International Journal of English Linguistics; Vol. 9, No. 2; 2019
ISSN 1923-869X E-ISSN 1923-8703
Published by Canadian Center of Science and Education

Proposing a Customised Method for Extratextual Documentative Annotation on Written Text Corpus

Niladri Sekhar Dash[1], Kesavan Vadakalur Elumalai[2], Mufleh Salem M. Alqahtani[2] & May Abdulaziz Abumelha[2]

[1] Linguistic Research Unit, Indian Statistical Institute, Kolkata, India

[2] Department of English Language and Literature, King Saud University, Riyadh, Saudi Arabia

Correspondence: Kesavan Vadakalur Elumalai, Department of English Language and Literature, King Saud University, Riyadh, Saudi Arabia. E-mail: ekesavan@ksu.edu.sa

Received: October 22, 2018 Accepted: November 15, 2018 Online Published: February 24, 2019

doi:10.5539/ijel.v9n2p99 URL: https://doi.org/10.5539/ijel.v9n2p99

Abstract

In this paper, we have made an attempt to portray a perceivable sketch of extratextual documentative annotation which, in the present frame of text annotation, is considered as one of the indispensable processes through which we can add representational information to the texts included in a written corpus. This becomes more important when a corpus is made with a large number of texts obtained from different genres and text types. To develop a workable frame for extratextual annotation, at each stage, we have broadly classified the existing processes of corpus annotation into two broad types. Moreover, we have tried to explain different layers that are embedded with extratextual annotation of texts as well as marked out the applications which can substantially enhance the accessibility of language data from a corpus for the works of text file management, information retrieval, lexical items extraction, and language processing. The techniques that we have proposed and described in this paper are unique in the sense that these are highly useful for expanding the utility of data of a written text corpus beyond the immediate horizons of language processing to the realms of theoretical, descriptive, and applied linguistics. In this paper, we have also argued that we should try to annotate all kinds of written text corpora so far developed in different natural languages at the extratextual level in a uniform manner so that the text samples stored in corpora can be uniformly used for various works of descriptive linguistics, theoretical linguistics, language technology, and applied linguistics including grammar writing, dictionary compilation, and language teaching. The annotation scheme proposed here is applied on a sample Bangla text corpus and we have noted that the accessibility of data and information from this kind of corpus is far easier than that of an un-annotated raw corpus.

Keywords: language resource, extratextual, intratextual, annotation, text, corpus, metadata, header file, documentative elements, interpretative elements, language technology

1. Introduction

The present scenario of growing diversion in the process of corpus generation and text annotation has given birth to several crucial issues that are directly interlinked to the rising quantity of text data, increasing varieties of text samples in machine-readable form, relaxation scale of the criterion designed for text and genre representation, the methods adopted in corpus annotation, and strategies followed in utilization of corpus data in varied linguistic and language technology works (Sinclair, 2004). All these issues have changed the primary notion of a 'corpus'. At present, the idea of 'corpus' transcends from the predetermined matrices of text categories, into an open-ended resource of natural language text collection in digital platform that has a strong possibility for regular updating of text database and the up-gradation of language flow with new sets of data in a centrally managed digital data archive (Leech, 1997). What we have clearly understood is that the different types of natural language texts that are now made available to us in digital version can hardly be analyzed and utilized in proper manners, unless these texts have been formatted properly with additional annotation and labels for easy access by both man and machine.

On the contrary, if these texts are not adequately annotated and labeled—both intratextually and extratextually—these resources may become less user-friendly because a corpus user may not be able to retrieve a suitable text, or a part of it, from the central digital archive and use it for particular research in question (Johansson, 1995). This clearly justifies our argument for developing a uniform method for extratextual documentative annotation of text corpora, although we know that some information is sometimes stored sporadically within Metadata normally attached to a corpus (Baker, 1997). A text corpus, which contains such additional layers of extratextual documentative information along with basic Metadata within its texture, is termed as **Extratextually Annotated Corpus (EAC).**

Keeping the multi-domain needs of annotated text corpora in mind, we have made attempt to portray the nature and form of extratextual documentative annotation that we need to apply on language corpora developed in an electronic version for enhancing their application utilities in various domains of applied linguistics, descriptive linguistics and language technology (Sinclair, 1994). For explicating various rubrics involved in the extratextual documentative annotation of written text corpora, we have discussed in brief about what is annotation with a reference to some of the theoretical issues of corpus annotation (Section 2); defined the notion of extratextual documentative annotation (Section 3); referred to some works on extratextual annotation (Section 4); discussed various types of extratextual annotation with a proposal for adding some new layers into the scheme (Section 5); and identified the usage of extratextually annotated text corpus in various domains of linguistics and language technology (Section 6). We have presented all the discussions with reference to their usages on a sample of modern Bangla written text corpus which has been developed in the TDIL (Technology Development for the Indian Languages) project of the Department of Electronics and Information Technology (Deity), Govt. of India.

2. What Is Corpus Annotation?

A written text corpus, when it is developed in a machine-readable form, generally appears in the raw state with a plain text format. On the other hand, an annotated text corpus is generated by way of attaching various tags of intralinguistic and extralinguistic information with the text samples (Leech, 2005; Berez & Gries, 2010). After annotation, a text corpus changes to a large extent in its character, content, form, and function (Aldebazal et al., 2009). The tags that are tagged to the texts usually increase the utility of a text corpus by way of providing specificity in text identification and accuracy in information extraction (O'Donnell, 1999). Thus, annotation makes a written text corpus accessible to the investigators interested in different kinds of research and application relating to linguistics and language technology (Archer, Culpeper, & Davies, 2008). The explicit information that is externally added in the form of tags to a text corpus, although no longer belongs to the textual body of a text, is an important repository of external information in the sense that it is able to provide relevant cues for retrieving implicit information stored within a corpus (deHaan, 1984).

According to some scholars, annotation of a text in a corpus is a practice of adding interpretative information to an existing database of spoken or/and written text samples of a language, by a set of predefined encodings that are normally attached to or interspersed with the *electronic representation* of the language text itself (Leech, 1993). A typical as well as widely familiar example of text corpus annotation is *grammatical annotation* (also known as *part-of-speech tagging*), where specific part-of-speech label or tag is attached to each and every word (by way of using some kind of attachment symbols such as *underline characters, slash characters,* and *special notations,* etc.), within a sentence to indicate the grammatical class or part-of-speech of the words in a particular sentential environment (Greene & Rubin, 1971).

