New perspectives for the analysis and formalisation of specialised genres: the GENTT proposal

Isabel García Izquierdo
Universitat Jaume I
GENTT group
igarcia@uji.es

Isabel García-Izquierdo és professora de lingüística aplicada a la traducció i d'espanyol per a traductors al Departament de Traducció i Comunicació de la Universitat Jaume I de Castelló. Des de l'any 2000, és la directora del Grup de Recerca de Gèneres Textuals per a la Traducció (GENTT, *www.gentt.uji.es*), que se centra en l'anàlisi multilingüe dels gèneres textuals en el marc de la comunicació especialitzada aplicada a la traducció. Ha publicat diversos llibres, així com uns quants articles nacionals i internacionals, relacionats amb la seva recerca. És l'autora d'*Análisis textual aplicado a la traducción* (2000), *Divulgación médica y traducción. El género información para pacientes* (2009) o *Competencia textual para la traducción* (2011).

Resum

Noves perspectives per a l'anàlisi i la formalització dels gèneres especialitzats: la proposta del GENTT

Aquest article examina el paper del gènere com a eina clau conceptual i metodològica per a l'anàlisi de les àrees de comunicació especialitzades. Es presta especial atenció als punts de convergència entre els estudis en lingüística i traducció, i els duts a terme en camps com ara la recuperació d'informació o la informació aplicada a les ciències en temes relacionats amb la formalització del text. Defensa un enfocament global polièdric dels gèneres especialitzats que en permetria la recuperació (semi)automàtica, així com una anàlisi dinàmica flexible, que és compatible amb les circumstàncies de la producció i l'ús.

Paraules clau: *gèneres textuals; comunicació especialitzada; traducció; recuperació de la informació*

Abstract

This paper examines the role played by genre as a key conceptual and methodological tool for analysing areas of specialised communication. Special attention is given to the points of convergence between studies in linguistics and translation and those conducted in fields like Information Retrieval or Applied Information Science on subjects related to textual formalisation. It defends a comprehensive multifaceted approach to specialised genres that allows their (semi-)automatic retrieval as well as a flexible dynamic analysis that is compatible with the circumstances of production and use.

Keywords: *textual genres; specialised communication; translation; information retrieval*

Terminàlia 3 (2011): 13-21 · DOI: 10.2436/20.2503.01.21
Data de recepció: 28/01/2011. Data d'acceptació: 10/03/2011
ISSN: 2013-6692 (impresa); 2013-6706 (electrònica) · *http://terminalia.iec.cat*

1 Introduction[1]

Recent research in the field of translation and applied languages in general has stressed the importance of the concept of *textual genre* as a conceptual and methodological tool that allows further advances to be made in the analysis of communication (and particularly specialised communication – García Izquierdo, 2007, 2009). It thus expands on the traditional view which focuses almost exclusively on terminology. The GENTT (Textual Genres for Translation, *www.gentt.uji.es*) research group based at Universitat Jaume I has taken this concept as the starting point for a multilingual research project that looks at the legal, medical and technical areas of specialised communication.

In addition, and as proof of the multidisciplinary nature of science, the latest research conducted both in the field of Natural Language Processing (NLP) on automatic classification or Information Retrieval (IR) and in the field of Applied Information Science underlines the value of the concept of *genre* in the (semi-)automatic retrieval of information.

Hence in this paper we review the points of agreement between the proposals put forward in the field of Translation and Applied Languages and those proposed in other fields, with the aim of identifying possible new methods of systematising and analysing textual genres that can help to extend the classical way of conceiving the analysis of specialised communication.

2 The GENTT approach to textual genre

As more and more research has been conducted by the GENTT group, a definition of the concept of *textual genre* has gradually been shaped, albeit in an eclectic way, based mainly on propositions borrowed from systemic functional linguistics, genre theory applied to translation and the sociology of professions (I. García Izquierdo, 2009). Although *genre* has been defined from a number of different perspectives in research conducted in linguistics and translation studies, it is generally agreed that it is a multifaceted concept that comprises formal, communicative and cognitive aspects.

Current research being carried out by the team analyses the communicative and formal aspects of the definition of genre above all.[2] Thus the research conducted by the GENTT group concentrates firstly on studying the formal aspects of genre and on systematising and analysing it from the point of view of linguistic transaction. Secondly it also aims to perform an analysis of a more sociological or socio-professional nature, which attempts to develop the communicative (and to an extent the cognitive) side of the concept and to incorporate the vision of the professionals who work with the genres under study. In this sense we could say that in our research we combine the two approaches to the study of genres that have, to date, proved to be the most fruitful, i.e. that of the *Australian School* (based on Systemic Functional Linguistics) and the *North American School* (see I. García Izquierdo, 2009).

