QUANTITATIVE DATA AND GRAPHICS ON LEXICAL SPECIFICITY AND INDEX OF READABILITY: THE CASE OF *WIKIPEDIA*

ANTONELLA ELIA
UNIVERSITÀ DEGLI STUDI DI NAPOLI "L'ORIENTALE"

Abstract
This paper is part of a wider corpus based study focused on Web encyclopedias (Elia 2008). It is built on and extends the comparative analysis of Emigh and Herring (2005). In particular, attention is focused on the English edition of *Wikipedia*. A quantitative analysis compares *Wikipedia* vs. *Britannica* encyclopedic entries. Linguistic features such as type/token ratio, word and sentence length, and Index of Readability are analyzed. The findings show to what extent collaboratively produced Wikipedia entries are readable and standardized in a way not very dissimilar from those produced by experts in the *Encyclopaedia Britannica Online*.

Keys words: Online Encyclopedias, Index of Redability, Discourse Analysis, *Wikipedia*, *Encyclopedia Britannica Online*, Quantitative Analysis

Résumé
Ce document fait partie d'un corpus d'études plus vaste sur les encyclopédies en ligne (Elia 2008) et il se base sur l'analyse comparative d'Emigh et Herring (2005). L'attention est centrée spécialement sur l'édition anglaise de Wikipedia. Une analyse quantitative compare les entrées des encyclopédies *Wikipedia* et *Britannica*. Les traits linguistiques comme le « type-token ratio », la longueur de mots et de la phrase, et l'index de lisibilité sont analysées. Les résultats montrent dans quelle mesure les entrées de Wikipedia produites en collaboration sont lisibles et normalisées d'une manière qui n'est pas très différentes des entrées produites par des experts de l'*Encyclopédie Britannica* en ligne.

Mots clé: Encyclopédies en Ligne, Index de Lisibilité**,** Analyse du Discours, *Wikipedia*, *Encyclopaedia Britannica Online*, Analyse Quantitative

1. Introduction

The advent of home computers has undermined shelf-load encyclopedias and door-to-door salespeople have become extinct as a working class. Encyclopedias, published as multi-volume sets of books for centuries, have been transformed since the 1990s into inexpensive CDs or DVDs integrating sounds, pictures, animation and text. Then, at the beginning of the new millennium, they migrated on the web. Portals, search engines and web directories have nowadays progressively transformed the way people search for information. Nevertheless, this has not made encyclopedias obsolete. On the contrary, reference works are needed more than ever to help searching and filtering in the jungle of the information overload. Entries to traditional encyclopedias, such as *Britannica*, are written by individual scholars, professionals, and experts whereas articles in *Wikipedia* are collaboratively written by volunteers who are, sometimes, anonymous contributors.

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 248-271
Recibido: 24/07/2009
Aceptación comunicada: 07/09/2009

The differences in the authorial and writing processes in the two above mentioned online encyclopedias have stimulated this paper and contributed to empirically identify the lexical specificity and index of readability of the two online reference works.

Lexical density (type/token ratio), word length and sentence length are the three factors which define the lexical specificity of a text thus influencing its linguistic formality. These three features have been investigated to define the lexical specificity of *Britannica* and *Wikipedia* in selected corpora.

With reference to lexical specificity, Chafe and Danielewicz (1987) claim that speakers tend to operate with a narrower range of lexical choices than writers while Biber (1988) states that a higher lexical specificity seems to be associated with formal written genre, marking a high density of information, by reflecting precise word choice and an exact presentation of informational content.

As mentioned above, the other aspect analyzed in this paper is that of Index of Readability. This index is designed to gauge the understandability of a text through the measurement of semantic (difficulty of words) and syntactic (difficulty of sentences) factors. Its output is a representation of the education grade level needed to comprehend a text. In order to calculate the Index of Readability different formulas can be used (Flesch-Kincaid Grade Level, Gunning-Fog Index, SMOG Index, Fry Readability Formula, and Coleman-Liau Index, etc.). The Gunning-Fog Index has been used for the analysis presented in this paper.

2. Research questions

The specific research questions which have been identified and for which this paper tries to provide the answer are the following:

1. to what extent are lexical density, word length and sentence length similar or different in the two online encyclopedias?

2. is Index of Readability, as a quantifiable parameter, equal or divergent in *Britannica* and *Wikipedia*? And if so, to what extent does it differ?

3. do the different authoring processes affect the style of the encyclopedic genre, as exemplified in *Wikipedia* and *Britannica Online*, in terms of lexical specificity and readability?

3. Expository writing

Expository writing, which is typical of the encyclopedic genre, is a mode of writing in which the purpose of the author is to inform, explain, describe or define the subject matter to the reader. According to Ball (1992), a well-written presentation is one that remains focused on its topic and provides facts in order to inform its reader. It should be unbiased and accurate, and it should use a scholarly third person tone. In addition, an expository text needs to encompass all aspects of the subject.

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 248-271
Recibido: 24/07/2009
Aceptación comunicada: 07/09/2009

In the creation of expository texts, writers cannot assume that readers have prior knowledge or former understanding of the topic that is discussed. An important point authors should always keep in mind is to use language that clearly shows what they are talking about. Since clarity requires a high degree of organization, one of the most important mechanisms used to improve the presentation of facts is to give the text a precise structure (definition, description, sequence, classification).

Written discourse, as opposed to the rapidity and dynamicity of spoken discourse, is static (Chafe and Danielewicz 1987). The effect of such diversity seems to generate differences in the linguistic production. One of the main aspects of linguistic production concerns vocabulary use. The different use of vocabulary can be empirically measured through lexical density (type/token ratio), word length and sentence length.

A high type/token ratio reflects the use of many different words in a text (vs. extensive repetition of relatively few words), representing a more careful word choice and a more precise presentation of informational content. Halliday (1985) also considers a high lexical density typical of formal writing.

Biber (1988), by analyzing his selected corpus, shows that Academic Prose (e.g. research papers, Ph.D. theses, academic reports, etc.) which he considers the most formal genre in the scale of informational production, has a lexical density of 50.6%.

Furthermore, Biber *et al.* (1998) claim that longer words and a high lexical density frequently co-occur in formal written genres. They affirm that longer words convey a more specific and specialized meaning than the shorter ones. In his *Multidimensional Analysis* Biber (1995) finds the average word length of Academic Prose to be of 4.8 characters. Zipf (1949), considered to be one of the first pioneers in linguistic quantitative analysis, shows that words become shorter in English when they are more general in meaning and more frequently used.

One of the most noticeable and consistent properties of formal (including academic) production is that text is produced in longer units. This happens because formal written texts can go through planning and editing. By contrast, involved production, mainly made up of informal, interactive and oral texts, should be characterized by shorter and simpler units. Chafe and Danielewics (1987) claim that writers, unlike speakers, have the time and leisure to perfect sentence structures and thus make them more coherent. There are many linguistic devices whose effect is to increase the size of written units, such as prepositional phrases, nominalizations, attributive adjectives, etc. According to Chafe and Danielewics, academic writing shows a relatively normal distribution of sentence lengths centred around an average of 24 words as if writers possessed an intuitive concept of *normal sentence* length. On the other hand, the average length of spoken utterances is of 18 words. Nevertheless, not all the linguists agree with the assumptions of Chafe and Danielewics. Ong (1982), Tannen (1989), Sheperd and Watters (1998) have interpreted the distributional pattern which conveys maximum content in the fewest word as marking an exact presentation of information.

