TOWARDS A FACETED TAXONOMY TO STRUCTURE WEB-GENRE CORPORA

Joseba Ezeiza Ramos
Xabier Payá Ruiz
Agurtzane Elordui Urkiza
Imanol Epelde Pagola
Universidad del País Vasco

Abstract: *The purpose of this paper is to contribute to the analysis of cyberjournalistic documents by proposing a taxonomy to structure web-genre corpora. It takes into account the peculiarities of this field, the new genres, their hybridization and complexness. In this sense, the taxonomy presented in this paper does not match a single theoretical framework, but it tries to gather the guidelines of various works intended to study online journalism and its genres. This theoretical flexibility is needed to set up a proposal good enough to suit the current needs of the area. The paper also describes the main axes of the taxonomy, defines its communication unit and remarks the values and limitations of such a work. Its result is a highly structured and document-oriented database, a tool that will enable users to understand the current trends, to create new hybrids, and to detect the changes that happen within this field that is widening the horizons of the usage of language.*

Key words: *Web-genres, applied linguistics, corpus-linguistics, multifaceted analysis, journalism.*

1. INTRODUCTION

Although the journalism and its continuous text production have been prominent study subjects, the influence of the Internet on the media has yet to be more systematically researched. The Web 2.0 era was foreseen as a huge revolution for the communication itself, but it is jet unknown the final result of the application of new communicative resources to journalism.

This article is intended to contribute to the analysis of cyberjournalistic documents, a field full of peculiarities, new genres and phenomena like hybridization. The contribution lies on a taxonomy to structure the web-genre corpora. It does not match a single theoretical framework, but it tries to gather the guidelines of various approaches intended to study online journalism and its genres. Several scholars have presented different perspectives to understand the reality of this area; some of them have even flagged up the difficulties arisen due to terminological diversity (Bhatia, 2002). Thus, a certain theoretical flexibility was needed to set up a proposal good enough to suit the current needs of the area.

The paper also describes the main axes of the taxonomy, defines its communication unit and remarks the values and limitations of such a work. Its result is a highly structured and document-oriented database, a tool that will enable users to understand the current trends, to create new hybrids, and to detect the changes that happen within this field that is widening the horizons of the usage of language.

The taxonomy has been developed in Basque language; it is hosted within the HIZLAN[1] platform (Ezeiza, 2009 & 2010) and has been developed under the shelter of the projects I + G DB (TTO: 2007.0077), EHLB (TTO: 2008.0368) HIZLAN (DIPE08/16) and EBALUA (EHU08/53),

[1] http://www.hizlan.org/db-hedabideak.

http://dx.doi.org/10.4995/rlyla.2011.899

in collaboration with the Ametzagaiña R & D, technological agent integrated within the Basque technology. The project involved academics and researchers from the departments of Basque Philology and Journalism at the University of the Basque Country and several researchers and computer technicians from Ametzagaiña R & D.

Even if the Internet is yet a new domain for the academic research, the diversity of articles published on issues related to the net remark the interest among different professional communities. The study of web-genres could make a huge contribution to a lot of extra-linguistic areas: the genre theory can improve the selection of the information, especially within wide document compilations; several enterprises analyse the genres to understand the inner communication of their own structures; the web has pushed the creation of new genres that might have not existed without the functions of the 2.0 era. In summary, the analysis of web-genres let us know about the creators and recipients of these repertoires, and also give us some clues to understand the nature of some nodes and links between websites (Montesi, 2005).

Other scholars state that web sites need a zero-to-multi genre classification-scheme, i.e. a scheme that accepts the genre-less classification as well as the multi-genre one, in addition to the traditional single-genre classification. More specifically, a zero-genre classification does not mean that a website does not have a genre, but it does not have a certain only one: this possibility accounts for those web pages that do not fit into any genre, and multi-genre classification is highly recommended when a website contains more than one genre. The genre hybridism and individualization (Lim Lee, & Kim, 2005) are two concepts that help understand the genre typology within the Internet.