Till date, we have come across several innovative and workable strategies and methods for annotation of a text within a corpus (Zinsmeister et al., 2008). In most cases, these methods are employed differently keeping in mind the different types of end application of the annotated texts. So far we have come across several types of text annotation, such as, grammatical annotation (Garside, Leech, & McEnery, 1997; Santorini, 1990; Brants et al., 2004; Schmid, 2008); syntactic annotation (Aldebazal et al., 2009); orthographic annotation (Leech, McEnery, & Wynne, 1997), prosodic annotation (Cox, 2011), semantic annotation (Rayson & Stevenson, 2008), discoursal annotation (Kipp, Neff, & Albrecht, 2007), anaphoric annotation (Lu, 2010), etymological annotation (Dash & Hussain, 2013), figurative annotation (Oostdijk & Boves, 2008), pragmatic annotation (Archer, Culpeper, & Davies, 2008); and sociopragmatic annotation (Archer & Culpeper, 2003), etc. Also we have come across a few text encoding techniques and standards, such as, *Constituent Likelihood Automatic Word-tagging System* (CLAWS) (Garside, 1987; McEnery & Hardie, 2011), *Text Encoding Initiative* (TEI) (Sperberg-McQueen & Burnard, 1994), *Hidden Markov Model* (HMM) (Kupiec, 1992), COCOA annotation method (McEnery & Wilson, 1996), *Dynamic Programming Method* (DPM) (DeRose, 1988), and EAGLE annotation guideline (Sinclair, 1996), etc. Besides, we have noted that there are several indigenous methods for corpus annotation which are different from encoding systems and models so far proposed for some corpora of advanced languages (Bird & Liberman, 2001; Johnston, 2013; Carletta et al., 2004; Xiao, 2008; Gilquin & Gries, 2009; Oostdijk & Boves, 2008; Thieberger & Berez, 2012).

However, we have not come across much discussion about the types and methods that we should adopt with regard to the annotation of extratextual information where non-textual data and information are to be annotated in the text in accordance with the Metadata generated from a particular text included in the corpus. Keeping this issue in view, in this paper, we intend to go beyond the levels proposed so far in search of an encoding technique that will help us define and explicate nature, process, and goal of extratextual documentative annotation, which in most cases, will provide us with information that is not found from other types of annotation. We also intend to divide the entire scheme of extratextual annotation into two broad types ((a) Text File and (b) Header File as well as a desire to explain the basic operation of the application of each type on a small sample of a written text corpus taken from a natural language. In our present scheme, the seven maxims of corpus annotation that have been proposed in Leech (1993) and supported by others (McEnery & Wilson, 1996) stand valid as we argue to adopt and implement these maxims in all kinds and types of annotation of a text corpus including the extratextual one.

3. What Is Extratextual Documentative Annotation?

Leech (1993) has proposed a scheme for theoretical-cum-application based distinction between the two types of information embedded within a piece of natural text:

(a) Representational Information (RI), and

(b) Interpretative Information (II).

According to the proposition of Leech (Leech, 1993), 'representational information' is the actual representation of a language text. It contains actual linguistic information relating to various linguistic items and elements that are used in a piece of text, namely, *letters, punctuations, words, phrases, clauses, sentences, tables, paragraphs, images, drawings, pictures and other textual and non-textual elements* which are available within a written or printed text corpus. On the other hand, 'interpretative information' is something, which is externally added to the basic textual elements, usually by a human text annotator who has a strong linguistic knowledge of the language she is annotating.

Based on the scheme of information embedding proposed in Leech (Leech, 1993), we can argue that an annotated text is actually embedded with four different types of information under the rubrics of two broad types as the following diagram shows (Figure 1).

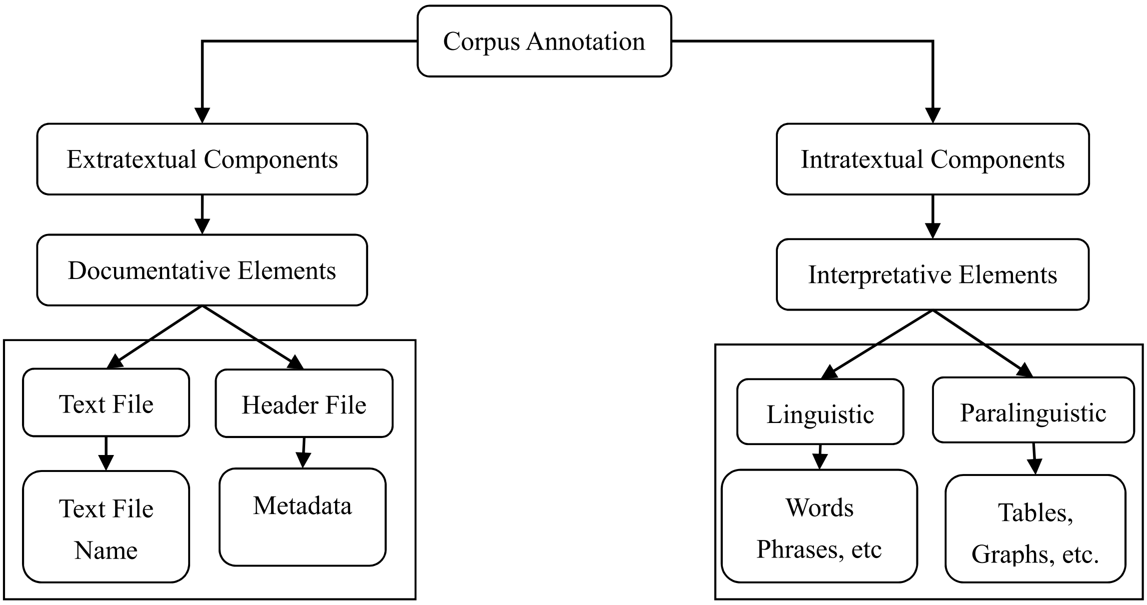

Figure 1. Components represented within an annotated text corpus

Based on the diagram given above (Figure 1), we can say that an annotated text corpus, in general, may contain two broad types of components: (a) extratextual components, and (b) intratextual components. The extratextual component is primarily documentative in form, nature, and function. Therefore, it can be further divided into two broad parts: (a) Text File, and (b) Header File. The intratextual component, on the other hand, is predominantly interpretative in form, nature, and function. Therefore, it can also have two broad parts: (a) linguistic component, and (b) paralinguistic component, which are available within the body of a written text corpus.