3.1 Encyclopedias as genres of expository writing

It is this author's point of view that the main peculiarity of the expository writing used in the encyclopedic genre seems to be its stylistic formality, objectivity and impersonality. The

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 248-271
Recibido: 24/07/2009
Aceptación comunicada: 07/09/2009

voice of the author(s) seems to disappear behind the presentation of facts and information (Elia 2008). Articles are written in the third person, are unsigned, highly informational, abstract in content, explicit and "context independent" (Hall 1976). Furthermore, as the central purpose of encyclopedias is pedagogical, the degree of readability of the entries should be very high. Encyclopedias should not be considered as student's manuals since their readers are an already educated public, a publique éclairé, as Diderot and D'Alembert say, curious and intelligent readers, as stated in the Preface of the Britannica (Pombo 2006). Heylighen and Dewaele (1999) also point out that expository and informational production is strictly related to the concepts of space and time. The wider the spatial setting between sender and receiver is, the smaller the shared context and higher the formality of the text produced will be. The same happens when the time span between sending and receiving is longer. The longer the time span, the more formal the text will be. In this case, less will remain of the original context in which the discourse has been produced, while a more explicit, precise and context-independent textual production will be needed. To conclude, audience size, different writers/readers' cultural backgrounds, settings of production and reception, time span and the need for understanding are factors which have to be considered if online encyclopedias are to be effective, fluently readable and easily comprehensible in order to fulfill their main educational and popular purpose. All the above mentioned variables influence the stylistic formality and the readability of encyclopedic texts.

3.2 Writing readable web encyclopedias

Since encyclopedias have a pedagogical and popular purpose, they should be written with readability in mind. The early research on readability was conducted only on traditional printed texts. Nowadays, there are new and additional elements which need to be taken into account when considering digital genres. When websites and, in this specific case, online encyclopedias are assessed, it is essential to consider online readability in terms of both online *content readability* and *web usability.*

Reading a text is mainly a left-brain activity. During the last thirty years, as well as traditional readability formulas connected to writing content, there has been great interest in the graphic aspects of writing that appeal also to the right side of the brain. They include editorial design, layout, symmetry, the use of illustrations, colours, blank spaces, graphs, bulleted lists, etc. They are essential factors which have to be considered to comprehensively understand the nature of a text.

Most online readers surf the web because they are looking for specific information, and they do not find it by reading a Web page word by word but rather by scanning the page for relevant items. For this reason it is important to make use of some basic stylistic conventions when writing and structuring webpages. The *Web Style Guide* site (Lynch *et al.* 2002) provides the following advice:

> - *Be frugal.* Do not use the first paragraph of each page to tell users what information they will find there. Instead, start with the information, written in a concise and factual prose style.
> - *Stick to the point.* Write in easily understood sentences. Steer clear of clever headings and catchy but meaningless phrases that users must think about and explore further to understand.

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 248-271
Recibido: 24/07/2009
Aceptación comunicada: 07/09/2009

- *Think globally*. Remember that you are designing documents for the World Wide Web and that your audience may not understand conventions specific to your little corner of the world. Also, avoid metaphors and puns that may make sense only in the context of your language and culture.

Web authors have also to bear in mind that reading from computer screens is more tiring for the eyes and about 25% slower than reading from printed papers (Nielsen 2006). Thus, the clearer the style of writing, the easier it will be for the site visitors to absorb what has been written on the webpage. Some techniques for using clear and simple language include, for example, the avoidance of slang or jargon expressions, the use of shorter words where possible, the avoidance of complex and ambiguous sentence structures, omission of needless words, inclusion of just one idea or concept per sentence, the use of active instead of passive verbs, and the organization and structuring of information in an orderly and logical way. Tailoring texts in less complicated sentences, using an objective language, and one that is at an appropriate reading level for the target audience improves textual readability.

For their linguistic peculiarities and communicative purpose, encyclopedias can be fully included in the category of informational production, as their main aim is to inform, to educate and to present facts and information in specific entries. Biber (1988), through his *Multidimensional Analysis*, has mapped linguistic feature patterns in different typologies of spoken and written English texts. He claims that expository texts are informational, detached, elaborated, highly explicit and context independent. They are characterized by the need for precise and dense packaging of information. Furthermore, he claims that frequent nouns and attribute adjectives, longer words and a high type/token ratio, can be associated with a high informational focus and a careful integration of information in a text, and a consequential formal expository register.

4. Index of readability

The key function of the encyclopedic genre is educational. Thus, articles should provide a general overview on a specific subject through an understandable and accessible expository style. Most of readers of paper encyclopedias seemed to be traditionally school learners, however since nowadays online readers of internet encyclopedias seem to collect a more varied audience, it is essential that texts be written in a clear, linear and comprehensible way in order to be easily understood and thus fulfill their primary pedagogical purpose.

MacCormick *et al*. (1982) tested many different encyclopedias (including Britannica) for readability. They proved that encyclopedias written by experts required high levels of reading skills in order to be easily understood. But, does *Encyclopaedia Britannica Online* carry on this tradition also in the new millennium? Is Britannica nowadays more popular and readable than thirty years ago? By contrast, to what extent does Wikipedian collaborative writing affect readability? When articles are edited in *Wikipedia*, their Index of Readability is not automatically tested, and this factor could generate articles of mixed readability levels which would presumably be extremely different from the highly monitored system used by *Encyclopaedia Britannica*. This hypothesis can be tested quantitatively by comparing the Index of Readability in *Wikipedia* and *Encyclopaedia Britannica*.

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 248-271
Recibido: 24/07/2009
Aceptación comunicada: 07/09/2009

Reading a text demands focus, word recognition, decoding linear processing, and prediction of outcomes. Merriam-Webster's dictionary (electronic edition, 2006) defines a *readable* text as one that is *easy to be read, interesting, agreeable, attractive in style and enjoyable*. It is clear that some textual qualitative factors (e.g. tone, complexity of ideas, page design, textual comprehensibility or obscurity, textual cohesion and coherence, interest, appeal and enjoyment aroused in the reader) cannot be measured through mathematical formulas. Readability formulas offer the opportunity to assess only the surface characteristics of texts. They evaluate features that can be subjected to mathematical computation such as semantic (the difficulty of words) and syntactic (the difficulty of sentences) factors. As already pointed out, lexical density, as well as word and sentence length influence stylistic formality. Complex texts often contain difficult and long words because they discuss abstract ideas, whereas easy texts use common and short words as they focus on concrete experiences. Although only sentence and word lengths and complexity of linguistic structures can be measured by *Readability* formulas, these semantic and syntactic factors are fundamental because they definitely affect text readability in a significant way. For this reason *Gunning Fog* Index, one of the most well known readability formulas, has been chosen among the many available indexes (e.g. *Dale-Chall, Flesch-Kincaid, Fry,* etc.), in order to assess the Index of Readability of *Britannica* and *Wikipedia* corpora.

4.1. The Gunning Fog formula

In 1952 Robert Gunning created one of the most popular readability formulas. It predicted the difficulty of a written passage with an 80% accuracy. The formula is the following:

$$0.4 * \left(\left(\frac{\text{words}}{\text{sentence}} \right) + 100 \left(\frac{\text{complex words}}{\text{words}} \right) \right)$$

This formula indicates the reading skill (based on grade level) necessary to understand a text on the first reading (of course, the lower the number, the more understandable the content will be to the reader). Gunning Fog formula is easy to apply. It is based on the calculation of (1) average sentence length and (2) the percentage of the polysyllabic words contained in a text. (1) and (2) have to be added and the sum must be multiplied by 0.4. The formula is an objective tool for measuring readability and it predicts quite satisfactorily the difficulty of a text.

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 248-271
Recibido: 24/07/2009
Aceptación comunicada: 07/09/2009

Fog Index Scores	
Score	**Resources**
6	TV guides, The Bible, Comic books
8	Reader's Digest, Ladies' Home Journal
8 - 10	Most popular novels
9	Reader's Digest
10	Time, Newsweek
11	Wall Street Journal
12	Atlantic Monthly
14	The Times, The Guardian
15 - 20	Academic papers

Table 1. Gunning Fog Index Scores

The Fog Index scores of some American resources are shown in Table 1. They have been provided by dr. Philip Chalmers[1], to help establish and assess the readability of textual documents. For example if a text has a score of 12, it means that it has the reading level of a U.S. high school senior. Texts designed for a wide audience generally require a *Fog Index* of less than 12. Score 17, for example, indicates a level of textual difficulty at postgraduate level.