Indeed, genres only make sense within the professional community that uses them. And that is why we are interested in studying not only aspects related to the (external) form, communication and cognitive interpretation by the speakers within the multilingual communities in which they develop, but also the socio-professional particularities of those communities. As stated by Hyland (2003: 21):

> The notion of discourse community foregrounds the socially situated nature of genre and helps illuminate something of what writers and readers bring to a text, implying a certain degree of inter-community diversity and intra-community homogeneity in generic forms. [...] Communities are where genres make sense; they are the systems where the multiple beliefs and practices of text users overlap and intersect. (Swales, 1998).

Consequently, for the members of GENTT genre is a category that can be applied to any sphere of communication because it is a collective product that results from each particular circumstance of communication. Any form of conventionalised and culturally determined text, regardless of the field (specialised or not) in which the communication takes place, can therefore be considered a genre (García Izquierdo [ed.], 2005). Nevertheless, the notion is especially significant in the fields of specialised communication, as has been made quite apparent by research in recent years. This is because there are certain variables which determine the way genres are organised that are more specific to this type of communication. Such variables refer essentially to the discipline involved, the degree of conceptualisation and abstraction and the scope of the communication (García Izquierdo, 2007: 122). We therefore position ourselves within the area of specialised translation (communication). In particular, we focus our interest on the legal, medical and technical fields from a multilingual perspective by working with texts in Spanish, Catalan, English, German and French.

This interest in the use of genres in professional practice has led us to develop a knowledge management tool that offers professionals a simple way to retrieve textual and contextual information. Thus the team is currently using a combination of quantitative (electronic corpus) and qualitative (surveys, interviews, focus group, etc.) methodologies (García Izquierdo, 2009 and forthcoming *a*); Muñoz, García Izquierdo and Montalt, forthcoming) to work on the electronic document management system called GENTT 3.0.

In this paper I will focus on this electronic document management tool to investigate its potential as a way of formalising genres.

A brief review of the history of the GENTT research team shows that it has been working since 2000 to create and improve its own management application: the software application called GENTT (which as we shall see below is soon to be released as version 3.0). This electronic document management tool is designed to be used to perform linguistic-terminological, phraseological, conceptual and contextual analyses of specialised texts belonging to different genres. In other words, it is a tool that takes into account the terminological and phraseological aspects as part of a complex analysis of genres and not (only) as a representative feature of the speciality.

The research concentrates on achieving an efficient exploitation of a corpus made up of texts classified according to their genre (we are therefore referring to a textual corpus containing complete comparable texts, with originals in the different working languages). As it stands at the moment, the computer system consists of a set of different tools and is intended to take care of *feeding* and managing the corpus, as well as allowing it to be searched and analysed. Nevertheless, this system works as an add-in for MS Word and this means that it is not very flexible when it comes to adapting it to new needs and keeping it growing. Likewise, there was a heterogeneity issue with the different parts of the system, which were largely dependent on the technologies each of them was based on, and this made the process of feeding and managing the corpus still more complicated. Furthermore, the current state of the system also made it somewhat more difficult to exploit the corpus using the different external analysis tools that were available due to the format in which the data were saved. Lastly, with version 2.0 it was difficult for outside collaborators to become involved in feeding and managing the corpus.

As a result, and despite the efficiency that the program had displayed on an internal level (i.e. to the research group), the decision was made to create a new version (version 3.0) that would improve the system and which would be able to cover both the current and future needs of the research project (see García Izquierdo, forthcoming *b*).

