Towards basic categories for describing properties of texts in a corpus

Serge Sharoff,
Centre for Translation Studies, University of Leeds,
s.sharoff@leeds.ac.uk

The following description is intended to be the standard for describing properties of texts in the majority of corpora to be developed at the University of Leeds with possible wider implications of establishing it as the standard for text encoding in other projects.

1 Introduction

There are several frameworks for describing properties of texts. In particular, Text Encoding Initiative (TEI) provides a very extensive set of tags and attributes for encoding text headers. However, many TEI tags are irrelevant for the purposes of corpus development, while the reduced set from the TEI-Lite guidelines is too narrow for corpus development, e.g. it leaves few options for describing the profile of texts. At the same time the TEI guidelines are not specific enough, because they lack a text typology proper, for instance, taxonomies of basic problem domains or properties of the intended audience. A version of text typology is offered by John Sinclair (1996) within the EAGLES (European Advisory Group on Language Engineering Standards) guidelines. However, unlike TEI it does define a set of tags and attributes. What is more important, it does not always deal with text types that are frequent in general-purpose corpora, such as types of newspaper texts or fiction.

Development of various corpora at the University of Leeds has led to identification of basic categories for describing text properties and the set of XML elements and attributes for encoding them in corpus headers. The proposed set of categories inherits the EAGLES guidelines and amends them on the basis of problems encountered in describing text collections, while every attempt has been made to borrow from the TEI guidelines the set of XML elements and attributes for encoding the categories.

A note on formatting. XML tags and attributes are given using the Arial font, e.g.

<teiHeader id="TRF021" type="text" lang="en">

This corresponds to a tag <teiHeader> with values of attributes id, type and lang. Possible values of an attribute are listed in the text using the | sign: lang="de|en|fr" meaning that the attribute can take one of the three values de, en or fr.

The aim of the study is to define the *minimal* subset for describing texts stored in a corpus using a TEI-compatible markup and a principled text typology. The full set of TEI tags can be used for corpus encoding, if necessary. On the other hand, the description does claim that it is suitable as the general framework for the majority of corpus development projects leaving the possibility to extend only its most delicate classifications.

2 The text typology

According to the TEI guidelines (Sperberg-McQueen, Burnard, 2001), a text stored in a corpus is described by its header (<teiHeader>). From the viewpoint of encoding texts in a corpus we need two obligatory elements in the header:

- a *file description*, tagged <fileDesc>, containing a full bibliographical description of the text … and
- a *text profile*, tagged <profileDesc>, containing classificatory and contextual information about the text

In addition to the text typology, the <profileDesc> section includes the definition of the language(s) used in the text. This is encoded by means of the <langUsage> element. Two other elements of the TEI header are <encodingDesc> and <revisionDesc>. However, information about the history of changes is not typically recorded for corpus files, while the principles of text encoding are (or should be) consistent for all files in a corpus.

The full bibliographical description of the text (<fileDesc>) is retained for documentation purposes. The obligatory bibliographical elements are the text title (<title>), the text size in words (<extent type="w">), and the text source (<sourceDesc>), while other elements, for instance, the author, publisher, publication date, edition, ISBN, etc, are optional. The reason for reducing the set of bibliographical elements is two-fold. First, many types of texts in a corpus lack a complete publication statement, for instance, texts existing only in the electronic form or spoken texts. Second, many types of corpus collection activity do not require the exact bibliographical information, and concentrate instead on the classification of texts to ensure the representativeness and balance of the corpus. It is more natural to consider information about the size of texts as a part the text typology, but the tag has been left in the file description section for the sake of compatibility with TEI.

The text typology proper is stored in the text profile section (mostly in the <textDesc> element) and is based on two text-internal (I) and three text-internal (E) parameters identified in the EAGLES classification (Sinclair, 1996):

E.1. origin – matters concerning the origin of the text that are thought to affect its structure or content.

E.2. state – matters concerning the appearance of the text, its layout and relation to non-textual matter, at the point when it is selected for the corpus.

E.3. aims – matters concerning the reason for making the text and the intended effect it is expected to have.

I.1. topic – the subject matter, knowledge domain(s) of the text.

I.2. style – the patterns of language that are thought to correlate with external parameters.

Categories and TEI tags for each EAGLES category are considered below. The only exception is the split of E.3 into two separate categories. The first describes the intended audience, while the second addresses the aims intended in making the text.

