1: Distribution of general semantic categories across the corpus



24 The statistical variations between "theory", "mediation", and "practice" illustrate the fact that genres (i.e. scientific articles, job advertisements, professional magazine articles) strongly orient lexico-grammatical choices. For example, the proportion of the "names and grammar" category in scientific articles and job advertisements is much higher than in professional magazines because scientific articles and job advertisements are partly based on references to authors in the former case and on references to company names and contacts in the latter.

4.1.2. Semantic categories hosting specialised items

25 All texts from the corpus, whether they are scientific articles, magazine articles or practical guidelines tend to deal with subjects which logically emphasise certain semantic categories. For example, the "general and abstract terms" category contains words referring to the state of "being" (e.g., "is", "are", "constitute", "exist"). However,

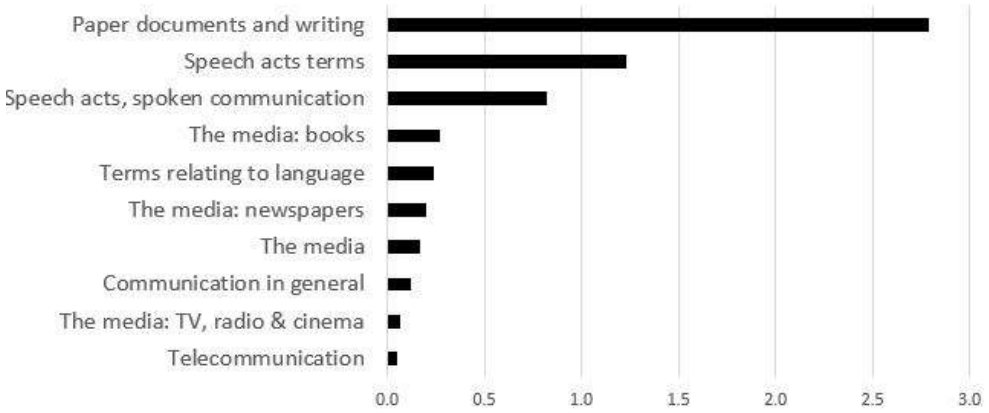
the same category also contains words referring to “getting and giving", which, in the RIM context, leads to the occurrence of technical terms. As figure 2 shows, the most frequent occurrences from the “getting and giving" category (tagged A9+ in the USAS category system) contain a great deal of RIM terms such as “retention” (n=1204), “collection” (n=168) or “retrieval” (n=118). They also contain frequent occurrences of words which are typically parts of specialised collocations referring to legal and business issues in the organisation of knowledge (e.g., “intellectual property rights”, “legal hold process in RIM policies”, “human resources”).[4]

Figure 2: “Term hosting” within a USAS semantic category. Sample of “getting and giving; possession" (A9+)

Word	Semtag	Frequency	Relative Frequency
retention	A9+	1204	0.24
have	A9+	907	0.18
storage	A9+	535	0.11
has	A9+	417	0.08
stored	A9+	322	0.07
resources	A9+	257	0.05
take	A9+	219	0.04
hold	A9+	173	0.04
accepted	A9+	169	0.03
collection	A9+	168	0.03
keep	A9+	162	0.03
get	A9+	147	0.03
had	A9+	135	0.03
maintain	A9+	131	0.03
retrieval	A9+	118	0.02
property	A9+	104	0.02

26 The “hosting” phenomenon may also be observed in other frequent categories. For example, in the “social actions” category, general English words such as “group” or “public” co-occur with a wide range of business-related terms such as “organization”, “corporate”, or “team”. In a similar way, the “psychological actions” category includes general words such as “think” or “consider” but also comprises concepts which are particularly salient in RIM such as “information”, “knowledge”, or “data”. Again, term hosting may be observed in the “communication” category in which the “paper documents and writing” part highly exceeds the others because it contains terms such as “documents” or “records” which are naturally central in the RIM domain (see figure 3). It is also worth noticing that the hosting phenomenon also occurs in less frequent and rather unexpected categories such as “emotions” in which general words such as “peace” or “calm” occur alongside metaphorical expressions referring to risk management situations such as “threat” (e.g., “security threats”) or “malicious” (e.g., “malicious code infections”). Therefore, the hosting of specialised items may be considered as a pervasive process affecting all kinds of semantic categories from the most expected ones to less expected areas of meaning. It therefore seems that although the USAS categories were originally conceived for dealing with “general English”, they can be used to retrieve a part of the specialised substance from the corpus.