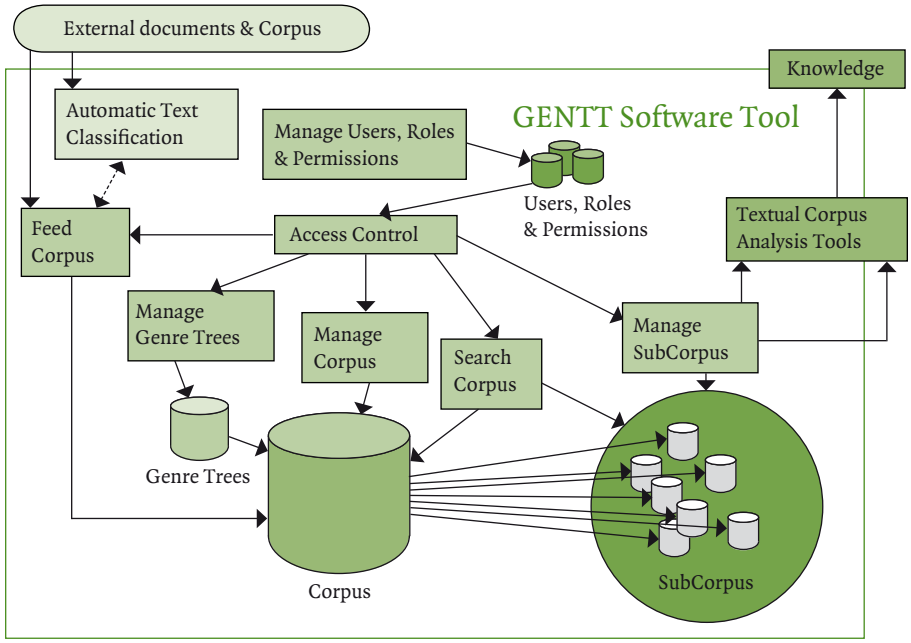
The main aim of the project to design version 3.0 of the GENTT management program was therefore to integrate and simplify the parts that make up the architecture of the previous system with web technology. This would simplify maintenance and make it easier for a network of geographically distributed collaborators to access it. Following the Web 2.0 philosophy, all of its parts will be integrated under one common technology and on a web-based platform. This will mean that the system can be accessed from any web browser that is connected to the Internet, thereby making it easier for a network of geographically distributed collaborators to feed and exploit the corpus.

The system is based on open standards, is flexible and is ready to be used with tools for analysing texts and textual corpora (stemmers, parsers, etc. like *TreeTagger*, *Freeling*, etc.). Moreover, it will also include documents containing help for users and simplify changing from one language to another on the interface.

At the time of writing, version 3.0 of the GENTT management application includes the following numbers of texts, which are being validated after migration from the old system to the new one:

Legal:	668 texts (1,579,910 words)
Medical:	296 texts (516,698 words)
Technical:	256 texts (402,953 words)
Total:	1220 texts (2,499,561 words)

The following diagram shows the general structure of the program.



A. Mascherpa (2009), previously unpublished

The utilities that the system currently offers can be summed up as follows (see also I. García Izquierdo, forthcoming *b*):

1. Text search based on the categories indicated by the headings of the columns, which are listed below, and the creation of specific subcorpora:

- Genre[3] (e.g. Fact Sheet for Patients, Informed Consent, Editorial, Patent, Agreement, Certificate, etc.)
- Name
  - Format: {Field: Jurídico (Legal), Médico (Medical), Técnico (Technical)} + [Number: 1..9999] + {language: C, E, I, A, F – Catalan, Spanish, English, German and French}
  - The name must be unique
- Title
- Copyright
- Status (original or quasi-original)
- Publisher
- Year of publication
- Language of the document
- Original format
- Text type (argumentation, exposition and instruction)
- Subject area (following the classification established by the Library of Congress)

2. Obtaining information about the context of the situation in which the texts are produced from the genre records incorporated into the system.
3. Kinship relation among the genres in the same area (*genre system*, Bazerman (1994: 79), or "... interdependent genres that are enacted in some typical sequence (or limited set of acceptable sequences) in relation to each other, and whose purpose and form typically interlock"); and with genres from other areas (*genre colony*, Bhatia (2004: 59), or "Groupings of closely related genres serving broadly similar communicative purposes, but not necessarily all the communicative purposes in cases where they serve more than one").
4. Keyword frequency analysis.
5. Concordance analysis.
6. Analysis of the terminology density of the texts (frequency of one class of words compared to others in certain genres).

In the medium term, as soon as the work on validating the corpus that is presently being carried out has finished and the management utilities have been updated in accordance with the results from a pilot study conducted in collaboration with three research groups from other universities, we hope the system will allow us to:

7. Create dictionaries from the terms found in the texts in the corpus.
8. Create terminological glossaries linked to genres, genre colonies or genre systems.

All this will be accomplished using general headers but without tagging the whole corpus. The idea is to allow each user to create his or her own subcorpus according to a number of search criteria and apply the utilities of basic management programs (length of words and phrases, terminology density, semantic web tags – e.g. rhetorical sequencing tags, etc.) to meet their own needs.

Nevertheless, as can be observed in the diagram of the structure of the program seen earlier, one of the medium-term goals would be for the program to be capable of semi-automatic identification and classification of the texts that are incorporated into the corpus. This would make it easier to offer professionals a far greater volume of analysed texts. And to do so it will be necessary to draw on findings from the latest research conducted in the fields of computing and document management.