2.1 E1. Origin

E1 is reflected in the TEI guidelines by several dozens of tags and attributes, including those coding the place of birth, the place of writing the text and foreign languages known by the author. The typology proposed by Sinclair is also quite elaborate. They are potentially relevant for describing text properties, but the very elaborate annotation scheme is not practical for a large corpus consisting of several thousands of documents. It is also unlikely that we can get much information about the author and other circumstances, for instance, when we develop a corpus of newspaper texts.

At the same time, both EAGLES and TEI guidelines miss the important issue of authorship, distinguishing texts created by explicitly named authors, texts attributed to a corporate body, and texts created by unknown authors. Corporate authorship assumes that the text represents the position of a corporate body and is typically subjected to external editing. It is the frequent case in coding user manuals, editorials, newswires, advertisements, etc (they typically lack the explicitly named author). Unnamed authors (in contrast to corporate "authors") speak for themselves, but we have no information about them. This is the frequent case in exchange on electronic forums, messages on notice boards, etc.

The minimal set of tags proposed for coding the origin of a written text includes:

1. information about the time of text creation is given within the <creation> element using the standard set of TEI ways for specifying the value attribute in the <date> element (normally, it is sufficient to give the year, but this can be the exact date or the period denoted by from and to attributes; the value of the exact attribute can be used to indicate the degree of precision attributed to the date).
2. information about the authorship is also given within the <creation> element using the <name> element (the named author) with the following authorship types

   type="single|mult|corporate|unknown" – corresponding to texts created by a single author, by several named co-authors (in this case there are several <person> tags with role="author" with respective sex and age attributes, see below), by a corporate author, or an unknown author;
3. information about the author as a person, if it is available, is given within the TEI element <particDesc> (the description of participants in interaction) using the <person> element with the following attributes:

   role="author" (role="speaker" is used for spoken texts, but it requires more information to be given about the speakers);

   sex="m|f" – the author's sex;

   age="child|teen|young|mid|senior" – author's age at the time of text creation (the set of the five values is sufficient for the majority of corpus types, but it can be extended, if necessary for a particular corpus);

A participant can be also described using the dialect and soc (social status) attributes available in the <person> element.

The experience of using the proposed set of categories for coding large corpora shows that typically this information is readily available. The only exceptions are the author's age and first language, which may require extra investigation, so the default choices are age="mid", i.e. 25-60 (the approximate age limits for the unmarked language), and that the author is a standard native speaker, unless there are reasons to believe otherwise, for example, in a corpus of teenage language or a corpus of FL learners.

2.2 E2. State

The primary classification of texts with respect to their physical appearance concerns the two standard speech modes: written and spoken. In addition to them Sinclair (1996) suggests to use the *electronic* mode "to emphasise that language transmitted in electronic media is not quite the same as the older established modes". In the current proposal the use of the electronic mode is restricted to electronic communication, such as emails, electronic forums or chat rooms, because they are similar to spoken communication modes in the spontaneity of production (like face-to-face or telephone conversations), but they lack prosodic information. Another mode (written-to-be-spoken) has been added from the experience in the BNC. The TEI tag for encoding this category is <channel> with the attribute mode="w|s|e|ws". The following taxonomy of written texts is a slightly edited version of the classification proposed by Sinclair (1996):

- printed (texts published for mass production):
  - books
  - news (broadsheet vs. tabloids vs. newswires)
  - magazines
  - ephemera
    - leaflets, pamphlets, brochures,
    - local flyers, junk mail
- typed material (reports and documentation)
- correspondence
  - official
  - personal
- manuscripts (actual handwritten texts, a very unlikely text type in modern corpora)

It is encoded using the taxonomy reference mechanism <catref target="X X" scheme="written" />.

2.3 E3.1. Audience

The TEI and EAGLES guidelines suggest compelementary classification criteria, from which we select a subset applicable to the majority of corpus projects. The audience is described within the <particDesc> element in parallel to information about the author. It is encoded using the <personGrp role="audience"> element with the following attributes:

size="private|small|medium|large|very-large" – this distinguishes between texts aimed at the private audience or audience measured in approximately 100s, 1,000s, 100,000s and millions;

sex="m|f|x" – the sex of the intended audience (the default value is x corresponding to the audience of any sex);

age="child|teen|adult|x" – the age of the intended audience (the default value is adult);

education="high|low|x" – some text are aimed specifically at the higher or lower educated audiences (the default value is x, which means that no preference can be given);

constituency="public|informed|professional" – distinguishing between the general public, informed lay people and professionals;

The classification of the audience with respect to its size may be different in very specific projects, for instance, in those concerned with the private audience or with minor languages (which have less than several million speakers, the notion of the audience size for them has to be scaled). The same applies to the audience age. For instance, projects collecting texts for children may classify their intended age in greater detail.