Figure 3: Distribution of semantic tags within the “communication” category

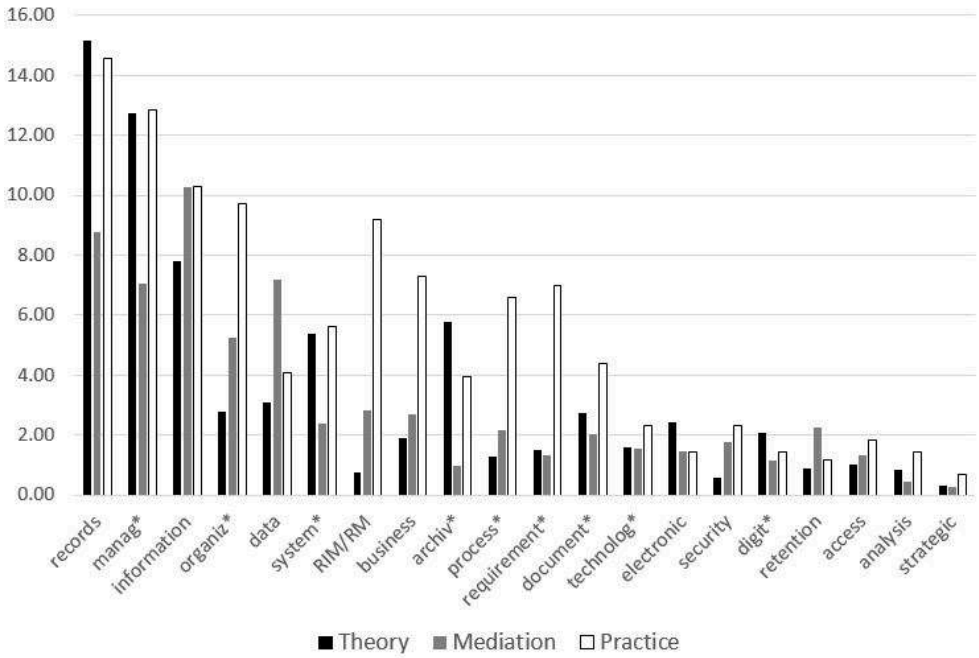


4.2. Thematic coherence

27 The section above may be considered as a first test assessing how specific our corpus is regarding general semantic categories. But we suggest that what makes a register specialised, thereby a specialised corpus, is not the mere orientation towards the intentions of the community, but also the fact that these specific meanings are realised consistently across the various parts of the register. In order to test the coherence, we conducted a second test by looking at whether all parts of the corpus showed similar keywords when they were compared with the BNC written corpus sample in Wmatrix. The procedure resulted in a list of several thousands of statistical keywords from which we selected the 500 most frequent cases from each part. We therefore obtained a list of 1,500 top keywords which we analysed in order to identify the ones that were common to all parts of the corpus. The method finally enabled us to identify a short list of twenty common words, which we considered as indicators of thematic coherence.

28 Figure 4 shows those twenty keywords and their frequency across the three subregisters. As one might expect, the list highlights the terms that are central in RIM such as “records”, “documents”, “management”, “information” or “data”. In a way, the terms describe the core concepts in the domain and this finding clearly illustrates the fact that RIM is essentially concerned with the organisation of information. The list also describes the contexts in which information is organised (e.g. business, technology) as well as scientists’ and professionals’ major priorities such as information access or data security.