3 Text Classification and Information Retrieval

Today, Text Classification (TC) has become a relevant field of analysis in the area of NLP. As stated by Civera (2008: 1):

> The purpose of TC is to convert an instrumental repository of documents into a structured one by automatically assigning documents to a predefined number of groups, in the case of text clustering, or to a set of predefining categories, in the case of text categorisation. Doing so, the tasks of storing, searching and browsing documents in this repository is significantly simplified.

Although research on the automatic classification of texts is not a recent development, the paradigm did undergo significant changes in the 90s. As a result, there are currently two approaches to achieving this automation: the automatic classification of texts based on expert systems and the use of automatic learning techniques (Sebastiani, 2002). The former is based on the extraction and formalisation of a series of rules that determine the categorisation grounded in the knowledge of a human expert, while the latter makes use of techniques that, based on a set of texts that have been classified by a human, train the machine and make it possible to obtain a decision model which is usually mathematical. Of these two options, the one based on automatic learning is the one that has become prevalent since the 90s because of the better results that are achieved (comparable to the efficiency of a human operator) and the reduction in the amount of time that needs to be spent by experts. The changes included in the GENTT 3.0 computer application also respond to the need to make it more flexible so that, whenever necessary, the techniques described above can be incorporated into the system.

In this context then, linguists, researchers working in translation studies, documentalists and computer experts defend the need to find useful systems for retrieving information (see García Izquierdo and Borja, 2009), especially in professional organisations (Crowston and Kwasnik; Tyrväinen and Päivärinta; etc.).

From the point of view of the documentalist, Izquierdo Alonso (2004) analyses the Documentary Form of Content (DFC), a discipline that is part of Documentary Textology and which is conceived as (2004: 35):

> [...] the study of the general principles governing the internal structure of informative messages and the delimitation of the nature, relations and functions of the different types of constituents that may be part of such a structure. (Our translation)

Thus Izquierdo Alonso addresses two aspects that can, in her opinion, provide an answer to the theoretical problems that arise in the field of automatic text processing, namely how the representation of knowledge is organised in the documents and how the "parts" of knowledge are distributed in them (2004: 33). In her view, current information retrieval techniques are still not sufficiently effective partly because of the lack of consistent theories about the structure and functions of discourse. Her study therefore focuses on rhetorical-functional relations or formal content spaces, according to which she proposes an ontology (ONTOFORM) as one of the possible models for representing content. She then takes this as the basis for her proposal for the creation of (2004: 46):

> A computer environment for the management of textual contents (GeConText) that allows effective knowledge structures to be generated from non-structured information (plain or raw text) and which is capable of categorising them according to parameters that are not based on just statistical analyses of IR but on discursive criteria from previous textual analyses. The system enables us to recognise the structure of a particular type of document automatically, model it by means of an ontology and navigate our way around the formal areas of the content in an XML environment. (Our translation)

Although the proposal grants more importance to the rhetorical aspects of the structuring of the contents (discourse acts, moves and sequences) and to pragmatic action, it is quite clearly similar to the GENTT approach. In fact, for Izquierdo Alonso (2004: 35) there are three aspects that determine the structure of the content, namely belonging to a genre, the choice of a particular form of text-type and the realisation of a specific textual structure to fulfil a particular communicative function. In my opinion, and starting from a comprehensive concept of genre as a place where all the formal, communicative and cognitive relations that characterise texts converge (García Izquierdo, 2002, 2007 and 2009), in actual fact both text typology and the realisation of textual structure depend on genre. Nevertheless, again the parallels between the approaches (both cases are indebted to the Sydney School) are obvious. In fact, Izquierdo Alonso (2004: 36) conceives the analysis of textual structure in a three-dimensional manner which, as we said earlier, is the same as the way it is understood in the research conducted by the GENTT group:

> From a functional and systemic dimension in which [...] we take into account semiotic interaction, pragmatic action and communicative-textual dimension, from psychocognitive and sociocultural factors. (Our translation)

One fact to be noted, therefore, is that genre is considered the central category of the analysis for the semi-automatic retrieval of information. Many studies in the field of IR or TC follow this same line. In this research, the concept of genre is used as the defining parameter because it allows the texts to be characterised in a different way from the propositional content, which has been the focus of most IR and TC studies and helps users to be more selective in the process of searching for information. Hence authors such as Yates and Orlikowski (2002), Tyrväinen and Päivärinta (1999), Crowston and Kwasnik (2004), Luštrek (2007) and Kim and Ross (2007) declare genre to be the starting point and fundamental pivot for understanding communicative situations, grouping the texts according to certain criteria and recognising them automatically. There is widespread agreement that, in addition to questions of form and content, purpose and function also need to be taken into account to be able to characterise genres in a suitable manner (Orlikowski and Yates, 1994). In this vein and following other studies that address the automatic analysis of genres, Kim and Ross (2007: 173) stated that:

> The definition of genre adopted by these researchers all rely on a combination of two notions: one of structure and one of function. *Structure* is defined by factors which are reflected in the visual layout of the document while function is defined by the intended purpose of the document. The two notions are closely related: the structure of the document is formed to optimise the function of the document within an environment, such as within the context of the community or event, in which the document is created.

For Crowston and Kwasnik (2004) genre identification is part of the larger process of IR, and more so since the advent of the Web, which has increased the need to find better methods of searching the vast stores of information that have become more easily accessible. In these authors' opinion (2004: 1), the IR process can be defined as follows:

> A user represents an information need by submitting a query to the system via an intermediating mechanism. The system searches through the document representation

> in its store and uses some form of matching to "retrieve" either the documents themselves, parts of the documents, or representations of the documents. These search results are then presented to the user for evaluation.

So by means of different structural features in most cases the documents are described by syntactic parsing and other types of tags. Yet many parts of this analysis still have to be performed manually. Furthermore, in the opinion of these authors even under the best of circumstances we rarely find a method that is 100% effective because matching users' needs to potential information in the system is a complicated task.[d] In this regard, the same authors (2004: 2) claim that it is possible to improve the representation of documents by incorporating non-thematic characteristics that signal their purpose (specifically, the genre) and, as a result, improve the stages of the IR process: "Knowledge of document genre may improve accuracy of relevance judgments that modern search engines make in order to rank order the search results."

Indeed, in the last years of the 20th century researchers focused on identifying characteristics that can be of use to discriminate between genres. As a result, this category has been defined in a number of different ways, according to the different research traditions. In Kwasnik and Crowston's (2005: 79) opinion, this is evidence of the richness of the concept and reflects its value as a lens through which an assortment of phenomena can be observed. Moreover (2005: 84), genres offer an efficient way of relating to documents in all the stages of their life cycle, from their creation and distribution to their storage, retrieval and use for new purposes. In the opinion of these authors "Genre acts as a template of attributes that are regular and can be systematically identified" (2005: 86).

In an earlier study, Crowston and Kwasnik (2004) claimed that in order to create a classification it is necessary to determine the scope and extension of the domain being classified, as well as the scale (granularity) of the entities and the terminology used to describe them. After that a conceptual structure is determined, because a classification is "a collection of such concepts that are related to each other through classificatory relationships" (p. 3). One important task in classifying genres is therefore to determine the unit because, as we saw earlier, genres can in turn be subclassified into subgenres (Bhatia, 1999), whose individual components join together to form a distinguishable, identifiable whole.

Thus Crowston and Kwasnik (2004) borrow an interesting concept from Erickson (2000), i.e. that of *genre ecologies*. This term refers to the fact that, like an organism in an ecological community, genres have effects on each other and sometimes depend on each other for their effectiveness. "[...] Genres exist in habitats or communities of practice" (2004: 88). This concept is clearly related to the concept of genre system, coined by Bazerman (1994). In this author's opinion, in every professional field there are what are known as *genre sets*, or types of texts produced by the professionals in those fields following specific patterns. Based on this idea, this author proposed (1994: 79 et seq.) the concept of *genre system* to designate interdependent genres "that appear as certain typical sequences, form relations with one another and have interacting purposes and forms", which again provides evidence of the relation between research in the two fields.

Once it has become clear that it is necessary to determine the scope of the domain of the genres, the method of classification will then have to be established. To date, and as pointed out by M. Luštrek (2007: 7), there have been three methods of automatic genre identification: the traditional one, the character-based one (only useful for certain categorisation tasks and with no evidence on which to be able to decide whether it is generally useful for identifying genres) and the visual one (although "Little information is available on visual methods", 2007). In his opinion (2007: 7), however, there is still very little information available about the identification of genres from very heterogeneous documents. Therefore, although most genre identification tasks could be carried out using traditional methods, a lot of work still needs to be done to overcome certain difficulties.

