Encoding Linguistic Corpora

Nancy IDE
Department of Computer Science
Vassar College
Poughkeepsie, New York 12604-0520
ide@cs.vassar.edu

Abstract

This paper describes the motivation and design of the Corpus Encoding Standard (CES) (Ide, *et al.,* (1996); Ide, 1998), an encoding standard for linguistic corpora intended to meet the need for the development of standardized encoding practices for linguistic corpora. The CES identifies a minimal encoding level that corpora must achieve to be considered standardized in terms of descriptive representation (marking of structural and linguistic information). It also provides encoding conventions for more extensive encoding and for linguistic annotation, as well as general architecture for representing corpora annotated for linguistic features. The CES has been developed taking into account several practical realities surrounding the encoding of corpora intended for use in language engineering research and applications. Full documentation of the standard is available on the World Wide Web at http://www.cs.vassar.edu/CES/.

Introduction

Today, corpora are considered to be indispensable to NLP work: they provide information for the creation of other resources (e.g., lexicons), enable the gathering of statistics on real-language use to inform theories and algorithms, and provide the raw materials for testing and training. Their importance is widely acknowledged: the creation of the Linguistic Data Consortium (LDC) in the United States and the European Language Resources Association (ELRA) in Europe shows the commitment of funding agencies on both sides of the Atlantic to gathering and distributing corpora for research use.

In addition to creating large-scale corpora, it is also necessary to develop standards for their encoding, in order to ensure their usability and, most importantly, reusability in corpus-based NLP work. Many freely available tools for language-related tasks such as segmentation, part of speech tagging, etc., exist, and even more in-house tools exist in labs and research centers. Input and output formats for these tools are rarely, if ever, compatible with each other, nor with the encoding formats in available corpora. Translation among formats is not a matter of simple transduction: sometimes the information needed by a tool does not exist in the data; sometimes it is not unambiguously translatable; sometimes the tool cannot retain information present in the original data and it is lost in processing. As a result, enormous amounts of research time and effort are currently spent massaging data and tools for compatibility. This in itself motivates establishing common encoding formats, to avoid redundant effort. This need has been acknowledged in Europe for several years, through efforts such as EAGLES. Recently, recognizing the amount of time and effort involved in creating and annotating corpora, this need has gained the attention of North American researchers and funders as well (see, in particular, the conclusions of an NSF-sponsored international workshop on the future directions of NLP research [Hovy and Ide, 1998]).

Designing a coherent encoding scheme is by no means trivial. It demands, first, the development of a sound model of the data to be represented and all its relevant features and attributes, as well as their structural, logical, linguistic, etc. relationships; together with consideration of processing needs. The format should provide for incremental encoding, allowing for enhancement of data with various kinds of annotation. Very few encoding formats have been designed with such considerations in view, resulting in the proliferation of a variety of encoding schemes (even within a common SGML/XML framework) which are, all too often, poorly designed and ultimately unsuitable for extensive use.

This paper describes the motivation and design of the Corpus Encoding Standard (CES) (Ide, *et al.,* (1996); Ide, 1998), an encoding standard for linguistic corpora intended to meet the need for the principled development of standardized encoding practices for linguistic corpora. The CES was initiated within the European projects

EAGLES (in particular, the EAGLES Text Representation subgroup) and Multext (EU-LRE), together with the Vassar/CNRS collaboration (supported by the U.S. National Science Foundation). The CES has so far been used in several pan-European corpus encoding projects, including PAROLE1 and TELRI2, as well as numerous smaller projects in both Europe and North America, and it has recently been adopted as a basis for the TIPSTER document attributes and annotation3.