Figure 4: Salient keywords shared by three subregisters (per thousand words)



29 These results also indicate sometimes strong variation patterns between subregisters. These patterns can be explained by the fact that each subregister is characterised by genres that highlight certain aspects of the professional domain. For example, the word "requirement" is particularly salient in the "practice" subregister because the word is almost systematically used in job advertisements and appears very regularly in the ISO norm. Another example is the term "record" which is much less frequent in the "mediation" subregister because professional magazine articles, which constitute this part of the corpus, deal with a greater variety of subjects than the scientific articles from the *Records Management Journal* or the genres from the "practice" subregister which tend to concentrate on the concept.

30 Broadening the scope of keyword observation to their collocates within a fairly broad span (3L, 3R)[5] enabled us to evidence how situated RIM discourse was. Our study of keyword collocations indeed highlights the existence of a specific professional universe at the crossroads of organizational science, technology, and information science (see table 2). The collocates lying outside this universe were actually very few and concerned very general notions such as "world" (e.g., "the world of professionals") or "understand" which was frequently found in the job advertisements where "understanding" certain concepts was presented as a pre-requisite for the jobs.

Table 2: The most frequent keyword collocates in the three subregisters (theory, mediation, and practice)

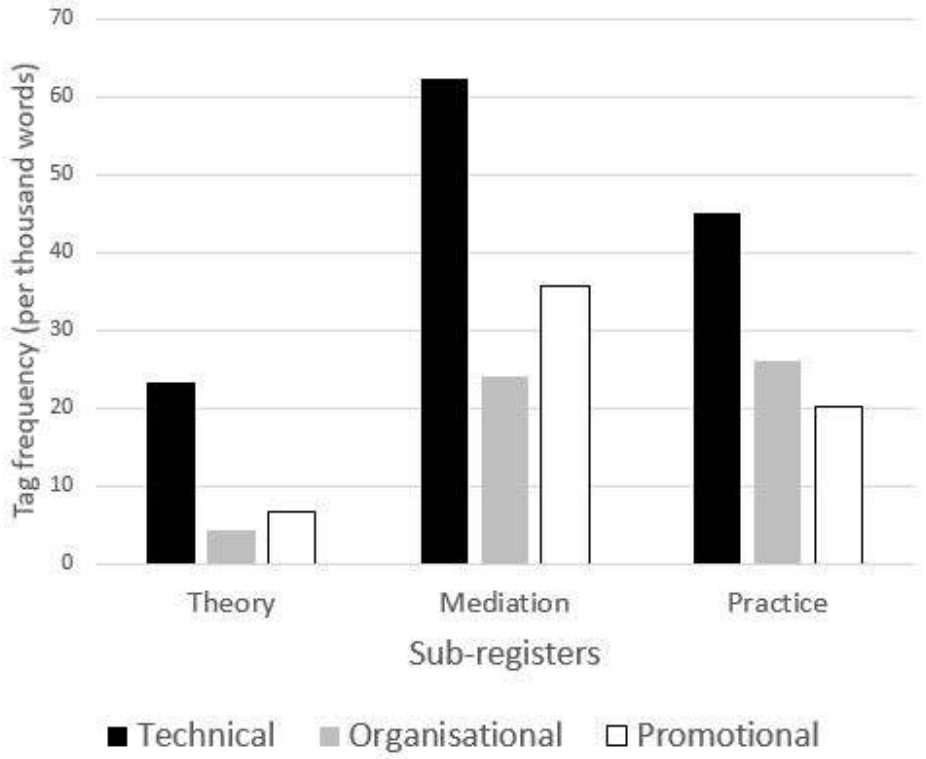
Keywords	Shared collocates
Business	Activities, environment, information, IT, legal, management, needs, operations, process, records, support, technology, unit
Data	Access, analysis, collect, electronic, information, management, processing, protection, quality, records, related, security, storage, systems, use

Information	Access, business, data, electronic, governance, ISO, manage, management, organization, organizational, personal, privacy, professionals, protection, records, security, system, technology, use
Management	Business, change, content, data, document, electronic, information, knowledge, practices, records, risks, senior, system
Organiz*	Business, information, management, needs, policies, process, records, use
Records	Access, archives, business, electronic, information, manag*, paper, part, policy, related, requirements, retention, system
System	Data, document, electronic, information, manag*, process, records, software, use