In computational linguistics, several attempts have tried to establish a standardised set of genres (Santini, Sharoff, Rehm, & Mehler, 2007). The automatic classification has been able to identify and classify the web-genres through computational linguistics, but they are rather intuitive. As far as it concerns the rest of approaches, there is no prominent unified model. One of the obstacles for that aim is the broad variety of terminology: the keywords vary more than the concepts do. It is crucial that a single terminology is established; otherwise, the taxonomy will come down to a mere folksonomy (Santini & Sharoff, 2009). In this sense, it is remarkable the effort made by some researchers to classify polymorphic web-documents according to the characteristics of their segments (Rosso, 2008).

The automatic genre classification has proved to be a reliable evidence to show that the linguistic data (textual characteristics) is more than important to classify the web-genres themselves (Lim Lee, & Kim, 2005). This statement underlines the necessity to identify all the linguistic characteristics of a text, in order to know more about each web-genre. For that purpose, we need structured corpora and efficient tools.

2. THEORETICAL AND METHODOLOGICAL GUIDELINES

The taxonomy of journalistic documents hereby proposed is based on a series of different approaches that all together could compose a powerful tool with uncountable possibilities. It takes particularly into account several theoretical and methodological proposals of genre theory, sociolinguistics and corpus linguistics, as well as some classification proposals for online journalist-genres.

In this first section, we try to outline the main contributions on which this proposal is based.

2.1. A multi-perspective view: the faceted analysis

One of those approaches lies on the genre theory, which is intended to analyse the structural essentials that systematically combine in the creation of a message; it also looks for patterns in collections of similar messages. Consequently, the main purpose of genre theory within the field of journalism lies on the representation of the apparently disordered realities of the world. At the same time, it tries to comprehend the personal aim of the author, as well as the publicly acknowledged communicative intentions. In this sense, the constant development, the recent spread of hybrids and the varying perspectives on and interpretations of the genres are relevant obstacles to consider. Based on the understanding of the three traditions of genre theory (Miller, 1984; Martin, 1993; Swales, 1990), Bhatia (Bhatia, 1993) put forward a comprehensive definition of genre:

> […] a recognisable communicative event characterised by a set of communicative purpose(s) identified and mutually understood by the members of the professional or academic community in which it regularly occurs. Most often it is highly structured and conventionalised with constraints on allowable contributions in terms of their intent, positioning, form and functional value. These constraints, however, are often exploited by expert members of the discourse community to achieve private intentions within the framework of socially recognised purpose(s).

Bhatia's definition of genre is relevant to the present study which has as its aim the discourse as a professional practice (journalism, in our case), as a social practice, as a genre and as a text. This multi-dimensional approach underlines the suitability of a polyhedral view that will make possible the investigation of instances of conventionalised textual artefacts in the context of specifically journalistic practices (as a social practice) in order to understand how members of media communities construct, interpret and use these genres (as a genre) to attain their community objectives and their rationale to write them the way they do (as a text).

The genre analysis is then a multifaceted (a multi-perspetive) description of the reality that composes a journalistic document. Between the perspectives or facets of this type of analysis, there is a kind of correlativity that justifies the interdisciplinarity of the analysed concept. The analysed documents are interdisciplinary units integrated by different aspects or facets, and each of the latter belongs to a certain level of the main analysis.

The genre is a class, an abstract model, a perspective that acts as a cast for the speaker and as a horizon of expectation to the recipients or to members of a specific discourse community, characterized by a number of conventional features. It differs from the concepts of *register* and *text types* but the confusion between genre and type in particular, is still common. Socio-cultural transformations can determine significant changes in textual patterns. Moreover, they show a tension between the restrictions imposed by the generic framework and the characteristic variation of a particular communicative situation. Among the disruptive elements in the current situation, we can find the remarkable emergence of the Internet and the multimodality, which have altered the traditional schemes, creating new genres and changing the established distinction between them (Calvi, 2010).

In summary, the notion of genre is a useful tool in linguistic analysis, even if they do not always match the extreme variety and complexity of actual products. Hence the need for multidimensional approaches that permit cover all genres from the same professional area in systems, colonies, nets, etc., and taking into account the ties of interdependence.