The traditional method is the one in which features are extracted from the texts and have algorithms assigned to them. Some of these algorithms will provide better results if the features have been selected properly. The characteristics to be taken into account may allow for several criteria:

a. *Surface features:* belonging to the surface of the text and which can be extracted easily without the need for complicated parsing. This involves function words, genre-specific words, phrases and punctuation marks;
b. *Structural features*: these require at least a tagger (POS), such as *TreeTagger*, for a part of the text. There is no evidence to show that more advanced linguistic analyses than the one provided by this program have any significant impact on the identification of genres (2007: 2). It is possible to study parts of speech (nouns, verbs, adjectives, etc.), phrases (frequencies and sizes), verb tenses, types of sentences, etc.;
c. *Presentation features*: HTML or TeX documents (Token type, graphic elements and links: images, tables, etc.); and
d. *Other features*: topic segments (TextTiling, splitting the text into multi-paragraphs using units of lexical cohesion to find segments with topical coherence); and URL.

Broadly speaking, to date four types of traditional features have been the most widely used for automatic classification:

(a) syntactic (parts of speech, e.g. adverbs, nouns, verbs and prepositions);
(b) lexical (terms of address, e.g. *Mr.*, *Mrs.*, *Ms.*; content words; most frequent words in a corpus, e.g. *the*, *of*, *and*, *a*);
(c) character-level (e.g. punctuation, character count, sentence count, word length in characters); and
(d) derivative (ratio measures, e.g. average words per sentence, average characters per word, type/token ratio).

These features, which are relatively easy to identify automatically in combination with automatic methods of machine learning, have been applied to distinguish between different types of documents and different genres and should allow us to extract information (semi-)automatically.

But what criteria should be used to train the machine to recognise genres? In Crowston and Kwasnik's opinion (2004: 2), the best method is to take a bottom-up approach in which the users' opinion is taken directly into account, with special attention given to the *facets*, or basic attributes, perceived by people. Thus these authors (2004: 4) suggest a facetted-analysis, similar to those conducted by Päivärinta (1999) and Tyrväinen and Päivärinta (1999). This is not really a different representational structure, but rather a distinct approach to the classification process. The notion of facet is based on the assumption that there is more than one way of viewing the world and that even those classifications that are perceived or seen as stable are in fact provisional and dynamic. The challenge is to construct classifications that are flexible and able to accommodate new phenomena. Facets can therefore be defined as attributes (M. Luštrek, 2007: 1) which will have to be selected, developed and applied to the analysis.

This facet-based approach is, in these authors' opinion, relatively hospitable (i.e. capable of accommodating new entities smoothly) and, above all, flexible. This means that it can allow us to discover interesting new associations but at the same time create fixed profiles if needed (2004: 6).

Kessler et al. (1997) also propose a classification of genres according to this criterion:

> We propose a theory of genres as bundles of *facets*, which correlate with various surface cues, and argue that genre detection based on surface cues is as successful as detection based on deeper structural properties.

In short, it seems that the different studies carried out in the field of Automatic Classification coincide in their interest in genres and in the relevance of a facetted classification, which to a certain extent goes beyond the traditional semi-automatic classifications.

4 The relation between TC's facetted classification and the GENTT proposal

We have just seen how several studies in the field of computer science and NLP recommend selecting certain facets that can help us to characterise genres in a semi-automatic manner. On this score the proposal is in agreement with the research conducted by the GENTT team, since each of the searches the system is required to carry out, as we saw earlier, focuses on attributes related with the genres that, in this case, make up the categories in the header of the texts in the corpus.

The aim is to go beyond traditional research on semi-automatic classification (which, as noted above, was essentially focused on morphosyntactic issues) and to incorporate other attributes, in addition to semantic and pragmatic analysis, that allow us to perform an analysis that is more closely adapted to the texts. This refers to aspects that are complementary to the form and propositional content, such as the language, the subject field, the specific use of the document (situational context) and its relationship with other texts. Hence the GENTT proposal addresses genre analysis using categories that have to do with the author, the source, the publisher, the year of publication, the language, the text type or the subject field, but also with the discipline that it belongs to and the context of production, all of which are aspects that could be considered to be more or less permanent attributes.

Therefore when enough subcorpora have been analysed and confirmation is obtained of the regularity of the attributes that have been used, we hope to be able to apply the GENTT program to the semi-automatic classification of texts. We trust that the development of the current system will allow information retrieval and automatic document classification techniques to be applied to the corpus, thereby improving the way it is fed, exploited and analysed. The tendency today in systems that have to manage and exploit large numbers of documents containing textual information is to automate parts of the constituent processes. Hence automation, or semi-automation, will be a prime objective for improving the system's performance.