4.3. Building specialised dimensions for the study of professional domains

31 The previous section has shown that general semantic categories may be used to retrieve a part of the specialised substance from the corpus. However, the method was an indirect one because the USAS categories were not originally conceived for the study of specialised varieties. Moreover, they say rather little about how the specialised meanings are actually structured. We therefore applied the three semantic dimensions presented in section 3.3 ("technical", "organisational" and "promotional") to a sample of our corpus. Figure 5 presents how the three dimensions are distributed across this sample. It shows that the three dimensions co-occur in all parts of the corpus sample, which confirms that these areas of meaning are probably the most relevant for the study of RIM. The three subregisters present, although in varying degrees, some organisational, technical, and promotional meanings which tends to confirm that the three dimensions regularly co-occur across the register.

Figure 5: Distribution of semantic dimensions across three subregisters of RIM



32 The examples below show that the dimensions co-occur even within short spans of text. They were taken from the professional magazine *Information Management*. Dimension 1 (organisational) is italicised. Dimension 2 (technical) is written in bold. Dimension 3 (promotional) is underlined.

> 1. Locate all **inactive *procurement* records** in the **storage repository** and in the *clerks' offices*
> 2. **Transfer the records for preservation into RIM repositories** and **conduct physical disposal of the records** whose **retention periods had expired**.
> 3. Clarify the **future retrieval process** in the **archival database**.
> 4. Train the *procurement employees* on the **practical RIM processes**.
> (Schmerbauch 2016: 37)

33 The dimensions also occur with a fairly high degree of density in the Practice subregisters as in the professional norm extract presented below.

> 1. The objective in issuing and implementing policies on **managing records** should be **creation, capture and management of authentic, reliable and useable records that possess integrity and support** and enable *business activity* for as long as they are required. [ISO 2016 : 8]
> 2. **Records professionals** or others responsible for **managing records** are responsible for **developing, implementing and maintaining metadata schemas and other controls**, in association with other *personnel*, such as *information technology professionals, business managers and legal professionals*. [ISO 2016: 9]
> 3. The training program should be ongoing and include training on **requirements, policies, practices, roles and responsibilities for managing records**, and should be addressed to all *members of management and personnel, as well as any other individuals responsible for any part of business activity* involving the **creation, capture and management of records**. [ISO 2016: 10]

34 Although scientific articles are generally presented as highly specialised texts, the Theory subregister contains comparatively fewer dimensions than the two other

categories. The relatively lower frequency can be explained by the fact that a great deal of the article genre actually contains much "academic material" such as long passages in which concepts are discussed academically and are not necessarily related to organisational matters or to promoting the profession. This observation leads us to say that the Mediation and the Practice subregisters are particularly relevant objects for observing RIM as a professional variety: not only do the three types of meanings co-occur very frequently, but, as the examples above indicate, they also mirror each other, especially at lexical, terminological, and phraseological levels. The mirroring effect between Mediation and Practice may in turn be explained by the fact that many texts from the professional press serve as user guides for implementing norms or for presenting business situations which involve professional profiles that are typically found in the job ad section of our corpus.

5. Conclusion

35 This article sought to gain insights into the discourses of the Records and Information Management professional culture and, more generally, refine our conception of what professional discourse implies in terms of corpus approach. As is the case for most studies in specialised varieties, the scope of the results is obviously limited by the size of the corpus and the fact that the represented genres are equally limited. However, the corpus has been structured in such a way that it may potentially include other text types whether written, spoken or hybrid.