No classification is final, but depends on the objectives of the analysis: to propose a complete taxonomy of the genres used in a professional sector is not the same as watching very closely the various joints of a specific genre.

In this sense, the socio-cognitive perspective (Bhatia, 1993) on discourse seems to offer a new view of the reality, as it encourages expert genre writers to appropriate rhetorical resources and other generic conventions across genres. Even if this kind of appropriation is optimally seen in other areas rather within journalism, the proliferation of new media and electronic modes of communication in public life has increased its frequency, motivating the creation of new genres or hybrids. However, these new hybrids may often have a considerable degree of creativity and innovation (Bhatia, 1993).

Particularly because of that hybrid character, a multi-perspective view is the best choice to establish the parameters of the taxonomy for journalistic documents. The interacting facets or views of discourse are not reciprocally exclusive, but fundamentally complementary to each other. Then, the goal one may decide to pursue will draw the best way to use the proposed tool.

2.2. Methodological perspective: based on the multi-dimensional model

As we said above, Bhatia (Bathia, 2002) claims that the research and interpretation of genre has to use a multi-perspective model that includes four dimensions. Two of them deal with the characteristics of the genres: the social dimension and the communicative dimension. The other two dimensions are related to the acquisition of knowledge: the socio-critical approach and the pedagogical approach. Therefore, according to Bhatia, to characterize the genre in its complexity, we should examine what social role it has, what are the linguistic elements that compose it, what kind of social experiences users have with these genres, and what type of learning do they develop around these genres. After all, the configuration of genre will be the result of the combination of all these variables. They are dynamic and interact with each other. The presence of these four dimensions legitimizes the four main currents of genre analysis (socio-critical, functional, textual and pedagogical), but their interaction is as evident as indisputable, and so it requires a new paradigm. The infrastructure provided by HIZLAN and the proposed taxonomy are more than justified within that paradigm or integrative vision: their design makes available to researchers two tools (conceptual and technical) that can be used in any of the four research currents. Consequently, the four dimensions are gathered in a single *physical* environment.

As Biber and Kurjian state (Biber & Kurjian, 2006), the Internet adds two new risks to the corpus analysis: on one hand, the net seems to be infinite, as huge as an ocean where it is impossible to calculate the totality of the information produced. On the other hand, it is not defined yet; we cannot foresee all the different texts available in the cyber-media. Hence too many things remain unpredictable, and we cannot design the most convenient characteristics and criteria to analyse the unknown quantity and diversity of texts. Nevertheless, if we design a genre typology for the Internet, the automatic searching systems will have easy interpretable results, if both the creation and searching of texts have been previously carried out by means of the same typology.

The multidimensional analysis (Biber & Kurjian, 2006) makes a distinction among the patterns applied to a text: genre, register and style are pre-established approaches that lie on different facets of the multifaceted analysis, as well as the variety. The variety is the type of language that is tied to a specific context (dialects, sociolects, genderlects...). The registers are the language types shaped by the situation, such as domain-specific language or computer-mediated communication.

A standardised approach will make easier the interpretation of a cyber-text, but if it does not foresee the range of linguistic or generic variations found on the web, any corpus constructed under this approach will not be representative of the reality. Documents that share the same topic do not necessarily belong to the same genre (Meyer zu Eissen & Stein, 2004). Moreover, the aforesaid variations, in addition to the purposes of the community and the author of a text, produce serious difficulties in order to select with security the genre of some cyber-texts.

In summary, the multidimensional model provides us the main guidelines to classify the cyber-texts into different aspects, and, consequently, the results of the correlations and the factor analysis will be highly reliable.

2.3. Analytical perspective: multilevel analysis

There are plenty of analyses that we can use in order to establish the main characteristics of a text. Critical and ethnographic analysis will let us know about the information related to the socio-cultural background of the text; textual analysis is the most suitable way to approach the cognitive and linguistic perspective of a discourse; corpus analysis, on the other hand, will mirror the quantitative use of linguistic resources within a text. The three of them and a few more are necessary tools to study the variability phenomena of genres, registers and styles in an integrated way.