Within the domain of intratextual component, the linguistic elements are purely linguistic items, such as, *letters, morphs, words, compounds, multiword units, punctuations, sentences, phrases,* etc. without which any language text is impossible to construct its theoretical identity or attest its functional relevance. Within the domain of paralinguistic elements, on the other hand, there are some components which are not linguistic elements. These are some other kinds of non-textual elements which are found to be used within a text, such as, *tables, figures, pictures, images, diagrams, flowcharts,* etc. These paralinguistic elements are usually removed from the digital version of a text corpus which is developed from printed and handwritten texts and contains only the samples of textual elements. However, these elements are easily found within a *multimodal corpus* which contains all kinds of linguistic and paralinguistic elements (Dash, 2008, pp. 66–67; Dash & Arulmozi, 2017, pp. 58–60).

The paralinguistic elements are usually 'marked-up' on a text using a set of predefined tags at the time corpus annotation. These are applied within a written text by way of some conventional sets of machine-readable labels as the indicators for introducing the text features like *chapters, paragraphs, segments, headings, titles, images, hyphenations, diagrams, pictures, tables, flowcharts,* etc. (Atkins, Clear, & Ostler, 1992). In case of spoken text corpora, these are introduced as indicators to tag the features that are often marked as salient features of a spoken texts but normally missing in a written text, such as, *hesitations, fillers, pauses, false starts, non-beginnings, abrupt endings, repetitions, gaps, utterance boundaries, suprasegmental properties, hedges,* etc. (Stenstrom, 1984; Kipp, Neff, & Albrecht, 2007).

In our argument, the intratextual information is indispensable in the sense that it is absolutely essential in the proper interpretation of texts that are stored in a corpus for better retrieval of relevant linguistic data and information essential in various linguistic tasks including language processing. The extratextual information, on the other hand, is required for successful identification of a text file stored in large text databases of a digital text archive. Since it maintains a direct link with the target text domains, without proper interpretation and understanding of extratextual elements, retrieval of linguistic information and data from a written text corpus will fail miserably or will become ruefully erroneous. To overcome this deficiency, we propose that extratextual components should include elaborate yet organized information relating to *title, author, time, gender, nativity, publisher, ethnicity, subject, domains* and other information, which may be stored as Metadata in a Header of a text file of a corpus. It is known to all that organized extratextual information is indispensable in almost all the works of *sociolinguistics, ethnolinguistics, ecolinguistics, sociology, demography, lexicography, language planning.* Such information also becomes useful for technical, legal, cultural, and similar other reasons.

In all practicalities, the extratextual annotation is of paramount importance to anyone engaged in the task of establishing a *Digitally Encoded Language Text Archive* (DELTA) for a natural language. The variables included in the extratextual annotation are considered as the most credential attributes in the context of corpus generation, corpus management, corpus classification, and extraction of data and information from corpora for linguistic and technology works. Since these extratextual variables are essentially non-linguistic in nature, these are often determined impromptu without delving into the textual components stored within a corpus, thereby it is ensured that no prior linguistic judgment is made at the time of designing the extratextual tagsets. The lists of attributes, which are relevant for extratextual annotation of a text corpus, are also relevant for the description of a particular language community from which the language corpus is sampled and produced (Atkins, Clear, & Ostler, 1992). These attributes are not only relevant for different types of language investigation, but also useful for designing methods for updating data archives with new types of text which are not yet classified within the list of attributes used in the scheme.

The importance of extratextual annotation is further envisaged, for example, within the domain of a *diachronic corpus*, which contains text samples dating over a long range of periods (Kyto, Rissanen, & Wright, 1994). In case of a diachronic corpus, extratextual annotation becomes useful to include information about the date(s) of origin of texts as well as other information relating to authorship, title of works, nationality of authors, source of availability of texts and so on within the Header File of the corpus (Leech & Fligelstone, 1992). This kind of extratextual information about the content of a diachronic corpus can be used for extracting data selectively from a particular corpus when it is stored within a large archive with corpora of other texts in the structured hierarchy (Archer & Culpeper, 2003). For example, we are interested to perform some operations on small samples of text taken from different periods of time, or on all the text samples from one particular period to understand the variations reflected in the usage of language across the ages or across different text types (Rissanen, 1989). In such a study, appropriate knowledge about the extratextual information of a corpus will be quite useful in the act of finding out relevant text samples from a DELTA. Thus, specified attributes of extratextual annotation can go beyond the immediate domains of descriptive linguistics to cater the varied needs of many other areas of applied linguistics, social sciences, and language technology. It is, therefore, rational to encode a text corpus in advance at the extratextual level making the tasks of corpus management, corpus search, and information retrieval a relatively easier and quicker exercise.

4. Early Works on Extratextual Annotation

The first two notable encoding projects (completed in 1977 & 1981, respectively) were those that were applied to the *Brown Corpus* (Francis, 1980) and the *LOB Corpus* (Gerside et al., 1987)—each one of which consisted of a little over than one million words. However, these were mainly elaborate schemes of grammatical annotation or part-of-speech tagging of the words that are included in the corpora, for which little attention was paid to the level of extratextual annotation on the corpus texts. Perhaps, our understanding about the varieties of annotation of that time was not advanced enough to visualize that a kind of extratextual annotation could be of indispensable significance in the act of language text archiving when corpus data would be of enormous volumes, varieties, and dimensions. At that time, we also failed to realize that the wide diversity and enormous amount of language data would eventually ask for a highly systematic and machine-friendly technique for data management in the act of quick and easy handling of language databases in the central digital archive. Nearly six decades passed away before we could realize the practical importance and relevance of extratextual annotation on corpus texts, and when we realized it, we started working in this direction in one or the other manner as far as we could manage without following any universal standard or approved norm.

Since a widely representative and universally approved standard for representing the extratextual information in a corpus text is yet to be developed and applied, people involved in different corpus developing groups and text annotating agencies have adopted different approaches and methods for serving their purposes. For instance, one well-known extratextual annotation scheme is known as the COCOA references, which has mostly been used by the developers of the *Longman-Lancaster Corpus* and the *Helsinki Corpus* (McEnery & Wilson, 1996, p. 26). As a standard convention, this particular annotation scheme contains a balanced set of angled brackets (< >) having two entities:

(a) A code symbolizing the name of a particular variable, and

(b) A string (or set of strings) that symbolizes the variable.

