

A Standard for Encoding Linguistic Corpora

Nancy Ide¹, Jean Veronis²

*Department of Computer Science, Vassar College, Poughkeepsie, New York 12601 USA
Laboratoire Parole et Langage/CNRS, Université de Provence, 29, Avenue Robert Schuman, Aix-en-Provence, France*

KEYWORDS: corpora, text encoding, SGML

AFFILIATION:¹Department of Computer Science, Vassar College USA
^{1, 2} Laboratoire Parole et Langage, CNRS, France

E-MAIL: ide@cs.vassar.edu
veronis@univ-aix.fr

FAX NUMBER: +1 914 437 7187

PHONE NUMBER: +1 914 437 5988

Abstract

The computational linguistics community has recently revived its interest in the use of empirical methods, thus creating a demand for large-scale corpora. Numerous data-gathering efforts exist on both sides of the Atlantic to provide wide-spread access to both mono- and bi-lingual resources of sufficient size and coverage for data-oriented work, including the U.S. Linguistic Data Consortium, the European Corpus Initiative (ECI), ICA-ME, the British National Corpus (BNC), and recently, the European Language Resources Association (ELRA). The rapid multiplication of such efforts has made it critical for the language engineering community to create a set of standards for encoding corpora.

The MULTEXT project and the EAGLES subgroup on Text Representation have joined efforts to develop a Corpus Encoding Standard (CES) optimally suited for use in corpus linguistics and language engineering applications, which can serve as a widely accepted set of encoding standards for European corpus work. The first goal is the identification of a minimal encoding level that corpora must achieve to be considered standardized in terms of descriptive representation (marking of structural and linguistic information) as well as general architecture (so as to be maximally suited for use in a text database). The standard also provides encoding conventions for more extensive encoding of linguistic corpora, and for linguistic annotation.

The CES is an application of SGML (ISO 8879: 1986, Information Processing—Text and Office Systems—Standard Generalized Markup Language). It is based on and in broad agreement with the

TEI Guidelines for Electronic Text Encoding and Interchange. The TEI Guidelines were expressly designed to be applicable across a broad range of applications and disciplines and therefore treat not only a vast array of textual phenomena, but are also designed with an eye toward the maximum of generality and flexibility. The CES, on the other hand, treats a specific domain and set of applications, and can therefore be more restrictive and prescriptive in its specifications. In addition, because the TEI is not complete, there are some areas of importance for corpus encoding that the TEI Guidelines do not cover. Therefore, the first major task in developing the CES has involved evaluating, adapting, selecting from, and extending the TEI Guidelines to meet the specific needs of corpus-based work.

Overview of the CES

In its present form, the CES provides the following:

- a set of metalanguage level recommendations (particular profile of SGML use, character sets, etc.);
- tagsets and a DTD for documentation of the encoded data;
- tagsets, DTDs, and recommendations for encoding textual data, including written texts across all genres, for the purposes of corpus-based work in language engineering.

The significant differences between the CES and the TEI are:

- The CES does not adopt the TEI strategy of building a single, large DTD using various modules (see Burnard, 1995). Instead, the CES comprises a set of individual DTDs for use with different documents.
- CES headers are stored independently of the text, in a central directory, in a separate SGML document with its own DTD.

Types of information covered by the CES

We distinguish three broad categories of information which are of direct relevance for the encoding of corpora for use in corpus linguistics.

A. Documentation

This includes global information about the text, its content, and its encoding. For example:

- bibliographic description of the document;
- documentation of character sets and entities;
- description of encoding conventions; etc.

B. Primary data

Within the primary data, we distinguish two types of information:

- Gross structure: This includes universal text elements down to the level of paragraph, which is the smallest unit that can be identified language-independently; for example:
 - structural units of text, such as volume, chapter, etc., down to the level of paragraph; also footnotes, titles, headings, tables, figures, etc.;
 - features of typography and layout, for previously printed texts: e.g., list item markers;
 - non-textual information (graphics, etc.). etc.
- Sub-paragraph structures: This includes elements appearing at the sub-paragraph level which are usually signalled (sometimes ambiguously) by typography in the text and which are language-dependent; for example:
 - orthographic sentences, quotations;
 - orthographic words;
 - abbreviations, names, dates, highlighted words; etc.

C. Linguistic annotation

This type of information enriches the text with the results of some linguistic analyses; most often in language engineering applications, such analysis is at the sub-paragraph level. For example:

- morphological information;
- syntactic information (e.g., part of speech, parser output);
- alignment of parallel texts;
- prosody markup; etc.

Encoding documentation

As noted above, the CES header is stored in a separate SGML document with its own DTD, and all text headers for the corpus are stored in a central directory together with a corpus header describing the corpus as a whole. This strategy has the following advantages:

- Parts or all of a corpus may be stored in different directories or in remote sites, while information about the component texts is retained in a single repository.
- The header can have a DTD which is different from the DTD for the text, which in turn
 - enables a modularity that SGML does not provide, so that it is possible to define the content of elements common to the header and text (e.g., title, author, etc.) in a way which is appropriate to each context, and so that changes to the same element in one context do not affect the other.
 - in those cases where it is appropriate, enables using the TEI header with a CES conformant text.