However, the aforementioned analyses have quite strict rules that build up hardly adaptable structures. This fact makes difficult the interconnection of all these approaches, unless we find a flexible system that will let us combine all of them in a simple way, such as the faceted or multilevel analysis. This ground-breaking system gives us the opportunity to register an unlimited set of characteristics that belong to different levels of the polyhedral approach, so we will be able to record a long list of contextual, functional, conceptual or structural characteristics in different levels.

The faceted analysis is a polyhedral approach: each facet or perspective represents a determined aspect of the communication product (concept, context, function and structure, as well as the document itself). The polyhedricity of the proposed taxonomy is not strict, but permits a partial analysis with not all the included facets, multiplying the potential research and study that could be carried out by using a single but multidimensional tool.

2.4. Procedural perspective: mapping-integrated taxonomy

Two main procedures mirror the web classification systems that we can find in the Internet: the taxonomy and the folksonomy. The latter has gained a big territory among the users, whereas the former is commonly used by the web-designers that structure the information shown on the websites.

Both classifications have a relevant presence within the Internet, and have advantages to analyse the production of journalistic cyber-documents. Taxonomy is controlled by experts, is precise, is closer to institutional organisations and provides successful results; however, it is static and requires big monetary investments. On the other hand, the folksonomy is established by the users, is dynamic, needs little investments and lies on the community; but it is not precise and has merely representative results for an academic research.

Moreover, the folksonomy is a powerful tool to encourage user-based social classifications. It could be impossible to get a similar productivity by structured means. But the potentiality of the users is a disadvantage too, as the results are not usable for an academic purpose: the homonyms, synonyms and lack of coherence or homogeneity are obstacles that make this mapping useless. However, it would be an unforgivable mistake if we didn´t take into account the folksonomy, as its actual strength and potential development are too significant to be disregarded. The folksonomy or social tagging system generates a scan-based navigation: the tags are used to classify, sort, search and find information while exploring the net. So, how could we take advantage of the positive points of both classifications, reducing their weaknesses?

As we are designing a precise classification system for journalistic cyber-documents, a tool with a scientific purpose, the base of the proposed classification should lie on the principles of the taxonomy. However, the clustering and mapping data, the inestimable amount of tags that somehow organise the information in the Internet will be integrated in the taxonomy as a part of the conceptual analysis. The facet concerning the content has a specific part intended to archive the tags assigned to the document; if there weren´t any, the user will have the opportunity to add them during the uploading process. Thus, the procedural perspective will be based on a mapping-integrated taxonomy.

2.5. Technical perspective: modular and document-oriented database

If we want to develop an operational taxonomy, it is necessary that the platform that hosts it must be very effective: a modular and document-oriented data-base. The user can create as many corpora and segments as needed. The contrastive analysis of these segments can provide meaningful data in accordance with the established dimensions and perspectives.

In traditional data-bases all documents are stored in the same structure. A single structure means a single operational data-base with the only options to save and retrieve texts. In some cases, there is the possibility to export the information into another format, but then changed files are unrecoverable. The HIZLAN system is a modular document-oriented data-base. The users can set up as many databases and segments as they desire within the system. Only authorised users have access to each segment and each has different privileges. That is one of the main innovations of the support.

Moreover, when searching for some data, a new segment is created, but the selected texts will remain in the corpora, the new and the original. Thus, comparative studies can be developed, among others. This new tool, then, promotes comparisons with high parametric precision.

2.6 Towards the classification of journalistic cyber-documents

The informative genre begins to change something as important as the authorship of the typical journalistic texts. The proliferation of media and the democratisation of the information give priority to narrations in first person. The subjectivity of the author shows that the audience prefers reliability and closeness to objectivity. Unsurprisingly, the interview is changing on the same way: the dialogic genre that used to focus on the interviewee is now a mixed genre that emphasizes the interviewer's comments too often. Nonetheless, the report represents the clearest example of mixture: it is now a news-feature-article-interview, a combination of different genres that join the same aim.

Conversely, genres are necessary, as Mijail Batjin states: “if we had to create them (the genres) while we are speaking, and build each linguistic situation by ourselves, communication would be almost impossible” (Rodríguez Betancourt, 2004). Genres are needed codes that let both the speakers and recipients foresee important information about the texts they are producing or receiving.