5. Wikipedia: a general overview

Nowadays *Britannica* and *Encarta* dominate the encyclopedia market due to their perceived high authorial quality. However, the claims *Britannica* and *Encarta* make about their products suggest more emphasis on their own marketing strategies than interest in the needs and demands of consumers (Panagiota 2002). Since the beginning of the new millennium, *Wikipedia*, a free, multilingual, web-based encyclopedia operated by *Wikipedia Foundation* represents one of the popular online phenomena. It is a freely available web-based, free-content, co-authored multilingual encyclopedic project, operated by the *Wikimedia Foundation* (Sloane 2007). As indicated on the *Wikipedia* website (http://www.wikipedia.org), in September 2009, *Wikipedia* had approximately more than 10 million articles in 253 languages. It can be considered one of the largest international virtual communities. The English edition, being made up of some 3,041,000 articles, is the largest edition and will most probably remain so in the future.

Looking at the recent statistics on the number of articles, the English edition of *Wikipedia* is over 20 times larger than *Britannica*. A key difference between the two encyclopedias mainly lies in article authorship. *Britannica*'s articles are generally written by recognized contributors, and are the product of an editorial staff and internal or external consultants. Moreover, most of *Britannica*'s contributors are experts in their field and some

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 248-271
Recibido: 24/07/2009
Aceptación comunicada: 07/09/2009

of them are also Nobel laureates. By contrast, the articles in *Wikipedia* are written by a community of volunteer editors with different levels of expertise. Most of these editors do not claim any particular expertise and many of them are anonymous, with no verifiable credentials. For this reason, it has been argued that *Wikipedia* cannot hope to compete with *Britannica* in accuracy (McHenry 2004). *Wikipedia* relies on the authority of peer-reviewed publications rather than on the personal authority of experts. It does not force its contributors to give their names to establish their identity. However, even if some contributors are authorities in their fields, *Wikipedia* only requires that information provided be supported by published and verifiable sources.

Wikipedia is based on a democratic and spontaneous system of information classification defined as "folksonomy. Emerging as an alternative to the more traditional taxonomy, it defines a practice of collaborative categorization using freely chosen keywords, where authors/readers cooperate spontaneously organizing information into categories.

Although collaborative creation and organization have been in practice since biblical times, with scribes transcribing and at the same time often editing, updating, interpreting or reinterpreting original texts, open access to large scale public collaborative content creation projects is a relatively recent phenomena (Stvilia *et al*. 2005).

Lowry *et al.*'s taxonomy (2006) has been used as a framework of reference to define *Collaborative Writing* (CW) in online encyclopedias and to identify the typology of writing carried out in *Wikipedia*. Lowry *et al.* (2006) define *Reactive Writing* (RW) as the strategy that occurs when writers create a document in real time, reacting and adjusting to each other's changes and additions without significant preplanning and explicit coordination. The term RW is used since written reaction may involve consensus or dispute, reflections or spontaneous contributions. For example, while some authors write a section in a Wikipedian entry, others may simultaneously review it and create new sections in response that may contradict or concur with the first authors' points of view. Advantages of *RW* include the possibility of building consensus through free expression and the development of creativity. The primary drawback of this strategy is that it makes coordination difficult and can cause difficulties with version control.

5.1 Norms of authoring

The *shared* control mode in *Wikipedia* offers all contributors simultaneous and equal access and writing privileges throughout the writing activity. This can be a highly effective, non-threatening form of control in groups that work face-to-face, engage in frequent communication and have high levels of trust. Nevertheless, this collaborative mode can lead to conflict in groups whose members work far away from one another (as sometimes happens in Wikipedian *edit wars*). One might usefully add the term *Massively Distributed Collaboration* to the definition of *Reactive Writing* as applied to *Wikipedia.* This term, used for the first time by Mitchell Kapor (2005), describes an emerging activity in content-creating virtual communities (e.g. mailing lists, blogs, wikis, etc.). This definition is nowadays applied in different domains (such as education, research, music, corporations, political action, etc.) and its main purpose is that of assembling a body of information that can be re-used later by the same contributors and by others.

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 248-271
Recibido: 24/07/2009
Aceptación comunicada: 07/09/2009

When writing in *Document Mode* pages, Wikipedians create and contribute to collaborative documents leaving their additions to wiki documents. The style used in encyclopedic entries (defined in *Wikipedia* as *articles,* henceforth WAs) is explicitly promoted in an official *Manual of Style*, which is the official framework of reference for all *Wikipedia's* contributors. According to *Wikipedia*'s *Manual of Style*, articles must firstly observe core principles of cooperation and objective writing based on three absolute and not negotiable principles, *Neutral Point of View* (NPOV), *Verifiability* (V) and *No Original Research* (NOR). These three policies determine the type and quality of material acceptable in the encyclopedic articles written in Document Mode. Document Mode WAs are coherent and self-contained. It seems that *Wikipedia*'s contributors follow the Manual of Style. Encyclopedic expository style is formal, in that it never makes use of first and second personal pronouns, acronyms, jargon expressions or neologisms, since their use is rigorously forbidden. Entries are impersonal and detached, accurate, rigid, refined and formal. Furthermore, they are unsigned, highly informative, objective, and respectful of stylistic conventions (Elia 2008). Emigh and Herring (2005) found that personal pronouns are avoided altogether in favor of impersonal constructions.

Wikipedia has attracted a widespread response, both positive and negative, from scholars. It has been faulted for lacking comprehensiveness and consistency, that are said to be endemic in volunteer–based projects that reflect the biases and interests of their contributors (Waldman 2004). Critics have also complained that being an open space that does not employ strict rules for citation of sources, *Wikipedia* is fundamentally unreliable (Schiff 2006). Others suggest that *Wikipedia* is reliable most of the time, but it is not always clear to what extent (Boyd 2005). Editors of traditional reference works, such as *Encyclopaedia Britannica,* have contested the project's utility and status as an encyclopedia (McHenry, 2004). Concerns have also been raised on the lack of accountability resulting from users' anonymity, the vulnerability to vandalism, and so forth. In addition, other critics claim that *Wikipedia*'s open structure makes it an easy target for advertisers (Sanger 2006). Ahrens (2006) has also noted the addition of news to articles by political organizations, including the U.S. House of Representatives. The most visible and public criticism of *Wikipedia* has been conveyed by Lanier (2006) who views *Wikipedia*'s growing importance as "recrudescence" of the concept of collective intelligence. According to Lanier, the idea that the greatest wisdom is collective wisdom can become a dangerous tool in the hands of any extreme ideology.

5.2 The Literature on Wikipedia

Positive appreciations have been expressed by contrast in the article published in the journal *Nature* by Giles (2006)*,* whose analysis results were widely seen as a validation of *Wikipedia*'s content and methods. Lih (2004) has studied *Wikipedia*'s content construction from the perspective of participatory journalism. Resnick *et al.* (2005) have highlighted the Wiki structure and its advantages in relation to other more traditional forms of online communication. Other approaches have stressed the productive power of Wiki discussions in collaborative knowledge creation (Lawler 2005; Shah 2005). In particular, Joseph Reagle (2006) has explored the character of "mutual aid" and interdependent decision-making

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 248-271
Recibido: 24/07/2009
Aceptación comunicada: 07/09/2009

within *Wikipedia* and Holloway *et al.* (2008) have reported a semantic analysis of it. The rapid growth of *Wikipedia* has also been the subject of study, as for example by Capocci *et al.* (2006), who have used the social network modeling of *Wikipedia* to predict its growth patterns.