For example, the code 'A' could be used to refer to a variable called 'Author', while the string would stand for the name of an author. Thus, the COCOA references, which indicated the name of the author of a passage of text, would look like the following:

<A CHARLES DICKENS>

<A WOLFGANG VON GOETHE>

<A HOMER>

A short representation of the COCOA format header that is used in the *Helsinki Corpus* is presented in the example (Table 1) given below (McEnery & Wilson, 1996, p. 31). It shows how the extratextual information could be encoded in a corpus text. The left-hand row represented the COCOA format header while the right-hand row represented the glossary provided for left-hand row (code 'N' represented the name of a text, while 'X' indicated that the information for that code was either irrelevant or unavailable).

Table 1. The COCOA format Header File for extratextual annotation

<BCEPRIV1>	Short descriptive code
<QE1XX CORP EBUEAM>	Text identifier
<N LET TO HUSBAND>	Name of the text
<EDA ELIZABETH>	Author's name
<C E1>	Sub-period
<O 1500-1570>	Date of composition
<MX>	Date of manuscript
<KX >	Relevance of the text
<D ENGLISH>	Dialect
<V PROSE>	Verse or prose
<T LET PRIV>	Text type
<G X>	Foreign or original
<F X>	Foreign or original
<W WRITTEN>	Written or spoken
<X FEMALE>	Sex of author
<Y X>	Age of author
<H HIGH>	Author's social status
<U X>	Audience description
<E INT UP>	Participant relationship
<J INTERACTIVE>	Interactive/non-interactive
<I INFORMAL>	Formal/informal
<Z X>	Prototypical text category
<S SAMPLE X>	Sample

The COCOA reference format, as it has been displayed above (Table 1), was, in essence, a kind of informal system which is used for encoding specific types of extratextual information (e.g., the *name of an author, date of composition, and title of a text,* etc.). It was comparatively simplified in format and information representation keeping in mind that the proposed annotation scheme would be applied more by human text users than by a machine. This, however, paved a way towards more formalised international standards of extratextual annotation, as it became instrumental in giving birth to a more general text annotation guidelines, namely, the *Text Encoding Initiative (TEI)* nurtured and maintained by the *Association for Literary and Linguistic Computing, the Association for Computers and the Humanities*, and the *Association for Computational Linguistics* with a clear aim at providing global standards for corpus annotation both at textual and extratextual level and generating machine-readable text interchange facilities.

In essence, the TEI has used a form of document markup system known as the *Standard Generalised Markup Language* (SGML-ISO, 1986), which is recognised as an international standard for corpus annotation because it provides clarity in the act of text understanding, simplicity in codes implementation, and rigorousness in the act of explicit encoding formalisation (Sperberg-McQueen & Burnard, 1994). Following this markup system, text corpora of different types are generated in all major advanced languages and these are annotated and used both in linguistic description and analysis as well as in complex machine learning algorithms (Goldfarb, 1990; Fierz & Grutter, 2000). The most unfortunate thing is that, till date, most of the Indian and Asian scholars have not even tried to apply the extratextual annotation markup system and properties on the corpora developed so far in electronic form in many of the Indian and Asian languages. Moreover, most of the scholars are still skeptical about the applicational relevance of the extratextually annotated corpora in the panoramic spectrum of varied linguistic works in the respective languages.

5. Types of Extratextual Annotation

In principle, the most rudimentary type of extratextual annotation is the one, which tells what type of a text it is. It also guides one who is looking for a particular type of text. Based on this principle we can depend sequentially on the following two kinds of extratextual elements:

(a) Text File Name, and

(b) Header File.

The Text File Name gives us important clues to know what kind of a text a Text File actually contains. However, in most cases, a Text File Name may provide us only a minuscule amount of information. The information about the nature of a text also consists of much more information than the title of a book and the name of an author can generate. For this kind of information, we have to refer to a Header File, which stores explicit and elaborate extratextual information of a text. Therefore, it may be argued that complete extratextual data and information of a text corpus is available partly from a Text File Name and partly from a Header File. In following subsections, we like to concentrate only on extratextual documentative annotation including two sub-domains: Text File Name, and Header File (i.e., Metadata). The other types of annotation (i.e., intratextual annotation) are kept out of the frame of the present discussion since these will reflect on different issues which are distantly linked with the present goal of the paper.

5.1 Text File Name: A Gateway

A text in a corpus is generally stored in a computer in the form of different text files under different head names, which are carefully designed after the major categories of subjects and the text disciplines are properly envisaged and formalized for the purpose of corpus generation and text classification. In a wider sense, the extratextual documentative annotation includes a Text File Name, which, at the outset, provides a vital clue to what kind of a text file one has to explore from the DELTA. Thus, a Text File Name becomes a gateway to the digital text database we call **Corpus Database**.

By virtue of its brevity (it has to be confined to a limited number of characters), the surface form of a Text File Name is highly elusive. For example, for the Indian languages corpora of the TDIL (Technology development for the Indian Languages) project, it was decided under a common consensus that the name of a text file should invariably be consistent for all the Indian languages. Also, it was agreed that the text file name must carry, within its length of 12 characters assigned for it, information of the following types:

(a) Text Category Information (TC)	: 2 characters (alphabet),
(b) Subject Category Information (SC)	: 2 characters (alphabet),
(c) Text Title Information (TT)	: 4 characters (alphabet),
(d) Dot Indicator (.)	: 1 character (symbol), and
(e) Number of text files generated	: 3 characters (number).

Thus, the surface representation of a Text File Name should be formed for the text files stored in the TDIL corpus in alphanumeric combination: CCCCCCCCC.DDD (C=Character, D=Digit).

5.2 Text Category Annotation

Taking all the five layers of information into consideration, the name of each text file (that contained large text databases) is formed in the following format:

(a) $[C_1C_{2[TC]}C_3C_{4[SC]}C_5C_6C_7C_{8\,[TT]}\,\cdot\,_{[Sym]}D_1D_2D_{3[Number]}]$

It shows how a Text File Name is constituted with 12 characters. The first 8 characters (shown by capital 'C' that stands for a *character* with a subscript digit signifying number of the character) constitute the name of the text file, while the dot (constituting a single character) marks the end of the name of a text file. The last 3 characters are actually cardinal numerals (displayed as D=digits along with its subscript number implying the value of a numeral) used to specify the number of the text file belonging to a particular text category. This is one of the methods that has been unanimously developed and used for the development of texts files in the Indian languages in the TDIL project (Dash, 2016). It should be clearly stated here that this method is not the only one method which has to be used for all types of corpus development activities. In fact, each corpus developer has the full freedom to develop a new method or adopt an existing one keeping in mind the nature of corpus generated as well as the possible applications of the corpus texts.