1 Goals

The CES is an application of SGML4 (ISO 8879:1986, Information Processing--Text and Office Systems--Standard Generalized Markup Language), conformant to the *TEI Guidelines for Electronic Text Encoding and Interchange* (Sperberg-McQueen and Burnard, 1994). The CES is designed for encoding corpora used as a resource across a broad range of language processing applications, including machine translation, information retrieval and extraction, lexicography, etc. Corpora are used primarily in these applications to gather real language evidence, both qualitative and quantitative; therefore the CES is designed to enable the common operations such as extraction of sub-corpora; sophisticated search and retrieval (e.g., collocation extraction, concordance generation, generation of lists of linguistic elements, etc.); and the generation of statistics (frequency information, averages, mutual information scores, etc.).

The design of the CES is motivated by three overall goals:

- to provide encoding specifications that include elements relevant to language processing work and reflect best practice in the field, and that are both flexible enough to accommodate the range of current use and precise enough to provide clear guidelines for encoders and annotators;
- to minimize the costs of corpus creation, annotation, and use;
- to provide specifications that ensure the maximum of usability and reusability with corpus-processing software and in integrated platforms.

Each of these goals is discussed more fully below. A full treatment of the CES encoding principles and goals can be found in Ide and Véronis (1993) and in the CES documentation (Ide, *et al.,* 1996).

1.1 Meeting the needs of corpus-based work

1.1.1 Adaptation of the TEI Guidelines

The CES uses the TEI scheme as a starting point. However, the *TEI Guidelines* are designed to be applicable across a broad range of applications and disciplines and therefore treat a vast array of textual phenomena beyond what is needed for a particular application. Therefore, the CES limits the TEI scheme to include only the sub-set of the TEI tagset relevant for corpus-based work. The CES also makes choices among TEI encoding options, constraining or simplifying the TEI specifications as appropriate; for example:

- element content models are substantially simplified, in order to maximize the ability to validate5 encoded documents;
- attributes and attribute values are constrained or extended to serve the needs of corpus-based applications;
- the TEI element and attribute class strategy is adopted, but the classes are simplified to form a shallow hierarchy with no overlaps;
- the *TEI Guidelines* are extended to meet the specific needs of corpus-based work; in particular:
  - addition of elements and DTD fragments for areas not covered by the TEI (e.g., detailed encoding of morpho-syntactic annotation)
  - specification of precise values for some attributes
  - specification of required, recommended, and optional elements to be marked
  - detailed semantics for elements relevant to language engineering (e.g., sentence, word, etc.)

Beyond this common basis, the CES diverges from the TEI in two major ways: first, in its data architecture and supporting set of DTDs (as opposed to the single, encompassing TEI DTD); and second, in its overall design philosophy. The TEI began development over ten years ago, when very few encoded texts existed; as a result,

---

1 <http://www2.echo.lu/langeng/en/le2/le-parole/le-parole.html>.

2 <http://www.ids-mannheim.de/telri/telri.html>.

3 http://crl.nmsu.edu/twg.annotation/

4 The CES is currently being updated for conformance to XML (The Extensible Markup Language).

5 Validation is the process by which software checks that the formal specifications of a Document Type Definition (DTD), which is a BNF description of legal tag syntaxare adhered to in a document's markup (e.g., tags are properly nested, appear in the correct order, contain all required tags; attributes appear when and only when they should, have valid values; etc.).

the *TEI Guidelines* were developed in large part in the absence of prior experience and practice. Given the benefit of the TEI experience and the initiation of many corpus-encoding projects over the past ten years, the CES can approach the development of a standard for encoding linguistic corpora from a different perspective. In particular, the CES is being developed incrementally, evolving as consensus and best practice emerge within the community. Accordingly, the CES so far includes specifications for basic document structure, sentences, words, and several other sub-paragraph elements, as well as encoding conventions for incorporating part of speech and alignment information. We are working on extending the CES for additional kinds of markup (terminology, speech, discourse, lexicons, etc.), but we rely on user input and established practice for continued development of the scheme.

1.1.2 Guidelines for encoding legacy data

Most data encoded for NLP work is adapted from legacy data, that is, pre-existing electronic data encoded in some arbitrary format (typically, word processor, typesetter, etc. formats intended for printing). This process, called *up-translation,* involves translating existing encoding, that describes the printed presentation of the text (e.g., font shifts, page breaks, etc.) into an encoding which is suitable as a basis for general use. The resulting encoding uses *descriptive* markup to identify the logical and structural parts of a text. Because up-translation is common and costly, the CES provides guidelines for the process.