Some scholars have already done ground-breaking research into the web-text production. However, it is not that much when it concerns cyber-journalism. Díaz Noci and Salaverría proposed the following classification in 2003 (Díaz Noci & Salaverría, 2003):

1. Informative genre (news).

2. Interpretative genre (reports).

3. Dialogic genre (interview).

4. Opinion genre (article).

5. Digital infography.

The proposal mirrors different criteria, as some of the options regard the content and others the form. This is due to a change on the genre theory applied to the cyber-documents, as Erickson observes (Erickson, 1999):

> As genre theory is applied to digital media rather than speech or writing, a couple of differences in emphasis have emerged. One of the chief differences is that those studying the digital medium are paying more attention to the role of technical features in shaping the evolution of digital genres.

The taxonomy of journalistic cyber-documents hereby proposed takes into account the role technical features have in the evolution of digital genres. It is then based on four main issues: firstly, it tends to analyse the rhetoric techniques that prevail in the text; secondly, it measures the virtuality of the hypertext used; thirdly, it takes into account the multimedia potentiality; and finally, it observes the interactivity.

In summary, rather than proposing a list of genres that can be considered static, a canonical classification with mandatory use, all the schemes just present all the elements that should be taken into account in order to establish the idiosyncrasy of a journalistic cyber-document, and then they see what elements appear together and how. Only that way we will be able to establish some general trends to approach a useful taxonomy, always open to any additions, changes and adjustments.

3. THE TAXONOMY

The taxonomy for journalistic cyber-documents aims to provide the main aspects that should be taken into account within the academic description of any kind of text that could appear in the area of journalism. For that purpose, we first need to establish the communication unit that will feed any database formed according to the following taxonomy. Then, we will go through the main aspects of the multifaceted analysis.

3.1 The communication unit

The work to find a communication unit that suits the flexible framework in which all taxonomic proposals are based is more than complicated; it is not easy to find a definition that encompasses the text of a micro-blogging platform and an audiovisual report from a digital newspaper. However, Bronckart allows us to approach the communication of cyber-journalism through this reflection (Bronckart, 2004):

> Every text is interdependent with the properties of the context in which it is produced; the actual text has a certain way of organising its referential content; it is composed of articulated sentences [...]. Each text, finally, uses mechanisms for contextualization and assumption of responsibilities set out to ensure its internal consistency. Thus, the notion of text production unit means any verbal message linguistically organized that tends to create in the recipient a coherence effect. And so this production unit can be regarded as a higher-ranked verbal communication unit.
>
> [...] In this sense, each text presents a unique set of individual characteristics and is therefore an object always unique. Consequently, the notion of singular or empirical text means a specific unit of verbal production, which necessarily depends on a genre, which is composed of various types of speech and also preserves traces of the decisions taken by the individual producer based on the particular communication situation where the text has been issued.

Based on the concepts of higher ranking and the empirical unit, this definition captures the essence of the cyber-journalistic communication, according to our view: The (multimedia) communication unit is a structured unit that has functional autonomy in its context in addition to conceptual uniformity, whose objective is to seek harmonization of various communication elements using different codes, resources and tools to provide the most appropriate environment and support to any information, participation and/or interactive process that the original issuer of the document wants to carry out. The communication unit will be considered multimedia when the resources used for such harmonization correspond to the audiovisual elements established in the Internet.

The communication unit, the document and the text are three different concepts. In some cases the three of them match but not always. The databases contain *documents*. These documents may be text only, or just communication units, or may not be either text or communication units (as defined above). Therefore, HIZLAN hosts three types of theoretical concepts: segments, documents and communication units.

The communication unit is a theoretical entity. The document is an empirical one. And the segment is a conventional unit. The theoretical entity allows us to define the document. The segment reflects the characteristics of the selected documents, as well as how to select and what metadata is saved.