Our experience with corpus coding shows that parameters of the audience present the biggest problems for coders describing text properties. The decision on the size, sex or education of the intended audience can be made only on the basis of subjective judgement, e.g. can we treat a cookery book as aimed at the female audience? This means that the inter-annotator agreement is quite low and cannot be used as the basis for a subcorpus selection. The problem with audience parameters is also corroborated by the assignment of the audience level codes in the BNC bibliographical database. The audience level for a propaganda leaflet from a brewing company (text A14) is treated as medium, while the audience level for a text from a car magazine (A6W) is low, but the values can be swapped over without any reservation. The parameter of audience constituency (based on EAGLES) looks more reliable and rarely causes a problem in description, but its original set of values, listing also students and specialists, may cause a confusion, so it has been narrowed down to only three values with professionals conflated with specialists and students with informed audience (education, as the most probable aim of production texts for students is described below in E3.2).

2.4 E3.2. Aims

The TEI guidelines provide the element <purpose>, which can be used for encoding the aims classes in the taxonomy <catref target="X X" scheme="aims" />. The original EAGLES scheme has been amended to take into account most frequent text types:

discussion – texts aimed at discussing a state of affairs (including typical newspaper articles, research papers, travel stories, etc); Sinclair proposes the following subtypes argument, position, polemic;

information –Sinclair (1996) restricts the category to reference compendia, while in corpora we find such subclasses as: reference, data (police reports, patents, summaries, etc), newswires (a Reuters message informing about an earthquake differs from a Guardian reportage about rescue efforts on the site, the latter is classified as discussion);

recommendation – recommendations differ from discussions as they provide an incentive for doing or abstaining from doing something; the proposed subclasses differ from the EAGLES set: review, advice, legal, advertisement;

recreation – the two important subclasses are fiction and nonfiction, with the following list of fiction subclasses: genfi, myst-crime-fi, scifi, histfi, adventurefi, lovefi, humorfi, drama, poetry (a modified version of the Brown Corpus list of the fiction genres); the list of nonfiction subclasses follows the EAGLES set: biography, autobiography, memoirs, letters-pub (the latter is the published variety of letters, typically from/to prominent persons);

instruction – with the subset textbook (types of textbooks are distinguished according to their audiences), manual (like flat-pack assembly, software or do-it-yourself manuals), practical-how-to (this category encodes more descriptive text varieties in comparison to manuals);

The TEI guidelines can also describe the level of the text factuality. If necessary, it may be encoded in the tag <factuality> with the attribute type="factual|mostly-factual|mostly-fictional|fictional".

2.5 I1. Domains

Sinclair (1996) mentions the frequent variation of topics within a single document or conversation and rejects the applicability of any general classification system (such as Dewey Decimal Classification). Instead, he lists domains considered in various classification and corpus studies and refers to the unsuitability of "trying to arrange a hierarchy of simple topic labels". However, in practical terms the list of 30 odd domains is too fine-grained. At the same time, development of a corpus in a specific domain may require a more delicate classification. Nevertheless such a classification should start from a node in the hierarchy. Even though any classification of topics is not complete and may be irrelevant for several project types, we risk proposing a set of general categories that can be extended for more delicate studies of domains. The eight first-level categories in the list below aim at the complete coverage of possible domains of corpus collection activity, while second-level categories are provisional and may be amended (or extended) in more delicate projects:

natsci	(mathematics, biology, physics, chemistry, geo, …)
appsci	(agriculture, medicine, ecology, engineering, computing, military, transport, …)
socsci	(law, history, philosophy, psychology, sociology, anthropology, language, education, …)
politics	(inner, world)
commerce	(finance, industry)
life	This is a general domain that is used for fiction, conversation, etc.
arts	(visual, literature, architecture, performing)
leisure	(sports, travels, entertainment, fashion…)

The TEI guidelines provide the element <domain>, which can be used for encoding the domain classes in the taxonomy <catref target="X X" scheme="domain" />.