In general, it is descriptive markup that is important for corpus-based research, and information about rendition in a printed original can or even must be ignored. For example, if the abbreviation for "number" (No.) is rendered sometimes with a superscripted "o" and sometimes not, a search on a document that retains the text in its original rendition will not identify the two as instances of the same linguistic element. However, for some applications it is necessary to retain certain information about printed rendering (e.g., in machine translation, where the resulting translated text must be rendered in the same fonts, etc.--but obviously not with the same line breaks--as the original). Therefore, the CES recommends retaining rendition information when it is cost-free or when it is required, but provides means to retain it while appropriately representing the content for the purposes of search and retrieval.

1.2 Minimizing costs

1.2.1 Minimizing creation costs

The minimization of the costs of corpus creation is a primary goal of the CES. The vast quantities of data involved and the difficulty (and cost) of up-translation into usable formats dictate that the CES be designed in such a way that this translation does not require prohibitively large amounts of manual intervention to achieve minimum conformance to the standard. However, the markup that is most desirable for the linguist is not achievable by fully automatic means. Therefore, a major feature of the CES is the provision for a series of increasingly refined encodings of text, beyond the minimum requirements.

The first level of encoding is the minimum level required to make the corpus (re)usable across all possible language processing applications. Encoding at this level includes elements often signalled by typography in the original (e.g., paragraph breaks) and is therefore achievable by fairly inexpensive, automated means. Successive encoding levels provide for increasing enhancement in the amount of encoded information and increasing precision in the identification of text elements. Automatic methods to achieve markup at each level are for the most part increasingly complex, and therefore more costly; the sequence is designed to accommodate a series of increasingly information-rich instantiations of the text at a minimum of cost. Section 3 outlines the precise requirements for each level of encoding defined by the CES.

1.2.2 Minimizing processing costs

The CES is also designed based on processing considerations and needs, such as the overhead of use of SGML mechanisms (e.g., entity replacement, use of optional features), as well as more complex textual phenomena such as linkage among elements and related information (for example, annotation, phonetic gloss, etc.), which can have more serious implications for processing (e.g., the use of inter-textual pointers demands that the entire corpus be available at all times for processing). It also considers processing demands of end-use, such as the ability to (efficiently) select texts according to user-specified criteria, etc.

1.2.3 A data architecture for corpus representation

There are additional problems involved in allowing for the simultaneous representation of, and selected access to, multiple *views* of a document, whereby it may be seen as a logical structure, a rhetorical structure, a linguistic object, a document database, etc., all of which are potentially conflicting in terms of well-formed, hierarchical markup. The CES addresses this problem as well as several others by defining a data architecture for corpora in which annotation information is not merged with the original, but rather retained in separate SGML documents (with different DTDs) and linked to the original or other annotation documents. This is opposed to the classical view of a document prepared for use in corpus-based research, in which annotation is added incrementally to the original as it is generated. The separation of original data and annotation is consistent with other recently developed data architecture models, such as the TIPSTER model.

The separate markup strategy is in essence a finely linked hypertext format where the links signify a semantic role rather than navigational options. That is, the links signify the locations where markup contained in a given annotation document would appear in the document to which it is linked. As such the annotation information comprises *remote markup* which is virtually added to the document to which it is linked. In principle, the two documents could be merged to form a single document containing all the markup in each. This approach has several advantages for corpus-based research:

- the base document may be read-only and/or very large, so copying it to introduce markup may be unacceptable;
- the markup may include multiple overlapping hierarchies;6
- it may be desirable to associate alternative annotations (e.g., part-of-speech annotation using several different schemes, or representing different phases of analysis) with the base document;
- it avoids the creation of potentially unwieldy documents;
- distribution of the base document may be controlled, but the markup is freely available.