3.2 The main axes of the taxonomy

This taxonomy requires three types of operations: classification, indexing and abstract. The classification, as an operation, is to gather information in different groups, but it can also be defined as a system to manage a segment of reality. The indexation translates the ideas in the document to a predetermined set of terms. Finally, according to the UNE 50-103:1990, [...] “the summary is the short, concise and objective representation of the content of a document”.

These three operations will be present in the five axes of the taxonomy. These axes are based on the main aspects that have been taken into account in the traditional views and are yet regarded as the fundaments of modern analysis. We are talking about the document, the content, the context, the function and the structure.

1. The document: This facet contains the essential information to identify a document in the database. As a testimony of a certain institution or individual, and as a information unit registered in different media of the digital communication, the aspect of the document has the following guidelines: to provide the document a name that serves as a basic identification; to load the entire file; to acknowledge the authorship of the document; the database automatically calculates the number of words in the entire text, if the document contains words; to indicate the publisher, signature or website that has published the document; the basic definition of medium (written, audio or audiovisual); and finally, this aspect gives the possibility of adding comments in the last section, in which various data can be annotated.

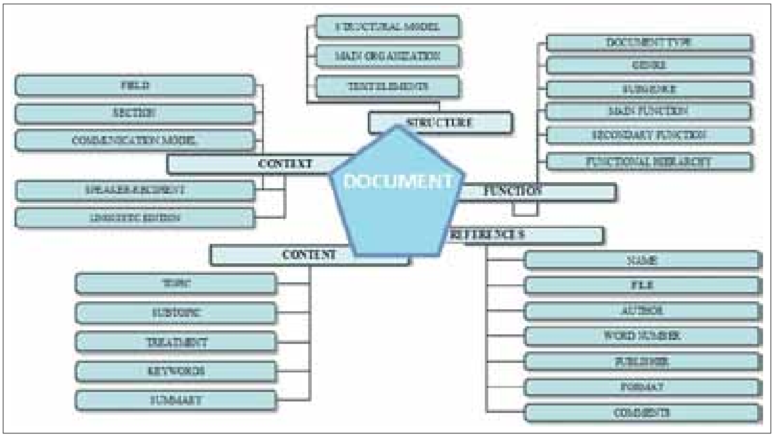
2. The function: This aspect refers to the set of relations established between two or more elements within the document. The role of media discourse in the network is built taking into account the following factors: the type of text, genre and subgenre, the primary function and secondary, as well as the functional hierarchy that relates the structure to its target.

3. The content: This aspect focuses on the conceptual analysis of the document, reviewing all those parameters that can influence the message of the document. The content could

be defined as the conceptual components in relation to the structure or logical units that build the sense of the speech. In this facet it is classified the general topic of the document (the first digit of the Universal Decimal Classification) and the specific classification (full classification with all the digits); then it defines the focus of the text in relation to the thematic treatment; it also annotates the tags or keywords of the document; and, finally, a larger subsection is intended to host a summary of the document.

4. The context: Although it is certainly difficult to define the set of relatively external elements of the documents in the Internet, we must make an effort to collect all those related to the context. The concepts of speaker and recipient, as well as the space-time relationship have little to do with those present in the traditional perspective. Therefore, the context is the meaning of a message, its relationship to other parts of the message, the environment in which the communication potentially occurs, and any perception that may be associated with communication. According to the peculiarities of journalism, the facet of the context marks the field and section in which the document was published; it will define the communication model taking into account the space-time relationship, the functionality and the dissemination, among others; it will also classify the roles of the speaker and recipient of the message; and, finally, it will show a final section created to annotate the name of the entity that has been responsible for the linguistic edition of the document.

5. The structure: This is the arrangement and order of content, the so-called *dispositio* in rhetorics. Textual organization, the role of the hypertext within the main structure and the analysis of the textual elements may have an added value: even if they may seem the simplest units to be classified, they will be the basis of interesting results that will increase the possibilities of any corpus.



Graphic 1. The scheme of the taxonomy.