2.6 I2. Styles

This is another "notorious" notion, because "Although a great deal is talked about style, and there are several parameters of organisation proposed in the literature, there are no agreed standards for any one parameter" (Sinclair, 1996). After reviewing conflicting proposals, he defines style as: "the way texts are internally differentiated other than by topic; mainly by the choice of the presence or absence of some of a large range of structural and lexical features, … e.g. verbs in the active or passive mood, politeness markers and mitigators". At the same time David Lee (2001) claims that "I believe there is actually more consensus on these issues than users of these terms themselves realise", adopts the analysis proposed by Sinclair and offers an example that differentiates styles from genres:

> So when we say of a text, "It has a very informal style," we are characterising not the *genre* to which it belongs, but rather the text producer's use of language in that particular instance (p. 45).

The preliminary classification that is offered by Sinclair and essentially adopted in Lee's analysis contains the following set of parameters: formality (formal vs. informal), preparation (considered vs. impromptu), communicative grouping (conversational group vs. speaker with the audience vs. remote audiences, e.g. radio, TV), and direction (one-way vs. interactive). However, with the exception of the first parameter the categories are applicable to the spoken language only. At the same time, our experience shows the need to distinguish between several classes within the formality parameter for the spoken language and redefine the formality classification for the written one. For nonfiction the set of style includes at least the following values:

neutral-style informal-style formal-style academic-style

On the other hand, fiction (or any piece of writing aimed at recreation) requires its own set of parameters, because there is no space for the formal or academic style in this genre, but there are specific distinctions of its own:

default-fiction-style

regional-style – a marked deviation from the received norm towards a dialect, e.g. Irvine Welsh;

lowly-style – an imitation of the spoken language used by a "lesser-educated" population, often slang, e.g. Henry Miller;

individual-style – a marked way of language use with significant deviations from the neutral style, this style is typically the result of linguistic or stylistic experiments, e.g. James Joyce.

Unlike many other parameters in the proposed classification, which are more or less language-independent, this set is relative to the literary culture of a particular language. Yet another set of parameters may be required for the spoken language in a given culture, cf. the issue of politeness in Japanese. The TEI guidelines offer no tags for describing the level of formality. Our suggestion is to define a flat taxonomy and encode the style as:

<catref target="X X" scheme="fiction-style|nonfiction-style" />

3 Experiments in encoding

3.1 Applications to existing corpus description schemes

Experiments on the application of the proposed scheme to code text types identified in the BNC, Reuters NewsML, as well as attempts to code samples from a corpus of British and Russian newspapers, cf. also (Santini, 2001) on the identification of newspaper genres, the Russian Reference Corpus, the corpus of modern Arabic developed in Leeds, etc.

The classification scheme of the BNC is well documented in the corpus files. It is also described in the bibliographical database created by Adam Kilgarriff (1995). It is no wonder that the classification proposed in the current paper covers the BNC codes, because both are based on the TEI, but the proposed classification scheme is richer, because it effectively distinguishes between text styles and goals of text production. For instance, the BNC coding uses identical codes for describing an article from The British Journal of Social Work (text GWJ) and an article on French smoking habits from the tabloid *Today* (CEK)[1]: both are published in periodicals and belong to the domain of humanities, there is a code distinguishing the audience level, but both texts are coded as medium (2).[2] In addition to these parameters, the proposed scheme codes the aims of text production (discussion, instruction, recommendation, etc), its style (neutral or academic) and circulation (very large vs. small). This helps in distinguishing such texts.

The classification scheme of Reuters NewsML is used for encoding the Reuters corpus, the complete collection of newswires for one year (Rose, et al, 2002). The classification of texts in the Reuters Corpus contains an impressive list of more than 800 industry codes, 126 topic codes, including subclasses of news from the business world and general topic codes (the latter are prefixed with G), and a list of about 370 regions, including international organisations. A typical news item is classified by several industry and topic codes, for instance, an article “Canada delivers war planes to Botswana” (21/12/96) is described in terms of topics as DEFENCE CONTRACTS (C331), CORPORATE/INDUSTRIAL (CCAT), GOVERNMENT/SOCIAL (GCAT) and DEFENCE (GDEF). Thus, the classification considers mostly informational properties (following the nature of texts) and is too fine-grained for linguistic-oriented corpus-development projects (though it is very useful for IR projects). In terms of the proposed classification, texts in the Reuters corpus have the following set of fixed parameters: author type (corporate), state (printed, newswires), and style (neutral). The parameters that vary depending on the message are the audience size (medium for financial news to very large for general topics), constituency (from professionals to public) and aims (even though the majority of texts are informational, some of them are aimed at discussion or recommendation).