36 Among the salient findings, the structure of our corpus has proved a very efficient way of organising the different discourse types circulating in the RIM domain. Three discourse types have been identified and may be considered as three planes of professional discourse in this field (see figure 6). Although they contain a rather limited number of texts and sources, we hypothesise that our corpus structure could be transferred to other professional domains since all professional domains generate practical activities, theoretical knowledge, and mediated forms of communication. The first category (Practice) covers a broad range of textual and discursive practices which may potentially include many other genres such as emails, telephone calls, meetings, job advertisements, financial reports and mission statements. These practices are defined by their communicative purposes and they directly contribute to the achievement of professional actions. The second category (Theories and Concepts) includes the areas where theories and concepts are discussed and are typically disseminated through journals. The third category (Mediation), includes the texts – written or spoken – in which professionals mediate expert knowledge. In this study, the category focused on professional magazines but many other types of practical literature may be considered such as professional blogs or micro-blogs on Twitter or on LinkedIn in which professionals offer guidelines and solutions to daily problems.

37 Among the other findings, our study shows that general semantic categories can help us retrieve a part of the specialised substance from the corpus. The categories precipitate at the contact of the specific context and yield what may be considered as specialised matter. In this professional context, the matter is the salience of subcategories that bear central meanings in the domain. Although these findings are interesting in themselves, they may also be misleading with specialised language varieties appearing as mere reactions to the context. General semantic categories also

say rather little about how the specialised variety is semantically organised. We therefore went back to the domain and created ad hoc conceptual tools based on how the professionals conceived their own specialised world. This world has been defined along three dimensions, namely "organisational", "technical", and "promotional" whose systematic combination in all parts of a random sample was considered as a strong defining criterion for characterising professional discourse. In our case, what makes RIM specialised is not the mere presence of "technical" vocabulary or that of extremely specific terms. Rather, the specialised character of the RIM register lies in the fact that these technical notions are systematically used in organisational contexts and are regularly promoted in order to serve intentions of the professional community.

Figure 6: Structure of specialised registers in professional discourse

38 Although these dimensions are clearly relevant, they remain very challenging when it comes to tagging the corpus. As we explained above, the tagging process was achieved manually by attributing a tag of the XML type to each part of the text that bore one or several semantic dimensions. Although this kind of tagging is very flexible and enables the identification of many features, it poses the problem of identifying where the specialised dimension starts in the clause and where it stops. As a consequence, approaching the dimensions quantitatively remains so far limited to comparing subregisters between them and, more importantly, to characterising the linguistic material they contain. As we have seen with the various examples presented above, technical and organisational material can be easily identified because it is essentially based on the lexicon that refers to "core competencies" or the words that describe the business world and its structure. The same, however, does not apply for the "promotional" dimension which deserves further developments. Indeed, the way professionals promote their domain ranges from very straightforward strategies (e.g., referring to training programs) to more subtle ones in which professionals legitimise

their work within the institution and society at large. These strategies include emphasising risks due to a lack of information governance in the company or showing the benefits of having one thus implying that information professionals should be hired. Some perspectives for this part of the study therefore include the characterisation of each dimension in the aim of retrieving the specialised substance more systematically.

BIBLIOGRAPHY

ANTHONY, Laurence. 2014. *Antconc 3.4. 3*. Tokyo: Waseda University.

ANTHONY, Laurence. 2018. *Introducing English for Specific Purposes*. London and New York: Routledge.

ASSOCIATION OF RECORDS MANAGERS AND ADMINISTRATORS (ARMA). 2016. *Glossary of Records Management and Information Governance. (5th edn.)*. Overland Park: Arma International.

ASSOCIATION OF RECORDS MANAGERS AND ADMINISTRATORS (ARMA). 2017. *Records and Information Management Core Competencies (2nd edn.)*. Overland Park: Arma International.

BHATIA, Vijay. 1993. Analysing Genre: Language Use in Professional Settings. New York: Routledge.

BHATIA, Vijay. 2002. "Professional discourse: Towards a multi-dimensional approach and shared practice". In CANDLIN, C. N. (Ed.), *Research and Practice in Professional Discourse*. Hong Kong: City University of Hong Kong Press, 39–60.

BHATIA, Vijay. 2017. "Analysing discourse variation in professional contexts". In BHATIA, V. & S. BREMNER (Eds.), *The Routledge Handbook of Language and Professional Communication*. New York: Routledge, 3–12.