The study that is most relevant to the research work presented in this paper was conducted by Emigh and Herring (2005) who made an interesting linguistic comparison between a traditional encyclopedia, *Columbia Encyclopedia*, whose articles are written by experts, and two community-based encyclopedias, *Wikipedia* and *Everything2*, that are authored by volunteers. Using corpus linguistic methods and factor analysis of word counts for features of formality and informality, Emigh and Herring showed that, in the presence of a greater degree of post-production editorial control afforded by the system, the formal and standardized language of the collaboratively-authored documents becomes more analogous to that found in traditional print encyclopedias. Thus, *Wikipedia* and Columbia Encyclopedia appear to be statistically indistinguishable in the use of a formal style measured by the use of nominalizations, suffixes and the non-use of personal pronouns and contractions. Moreover, it was suggested that users who faithfully appropriate the *Wikipedia* system create homogeneous entries, which is at odds with the goal of open-access authoring environments (that goal being to create diverse content). Their findings shed light on how users, acting through mechanisms provided by the system, can shape features of content in particular ways. They concluded by identifying sub-genres of web-based collaborative authoring environments based on the technical affordances of such environments.

6. Methodology

This paper presents a corpus based study which, by way of a descriptive approach, empirically analyses and compares two different subcorpora. The overall corpus of reference consists of 638,740 tokens. It comprises two subcorpora: *Britannica* and *Wikipedia* encyclopedic subcorpora with respectively, 247,103 and 391,637 tokens.

The Encyclopedic corpus is made up of an equal number of articles that are collected from both the websites of *Encyclopaedia Britannica* and *Wikipedia*. The two subcorpora include one hundred articles randomly selected from the ten categories of *Wikipedia* folksonomy and by the one hundred equivalent articles found in *Encyclopaedia Britannica Online*. The selection includes encyclopedic articles of different quality and at different evolution stages as testified by the identification label used by the Wikipedian department of the *Heraldry and Vexillology* (which assesses the quality of *Wikipedia*'s articles). At the end of 2007, the 100 selected articles were distributed as follows: forty-eight articles belonged to the ★*FA Class* (featured articles – the best ones), twenty-two of them to the *A-class* (articles with well written texts and contents), sixteen to the ⊕ *GA Class* (good article) and fourteen articles to the *B-Class* (articles to be improved). In the selection of the 100 articles those belonging to the *Start* and *Stub Class* because too short and incomplete and thus incomparable have been excluded. The former class collects articles too weak in many areas and which do not provide fundamental information, while the latter gathers articles too short to provide encyclopedic coverage of the subject. Despite the persistent changes in *Wikipedia*

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 248-271
Recibido: 24/07/2009
Aceptación comunicada: 07/09/2009

folksonomy, its basic taxonomy is not very dissimilar from *Britannica*'s which has been, since the beginning of this study, consistently structured in ten steady 'subjects' and subdivided into more specific and equally stable subcategories.

More specifically, in April 2007 *Wikipedia*, was articulated in the following ten categories: Art, Biography, Culture, Society, Geography, History, Mathematics, Philosophy, Science and Technology. Entries to *Britannica* were (and still are) organized in ten similar categories: Arts and Literature, The Earth and Geography, Health and Medicine, Philosophy and Religion, Sport and Recreation, Science and Mathematics, Life, Society, Technology and History.

The folksonomy of *Wikipedia*, as well as its encyclopedic articles are dynamic and, thus, in constant evolution. This aspect has obviously represented a critical point in data collection and cataloguing. For example *Wikipedia*'s Folksonomy was made up of ten categories when the first survey was carried out (April 2007). During later years some slight changes were introduced in the name categories and furthermore the ten categories became twelve in November 2008 and still are at the time of writing of this article (March 2009) (Figure 1).

Reference to the original classification system (2007) has been maintained during the present investigation for the following different reasons; firstly, because there was a natural numerical and topical matching with the ten *Britannica* categories, secondly because the evolution of *Wikipedia*'s Folksonomy is extremely fluid and too fast to be constantly followed and, above all, because it was not influential to the objectives of the present work. Ten sample articles have been randomly selected from each category and analysed. The advantage of representativeness and generalization was offered by the random technique. The choice of the same number of articles taken from the two encyclopedias has given topical coherence to the present investigation. The two hundred articles which were chosen are shown in Table 2. The first phase of this study was devoted to data collection and rationalization and to corpus definition in order to create a coherent and representative corpus made up of a collection of encyclopedic articles which were cleaned out by removing information irrelevant to the specific content body (e.g. index of contents, graphs, photos and tables, references and extra links). Most of the quantitative analyses (type/token ratio, word and sentence length, etc.) were carried out mainly by using *Antconc* (a concordancer program developed by Laurence Anthony, at the School of Science and Engineering, Waseda University in Japan).

Figure 1. *Wikipedia* categories

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 248-271
Recibido: 24/07/2009
Aceptación comunicada: 07/09/2009

Arts	Biography	Culture	Society	Geography
Cinemascope	Beatles	Diaspora	Alcoholism	Barcelona
Colosseum	Benjamin Franklin	Fairy tale	Euro	Bermuda triangle
Graffiti	Bill Gates	Flag	Feminism	Gobi desert
Holography	Albert Einstein	Geisha	Homosexuality	Hydrography
Proscenium	Fred Astaire	Jazz Dance	Women's suffrage	Himalaya
Jazz	James Dean	Pizza	Poverty	Klodzko
Madonna	Karl Marx	Romanticism	Racism	London
Polka	Adam Smith	Superstition	Tamil	Piccadilly Circus
U2	Vittorio Alfieri	Tea	Terrorism	San Josè
Wind rose	C. Columbus	Walt Disney	Zulu	Weather

History	Mathematics	Philosophy	Science	Technology
Anne Frank	Boolean algebra	Agnosticism	AIDS	Balloon
Aztec	Catastrophe theory	Aristotle	Big Bang	Gasoline
Silvio Berlusconi	Cryptography	Francis Bacon	Heart	Internet
Tony Blair	Graph theory	Epistemology	Neuron	Jet engine
Brit. East India C.	Matrix	Michel Foucault	Nuclear weapon	Microprocessor
Wars of the Roses	Numerical analysis	Frankfurt school	Pneumonia	Microsoft
Ku Klux Klan	Pythagorean theorem	Philosophy of mind	Royal Astr. Soc.	Radar
Garibaldi	Quantum number	Skepticism	Sars	Typewriter
French Revolution	Real number	Thomas Huxley	Solar energy	Virtual Reality
George Bush	Vector space	Wittgenstein	Turquoise	World Wide Web

Table 2. Articles in *Wikipedia* corpus

Antonc is a lexical analysis tool which can be used to search for keywords, perform concordance searches, etc. It has been mainly used to create word lists useful to compute the frequency of the linguistic classes considered functional for the purpose of this research. As a further control on the results the concordancer program *WordSmith Tools* (a proprietary software developed by Mike Scott at the Oxford University Press) has also been used.

Since the linguistic investigation is mainly frequency based, the count of occurrences was standardized to make the quantitative findings comparable. Standardization of frequency count was made following Biber's theory (1998: 263), which demonstrates that raw frequency counts are not directly comparable when textual units have different lengths.

In this study, standardization was made on the basis of 100 words per text. The selection of the 100 words has been determined by the length of the shortest article found. To this purpose the following formula was applied:

$$\textbf{Count of Occurrences : Tokens = X : 100}$$
$$\textbf{X} = \frac{\textbf{Count of Occurrences * 100}}{\textbf{Tokens}}$$

Thus, *absolute frequencies* (total occurrences) have been multiplied for the basis chosen for standardization (100 words) and then divided by the total number of words in the text (total tokens).

In calculating lexical density it has been noted that the ratio decreases as the number of words in a sample increases, therefore the ratio of text with different length is not comparable (Chafe 1987; Biber 1988). Many of the different words used in the first 100 words of a text are then repeated (Biber 1988), consequently in each additional 100 words the number of new types decreases since the relationship between text length and unique words (tokens/types) is not proportional. In fact, when the length of encyclopedic articles varies widely, as it frequently occurs in the specific encyclopedic entries analysed, the raw lexical density will appear to be much higher in the shorter text. Thus, when calculating lexical density in the microanalysis phase, and in order to have authentic and comparable data, the length of the longer article was reduced to the length of the shorter one.