5 Conclusion

In sum, from the considerations outlined above it becomes clear that the concept of genre is a valuable tool for document management (retrieval, writing, translation, etc.) and that it is important to characterise this notion, especially when applied to specialised fields, by means of a variety of attributes (*facets*) that go beyond the traditional conception. This means that they are not focused exclusively on content or on lexical items or terminology and thus allow us to address analysis, according to the circumstances, in a flexible dynamic way that takes reality itself into account.

References

BAZERMAN, Charles (1994). "Systems of Genres and the Enactment of Social Intentions", in FREEDMAN, Aviva and MEDWAY, Peter (eds.). *Genre and the New Rhetoric*. Abingdon: Taylor & Francis Group, p. 79 et seq.

BHATIA, Vijay K. (1999). "Integrating Products, Processes, Purposes and Participants in Professional Writing", in CANDLIN, Christopher and HYLAND, Ken (eds.). *Writing: Texts, Processes and Practices*, London: Longman, 21-40.

BHATIA, Vijay K. (2004). *Worlds of Written Discourse: A Genre-Based View*. London: Continuum International.

BORJA, Anabel; GARCÍA IZQUIERDO, Isabel and MONTALT, Vicent (2009). "Research Methodology in Specialized Genres for Translation Purposes". *The Interpreter and Translator Trainer (ITT), vol. 3, no. 1*, 57-79.

CIVERA SAIZ, Jorge (2008). *Novel statistical approaches to text classification, machine translation and computer-assisted translation*. Unpublished doctoral thesis. Valencia: Universitat Politècnica de València.

CROWSTON, Kevin and KWASNIK, Barbara (2004). "A Framework for Creating a Facetted Classification for Genres: Addressing Issues of Multidimensionality". *Proceedings of the 37th Hawaii International Conference on System Sciences*, Los Alamitos: IEEE Computer Society Press, 1-9.

ERICKSON, Thomas (2000). "Making sense of Computer.Mediated Systems as Genre Ecologies". 33rd Hawaii International Conference on System Sciences, vol. 3, p. 3011.

GARCÍA IZQUIERDO, Isabel (2002). "El género: plataforma de confluencia de nociones fundamentales en didáctica de la traducción". *Discursos, Série Estudos de Traduçao, 2*, Lisbon: Universidade Aberta, 13-21.

GARCÍA IZQUIERDO, Isabel (ed.) (2005). *El género textual y la traducción. Reflexiones teóricas y aplicaciones pedagógicas.* Bern: Peter Lang.

GARCÍA IZQUIERDO, Isabel (2007). "Los géneros y las lenguas de especialidad", in ALCARAZ, Enrique. (ed.). *Las lenguas profesionales y académicas.* Barcelona-Alicante: Ariel-IULMA, 119-125.

GARCÍA IZQUIERDO, Isabel (2009). *Información médica y traducción. El género Información para pacientes*, Bern: Peter Lang.

GARCÍA IZQUIERDO, Isabel (2011). *Competencia textual para la traducción*, Valencia: Tirant lo Blanch.

GARCÍA IZQUIERDO, Isabel (forthcoming *a*). "La investigación cualitativa en traducción especializada: una mirada a los ámbitos socioprofesionales". Actas de IV Congreso de la AIETI, "Traducir en la frontera". Vigo: Universidade de Vigo.

GARCÍA IZQUIERDO, Isabel (forthcoming *b*). "Investigating Professional Languages through Genres", in SUAU, Francisca and PENNOCK, Barry (eds.). *Interdisciplinarity and languages: current issues in research, teaching and professional applications and ICT. Selected Papers*, Bern: Peter Lang.

GARCÍA IZQUIERDO, Isabel and BORJA, Anabel (2009). "La gestión de la documentación multilingüe en entornos profesionales: propuesta de formalización". *LYNX. Panorámica de estudios lingüísticos*. Initial article, 1-27.

GIL LEIVA, Isidoro and RODRÍGUEZ MUÑOZ, José Vicente (1996). "El procesamiento del lenguaje natural aplicado al análisis del contenido de los documentos". *Revista general de Información y Documentación, vol. 6, no. 2*, 205-217.

GUARINO, Nicola (1997). "Semantic matching: Formal ontological distinctions for information organization, Extraction and Integration". Lecture Notes in Computer Science 1299, Berlin: Springer, 139-170.

HYLAND, Ken (2003). "Genre-based pedagogies: A social response to process". *Journal of Second Language Writing*, vol. 12, issue 1, 17-29.

IZQUIERDO ALONSO, Mónica (2004). "Nuevos retos en el análisis documental de contenido: la gestión de la forma documental del contenido". *Scire, vol. 10, no. 1*, 31-50.