3.2 Examples of encoding of text types

Classes from taxonomies are assigned using the <catref> tag. This allows tagging with multiple classes (this is unlike attributes that require selection of a single value). The following generic structure describes written texts:

```
<teiHeader id="ID" target="filename" type="text" lang="en">
  <fileDesc>
        <titleStmt>
```

[1] Actually the BNC contains the complete content of the two text sources, so it does not distinguish codes for separate articles.

[2] This is another example of the problem with coding the audience level.

```
                <title>title</title>
                <author>name</author>
                <extent type="w">document length in words</extent>
        </titleStmt>
        <publicationStmt>                              // only for published texts and if necessary
                <publisher>publisher</publisher>
                <editor>X</editor><bookTitle>X</bookTitle>     // for collections if necessary
                <date>year</date>
        </publicationStmt>
        <sourceDesc>
                <respStmt><resp>the statement of responsibility</resp></respStmt>
                <address>a reference to the source of the document, e.g. URL</address>
        </sourceDesc>
  </fileDesc>
  <profileDesc>
        <creation><date value=”yyyy-mm-dd”/><name type="type"/></creation>
        <particDesc><person role="author" type="type" sex="m|f" age="type"/>
                <personGrp role="audience" sex="m|f|x" size="type" age="adult|child|teen|x"/>
        </particDesc>
        <textDesc>
                <channel mode="s|w|ws|e"><catref target="X X" scheme="spoken|written"/>
                </channel>
                <factuality type="fiction|mostly fiction|mostly factual|factual"/>
                <purpose><catref target="X X" scheme="aims"></purpose>
                <domain><catref target="X X" scheme="domain"></domain>
                <catref target="X X" scheme="fiction-style|nonfiction-style">
        </textDesc>
  </profileDesc>
</teiHeader>
```

3.2.1 A paper from the Guardian

```
<teiHeader id="BR-GU-001" target="BR-GU-001.xml" type="text" lang="en">
  <fileDesc>
        <titleStmt>
                <title>Movie stars get hung up on KGB's anti-hangover drug</title>
                <author>Nick Paton Walsh</author>
                <extent type="w">293</extent>
        </titleStmt>
        <sourceDesc>
                <respStmt><resp>Guardian</resp></respStmt>
                <address>http://www.guardian.co.uk/print/0,3858,4759175-103610,00.html</address>
        </sourceDesc>
  </fileDesc>
  <profileDesc>
        <creation><date value=”2003-09-23”/><name type="sole"/></creation>
        <particDesc><person role="author" sex="m" age="mid"/>
              <personGrp role="audience" sex="x" size="large" age="adult" education=“x“
                  constituency=”public”/>
        </particDesc>
        <textDesc>
                <channel mode="w"><catref target="printed news broadsheet" scheme="written"/>
                </channel>
                <factuality type="factual"/>
                <purpose><catref target="discussion recreation" scheme="aims"></purpose>
                <domain><catref target="leisure entertainment" scheme="domain"></domain>
                <catref target="neutral" scheme="nonfiction-style">
        </textDesc>
  </profileDesc>
</teiHeader>
```

3.2.2 A research paper by Demetriou and Atwell

```
<teiHeader id="SCI-AC-001" target="SCI-AC-001.xml" type="text" lang="en">
  <fileDesc>
        <titleStmt>
                <title>A domain-independent semantic tagger for the study of meaning associations in
English text </title>
                <author>Demetriou, G and Atwell, E.</author>
                <extent type="w">3842</extent>
        </titleStmt>
        <sourceDesc>
                <respStmt><resp>Link from Eric Atwell’s homepage</resp></respStmt>
                <address>http://www.comp.leeds.ac.uk/eric/iwcs.ps</address>
        </sourceDesc>
  </fileDesc>
  <profileDesc>
        <creation> <date value=”2000”/> <name type="mult"/></creation>
        <particDesc><person role="author" sex="m" age="mid"/>
                <person role="author" type="mult" sex="m" age="young"/>
                <personGrp role="audience" sex="x" size="small" age="adult" education=“high“
                  constituency=”professional”/>
        </particDesc>
        <textDesc>
                <channel mode="w"><catref target="printed books" scheme="written"/>
                </channel>
                <factuality type="factual"/>
                <purpose><catref target="discussion" scheme="aims"></purpose>
                <domain><catref target="appsci computing language" scheme="domain"></domain>
                <catref target="academic-style" scheme="nonfiction-style">
        </textDesc>
  </profileDesc>
</teiHeader>
```