The Index of Readability has been obtained by using the online text analysis tool *Textalyser,* hosted on the *lexicool.com* website, which has automatically computed the Gunning Fog Index of Readability for both *Britannica* and *Wikipedia* articles. The purpose of this specific investigation was to ascertain whether if the web readers of online encyclopedias had the reading skills necessary to easily understand its content. The measurement of course has also served the purpose of comparing the two resulting scores in order to verify discrepancies or similarities.

A specific microanalysis has been conducted separately on each of the 200 articles of the encyclopedic corpus in order to define the Index of Readability of every single analysed article. Then, the average Index of Readability was also generated for each specific encyclopedic subcorpus. In addition, the Microsoft program *Excel* has allowed the creation of dynamic data sheets and automatic classification and updating in case of data variation. Although *Wordsmith tools*, *Antconc* and *Lexicool.com* have allowed the measurement of the selected linguistic features by means of a statistical approach and have been very useful to facilitate the quantitative analysis, it has often been necessary to supplement the automatic analysis with a manual inspection to evaluate information in context.

7. Results

Type/token ratio, word length and sentence length are the three elements which were investigated to define *lexical specificity* in the *Wikipedia* and *Britannica* corpora. Lexical density has been measured through the type/token ratio. For example, the article *Graffiti* in

Britannica contains 406 tokens and 224 types. The type/token ratio is 224/406 and the raw lexical density is 55.2%. On the other hand, the parallel article in *Wikipedia* contains 4141 tokens and 1488 types. The type/token ratio is 1488/4141, and the resulting raw lexical density is 35.9% (Table 3), while the standardized lexical density is 55.2 in *Britannica* and 52.5 in *Wikipedia*.

Graffiti	tokens	types	Raw type/token ratio	Raw lexical density %	Standardized type/token ratio	Standardized lexical density %
Britannica	406	224	224	55.2	224	55.2
			406		406	100
Wikipedia	4141	1488	1488	35.9	224	52.5
			4141		427	100

Table 3. Graffiti article's Lexical Density

The results show that when lexical density has been calculated on similar samples, the standardized types/tokens ratio tends to be very similar in the two encyclopedias.

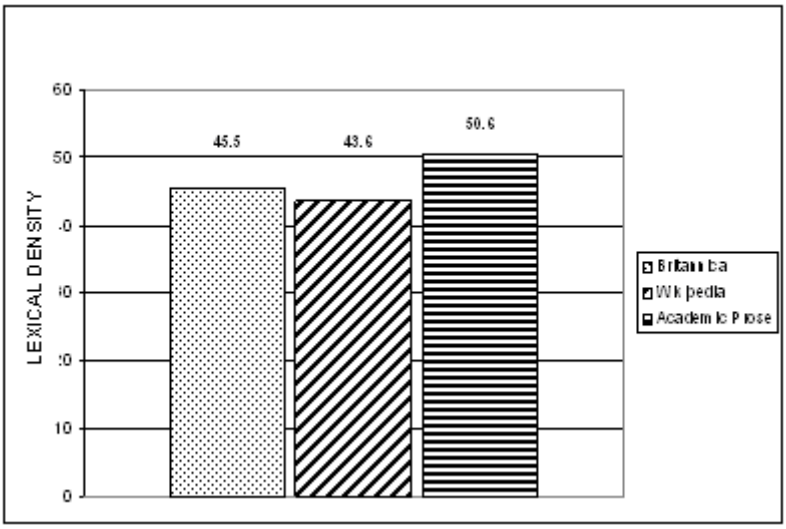

Figure 2. *Britannica*, *Wikipedia*, Academic Prose: Lexical Density

The micro analysis has thus highlighted the fact that the difference between the standardized lexical density of each pair of encyclopedic articles is similar in most cases. Except for 15 articles (out of 100) standardized lexical density is always slightly higher in *Britannica*. The findings of the micro analysis have been confirmed by the results of the macro analysis since the total standardized lexical density proves to be similar in the two corpora. More specifically, total standardized lexical density is 45.5% in *Britannica*'s and 43.6% in *Wikipedia*'s corpus. According to Biber's (1998), Halliday's (1985) and Chafe's (1987) theories, this means that lexical variety in the two encyclopedias is very similar. However, as the percentage is higher in *Britannica*, this data confirms the slight predominance of a linguistic variety (and consequently a more formal expository register) in *Britannica* when

compared to *Wikipedia*. Nevertheless, compared to the previously mentioned Biber's findings on academic prose, these results are lower. (Figure 2).

Hence, the formality of the encyclopedic genre, according to this specific linguistic feature, is not very different from, although measurably lower than, Academic Prose.

As far as average word length is concerned and on the basis of the theoretical assumptions mentioned above (Biber *et al.* 1998; Zipf 1949) a difference in the formal expository style has been detected by the measurement of word length in *Britannica* and *Wikipedia*. The measurement of the average word length has revealed that words in the two corpora have an equal average number of characters. The range goes from a minimum value of 3.9 (*Matrix* article) to a maximum of 6.7 (*Microprocessor* article) in *Britannica*, and from 4.4 (*Vector Space* article) to 6.1 (*Hydrography* article) in *Wikipedia*. Most of the articles (92/100 in *Britannica* and 81/100 in *Wikipedia*) have an average word length of 5 characters which corresponds to two/three syllables per word. More specifically, the average word length is 5.3 characters in *Britannica* and 5.2 in *Wikipedia*. This data reveals again that the results for the two corpora are very close.

Close average word length in *Wikipedia*, *Britannica* and Academic Prose has also been detected. These results are similar to the findings of Emigh and Herring (2005) who discovered an average word length of 5.04 in *Wikipedia* and 5.28 in Columbia Encyclopedia.

While lexical density may be lower, average word length proves to be slightly higher in encyclopedias than in Academic Prose.

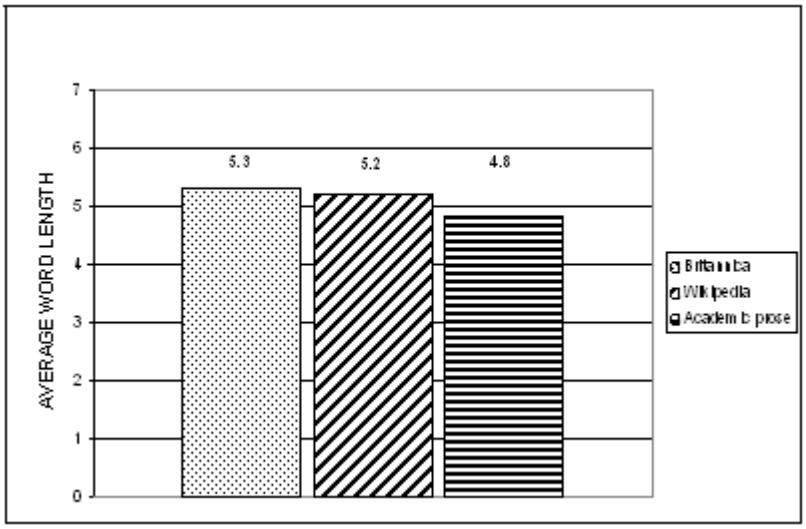

Figure 3. Average Word Length

Mot likely, the main reason for longer words (although minimal) found in encyclopedias is due to the pedagogical need of clarity, exactness and precision in the information delivery and to the need for conciseness in encyclopedic entries, which gives rise to more information packed into fewer longer words (Chafe 1987). Figure 3 shows a comparison for the average word length in *Britannica*, *Wikipedia* and Academic Prose.

On the basis of the mentioned divergent theoretical assumptions, (Ong 1982; Tannen 1989; Sheperd and Watters 1998) *Britannica*'s vs. *Wikipedia*'s sentence length has been

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 248-271
Recibido: 24/07/2009
Aceptación comunicada: 07/09/2009

measured. Apart from the different theoretical positions, what has been interesting for the purpose of this study is that sentence length appears to be very close in the two encyclopedias and not very different from the average sentence length of academic written texts, which Chafe (1987), as previously mentioned, found to be typically of 24 words. As shown in Figure 4 the average sentence length has proved to be very similar in the two corpora, although slightly longer sentences have been found in *Wikipedia* (22.09 words per sentence) than in *Britannica* (22.05 words per sentence).