2 CES Overview

The development of the CES involves the following steps: (1) analysis of the needs of corpus-based NLP research, both in terms of the kinds and degree of annotation required and the requirements for efficient processing, accessibility, etc.; and (2) analysis of general properties and configuration of corpora, the relevant structural and logical features of component text types, and the design of encoding mechanisms that can represent all required elements and features while accommodating the requirements determined in (1).

The CES applies to monolingual corpora including texts from a variety of western and eastern European languages, as well as multi-lingual corpora and parallel corpora comprising texts in any of these languages. The term "corpus" here refers to any collection of linguistic data, whether or not it is selected or structured according to some design criteria. According to this definition, a corpus can potentially contain any text type, including not only prose, newspapers, as well as poetry, drama, etc., but also word lists, dictionaries, etc. The CES is also intended to cover transcribed spoken data.

The CES distinguishes primary data, which is "unannotated" data in electronic form, most often originally created for non-linguistic purposes such as publishing, broadcasting, etc.; and linguistic annotation, which comprises information generated and added to the primary data as a result of some linguistic analysis. The CES covers the encoding of objects in the primary data that are seen to be relevant to corpus-based work in language engineering research and applications, including:

(1) Document-wide markup:
  - bibliographic description of the document, encoding description, etc.

(2) Gross structural markup:
  - structural units of text, such as volume, chapter, etc., down to the level of paragraph; also footnotes, titles, headings, tables, figures, etc.
  - normalization to recommended character sets and entities

(3) Markup for sub-paragraph structures:
  - sentences, quotations
  - words
  - abbreviations, names, dates, terms, cited words, etc.

In addition, the CES covers encoding conventions for linguistic annotation of text and speech, currently including morpho-syntactic tagging and parallel text alignment. We intend to extend the CES in the near future to cover speech annotation, including prosody, phonetic transcription, alignment of levels of speech analysis, etc.; discourse elements; terminology; and lexicon encoding.

Markup types (2) and (3) above include text elements down to the level of paragraph, which is the smallest unit that can be identified language-independently, as well as sub-paragraph

6 For example, lines and sentences in poerty, transcriptions of multi-party dialogues, multi-media corpora, etc.

structures which are usually signaled (sometimes ambiguously) by typography in the text and which are language-dependent. Document-wide markup and markup for linguistic annotation provide "extra-textual" information: the former provides information about the provenance, form, content and encoding of the text, and the latter enriches the text with the results of some linguistic analysis. As such, both add information about the text rather than identify constituent elements.

The CES is intended to cover those areas of corpus encoding on which there exists consensus among the language engineering community, or on which consensus can be easily achieved. Areas where no consensus can be reached (for example, sense tagging) are not treated at this time.

3 Levels of Conformance

The CES provides a TEI-conformant Document Type Definition (DTD) for three levels of encoding for primary data together with its documentation (the "cesDoc DTD"):

Level 1 : the minimum encoding level required for CES conformance, requiring markup for gross document structure (major text divisions), down to the level of the paragraph. Specifically, the following must be fulfilled:

- The document validates against the cesDoc DTD, using an SGML parser such as sgmls.
- The header provides a full description of all encoding formats utilized in the document.
- The document does not contain foreign markup.
- CES-conformant encoding to the paragraph level is included. However, note that for Level 1 CES conformance, paragraph-level markup need not be refined. For example, via automatic means all carriage returns may be changed to `<p>` (paragraph) tags; identification of instances where the carriage return signals a list, a long quote, etc. is not required.

It is also recommended that there should be no information loss for sub-paragraph elements. Sub-paragraph elements identified in the original by special typography not directly representable in the SGML encoded version (e.g., distinction by font such as italics, vs. distinction by capital letters or quote marks, which is directly representable in the encoded version) should be marked, typically using a `<hi>` ("highlighted") tag.