4. THE CONTRIBUTION OF THE TAXONOMY

The new forms and modes of communication in the Internet have generated a renewed interest of linguists and other professionals for the usages of language in the new media (Díaz Noci, 2001; Ferris, 2002, Hasan & Martin, 2002; Manchón, 2003; García, 2005; Lamarca, 2006; Canhavilas, 2007, Franco, 2009; Yus, 2010). A line of work that seems to be productive in this area is based on the study of corpus (Reppen et al., 2002, Lim et al., 2004, Biber & Kurjian, 2006;

Hund et al., 2007, Meyer & Stein, 2009; Renouf & Kehoe 2009). Precisely HIZLAN, the platform based on the proposed taxonomy, is a web platform designed and developed for this purpose.

It has four integrated databases that can be managed from the web: a bibliographic database, a document database, a lexical one and one related to the style. The document database can generate and flexibly manage any number of corpora of text documents, hypertext, multimedia and hypermedia in different formats. And, thanks to the faceted structure of its taxonomy, it provides a very precise characterization of the documents hosted on the database.

For that purpose, it has several instruments that can obtain relevant linguistic information, among which a search engine for words, a search engine for five-word groups, another one for morphological categories, a calculating device, another comparative one for the frequency of use, a search tool for lexical combinations of two or three elements, a search engine for syntactic patterns, and a tool to discriminate the most representative lexicon. All these tools operate either on the body as a whole or on a specific collection of documents that share contextual features (area of production, mode of communication...), the subject (topic, subtopic, treatment of the subject...), the function (document type, genre, subgenre...), the structure (superstructure, macrostructure, microstructure...) or any combination of traits that are considered relevant. Then, it is really easy to perform comparative analysis, precisely stratified and very detailed.

The Basque version of HIZLAN[2] currently includes only documents written in Basque, but in the future the development of a multilingual version could be raised. In any case, the taxonomic structure is independent of this variable and can be transferred (either full or partial) to any other tool, language or project that is interested in studying the linguistic features of the (cyber-)communication.

The fact of creating and using this taxonomy, then, can be very useful for many applications. It may allow a deep analysis of the opposition, hybridization, or evolution that is happening between traditional journalism and cyber-journalism, particularly in regard to the usage of language. Faceted analysis can accurately notice the contrasts that happen in the press today.

This taxonomy may be a new source of empirical linguistic data, a reliable source to take into account to make decisions or to launch projects related to a standardization process, for instance. The same database that could be used for large research work can also be a didactic tool in all a lot of degrees: Sciences of Communication, Translation and Linguistics... Its multiple functions may reflect the usage of language in the media, the reality of electronic writing and its criteria, the presence of adaptation...

Through the network, it can be used to create new applications for the Internet, in order to develop the semantic web, to facilitate the search and retrieval of documents, and above all, it will allow us to exploit the web as a corpus in an effective way.

REFERENCES

Bhatia, V.K. (1993). *Analysing genre; Language use in professional settings.* London: Longman.

Bhatia, V.K. (2002). Applied genre analysis: a multi-perspective model. *Asociación Europea de Lenguas para Fines Específicos* 3-19.

Biber, D., & Kurjian, J. (2006). *Towards a taxonomy of web registers and text types: a multi-dimensional analysis*. http://www.ingentaconnect.com/content/rodopi/lang/2006/00000059/00000001/art00007.

[2] http://www.hizlan.org/db-hedabideak.

Bronckart, J.-P. (2004). *Actividad verbal, textos y discursos. Por un interaccionismo socio-discursivo.* Madrid: Fundación Infancia y Aprendizaje.

Calvi, M.V. (2010). Los géneros discursivos en la lengua del turismo: una propuesta de clasificación. *Asociación Europea de Lenguas para Fines Específicos* 9-32.

Canhavilas, J. (2007). *Webnoticia. Propuesta de modelo periodístico para la www.* http://www.livroslabcom.ubi.pt/pdfs/canavilhas-webnoticia-final.pdf

Ciapuscio, G. (2003). *Textos especializados y terminología.* Barcelona: Institut Universitari de Lingüística Aplicada.

Díaz Noci, J. 2001. *La escritura digital. Hipertexto y construcción del discurso informativo en el periodismo electrónico.* Leioa: UPV/EHU.