BHATIA, Vijay & Stephen BREMNER (Eds.). 2017. *The Routledge Handbook of Language and Professional Communication.* New York: Routledge.

BIBER, Douglas & Susan CONRAD. 2009. *Register, Genre and Style*. Cambridge: Cambridge University Press.

BUCKLAND, Michael K. & Ziming LEE. 1995. "History of information science". *Annual Review of Information Science and Technology* 30, 385–416.

CAMICIOTTOLI, Belinda C. 2010. "Discourse connectives in genres of financial disclosure: Earnings presentations vs. earnings releases". *Journal of Pragmatics* 42/3, 650–663. <doi.org/10.1016/j.pragma.2009.07.007>.

CANDLIN, Christopher N. & Yves MALEY. 1997. "Intertextuality and interdiscursivity in the discourse of alternative dispute resolution". In GUNNARSSON, B.-L., P. LINELL & B. NORDBERG (Eds.), *The Construction of Professional Discourse*. London and New York: Longman, 201–222.

CHENG, Winnie. 2017. "Corpus analyses of professional discourse". In BHATIA, V. and S. BREMMER (Eds), *The Routledge Handbook of Language and Professional Communication*. New York: Routledge, 13–25.

CHO, Hyeyoung & Hyunsook YOON. 2013. "A corpus-assisted comparative genre analysis of corporate earnings calls between Korean and native English speakers". *English for Specific Purposes* 32/3, 170–185. <doi.org/10.1016/j.esp.2013.03.001>.

COGO, Alessia & Martin DEWEY. 2012. *Analysing English as a Lingua Franca: A Corpus-Driven Investigation*. London and New York: Continuum.

COX, Richard J. 2000. Closing an Era. Historical Perspectives on Modern Archives and Records Management. Westport, CT: Greenwood Press.

DREW, Paul & John HERITAGE. 1992. "Analyzing talk at work: An introduction". In DREW, P. & J. HERITAGE (Eds.), *Talk at Work: Interactions in Institutional Settings*. Cambridge: Cambridge University Press, 3–65.

FLOWERDEW, John & Maurizio GOTTI (Eds.). 2006. *Studies in Specialized Discourse*. Bern: Peter Lang.

FLOWERDEW, Lynne. 2008. "Corpora and context in professional writing". In Flowerdew, J., V. BHATIA & R. H. JONES (Eds.), *Advances in Discourse Studies*. London and New York: Routledge, 115–127.

FUERTES-OLIVERA, Pedro. A. 2007. "A corpus-based view of lexical gender in written business English". *English for Specific Purposes* 26/2, 219–234. <doi.org/10.1016/j.esp.2006.07.001>.

GILLAERTS, Paul. 2012. "Email use in a Belgian company: Looking for the hybridity of the genre". In GILLAERTS, P. *et al*. *Researching Discourse in Business Genres: Cases and Corpora*. Peter Lang, linguistic Insights, 15–32.

GILLAERTS, Paul & Maurizio GOTTI (Eds.). 2005. *Genre Variation in Business Letters*. Bern: Peter Lang.

GILLAERTS, Paul, Elizabeth DE GROOT, Sylvain DIELTJENS, Prescilla HEYNDERICKX & Geert JACOBS (Eds.). 2012. *Researching Discourse in Business Genres: Cases and Corpora*. Bern: Peter Lang.

GIMENEZ, Julio. 2000. "Business e-mail communication: Some emerging tendencies in register". *English for Specific Purposes* 19/3, 237–251. <doi.org/10.1016/S0889-4906(98)00030-1>.

GREER, Roger C., Robert J. GROVER & Susan G. FOWLER. 2013. *Introduction to Library and Information Science*. Exeter: Libraries Unlimited.

GUNNARSSON, Britt-Louise. 2009. *Professional Discourse*. London and New York: Continuum.