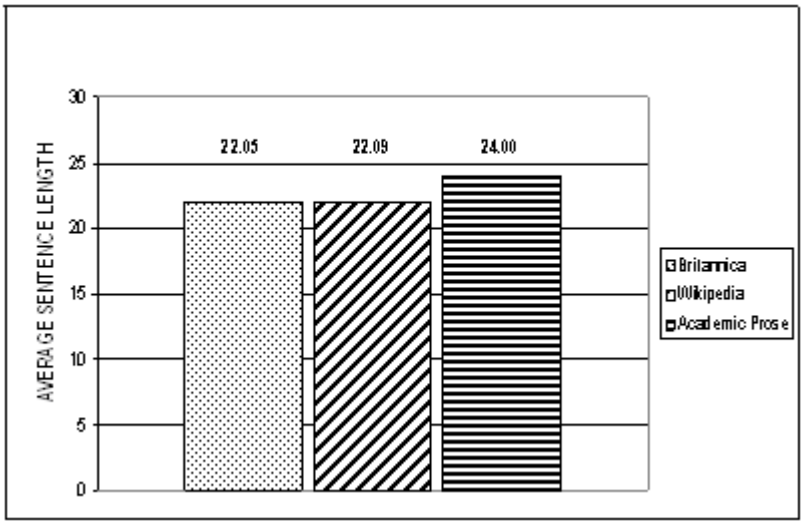

Figure 4. Average sentence length

However, the difference can be considered insignificant and the formality of the expository style, according to the selected criteria, is confirmed in the two encyclopedias by similar average sentence length. The micro analysis proves that the range from the minimum and the maximum number of words per sentence is very close. In fact, the shortest sentence has 14.4 words (*Vittorio Alfieri*), whereas the longest one 32.8 (*Racism*) in *Britannica* corpus.

By contrast, the shortest sentence has 14,7 words (*Geisha*) and the longest 35.8 (*Microsoft Corporation*) in *Wikipedia*.

The contrastive analysis has shown that there is an average difference of just two words (24 vs. 22 words) per sentence. Thus, encyclopedias make use of shorter sentences when compared to academic texts.

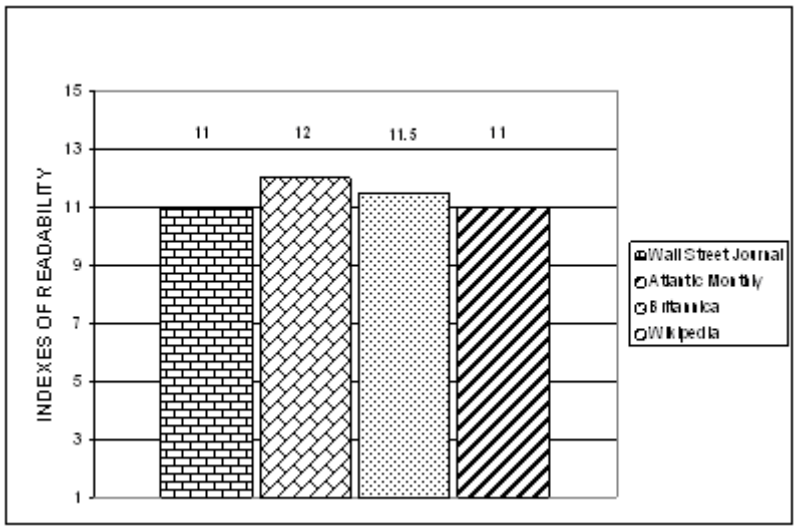

Figure 5. Different Indexes of Readability

The micro/macroscopic analysis seems to confirm, once again, that the formality of the two corpora is very similar and not far from the formal style conveyed in academic texts.The microanalysis on 200 encyclopedic articles has revealed, in most cases, very similar average Indexes of Readability in the two corpora (data shown in Table 4). The final average score is 11.5 in *Britannica* and 11 in *Wikipedia* (Figure 5).

According to Robert Gunning, texts designed for a wide audience and with a popular purpose in mind generally require a Fog index below 12.

Britannica and *Wikipedia* have revealed an average score (calculated on 100 articles from each corpus) very close to the *Index of Readability* of some of the most popular American magazines such as: *Time, Newsweek* (10), *Wall Street Journal* (11) and *Atlantic Monthly* (12) (Figure 5). Thus, the two encyclopedias seem to have successfully passed the test since the respective scores demonstrate that they should be comprehensible to a wide audience.

INDEX OF READABILITY			
	ARTICLE TITLE	BRIT.	WIKI.
ARTS	Cinemascope	11,6	13,3
	Colosseum	10,6	9,4
	Graffiti	12,2	6,6
	Holography	11,9	11,6
	Proscenium	12,7	8,0
	Jazz	14,0	11,5
	Madonna	12,0	9,5
	Polka	8,2	9,1
	U2	10,3	9,6
	Wind rose	10,6	10,4
BIOGRAPHY	Beatles	11,6	11,7
	Benjamin Franklin	11,5	10,9
	Bill Gates	12,4	10,0
	Albert Einstein	12,4	10,5
	Fred Astaire	10,5	10,2
	James Dean	12,8	10,9
	Karl Marx	10,2	12,9
	Adam Smith	14,6	11,6
	Vittorio Alfieri	8,4	12,5
	Christopher Columbus	10,7	10,4
CULTURE	Diaspora	13,1	12,0
	Fairy tale	11,6	11,9
	Flag	9,0	9,7
	Geisha	13,3	9,7
	Jazz Dance	9,8	8,1
	Pizza	8,4	8,9
	Romanticism	12,1	13,9
	Superstition	10,8	11,3
	Tea	10,4	9,6
	Walt Disney	12,2	11,1
SOCIETY	Alcoholism	12,0	12,8
	Euro	10,7	13,3
	Feminism	11,7	12,0
	Homosexuality	14,5	13,3
	Women's suffrage	10,6	10,4
	Poverty	14,5	10,7
	Racism	15,8	11,9
	Tamil	12,7	9,6
	Terrorism	10,9	11,7
	Zulu	10,4	8,6
GEOGRAPHY	Barcelona	10,3	11,7
	Bermuda triangle	8,5	11,1
	Gobi desert	9,4	11,0
	Hydrography	11,7	15,4
	Himalaya	10,4	8,9
	Ischia	7,7	9,5
	London	10,7	10,5
	Piccadilly Circus	10,2	11,6
	San Josè	9,8	9,1
	Weather	12,1	11,3

	ARTICLE TITLE	BRIT.	WIKI.
HISTORY	Anne Frank	11,0	11,2
	Aztec	10,5	10,5
	Silvio Berlusconi	11,5	13,3
	Tony Blair	7,5	11,7
	British East India Company	12,6	10,8
	George Bush	8,9	9,6
	French Revolution	12,0	11,3
	Giuseppe Garibaldi	12,1	10,7
	Ku Klux Klan	11,1	11,2
	Wars of the Roses	8,0	11,2
MATHEMATICS	Boolean algebra	14,2	8,0
	Catastrophe theory	15,3	15,0
	Cryptography	12,5	13,3
	Graph theory	10,7	10,8
	Matrix	9,9	9,0
	Numerical analysis	13,8	12,1
	Pythagorean theorem	11,3	10,0
	Quantum number	11,5	8,5
	Real number	12,6	9,3
	Vector space	15,4	7,3
PHILOSOPHY	Agnosticism	13,8	11,0
	Aristotle	13,4	11,6
	Francis Bacon	12,6	10,2
	Epistemology	10,7	10,7
	Michel Foucault	12,8	11,4
	Frankfurt school	11,9	15,3
	Philosophy of mind	13,1	11,8
	Skepticism	11,3	12,1
	Thomas Huxley	11,8	10,6
	Wittgenstein	11,8	10,2
SCIENCE	AIDS	14,7	11,4
	Big Bang	11,6	12,8
	Heart	9,5	9,6
	Neuron	7,2	10,9
	Nuclear weapon	9,6	11,8
	Pneumonia	11,3	11,7
	Royal Astronomical Society	12,7	11,7
	Sars	12,7	12,0
	Solar energy	11,0	10,6
	Turquoise	10,0	10,8
TECHOLOGY	Balloon	7,6	10,4
	Gasoline	15,0	10,0
	Internet	13,0	10,7
	Jet engine	13,0	11,4
	Microprocessor	11,5	11,8
	Microsoft Corporation	10,6	10,8
	Radar	11,3	11,3
	Typewriter	12,1	12,3
	Virtual Reality	13,2	13,7
	World Wide Web	10,3	10,3

Table 4. *Britannica* vs. *Wikipedia*'s articles: Index of Readability

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 248-271
Recibido: 24/07/2009
Aceptación comunicada: 07/09/2009

Table 5 shows the numerical distribution of Indexes of Readability in *Britannica* and *Wikipedia* articles. The first column indicates the Index of Readability (from grade 6 to 15), while the second and the third column indicate the number of analysed articles in that specific range.