Again, the first eight characters of the Text File Name are further classified in such a manner that they can be the best representative of the text and the subject category as well as for the title of a text. Therefore, it is planned in such a way that the first two characters are used for a code for the text category; next two characters are reserved for a code for the subject category, and the last four characters are used to represent the abbreviated form of the title of a text. An example is furnished below from the Bangla text file to elucidate the basic idea that has been adopted for a Text File Name for the purpose of documentative annotation of the texts of the TDIL corpus of the Indian languages:

(b) Text File Name: <LTFCKLBL.005>

The example given above contains, in total, 12 characters. The first 2 characters <LT> constitute the abbreviated code for the text category [Literature]; the next 2 characters <FC> constitute the domain of the subject category [Fiction]; and the last 4 characters <KLBL> constitute a code for representing the title of the source text from where the sample texts are collected.

This process was useful for accommodating the names of the major text categories selected for data collection for developing the TDIL Bangla corpus within the Text File Name. It also helped to divide the texts belonging to various subjects under the 6 major text categories mentioned below. In fact, the grouping of the source documents under six text categories along with their extratextual encodings minimized the effort in searching out the subject categories as well as the actual text files required for particular investigation (Table 2).

Table 2. Six major text categories and their codes used in the TDIL corpus

No. of domains	Text Categories considered for the TDIL Corpus	Codes Used
(1)	[Texts procured from the domain of Literature]	<LT>
(2)	[Texts procured from the domain of Social Sciences]	<SS>
(3)	[Texts procured from the domain of Natural Sciences]	<NS>
(4)	[Texts procured from the domain of Commerce]	<CC>
(5)	[Texts procured from the domain of Mass Media]	<MM>
(6)	[Texts procured from the domain of other domains]	<OR>

Based on the content or subject matter, a text is put into one of the six major text categories stated above (Table 2). Moreover, each text category is designated in such a way that it includes several subject categories following a simple *tree-branching technique* of text classification, where texts of different subjects and disciplines are also divided into different text categories. For instance, texts belonging to sub-disciplines like *fiction, short story, essay, painting, drawing, music, sculpture*, etc. are brought under the head text category: [Literature] <LT>. Similarly, the subject domains like *political science, history, philosophy, education*, *religion,* etc. are brought under the head text category: [Social Science] <SS>; the text category [Natural Science] <NS> includes subject categories such as *physics, chemistry, biology, mathematics, geography,* and others; text category [Commerce] <CC> includes subject categories like *accountancy, banking,* and *business,* etc.; and the text category [Mass Media] <MM> includes subject categories such as, *newspapers, periodicals, advertisements*, *notices, magazines,* etc. Finally, the text category [Others] <OR> includes those subject materials that do not belong to any of the text categories stated above. It contains representative text samples collected from *administrative documents, legislative proceedings, government circulars, legal documents,* etc. This tree-branching technique that we have applied for the purpose of text collection provides some important functional advantages over other techniques of text classification. Above all, it serves the primary purpose of proper representation of texts within a corpus (Biber, 1993). For instance, if we look at the diagram given below, we can note that by using the tree-branching technique, from the name of a text file, we can easily tell to which text category a text sample, which is taken into the corpus, actually belongs (Figure 2).

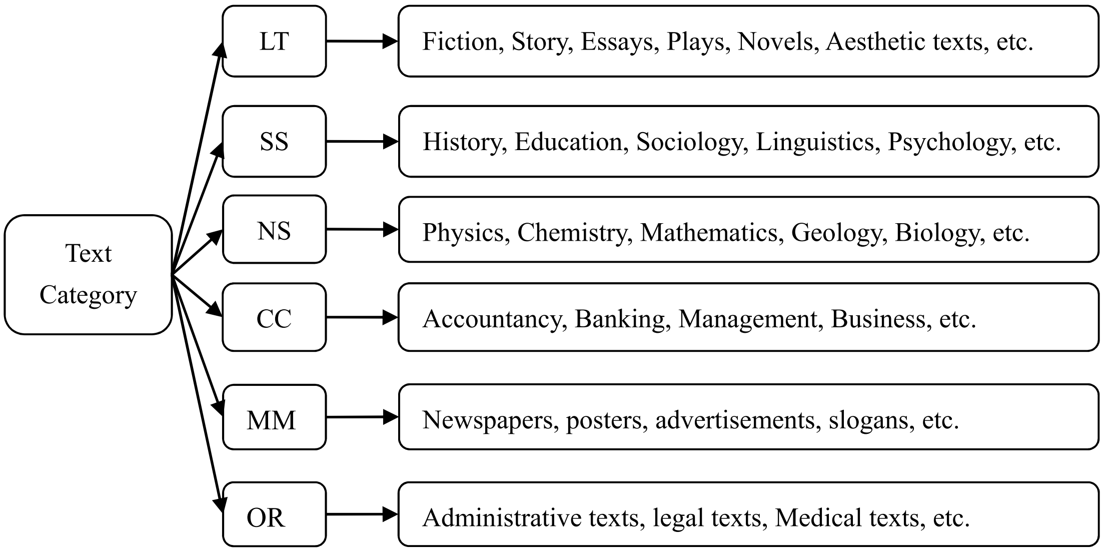
Text Category
LT
Fiction, Story, Essays, Plays, Novels, Aesthetic texts, etc.
SS
History, Education, Sociology, Linguistics, Psychology, etc.
NS
Physics, Chemistry, Mathematics, Geology, Biology, etc.
CC
Accountancy, Banking, Management, Business, etc.
MM
Newspapers, posters, advertisements, slogans, etc.
OR
Administrative texts, legal texts, Medical texts, etc.

Figure 2. Text classification scheme used for developing the TDIL corpus

The most important thing to be noted here is that the names of different text categories have been confined within two characters only. Therefore, the main challenge was to abbreviate the names of the text categories into 2-letter codes in such a manner that no code should replicate to other code. If it does, it will lose its unique code identity as it will generate identical codes of the subject categories which will eventually generate severe confusion in text identification and processing.

5.3 Subject Category Annotation

For each subject category, a unique 2-letter code has been assigned immediately after the codes are assigned to each text category (Table 3). This method easily makes the name of a text file different from the others and it becomes much easier for handling the text files in corpus processing tasks. In the table given below (Table 3) the process of subject category encoding technique is exhibited, which provides highly systematic and effective means for data storage in a corpus.