HALLIDAY, Michael. 1994. "Language as social semiotic". In MAYBIN, J. (Ed.), *Language and Literacy in Social Practice*. Clevedon: The Open University, 23–43.

HALLIDAY, Michael & Ruqaiya HASAN. 1976. *Cohesion in English*. London and New York: Routledge.

HANDFORD, Michael & Petr MATOUS. 2011. "Lexicogrammar in the international construction industry: A corpus-based case study of Japanese–Hong-Kongese on-site interactions in English". *English for Specific Purposes* 30/2, 87–100. <doi.org/10.1016/j.esp.2010.12.002>.

ISO. 2016. Information and Documentation – Records management: Concepts and Principles. Geneva: International Organization for Standardization.

KANKAANRANTA, Anna & Brigitte PLANKEN. 2010. "Belf competence as business knowledge of internationally operating business professionals". *The Journal of Business Communication* 47/4, 380–407. <doi.org/10.1177/002194361037730>.

KESSLER, Greg. 2010. "Virtual business: An Enron email corpus study". *Journal of Pragmatics* 42/1, 262–270. <doi.org/10.1016/j.pragma.2009.05.015>.

KOESTER, Almut. 2006. *Investigating workplace discourse*. London and New York: Routledge.

KOESTER, A. 2010. "Building small specialised corpora". In O'KEEFE, Anne & Michael. MCCARTHY (Eds.), *The Routledge Handbook of Corpus Linguistics* (volume 1). London and New York: Routledge, 66–79.

KVALNES, Florence H. 1999. "The history of managing technical information at Dupont". *Science Information Systems*, 107–114.

MAIER, Paula. 1992. "Politeness strategies in business letters by native and non-native English speakers". *English for Specific Purposes* 11/3, 189–205. <doi.org/10.1016/S0889-4906(05)80009-2>.

MCENERY, Tony & Andrew. HARDIE. 2011. *Corpus Linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

MEES, Kenneth. 1920. *The Organization of Industrial Scientific Research*. New York: McGraw-Hill Book Company, Incorporated.

MILLOT, Philippe. 2017. "Inclusivity and exclusivity in English as a Business Lingua Franca: The expression of a professional voice in email communication". *English for Specific Purposes* 46, 59–71. <doi.org/10.1016/j.esp.2016.12.001>.

MILLOT, Philippe. 2018. « L'évolution terminologique des sciences de l'information: Étude sur corpus diachronique d'un nouveau domaine spécialisé ». *Textes et Contextes* 13-2. <http://preo.u-bourgogne.fr/textesetcontextes/index.php?id=2318>.

NELSON, Mike. 2006. "Semantic associations in business English: A corpus-based analysis". *English for Specific Purposes* 25/2, 217–234. doi.org/10.1016/j.esp.2005.02.008>.

NICKERSON, Catherine. 2000. Playing the Corporate Language Game: An Investigation of the genres and discourse strategies in English used by Dutch writers working in multinational corporations. Amsterdam: Rodopi.

RAYSON, Paul. 2008. "From key words to key semantic domains". *International Journal of Corpus Linguistics* 13/4, 519–549. <doi.org/10.1075/ijcl.13.4.06ray>.

RUTHERFORD, Brian. A. 2005. "Genre analysis of corporate annual report narratives: A corpus linguistics-based approach". *The Journal of Business Communication* 42/4, 349–378. <doi.org/10.1177/0021943605279244>.

SACKS, Harvey. 1972. "An initial investigation of the usability of conversational data for doing sociology". In SUDNOW, D. (Ed.), *Studies in Social Interaction*. The Free Press, 31–73.

SACKS, Harvey & Harold GARFINKEL. 1970. "On formal structures of practical actions". In MCKINNEY, J. C. & E. A. TIRYAKIAN (Eds), *Theoretical Sociology*. New York: Appleton Century Crofts, 338–366.

SCOTT, Mike & Christopher TRIBBLE. 2006. *Textual Patterns: Key Words and Corpus Analysis in Language Education* (Vol. 22). Amsterdam and New York: John Benjamins Publishing.