Index of Readability	Britannica	Wikipedia
6	**0**	**1**
7	**4**	**1**
8	**6**	**7**
9	**7**	**14**
10	**22**	**27**
11	**21**	**30**
12	**21**	**10**
13	**10**	**7**
14	**6**	**0**
15	**3**	**3**

Table 5. Gunning Fog Index Score

As the data prove, about 65% of articles in the two encyclopedic corpora have an Index of Readability between 10-12, consequently this range can be considered the most important one.

In particular, 17% of articles in *Britannica* vs. 23% in *Wikipedia* are easily readable and understandable having an *Index of Readability* from 6 to 9, whereas 19% of articles in *Britannica* vs. 10% in *Wikipedia* have a more complex *Index of Readability* (from 13 to 15). These data show, firstly and independently of the individual or collaborative writing technique adopted by *Britannica* encyclopedists or Wikipedians, that the *Index of Readability* is not homogeneously distributed in both corpora. Secondly, the average *Index of Readability* of the two encyclopedic corpora (11,5 BAs vs. 11 WAs) suggests that *Wikipedia* articles should be slightly simpler to be read and understand. In detail, the highest score (15.8) has been found in the article *Racism* in *Britannica*, and the lowest in the article *Graffiti* (6.6) in *Wikipedia*. The latter article's equivalent in *Britannica* has recorded a score of 12.2. This is the only case in which a marked score variation has been detected. Excluding these exceptions, all the remaining articles have recorded very similar Indexes of Readability.

Since according to Gunning Fog *Index of Readability* the Fog Index required to have a reading level of a U.S. high school senior is 12, then *Britannica* and *Wikipedia* articles are easily understandable also by less educated readers (11-11.5) and both encyclopedias should succeed in fulfilling their primary educational and popular purpose.

8. Conclusions

The purpose of this study has been the identification of some linguistic parameters of variation, and to specify the linguistic similarities and differences between *Britannica* and

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 248-271
Recibido: 24/07/2009
Aceptación comunicada: 07/09/2009

Wikipedia encyclopedic expository style in order to map online intra-genre variations according to the following selected criteria: word length, sentence length, type/token ratio and Index of Readability. Some linguistic features, which have been identified as typical of informational production and specifically of encyclopedias, have also been analyzed in this study. This paper is built on and extends the comparative analysis of Emigh and Herring (2005) lending further empirical support to their earlier findings which showed that *Wikipedia* is not statistically distinguishable from Columbia Encyclopedia in some features. Thus, the results from this study seem to support and confirm their conclusions in terms of the different linguistic features measured and according to a different object of comparison: *Britannica Online*. In the case of *Wikipedia* this means that although authorship is no longer individual but shared with other writers, the maturity of their style, at least from the quantitative analysis point of view, is not dissimilar from that of online proprietary encyclopedias.

Even though *Britannica*'s production is, according to the selected criteria, more formal than that of *Wikipedia*, the stylistic difference is not as marked as expected. If *Britannica* is considered the best and most refined example of encyclopedias in the English speaking world, the different individual vs. collaborative authorial production, the copyright vs. copyleft license and, finally, the different authorship (professional paid writers vs. volunteer and anonymous amateurs), are controlled variables which do not significantly invalidate *Wikipedia*'s "fair" and correct formal production.

The main aim of encyclopedias is to inform, to educate and to present facts and information in specific entries. As informational texts, the presentation of encyclopedic information is packed within textual units which make use of an explicit formal expository style. A micro/macroscopic contrastive analysis has been carried out for the purpose of defining, through a frequency criterion, the incidence of the selected features on the encyclopedic expository style, highlighting similarities and differences in the two corpora. More specifically, word length is (in characters) 5.3 (*Britannica*) vs. 5.2 (*Wikipedia*), sentence length (tokens) is 22.05 (*Britannica*) vs. 22.09 (*Wikipedia*), lexical density is 45.5 (*Britannica*) vs. 43.6 (*Wikipedia*), while the Index of Readability is 11.5 (*Britannica*) vs. 11 (*Wikipedia*).

The data reported, clearly demonstrate that while the results are not very dissimilar, except for sentence length (22.05 words BAs vs. 22.09 WAs) they are constantly slightly higher in *Britannica*.

Although visual and audio media are included on encyclopedic web pages, their primary mode of communication is through written text. According to frequency criteria, one of the peculiarities of encyclopedic expository style, inside the more general class of informational production, is that it is very close to academic prose. The use of similar word and sentence length, as well as a high lexical density, associates the two textual productions. In this study, it has been shown that the average word length is 2-3 syllables (5.3 characters per word BAs vs. 5.2 WAs). Furthermore, sentence length does not appear to be very short (22.05 words BAs vs. 22.09 words WAs), being very similar to the average sentence length of academic writings (24 words).

Index of Readability analysis has confirmed that both *Britannica* and *Wikipedia* make use of an expository style immediately comprehensible to non-specialist readers. This

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 248-271
Recibido: 24/07/2009
Aceptación comunicada: 07/09/2009

indicates that traditional encyclopedia companies have adapted their products to capture the attention of a global, multi-cultural, web audience. Nevertheless, the basic features of expository style have not been betrayed or sacrificed in both copyrighted (*Britannica Online*) and *copyleft* textual environments (*Wikipedia*). In conclusion, for different reasons, entries of online encyclopedias are nowadays more readable, clear and comprehensible than the printed version of 30 years ago. Compared to the previous printed editions of *Britannica*, the less "formal" style of *Wikipedia* is probably due to a more informal mode of communication which is stylistically dominant on the Web, and to a less "controlled" *Massively Distributed Collaboration* which globally involves the web writers/readers scattered all over the world. The less "dignified" collaborative writing of *Wikipedia* may also have affected *Britannica*, which has gradually shortened and simplified its entries to make them more accessible and to expand its global market.

It is conceivable that the articles' quality, reliability, verifiability and formality of *Wikipedia* articles will actually increase over the coming years. *Wikipedia* content is getting constantly better as people go back again and again to old articles improving their quality, something which will increasingly come to the experts' attention. In the beginning, *Wikipedia* had a number of limited participating experts, but it has since attracted a higher number of graduate students, professors and professionals and it will probably attract the attention of many more experts in the near future. As the *Wikipedia* project improves and becomes better known, it is reasonable to expect that it will obtain wider academic recognition as many American institutions have already done. It is also reasonable to suppose that, in the coming years increasing numbers of academics will take part in the project seeing the increasing value of being associated with it. After all, many online courses, which can be read free of charge, demonstrate a very encouraging enthusiasm on the part of distinguished academics, to associate themselves with imparting free knowledge.

The Linguist List, the world's largest online linguistic resource started a "*Wikipedia* Update Project" in mid-June 2007. Recently O'Donnell (2007), reporting on the *Wikipedia* phenomenon, has suggested that academics need to accept *Wikipedia* open-based collaborative model and view further contributions to it as a unique form of community service scholarship. He claimed: "We are in a position to contribute to the construction of individual articles in a uniquely positive way by taking the time to help clean up and provide balance to entries in our professional areas of interest."