Table 3. Coding of subject categories of the texts in the TDIL corpus

Fiction	<FC>	Short story	<ST>	Mathematics	<MT>
Travelogue	<TL>	Letters	<TS>	Geology	<GL>
Song/Music	<SM>	Criticism	<CR>	Geography	<GR>
Dance	<DN>	Child Literature	<CL>	Chemistry	<CH>
Diary	<DR>	Drawing	<DW>	Banking	<BK>
Drama/play	<DP>	Essay	<ES>	Criminology	<CR>
Humour	<HR>	Folk literature	<FL>	Accountancy	<AC>
Hobby	<HB>	Biography	<BG>	Magazine	<MZ>
Anthropology	<AN>	Archaeology	<AR>	Administrative	<AD>
Astrology	<AS>	Psychology	<PS>	Legal documents	<LD>
Philosophy	<PH>	Political Science	<PL>	Photography	<PG>
Religion	<RL>	Sociology	<SL>	Printing technology	<PT>
Journalism	<JL>	Library Science	<LS>	Computer Engineering	<CE>
Linguistics	<LN>	Law and Order	<LW>	Cinematography	<CG>
Mythology	<MY>	Education	<ED>	Business Management	<BM>
Economics	<EC>	Games & sports	<GS>	Advertisement	<AD>
History	<HS>	Astronomy	<AS>	Newspapers	<NW>
Ayurveda	<AV>	Agriculture	<AG>	Pamphlets	<PP>
Architecture	<AR>	Botany	<BT>	Public Administration	<PA>
Biology	<BL>	Physics	<PS>	Legislative Proceeding	<LP>
Statistics	<ST>	Sculpture	<SC>	Sexology	<SX>
Zoology	<ZO>	Veterinary	<VT>	Medicine	<MD>

5.4 Title of Text Annotation

It was clearly understood that the subjects selected for data input should have its proper representation within a Text File Name so that one can easily understand, after seeing a Text File Name, to which title the text actually belongs. For obvious reasons, the whole title of a text cannot be stored within a Text File Name. Therefore, the best option is to abbreviate the title in such a manner that it can be suitably embedded within the code of four characters. And at the same time, a corpus user, by a single look, can have clear information to understand that to which text the title belongs. The best possible way to create an acrostic term is to use the four characters allotted for the title of the text. For instance, in the Bangla text corpus, the code <KLBL> supplies the necessary information to a corpus user to understand that the title of the text is a fiction entitled [kālbelā "Bad Time"].

This technique of encoding of the title name is, however, not a very accurate practice, as in most cases, corpus users may deduct a wrong name or a different name from the codes used. The possible solution to this problem is to tabulate elaborately the titles and their respective codes beforehand (at the time of language data input or text generation) in a separate file. Thus, if there is any doubt, a text user (a native or a non-native one) can easily verify, if required, to know if the particular code is used for a particular title of a text found in the language.

The 3-digit extension after the dot (.) indicates further particularization of a Text File Name. The digits stand for the number of files generated from a particular title of a text. Experiences gathered from the TDIL project of the Bangla text corpus generation (Dash, 2000; Dash, 2007) have shown that even from a fiction containing one million words, one cannot create more than 100 text files, if each text file contains data of nearly 10,000 words. It would have been sufficient to allow only two digits for the total number of files generated from a text.

However, to be on the safe side it is always better to allow three digits because some books may have the much larger amount of data, which would exceed the limit of 100 text files, but definitely not the limit of 999 text files.

Looking at the last three digits we can easily understand how many text files are actually created from a particular title. Also, it provides information to the users regarding a particular text file under scrutiny. For instance, the extension <.005> implies that the text file is fifth in the serial number created from a particular title. Besides, if a user wants to explore texts belonging to the last part of the title, she can easily do that work by accessing the text files tagged with last digits of the series.

5.5 Header File: A Documentary Safeguard

In general, a Header File contains extratextual data and information of the followings: name of the author, the gender of the author, the nationality of the author, the age of the author, year of first publication, the name of the publisher, place of publication, edition used in the corpus, type of text, etc. Some ideas may be obtained from the following table that shows how different kinds of extratextual information is recorded in a header file (Table 4).

Table 4. Extratextual information in the Header File

<Header File>	Metadata Information
<Title>	Śāmba
<Language >	Bengali
<Genre>	Written Text
<Text Category>	Literature
<Subject Category>	Fiction
<Text Type>	Imaginative
<Source Type>	Book
<Publication Year>	1978
<Edition>	First
<Volume>	Single
<Issue>	Not Applicable
<Publisher>	Ananda Publishers
<Publication Place>	Kolkata, India
<Author>	Kālkuṭ
<Gender>	Male
<Age>	60+
<Nationality>	Indian
<Total Words>	5120
<Text>	[Sample Text]

In essence, a Header File carries the basic type of additional information, which actually informs one to what kind of text one should look for in the text files (McEnery & Wilson, 1996, p. 30). This basic information about the nature of the text can often be much more detailed than simply giving the title and author name. Sometimes, it will tell us more about the content of a text file, as it may give us information about the age of the author, sex of the author, date the text was published or written first, variety of language (e.g., *American English, British English, Australian English, Indian English,* etc.), a broad subject domain (e.g., *science, literature, religion, culture,* etc.), type of text (e.g., *informative, imaginative*), and so on. This kind of information may be used for various studies of descriptive linguistics, sociolinguistics, stylistics, psycholinguistics, language teaching, and discourse analysis as well as for solving the conflicts of copyright and legal rights. For example, if we look at the metadata stored in the Header File (Table 5), we can easily retrieve a large load of extratextual information, which is hardly available from the text itself.

Table 5. The format of a Header File and text used in the TDIL corpus

Metadata	<Title : Śāmba>	<Language : Bangla>,	<Genre : Written>,
	<TC : LIT>,	<SC : Fiction>,	<TT : Imaginative>,
	<ST : Book>,	<Year : 1978>,	<Edition : First>,
	<Volume : Single>,	<Issue : 0>,	<Publisher : Ananda>,
	<Place : Kolkata>,	<Author : Kālkut>,	<Gender : Male>,
	<Age : 60+>,	<Nationality : Indian>,	<Words : 5120>
Text	[marite cāhi nā āmi sundar bhūbane. kathāṭā āj anya ekṭi kathār khei dhariye dila. dhariye deoyā khei kathāṭi abiśyi biparīt. nā-te āche hynā. bhramite cāhi āmi sundar bhūbane.]		