Díaz Noci, J. & Salaverría, R. (2003). *Manual de Redacción Ciberperiodística.* Barcelona: Ariel Comunicación.

Erickson, T. (1999). Rhyme and Punishment: The Creation and Enforcement of Conventions in an On-Line Participatory Limerick Genre. *Proceedings of the Thirty-second Hawaii International Conference on System Sciences.* Maui, Hawaii.

Ezeiza, J. (2009). Herramientas para la compilación, estudio y gestión de la producción lingüística en la universidad: una aproximación didáctica y social. In: Caridad de Otto, E. & López de Vergara (comp.). *Las lenguas para fines específicos ante el reto de la Convergencia Europea.* La Laguna: Universidad de la Laguna, 553-567.

Ezeiza, J. (2010). "DB (Dokumentu Biltegia): Corpus akademikoak sortzeko eta kudeatzeko azpiegitura teknologikoa". In: Alberdi, X. & Salaburu, P. *Euskararen garapena esparru akademikoetan.* UPV/EHU: Leioa: 168-190.

Franco, G. (2009). *Cómo escribir para la web. Base para la discusión y construcción de manuales de redacción "on line".* Texas: Centro Knight de la Universidad de Texas.

Ferris, S. (2002). "Writing Electronically: The Effects of Computers on Traditional Writing". *Journal of Electronic Publishing*. http://quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;view=text;rgn=main;idno=3336451.0008.104

García, J. C., (2005). "Escribir para la red". In http://usalo.es/imagenes/Escribir_para_la_red.pdf

Hassan, Y. & Martín, F. (2002). "Escritura Hipertextual". *No Solo Usabilidad journal*, 1. http://www.nosolousabilidad.com/articulos/escritura_hipertextual.htm

Hund, M., al. (2007). *Corpus linguistics and the web.* Amsterdam–New York: Rodopi

Lamarca, Mª J. (2006). *Hipertexto: el nuevo concepto de documento en la cultura de la imagen.* Madril: Universidad Complutense de Madrid. In http://www.hipertexto.info

Lim, C. S., Lee, K. L., & Kim, G. C. (2004)."Multiple sets of features for automatic genre classification of web documents". *Elsevier; Information Processing and Management* 1263-1276.

Manchón, E. (2003). *Escribir y redactar contenidos para internet.* Avalaible at: http://alzado.org/articulo.php?id_art=54

Martin, J. R. (1993). *A Contextual Theory of Language. In The Powers of Literacy -- A Genre Approach to Teaching Writing.* Pittsburgh: University of Pittsburgh Press.

Meyer zu Eissen S. and Stein B. (2004), "Genre Classification of Web Pages: User Study and Feasibility Analysis". In: Biundo S., Fruhwirth T., Palm G. (eds.), *Advances in Artificial Intelligence.* Berlin: Springer, 256-269.

Miller, C. R. (1984). "Genre as social action". *Quarterly Journal of Speech* 151-167.

Montesi, M. (2005). "Géneros web: líneas de investigación". *El profesional de la información* 17-5.

Renouf, A., Kehoe, A. (2009). *Corpus linguistics: refinements and reassessments.* New York: Rodopi.

Reppen, R., al., (2002). *Using corpora to explore linguistic variation.* Amsterdam: John Benjamins

Rodríguez Betancourt, M. (2004). "Géneros periodísticos: para arropar su hibridez". *Estudios sobre el Mensaje Periodístico* 319-328.

Rosso, M. (2008). "User-Based Identification of Web Genres". *Journal of the American Society for Information Science and Technology* 1053-1072.

Santini, M., & Sharoff, S. (2009). “Web Genre Benchmark Under Construction”. *Journal for Language Technology and Computational Linguistics* 129-145.

Santini, M., Sharoff, S., Rehm, G., & Mehler, A. (2007-10-4). *Web Genre Wiki.* In http://www.webgenrewiki.org

Swales, J. M. (1990). *Genre Analysis - English in Academic and Research Settings.* Cambridge: Cambridge University Press.

Yus, F., (2010). *Ciberpragmática 2.0. Nuevos usos del lenguaje en Internet.* Madrid: Ariel.