SCHMERBAUCH, Mike. 2016. "Establishing a records appraisal workflow". *Information Management* November/ December, 36–38.

SOMEYA, Yasumasa. 1999. A Corpus-Based Study of Lexical and Grammatical Features of Written Business English. Unpublished MA thesis. Tokyo: University of Tokyo.

SWALES, Jim and Priscilla S. ROGERS. 1995. "Discourse and the projection of corporate culture: The mission statement". *Discourse & Society* 6/2, 223–242. <https://www.jstor.org/stable/42887976>.

VAN DER YEUGHT, Michel. 2016. "A proposal to establish epistemological foundations for the study of specialised languages". *ASp* 69, 41–63. <DOI : 10.4000/asp.4788>.

YATES, Joanne & Wanda ORLIKOWSKI. 1992. "Genres of organizational communication: A structurational approach to studying communication and media". *Academy of Management Review* 17/2, 299–326. <https//doi.org/10.5465/amr.1992.4279545>.

NOTES

1. However, it is interesting to note that the Information and Records Management Society in Britain can boast approximately 2,000 members and that the number of members of its American counterpart, the Association of Records Managers and Administrators (ARMA), is about tenfold.
2. Those issues were available online and they represent a sample covering contemporary issues.
3. The introduction was not published but it may be downloaded as a pdf file from <http://www.comp.lancs.ac.uk/ucrel/usas/>.
4. The expressions were retrieved from the Mediation part of the corpus.
5. In corpus linguistics, word spans are expressed in numbers of words to the left (L) and to the right (R) of a central word or "node".

ABSTRACTS

This article seeks to gain insights into the discourses of the Records and Information Management (RIM) professional culture and, more generally, refine our conception of what professional discourse implies in terms of corpus approach. We start with an overview of the long tradition of professional discourse analysis (PDA) by focusing on the specialised nature of this discourse type and by overviewing some corpora which have been designed for PDA. We proceed with a method for compiling and analysing a corpus in the domain of RIM, more particularly in the aim of retrieving the "specialised substance" of the corpus. This substance is evidenced through three micro-studies: the analysis of a specialisation process of general semantic categories, semantic coherence analysis, and the making of new, ad hoc semantic dimensions. Firstly our results show that RIM may be considered as a specialised register in that the corpus presents a coherent whole of meanings and text types. Secondly, they show that comparing a professional variety with general usage only provides a partial view of specialisation and that the creation of ad hoc, specialised, semantic dimensions should be considered.

Cet article vise à approfondir notre connaissance des discours professionnels dans le domaine de la gestion de l'information et de la documentation, à partir d'un corpus spécialisé. Après un bref résumé de la longue tradition des travaux en analyse des discours professionnels, nous présentons une méthode de compilation et d'analyse d'un corpus dans le domaine de la gestion de l'information, notamment dans le but d'identifier sa « matière spécialisée ». Cette matière est mise en évidence par trois micro-analyses : celle du processus de spécialisation des catégories sémantiques générales de l'anglais, celle de la cohérence sémantique, et celle de dimensions sémantiques spécifiques. Nos résultats montrent tout d'abord que la gestion de l'information peut être considérée comme un registre spécialisé de l'anglais. Ils montrent ensuite que la comparaison d'une variété professionnelle avec l'usage général fournit une vue partielle du spécialisé-professionnel et qu'il convient d'envisager la création de dimensions sémantiques ad hoc.

INDEX

Mots-clés: analyse de corpus, milieu professionnel, dimension sémantique, gestion de l'information et de la documentation
Keywords: corpus analysis, professional setting, semantic dimension, specialised varietiy, records and information management

AUTHOR

PHILIPPE MILLOT

Philippe Millot is a senior lecturer in English for specific purposes at the faculty of languages of the University of Lyon (France) and he is a member of its Linguistics Research Centre. His research deals with the analysis of specialised varieties, more particularly in professional and occupational contexts, as well as corpus linguistics. He also teaches ESP to students who specialise in records and information management. philippe.millot@univ-lyon3.fr