A Header File, when it is provided with information relating to these fields, can easily be accessed for retrieving required data from the text itself based on the attributes it includes. Also, a computer program can retrieve this information through its elegant searching and sorting algorithms running on the corpus database. For instance, within a corpus database if one wants to refer to the texts written by female writers only, one can easily do this by running a search program on the lists of Header Files, which can refer to the variables where the 'gender of the author' (say, X) is equal to the code 'Female'. Thus, information fields provided within the Header File can be utilized in a uniform manner for an information retrieval system that is engaged in searching and sorting all the extratextual variables, thereby, the texts. Similarly, a text investigator can retrieve language data belonging to a particular year, text type, genre, or author from the large list of corpus database files based on the extratextual documentative information stored in the Header File of a corpus (Oostdijk & deHaan, 1994).

6. Conclusion

When a language text database is generated through hundreds of corpus text files, a bunch of workable keys is required for unlocking corpus for collecting the intralinguistic and extralinguistic information the possession of which can enrich the faculty of the observational, descriptive, and explanatory adequacy of the language users. Here lies the practical utility of Text File Name as well as Header File, which combines and contribute to constituting the concept of extratextual documentative annotation.

Apart from pure language data, a written corpus carries some additional information relating to intralinguistic and extralinguistic domains. Since information of these domains is varied in nature, it is better to devise strategies by which we can store and retrieve various intralinguistic and extralinguistic data and information from the Text File Name and the Header File of a text rather than entering into the text database itself and groping in the dark. In this case, the extratextual documentative annotation is, perhaps, the best strategy for language investigators as well as for a computer system that is engaged in the quick and accurate retrieval of text files from a large data archive.

We can conclude that extratextual documentative annotation is quite useful in text retrieval, data mining, and text classification activities, since the quick reference to extratextual annotation increases precession of a computer system in data mining, text identification, and data retrieval. Quick, easy, and accurate methods of identification and extraction of texts from a DELTA are the most useful strategies in the utilization of corpus data in language description, analysis, and application in all branches of linguistics.

Acknowledgments

The authors extend their thanks and appreciation to the Deanship of Scientific Research and the Research Centre, College of Arts at King Saud University, Riyadh, Kingdom of Saudi Arabia, for funding this research work.

References

Aldebazal, I., Aranzabe, M. J., Arriola, J. M., & Dias de Ilarraza, A. (2009). Syntactic annotation in the reference Corpus for the processing of Basque (EPEC): theoretical and practical issues. *Corpus Linguistics and Linguistic Theory*, *5*(2), 241–269.

Archer, D., & Culpeper, J. (2003). Sociopragmatic annotation: New directions and possibilities in historical corpus linguistics. In A. Wilson, P. Rayson & T. McEnery (Eds.), *Corpus Linguistics by the Lune: Studies in honor of Geoffrey Leech* (pp. 37–58). Frankfurt: Peter Lang.

Archer, D., Culpeper, J., & Davies, M. (2008). Pragmatic annotation. In A. Ludeling & M. Kyto (Eds.), *Corpus Linguistics: An International Handbook* (pp. 613–642). Berlin: Walter de Gruyte.

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing*, *7*(1), 1–16. https://doi.org/10.1093/llc/7.1.1

Atwell, E., Demetriou, G., Hughes, J., Schiffrin, A., Souter, C., & Wilcock, S. (2000). A comparative evaluation of modern English corpus grammatical annotation schemes. *International Computer Archive of Modern English Journal*, *24*, 7–23.

Baker, P. (1997). Consistency and accuracy in correcting automatically tagged corpora. In R. Garside, G. Leech & A. McEnery (Eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora* (pp. 243–250). London, Longman.

Berez, A. L., & Gries, S. T. (2010). Correlates to middle marking in Dena'ina iterative verbs. *International Journal of American Linguistics*, *76*(1), 145–165. https://doi.org/10.1086/652757

Biber, D. (1993). Representativeness in corpus design. *Literary, and Linguistic Computing*, *8*(4), 243–257.

https://doi.org/10.1093/llc/8.4.243

Bird, S., & Liberman, M. (2001) A formal framework for linguistic annotation. *Speech Communication*, *33*(1–2), 23–60. https://doi.org/10.1016/S0167-6393(00)00068-6

Brants, S., Dipper, S., Eisenberg, P., Hansen, S., Konig, E., Lezius, W., Rohrer, C., Smith, G., & Uszkoreit, H. (2004). TIGER: linguistic interpretation of a German Corpus. *Research on Language and Computation*, *2*(4), 597–620. https://doi.org/10.1007/s11168-004-7431-3

Carletta, J., McKelvie, D., Isard, A., Mengel, A., Klein, M., & Moller, M. B. (2004). A generic approach to software support for linguistic annotation using XML. In G. Sampson & D. McCarthy (Eds.), *Corpus Linguistics: Readings in a Widening Discipline* (pp. 449–459). London: Continuum.

Cox, C. (2011). Corpus linguistics and language documentation: challenges for collaboration. In J. Newman, R. Harald Baayen & S. Rice (Eds.), *Corpus-based Studies in Language Use, Language Learning, and Language Documentation* (pp. 239–264). Amsterdam: Rodopi.

Dash, N. S. (2007). Indian scenario in language corpus generation. In N. S. Dash, P. Dasgupta & P. Sarkar (Eds.) *Rainbow of Linguistics* (vol. I., pp. 129–162). Kolkata: T. Media Publication.

Dash, N. S. (2008). *Corpus Linguistics: An Introduction*. New Delhi: Pearson Education-Longman.

Dash, N. S. (2016). The history and methodologies of corpus development research in India. In H. H. Hock & E. Bashir (Eds.), *The Languages and Linguistics of South Asia* (vol. 7, pp. 740–748). Berlin: Mouton de Gruyter.

Dash, N. S., & Arulmozi, S. (2017). *History, Features, and Typology of Language Corpora*. Singapore: Springer Nature.

Dash, N. S., & Chaudhuri, B. B. (2000). The process of designing a multidisciplinary monolingual sample corpus. *International Journal of Corpus Linguistics*, *5*(2), 179–197. https://doi.org/10.1075/ijcl.5.2.05das

Dash, N. S., & Hussain, M. M. (2013). *Designing a Generic Scheme for Etymological Annotation: A New Type of Language Corpora Annotation* (pp. 64–71). Proceedings of the ALR-11 and 6th International Joint Conference on Natural Language Processing, Nagoya Congress Centre, Nagoya, Japan, 14–18 October 2013.

deHaan, P. (1984). Problem-oriented tagging of English corpus data. In J. Aarts & W. Meijs (Eds.), *Corpus Linguistics* (pp. 123–139). Amsterdam: Rodopi.