Level 2 : requires that paragraph level elements are correctly marked, and (where possible) the function of rendition information at the sub-paragraph level is determined and elements marked accordingly. Specific requirements are:

- The requirements for a Level 1 document are satisfied.
- If a sub-paragraph element is marked, every occurrence of that element has been identified and marked in the text.
- SGML entities replace all special characters (e.g., `—`, `£`, etc.).
- Quotation marks are removed and either replaced by appropriate standard SGML entities, or represented in a *rend* attribute on a `<q>` or `<quote>` tag.
- The document validates against the cesDoc DTD, using an SGML parser such as sgmls.

It is further recommended that all paragraph level elements (lists, quotes, etc.) are correctly identified, and, where possible, `<hi>` tags are resolved to more precise tags (foreign, term, etc.)

Level 3 : the most restrictive and refined level of markup for primary data. It places additional constraints on the encoding of s-units and quoted dialogue, and demands more sub-paragraph level tagging. Conformance to this level demands:

- Requirements for a Level 2 document are satisfied.
- All paragraph level elements (lists, quotes, etc.) are correctly identified
- Where possible, `<hi>` tags are resolved to more precise tags (foreign, term, etc.)
- The following sub-paragraph elements have been identified and marked (either with explicit tags such as `<abbr>`, `<num>`, etc. or with user-defined morpho-syntactic tags.
  - abbreviations
  - numbers
  - names
  - foreign words and phrases
- Where s-units and dialogue are tagged, the `<p> - <s> - <q>` hierarchy must be followed.
- The encoding for all elements including and below the level of the paragraph has been validated for a 10 percent sample of the text. Note: this does not include morpho-syntactic tagging, if present.
- The document validates against the cesDoc DTD, using an SGML parser such as sgmls.

4 Data Architecture

The CES adopts a strategy whereby annotation information is not merged with the original, but rather retained in separate SGML documents (with different DTDs) and linked to the original or other annotation documents. Linkage between original and annotation documents is accomplished using the TEI addressing mechanisms for element linkage. The CES linkage specifications are currently being updated to conform to XML (Mater & DeRose, 1998).

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 or another annotation document.

All annotation documents are linked to the SGML original (containing the primary data) or other annotation documents using one-way links. The exception is output of the aligner for parallel texts, which consists of an SGML document containing only two-way links associating locations in two documents in different languages. The two linked documents are two documents containing the relevant structural information, such as sentence or word boundaries. The overall architecture is given in Figure 1.

Figure 1. CES data architecture

5 The CES DTDs

Because the CES is an application of SGML, document structure is defined using a context free grammar in a *document type definition (DTD)*. At present, the CES provides three different TEI customizations, each instantiated using the TEI.2 DTD and the appropriate TEI customization files, for use with different documents. For convenience, a version of each of these three TEI instantiations is provided as a stand-alone DTD, together with a means to browse the element tree as a hypertext document.

5.1 The cesDoc DTD

The cesDoc DTD is used to encode primary documents, including texts with gross structural markup only to texts heavily and consistently marked for elements of relevance for corpus-based work. It defines the required structure for marking Level 1 conformant documents down to the paragraph level. It also defines additional elements at the sub-paragraph level which may appear, but are not required, in a Level 1 encoding, and which are used in Level 2 and Level 3 encodings.

There are five main categories of sub-paragraph elements:

〈 linguistic elements;

〈 elements indicating editorial changes to the original text;

〈 the `<hi>` element for marking typographically distinct words or phrases, especially when the purpose of the highlighting is not yet determined;

〈 elements for identifying s-units (typically orthographic sentences) and quoted dialogue;

〈 elements for pointing and reference.

There have been two main defining forces behind the choice of linguistic elements:

(1) the needs of corpus-annotation tools, such as morpho-syntactic taggers, whose performance can often be improved by pre-identification of elements such as names, addresses, title, dates, measures, foreign words and phrases, etc.

(2) the need to identify objects which have intrinsic linguistic interest, or are often useful for the purposes of translation, text alignment, etc., such as abbreviations, names, terms, linguistically distinct words and phrases, etc.