3.2.3 A weblog by Dug Falby

```
<teiHeader id="WEB-BLOG-001" target="WEB-BLOG-001.xml" type="text" lang="en">
  <fileDesc>
        <titleStmt>
                <title>A donkey on the edge</title>
                <author>Dug Falby </author>
                <extent type="w">2010</extent>
        </titleStmt>
        <sourceDesc>
                <respStmt><resp>Weblog page taken 01/10/2003</resp></respStmt>
                <address>http://www.donkeyontheedge.com/2.html</address>
        </sourceDesc>
  </fileDesc>
  <profileDesc>
        <creation><date value=”2003-08”/><name type="sole" /></creation>
```

```
          <particDesc><person role="author" sex="m" age="mid"/>
                    <personGrp role="audience" sex="x" size="medium" age="adult" education="high"
                         constituency="public"/>
          </particDesc>
          <textDesc>
                        <channel mode="e"/>
                        <factuality type="mostly-factual"/>
                        <purpose><catref target="discussion recreation" scheme="aims"></purpose>
                        <domain><catref target="life leisure travel" scheme="domain"></domain>
                        <catref target="informal-style" scheme="nonfiction-style">
          </textDesc>
   </profileDesc>
</teiHeader>
```

4 The organisation of documents

It is quite probable that documents (especially if they are POS-tagged or annotated in any other way) are stored separately from the headers. In order to link headers to filenames, the latter are stored in the headers in the target attribute.

If documents are also annotated according to TEI guidelines, we can adopt a subset of tags that can be added without extra processing (the # sign before a tag indicates that the tag is unlikely to occur in a large corpus):

```
<text><body>
# <div>                    // division, e.g. a chapter
 <p>
 <s>
 # <cl type="AnyClauseType">
 # <phr type="AnyPhraseType">
 <w> word
 <ana lemma="xx" pos="yy">analysis</ana>    // this allows us to add ambiguous analyses
 <ana lemma="xx1" pos="yy1">ambiguity</ana>
 </w>
 #</phr>
 #</cl>
 </s>
 </p>
 #</div>
</body></text>
```

5 Conclusions

Use of the unified set of categories for describing properties of texts has two main advantages.

First, it helps to position a corpus under development with respect to a reference corpus covering all possible features by explicit selection of a subset of features to be considered in the study, e.g. we are going to create a corpus in a specific applied domain (medicine) consisting of texts aimed at a general audience, but allow a variation with respect to text aims (information, discussion, recommendation, instruction), audience size, text size, etc. Currently a corpus collection activity starts with defining an *ad hoc* set of categories aimed at a specific project and the classification resulted from the project is not compatible with anything else, e.g. (Santini, 2001). The existing standards, like EAGLES and TEI, are rarely used, first because of their all-encompassing nature (in Sinclair’s words they are “both exhaustive and exhausting to contemplate”), and second because they do not address specific aspects of corpus collection. The proposed classification fits between the two extremes: it is principled and suitable for the majority of projects and at the same time it is manageable in its size and the number of categories used. Also a corpus using the proposed scheme benefits from constraining the space for defining texts types to be included in the corpus.

Second, the standard scheme provides the possibility to design corpus management software that is aware of the text typology to select subcorpora according to text properties and reuse the software across corpus development projects.

I hope that the set of categories proposed in the current document can serve as the basis for collaborative corpus development activity.

References

Kilgarriff, A., (1995). The BNC bibliographical database. ftp://ftp.itri.bton.ac.uk/bnc/bib-dbase

Lee, D. (2001). Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology* Vol. 5, No. 3, September 2001, pp. 37-72. http://llt.msu.edu/vol5num3/pdf/lee.pdf

T.G. Rose, M. Stevenson and M. Whitehead, (2002) The Reuters Corpus Volume 1 – from Yesterday's News to Tomorrow's Language Resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria, 29-31 May 2002. http://about.reuters.com/researchandstandards/corpus/LREC_camera_ready.pdf

Santini, M (2001) Text typology and statistics. Explorations in Italian press subgenres, *Italian Journal of Linguistics/Rivista di linguistica*, Volume 13, Issue 2, 2001, p. 339-374. http://www.itri.brighton.ac.uk/~Marina.Santini/articolo_bertinetto.pdf

Sinclair, J. (1996). *Preliminary recommendations on text typology*. EAGLES Document EAG-TCWG-TTYP/P. http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html

Sperberg-McQueen, C. M., Burnard, L. (eds.) (2001). *Guidelines for Electronic Text Encoding and Interchange*. http://www.hcu.ox.ac.uk/TEI/P4X/index.html