KESSLER, Brett; NUNBERG, Geoffrey and SCHÜTZE, Hinrich (1997). "Automatic Classification of text genre". In Proceedings of 35th ACl/8th EACL. Madrid, p. 32-38.

KIM, Yunhyong and ROSS, Seamus (2007). "Detecting Family resemblance: Automated Genre Classification". *Data Science Journal, vol. 6*, Supplement, 172-183.

KOPPEL, Moshe et al. (2003). "A Corpus-independent Feature Set for Style-Based Text Categorization". *Proceedings of Workshop on Computational Approaches to Style Analysis and Synthesis*, IJCAI, Acapulco, Mexico. Electronic Edition.

KWASNIK, Barbara and CROWSTON, Kevin (2005). "Genres of Digital Documents". Introduction to Special Issue of *Information, Technology & People, vol. 18 (2)*, 76-88.

LUŠTREK, Mitja (2007). "Overview of Automatic Genre Identification". Technical report IJS-DP-9735, derived from a report prepared for Alvis, a European FP6 STREP [*http://www.alvis.info*], 1-8.

MASCHERPA, Alessandro (2009). Informe de tareas realizadas en la plaza *Gerónimo Forteza* de técnico en Informática del grupo GENTT, Castellón: Universitat Jaume I. Unpublished.

MONTALT, Vicent; EZPELETA, Pilar and GARCÍA IZQUIERDO, Isabel (2008). "The acquisition of Translation Competence through Textual Genre", *Translation Journal*, volume 14, nº2, (Issue october)

MUÑOZ, Ana; GARCÍA IZQUIERDO, Isabel and MONTALT, Vincent (forthcoming). "La investigación socio-profesional y la competencia traductora aplicadas a la pedagogía de la traducción médica". *Actas de IV Congreso de la AIETI, "Traducir en la frontera"*, October 2009, Vigo: Universidade de Vigo.

ORLIKOWSKI, Wanda and YATES, Joan (1994). "Genre Repertoire: The Structuring of Communicative Practices in Organizations". *Administrative Science Quarterly, 39*, 541-574.

PÄIVÄRINTA, Tero (1999). "A genre approach to Applying Critical Social Theory to Information Systems Development". In C. Gilson (ed.). Proceedings of the 1st Critical Management Studies Conference, Information Technology and Critical Theory–stream, Manchester.
<*http://www.mngt.waikato.ac.nz/ejrot/cmsconference/cmsdefault.htm*>

SEBASTIANI, Fabrizio (2002). "Machine learning in automated text categorization". *ACM Computing Surveys, vol. 34, Issue 1*, 1-47.

SWALES, John M. (1998). *Other floor, other voices: A Textography of a Small University Building*, Mahwah, NJ: Erlbaum.

TYRVÄINEN, Pasi and PÄIVÄRINTA, Tero (1999). "On rethinking organizational document genres for electronic document management". Proceedings of the 32nd Annual Hawaii International Conference on System Sciences, Los Alamitos: IEEE Computer Society Press.

YATES, JoAnne and ORLIKOWSKI, Wanda (2002). "Genre Systems: Structuring Interaction through Communicative Norms". *Journal of Business Communication, 39 (1)*, 13-35.

Notes

1. This research was made possible thanks to funding from the Spanish Ministry of Science and Technology (HUM2006-05581/FILO), the Fundación Bancaja-UJI (P112008-18) and the Spanish Ministry of Science and Innovation (FFI2009-08531).

2. Research work has, nevertheless, begun on the cognitive aspect, mainly with regard to the understanding of genres and the determination of their constituent features, both in teaching and in research (in this case linked to the acquisition of what is known as *translator competence*. See Montalt, Ezpeleta and García Izquierdo, 2008, and Borja, García Izquierdo and Montalt, 2009).

3. Genre trees are important in the application, since at present the main criterion for classifying the text corpus is dependent on them. It is in fact a linguistic domain ontology (Guarino, 1997) in which each genre depends on a macro-genre (abstract category) and can have different subgenres (and sub-subgenres) assigned to it. The difficulties that arise when trying to classify some areas, such as the legal domain, and the lack of symmetry between the disciplinary areas under study have led us to consider the possibility of incorporating numeric tags into the texts, which make them independent from the macro-genre. This is a possibility we are working on at the moment.

4. This opinion was also shared by Gil Leiva and Rodríguez Muñoz (1996: 204), for whom NLP is a complex field that entails a number of difficulties.