Published at: Language Resources and Evaluation Conference, May 2004, Lisbon, Portugal

Towards basic categories for describing properties of texts in a corpus

Serge Sharoff

Centre for Translation Studies
School of Modern Languages and Cultures
University of Leeds, Leeds, LS2 9JT, UK
s.sharoff@leeds.ac.uk

Abstract

The paper discusses the basic principles for describing properties of texts to be stored in a corpus and suggests the standard that is used in the majority of corpora developed at the University of Leeds and can be potentially employed for describing texts in any corpus collecting activity. The standard defines the minimal subset of tags and attributes that are necessary for describing texts stored in a corpus. The proposed text typology 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.

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, for instance, because they are aimed at library activities, while the reduced set from the TEI-Lite guidelines is too narrow, e.g. it leaves few options for describing the profile of texts. At the same time, even though the TEI guidelines are huge, they are not specific enough, because they lack a text typology proper, for instance, they do not suggest the taxonomy of basic problem domains or the set of properties of the intended audience. A version of a text typology is offered by John Sinclair (1996) within the EAGLES (European Advisory Group on Language Engineering Standards) guidelines. However, unlike TEI it does not 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, so it requires an extension.

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. XCES (XML Corpus Encoding Standard) offers another set of XML elements, including those used for describing text "headers". The XCES set is compatible with the TEI guidelines, but it does not add any extra text typology scheme, so the description below concentrates on the use of TEI and EAGLES.

The aim of the study is to define the *minimal* subset of tags and attributes that are necessary 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 metatextual annotation in the majority of corpus development projects leaving the possibility to extend only its most delicate classifications. The present text discusses only the basic parameters for describing texts in the corpus. The complete version of the guidelines, including exact names of tags and values, as well as elaborate examples, is available from (Sharoff, 2003).

2. The text typology

2.1. Preliminary considerations

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 …
- a *text profile*, tagged <profileDesc>, containing classificatory and contextual information about the text.

The bibliographical description of the text is retained for documentation purposes, but the set of obligatory bibliographical elements is reduced to the text title, the text size in words, and the source from which the text was received, while other elements possible in TEI, 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.

Below I consider specific encoding guidelines according to categories defined in EAGLES. 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.

One important issue concerns the representation of the ambiguity of parameters: some categories are mutually exclusive and allow only one reasonable value from the set (for instance, the sex of the author, if it is known), while others allow several interpretations (for instance, the text topic, when political, economical and medical issues are discussed in one text). It is natural to describe parameters of the first type in terms of values of attributes, for instance, sex="m|f|u", meaning that the attribute can take *one* of the three values m, f or u. For parameters of the second type, TEI provides the possibility to define taxonomies and refer to them using catref tags, for instance, <catref target="appsci politics" scheme="topic" />

2.2. 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, for instance, foreign languages known by the author and circumstances of text production, 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:

- information about the time of text creation (it is sufficient to give the year or the period of several years)
- information about the authorship with the following authorship types: single – created by a single author, mult – by several named co-authors, corporate – by a corporate author, unknown – by an unknown author;
- information about the author as a person, if it is available, is given within the TEI element <particDesc>, including the following attributes: role="author|speaker", sex="m|f", and age.

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. Typically, it is enough to keep the default choices age="mid", i.e. 25-60 (the approximate age limits for the unmarked language), and assume 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, or a corpus of dialects.

2.3. 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".

Written texts can be classified into printed texts (published for mass production), typed materials (reports and documentation), and correspondence (official and personal). Sinclair (1996) distinguishes between four types of printed texts, such as books, newspapers, magazines and ephemera, however, it seems sensible to have a class for news with further distinction between broadsheet, tabloids and newswires.

2.4. E3.1. Audience

In this respect both TEI and EAGLES guidelines suggest compelementary classification criteria, which are too diverse for the majority of corpus projects. In our projects we adopted a subset to which both classifications contribute. First, we encode the size of the audience, distinguishing between texts aimed at the private audience or the public audience measured in approximately 100s, 1,000s, 100,000s and millions. For some applications, we will need to distinguish the sex and/or age of the intended audience, for instance, child, teen, adult, senior. Finally, we should distinguish between two parameters related to the level of education of the intended audience:

- education in general, coded as high, low or 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)
- audience constituency with respect to specific profession, coded as: public, informed or 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 in-

stance, 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 a 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 audience level codes in the BNC bibliographical database (audience level coding in the BNC roughly corresponds to the general audience education in the proposed scheme). As the result, the audience level for a propaganda leaflet from a brewing company (text A14 in the BNC) 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 purpose of production texts aimed for students is described below in E3.2).

2.5. E3.2. Aims

Aims of text production are not mentioned in either the TEI guidelines or the BNC, however, they are important for corpus development. The EAGLES guidelines pay more attention to them, but the original EAGLES scheme is not well documented and needs amendments 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, but does not elaborate on their specific properties;
- 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: 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);

2.6. 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 all possible domains of corpus collection activity, while second-level categories are provisional and may be amended (or extended) in more delicate projects:

natsci (maths, 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 labels associated with states whenever possible follow the practice of the domain codes used in the BNC (Kilgarriff, 1995), but some have been changed to reflect additional dimensions of classification, such as the goals of text production, or to generalise over topics.

2.7. 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 (Lee, 2001: 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 of formality for the spoken language and redefine the formality classification for the written one to take into account the variety of styles in fiction (such as lowly or regional) and nonfiction, this includes a classification of styles in the most frequent registers (such as academic or informal).

3. Experiments in encoding

We made an experiment comparing the proposed scheme to codes identified in the BNC, Reuters NewsML and the identification of newspaper genres from (Santini, 2001). We also made several experiments on coding samples from a corpus of British and Russian newspapers, the Russian Reference Corpus, the corpus of modern Arabic (the corpora under development in Leeds).

The classification scheme of the BNC is well documented in the corpus files. We used its snapshot as described in the BNC bibliographical database (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. However, the proposed classification scheme offers more choices, 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) to distinguishing between 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, CORPORATE/ INDUSTRIAL, GOVERNMENT/SOCIAL and DEFENCE. 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).

4. 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. 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.

5. 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* 5(3), 37--72. http://llt.msu.edu/vol5num3/pdf/lee.pdf

Rose, T., Stevenson, M., & Whitehead, M. (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.

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

Sharoff, S (2003). The guidelines for describing properties of texts stored in a corpus. UoL Technical Report. URL http://www.comp.leeds.ac.uk/ssharoff/texts/text-typology.pdf

Sharoff, S (2004). Methods and tools for development of the Russian Reference Corpus. In D. Archer, A. Wilson, P. Rayson (eds.) *Corpus Linguistics Around the World*. Amsterdam: Rodopi.

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

XCES, (2002). XML Corpus Encoding Standard, version 0.2. http://www.cs.vassar.edu/XCES/

[1] Actually the BNC contains the complete content of the two text sources and does not use codes for separate articles.

[2] This is another example of the problem with the interannotator agreement, when coding the audience level.