In conclusion, a new project that deserves to be mentioned is the *Citizendium Project*. It is a Citizens Compendium of Everything launched by Larry Sanger, co-founder of *Wikipedia*, with Jimmy Wales, in March, 2007. The project, considered by Jim Giles the academic rival of *Wikipedia*, has been initially described as a progressive fork of *Wikipedia*, a mirror of the *Wikipedia* site which allows anyone to contribute changes to articles, merging public participation with "gentle expert guidance". The final aim of the *Citizendium* is to improve the *Wikipedia* model by requiring all contributors to use their real names, by strictly moderating the project for unprofessional behaviour. It contained more than 10,000 articles up to March 2009. What will this new anti-populist project mean? How much and how will the quality and reliability of the information provided progress? To what extent will the formal style of encyclopedic articles improve in the future? Will a new cultured community discourse come to life as more scholars, professionals, educators and

Revista Electrónica de Lingüística Aplicada (ISSN 1885-9089)
2009, Número 8, páginas 248-271
Recibido: 24/07/2009
Aceptación comunicada: 07/09/2009

experts contribute and edit topics? It would indeed be interesting to investigate these phenomena in future research.

References

Ahrens, F. 2006. "Death by Wikipedia: The Kenneth Lay Chronicles" *Washington Post.com.* [Available at http://www.washingtonpost.com/wpdyn/content/article/2006/07/08/AR2006070800135.html

Ball, A.F. 1992 "Cultural preference and the expository writing of African-American adolescents". *Written Communication,* Vol. 9 , No. 4, 501-532.

Biber, D. 1988. *Variation across speech and writing.* Cambridge: Cambridge University Press.

Biber, D. 1995. "*Dimensions of register variation – across linguistic comparison"*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S., Reppen, R. 1998. "*Corpus linguistics: investigating language structure and use"*. Cambridge: Cambridge University Press.

Boyd, H. D. 4 January 2005. "Academia and Wikipedia - Many-to-Many". *Corante Blog* . [Available at http://many.corante.com/archives/2005/01/04/academia_and_wikipedia.php]

Capocci, A., Servedio, V., Colaiori, F., Buriol, L. S., Donato, D., Leopardi, S., Caldarelli, G. 2006. Preferential attachment in the growth of social networks: the case of Wikipedia. In *Physical Review*. [Available at http://www.inf.ufrgs.br/~buriol/papers/Physical_Review_E_06.pdf]

Chafe, W., Danielewicz, J. 1987. "Properties of spoken and written language". Horowitz, R., Samuels, S. J. (Eds.), *Comprehending Oral and Written Language.* New York: Academic Press, 83-113

Emigh, W., Herring, S. 2005. "Collaborative Authoring on the Web: A Genre Analysis of Online Encyclopedias". *Proceedings of the Thirty-Eighth Hawai'i International Conference on System Sciences, HICSS-38,* Los Alamitos: IEEE Press.

Giles, J. 2006 "Internet encyclopedias go head to head". *News@nature.com* [Available at http://www.nature.com/doifinder/10.1038/438900a]

Halliday, M. A. 1985. *Spoken and Written Language*. Oxford: Oxford University Press.

Hall, E. T. 1976. *Beyond Culture*. New York: Doubleday.

Heylighen, F., Dewaele, J. M. 1999. *Formality of Language: definition, measurement and behavioural determinants,* Leo Apostel University of Brussels. [Available at http://pcp.lanl.gov/Papers/Formality.pdf]

Hammonds, T., et *al.* 2006 "Social bookmarking tools (I), a general review." *D-Lib Magazine* 11(4). [Available at http://www.dlib.org/dlib/april05/hammond/04hammond.html]

Holloway, T., Bozicevic M., Börner K. 2005. "Analyzing and visualizing the Semantic Coverage of Wikipedia and Its Authors". *Complexity*, *Special issue on*

Understanding Complex Systems. [Available at Available at http://arxiv.org/ftp/cs/papers/0512/0512085.pdf]
Kapor, M. 9 November 2005. "Content Creation by Massively Distributed Collaboration*"*. UC Berkley School of Information. [Available at http://www.ischool.berkeley.edu/about/events/dls11092005]
Lanier, J. 2006. "On Digital Maoism: The Hazards of the New Online Collectivism". *Edge*, 183. [Available at http://www.edge.org/documents/archive/edge183.html]
Lawler, C. 2005. *Wikipedia as a Learning Community.* Master Thesis. Manchester: University of Manchester.
Lih, A. 2004. "Wikipedia as participatory journalism: reliable sources? Metrics for evaluating collaborative media as a news resource". *Proceedings of 5th International Symposium on Online Journalism*. [Available at http://jmsc.hku.hk/faculty/alih/publications/utaustin-2004-wikipedia-rc2.pdf]
Lynch, P. J., Horton, S. 2002. *Web Style Guide* [Available at. http://webstyleguide.com/]
Lowry, P.B., Curtis, A., Lowry, M. R. 2004. "Building a taxonomy and nomenclature of collaborative writing to improve interdisciplinary research and practice". *The Journal of Business Communication* 41, 66. [Available at http://www.questia.com/googleScholar.qst?docId=5002069499]
MacCormick, K., Pursel, J. 1982. "A Comparison of the Readability of the Academic American Encyclopedia, the Encyclopedia Britannica, and World book". *Journal of Reading.* Vol. 25, No.4, 322-25.
McHenry, R. 2004. "The Faith-Based Encyclopedia". *TCS Daily*. [Available at http://www.tcsdaily.com/article.aspx?id=111504A]
Ong, W. J. 1982. *Orality and Literacy.* London: Routledge.
Technical Report ISRN UIUCLIS-- 2005/2+ CSCW. [Available at http://www.isrl.uiuc.edu/~stvilia/papers/qualWiki.pdf]
Panagiota, A. 2002. "To wire or not to wire? Encyclopaedia Britannica versus Microsoft Encarta". *Educational Technology and Society* 5(1). [Available at http://www.ifets.info/journals/5_1/alevizou.html]
Pombo, O., Guerreiro A., Alexandre, A. F. 2006. *Enciclopédia e Hipertexto*. Lisboa: Editora Duarte Reis.
Reagle, J. 2006. *A Case of Mutual Aid: Wikipedia, Politeness, and Perspective Taking*. [Available at http://reagle.org/joseph/2004/agree/wikip-agree.html]
Resnick P., Hansen D., Riedl J., Terveen L., Ackerman M. 2005. "Beyond Threaded Conversation". *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems HCI'05.* Portland, OR, 2-7 April 2005.
Sanger, L. 15 September 2006. "Toward a New Compendium of Knowledge". *Citizendium.org*. [Available at http://www.citizendium.org/essay.html]
Schiff, S. 2006. "Know It All", *The New Yorker*.
Shah, S. 2005. "Productive Controversy". *Proceedings of the Wikimania'05* , Frankfurt am Main, Germany, August 4-7, 2005.
Shepherd, M., Watters, C.R. 1998. "The evolution of cybergenres". *Proceedings of the Thirty-First Annual Hawaii International Conference on System Sciences* (HICSS '98). Hawaii, Vol. 2, 97-109.

Sloane, J. 2007. "Wikimedia Foundation Moving To San Francisco". *Wired News,* 10 October 2007. [Available at http://blog.wired.com/business/2007/10/wikimedia-found.html]
Stvilia, B., Twidale, M.B., Gasser, L., Smith, L.C. 2005. "Information Quality Discussion in Wikipedia".
Tannen, D. 1989. *Talking voices: Repetition, dialogue, and imagery in conversational discourse.* Cambridge: Cambridge University Press.
Waldman, S. 2004. "Who knows?" *The Guardian*. [Available at http://www.guardian.co.uk/guardian/]
Zapf, G.K. 1949. *Human Behaviour and the principles of least effort*. Cambridge, Mass: Addison-Wesley.