- can facilitate processing by corpus-handling tools, for which the header is often irrelevant, since the text can be easily handled separately.
- conversely, it enables using the CES header with an SGML-encoded text which is not itself CES conformant; this is advantageous in the early stages of corpus preparation, where the text may temporarily be in a freer SGML format such as Rainbow, TEI Lite, FORMEX, etc.
- The user does not necessarily need to know where a corpus or text is stored to access it.

The CES header, which is now very nearly a subset of the TEI header, is an area which needs more development. We see that it will eventually be possible to provide for precise pieces of information in a rigid structure, tailor-made to suit the needs of corpus work, that will facilitate retrieval. We also see the need to provide additional fields for the headers of annotation data. By the time of the ALLC/ACH conference the CES header should be more fully developed along these lines.

Encoding primary data

The CES has also been developed taking into account several practical realities surrounding the encoding of corpora intended for use in linguistic research and applications. In particular, at the present time and for the foreseeable future, the majority of corpora will be adapted from legacy data, that is, pre-existing electronic data encoded in some arbitrary format (typically, word processor, typesetter, etc. formats intended for printing). The vast quantities of data involved and the difficulty (and cost) of the translation into usable formats imply that the CES must be designed in such a way that this translation does not require prohibitively large amounts of manual intervention. In many instances, the markup that would be most desirable for the linguist is not achievable by automatic means. Therefore, the CES provides a series of Document Type Definitions (DTDs) for various levels of primary data encoding, corresponding to increasing enhancement in the amount of encoded information and increasing precision in the identification of text elements. Among these levels, the CES identifies a minimum level of encoding required to make the corpus (re)usable across all possible language engineering applications.

The development of this part of the CES has demanded the most detailed consideration and will receive ample treatment in the presentation. Briefly, it has required:

- Identification of those elements which are automatically retrievable from legacy data, and a mapping among elements according

to increasing degrees of refinement (and, usually, cost of capture)—for example, italics can be automatically transduced to <hi> (highlighted); it usually requires human intervention to determine that the highlighting signifies a foreign word (<foreign>), and even more work may be required to ascertain that the element is a technical term (<term>); thus we can identify a sequence of increasingly precise encodings <hi> -> <foreign> -> <term> , which in turn enables defining the minimum, recommended, and desirable encodings achievable.

- Provision of a precise semantics for tag content, and in particular for those elements of special interest for corpus linguistics (sentence, word, etc.).
- Identification and provision of encoding specifications for those elements which comprise unbreakable “tokens” (names, dates, etc.), based on linguistic criteria as well as processing needs.
- specifications for encoding dialogue, and especially for handling the overlapping hierarchies that sometimes exist between dialogue and sentence markup.

Encoding linguistic annotation

The CES provides a set of DTDs for encoding linguistic analyses commonly associated with texts in language engineering, currently including:

- Segmentation of the text into sentences and words (tokens);
- Morpho-syntactic tagging;
- Parallel text alignment.

The CES adopts a strategy whereby annotation information is not merged with the original, but rather retained in separate SGML documents and linked to the original or other annotation documents. Linkage between documents is based on the HyTime-based TEI addressing mechanisms. This approach has several advantages for corpus-based research:

- It avoids the creation of potentially unwieldy documents—envison, in a worst case, a single document containing segmentation and part of speech markup, plus markup for alignment with translations in several languages, plus alignment with the speech recording, plus variant part of speech taggings from several taggers, etc.
- The original or hub document remains stable and is not modified by any process which may add annotation.
- It avoids problems with markup containing overlapping hierarchies.

- Different versions of the same kind of annotation (e.g., different POS annotation) can be associated with the text.
- Annotation can be accomplished by associating the SGML original or other annotation documents with other, pre-existing documents—e.g., instead of generating a document containing morphosyntactic markup and linking it to the original, links could be made directly with lexicon entries.

Thus, in our scheme the hyper-document comprising each text in the corpus and its annotations consists of several documents. The base or “hub” document is the unannotated document containing only primary data markup. The hub document is “read only” and is not modified in the annotation process. Each annotation document is a proper SGML document with a DTD, containing annotation information linked to its appropriate location in the hub document. The precise data architecture, including linking mechanisms, etc. will be described in full in the presentation.

Conclusion

The CES exists in a first draft, and the standard will continue to evolve on the basis of input and feedback from users. We see this process as essential; it is not possible to develop a priori a standard which can address every need for corpus-based work. It is also necessary to allow for the continued development of the CES even while large amounts of text are being encoded according to its specifications. Therefore, the CES is being developed “bottom-up”, beginning with relatively minimal specifications to which we can easily add, rather than attempting to be comprehensive at the outset. In principle, earlier versions of the CES will be upwardly compatible with later versions, so that texts encoded using earlier versions are not rendered obsolete.

At the ALLC/ACH conference in Paris in 1994, we presented a paper outlining a broad set of encoding principles for the design of an encoding scheme suited to linguistic corpora. This paper is intended as a complement to the earlier paper, by giving the details of the encoding scheme that has been built on those principles. We will provide a fuller overview in the presentation of various technical details and considerations for the encoding of linguistic corpora, not covered here due to lack of space.