DeRose, S. J. (1988). Grammatical category disambiguation by statistical optimization. *Computational Linguistics*, *14*(1), 31–39.

Fierz, W., & Grutter, R. (2000). The SGML Standardization Framework and the Introduction of XML. *Journal of Medical Internet Research*, Jun 30. https://doi.org/10.2196/jmir.2.2.e12

Francis, W. N. (1980). A tagged corpus: problems and prospects. In S. Greenbaum, G. Leech & J. Svartvik (Eds.), *Studies in English Linguistics: In Honour of Randolph Quirk* (pp. 192–209). London: Longman.

Garside, R. (1987). The CLAWS word-tagging system. In R. Garside, G. Leech & G. Sampson (Eds.), *The Computational Analysis of English: a corpus-based approach* (pp. 30–41). London: Longman.

Gilquin, G., & Gries, S. T. (2009). Corpora and experimental methods: a state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, *5*(1), 1–26. https://doi.org/10.1515/CLLT.2009.001

Goldfarb, C. F. (1990). *The SGML Handbook*. London: Clarendon Press.

Greene, B., & Rubin, G. (1971). *Automatic Grammatical Tagging of English*. Technical Report, Department of Linguistics, Brown University, Rhode Island (Handout).

Johansson, S. (1995). The encoding of spoken texts. *Computers and the Humanities*, *29*(1), 149–158. https://doi.org/10.1007/BF01830708

Johnston, T. (2013). *Auslan Corpus Annotation Guidelines*. Sidney: Macquarie University.

Kipp, M., Neff, M., & Albrecht, I. (2007). An annotation scheme for conversational gesture: how to economically capture timing and form. *Language Resources and Evaluation*, *41*(3/4), 325–339. https://doi.org/10.1007/s10579-007-9053-5

Kupiec, J. (1992). Robust part-of-speech tagging using a hidden Markov model. *Computer Speech and Language*, *6*(1), 3–15. https://doi.org/10.1016/0885-2308(92)90019-Z

Kyto, M., Rissanen, M., & Wright, S. (1994). *Corpora across the Centuries*. Amsterdam: Rodopi.

Leech, G. (1993). Corpus annotation schemes. *Literary and Linguistic Computing*, *8*(4), 275–281. https://doi.org/10.1093/llc/8.4.275

Leech, G. (1997). Introducing Corpus Annotation. In R. Garside, G. Leech & A. McEnery (Eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora* (pp. 1–18). London, Longman.

Leech, G. (2005). Adding Linguistic Annotation. In M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 17–29). Oxford: Oxbow Books.

Leech, G., & Fligelstone, S. (1992). Computers and corpora analysis. In C. S. Butler (Ed.), *Computers and Written Texts* (pp. 115–140). Oxford: Blackwell.

Leech, G., McEnery, T., & Wynne, M. (1997). Further levels of annotation. In R. Garside, G. Leech & T. McEnery (Eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora* (pp. 85–101). London: Longman.

Lu, H. C. (2010). An annotated Taiwanese learners' Corpus of Spanish: CATE. *Corpus Linguistics and Linguistic Theory*, *6*(2), 297–300. https://doi.org/10.1515/cllt.2010.011

McEnery, T., & Hardie, A. (2011). *Corpus Linguistics: Method, Theory, and Practice.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511981395

McEnery, T., & Wilson, A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

O'Donnell, M. B. (1999). The Use of Annotated Corpora for New Testament Discourse Analysis: A Survey of Current Practice and Future Prospects. In S. E. Porter & J. T. Reed (Eds.), *Discourse Analysis and the New Testament: Results and Applications* (pp. 71–117). Sheffield: Sheffield Academic Press.

Oostdijk, N., & deHaan, P. (1994). *Corpus-Based Research into Language*. Amsterdam: Rodopi.

Oostdijk, N., & Boves, L. (2008). Pre-processing speech corpora. In A. Ludeling & M. Kyto (Eds.), *Corpus Linguistics: An International Handbook* (vol.1, pp. 642–663). Berlin: Walter de Gruyter.

Rayson, P., & Stevenson, M. (2008). Sense and semantic tagging. In A. Ludeling & M. Kyto (Eds.), *Corpus Linguistics: An International Handbook* (pp. 564–579). Berlin, Walter de Gruyter.

Rissanen, M. (1989). Three problems connected with the use of diachronic corpora. *International Computer Archive of Modern English Journal*, *13*(1), 16–19.

Santorini, B. (1990). *Part-of-Speech Tagging Guidelines for the Penn Treebank Project* (3rd revision, 2nd printing). ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz

Schmid, H. (2008). Tokenizing and part-of-speech tagging. In A. Ludeling & M. Kyto (Eds.), *Corpus Linguistics: An International Handbook* (vol. 1, pp. 527–551). Berlin: Walter de Gruyter.

Sinclair, J. M. (1994). Trust the text. In M. Coulthard (Ed.), *Advances in written text analysis* (pp. 12–25). London: Routledge.

Sinclair, J. M. (1996). *EAGLES Preliminary recommendations on Corpus Typology*. Retrieved from http://www.ilc.pi.cnr.it/EAGLES96/corpustyp/corpustyp.html

Sinclair, J. M. (2004). *Trust the Text: Language, Corpus, and Discourse.* London: Routledge.

Sperberg-McQueen, C. M., & Burnard, L. (1994). *Guidelines for Electronic Text Encoding and Interchange*. Chicago and Oxford: ACH, ALLC, and AC.

Stenstrom, A.-B. (1984). Discourse tags. In J. Aarts & W. Meijs (Eds.), *Corpus Linguistics: Recent Developments in the use of Computer Corpora in English Language Research* (pp. 65–81). Amsterdam: Rodopi.

Thieberger, N., & Berez, A. L. (2012). Linguistic data management. In N. Thieberger (Ed.), *Oxford Handbook of Linguistic Fieldwork* (pp. 90–118). Oxford: Oxford University Press.

Xiao, R. (2008). Theory-driven corpus research: using corpora to inform aspect theory. In A. Ludeling & M. Kyto (Eds.), *Corpus Linguistics: An International Handbook* (vol. 2, pp. 987–1008). Berlin: Walter de Gruyter.

Zinsmeister, H., Hinrichs, E., Kubler, S., & Witt, A. (2008). Linguistically annotated corpora: quality assurance, reusability, and sustainability. In A. Ludeling & M. Kyto (Eds.), *Corpus Linguistics: An International Handbook* (vol. 1, pp. 759–776). Berlin: Walter de Gruyter.

Copyrights