The CES documentation provides an informal semantics for tags used in the cesDoc DTD, especially sub-paragraph linguistic elements. For example, the CES provides precise description of the textual phenomena that should be marked with `<name>` tags (e.g., do not tag laws named after people, etc.). The documentation also includes specifications for the format of such encoding. For example, titles and roles (e.g., "President" in "President Clinton) should not be included inside the `<name>` tag, punctuation not a part of the name is not enclosed in the `<name>` tag (e.g., `"President <name type=person> Clinton</name>,"`), etc. In addition, precise rules for handling punctuation in abbreviations, sentences, quotations, as well as apostrophes, etc., are provided, as well as a hierarchical referencing system used to generate distinct identifiers (SGML *id's*) for structural elements such as chapters, paragraphs, sentences, and words.

In general, the rules for encoding sub-paragraph elements are driven by two considerations:

(1) *Retrieval:* it is essential that items marked with like tags in a document represent the same kind of object. Therefore, while "Clinton" in a phrase such as "President Clinton today said…" is marked as a name, it is not marked as a name in the phrase "the Clinton doctrine".

(2) *Processing needs:* There is a small class of tags which mark the presence of tokens that have been isolated and classified by the encoder, e.g., abbreviations, names, dates, numbers, terms, etc. For many language processing tools, when such an element is identified in the input stream, it is not desirable to further tokenize the string inside the tag; rather, the string inside the tag can be regarded as a single token (possibly with the type indicated by the tag name). For example, in some languages it may be possible for lexical lookup routines and morpho-syntactic taggers to assume that an element with the tag `<name>` is a single token with the grammatical category PROPER NOUN. Therefore, adjectival forms in English (e.g., "Estonian") are not marked as names; generally, for any language, only nouns or noun phrases are marked as names. Similarly, for language processing purposes "Big Brother" can be regarded as a single token instead of two distinct tokens; if marked with a `<name>` tag, processing software may opt to avoid further tokenization of the marked entity. Based on this possibility, punctuation that is not a part of the token is not included inside the tag; in English, possessives are marked by placing the "'s" outside the tag, etc.

The CES recommends that linguistic annotation be encoded in a separate SGML document with its own DTD, which is linked to the primary data. However, for some applications it is still desirable to retain morpho-syntactic annotation in the same SGML document as the primary data. Therefore, the CES provides means to accomplish this in-file tagging. To implement it, a pre-defined module containing all the required definitions for the morpho-syntactic information is brought in at the beginning of the document.

5.2 The cesAna DTD

The cesAna DTD is used for segmentation and grammatical annotation, including:

〈 sentence boundary markup
〈 tokens, each of which consists of the following:
〈 the orthographic form of the token as it appears in the corpus
〈 grammatical annotation, comprising one or more sets of the following:
〈 the base form (lemma)
〈 a morpho-syntactic specification
〈 a corpus tag

Allowing more than one possible set of grammatical annotation enables representing data for which lexical lookup or some other morpho-syntactic analysis has been performed, but which has not been disambiguated. When disambiguation has been accomplished, an optional element can be included containing the disambiguated form.

The structure of the DTD constituents is based on the overall principle that one or more "chunks" of a text may be included in the annotation document. These chunks may correspond to parts of the document extracted at different times for annotation, or simply to some subset of the text that has been extracted for analysis. For example, it is likely that within any text, only the paragraph content will undergo morpho-syntactic analysis, and titles, footnotes, captions, long quotations, etc. will be omitted or analyzed separately.

The following example, which shows the annotation for the first word ("le" in French) of a primary data document stored in a file called "MyText1", shows the use of many of the options provided in the cesAna DTD. This set of annotation data could be the final result after tokenization, segmentation, lexical lookup or morpho-syntactic analysis, and part of speech disambiguation. All the original options for morpho-syntactic class are retained here, and the disambiguated tag is provided in the `<disamb>` element.

```
<!doctype cesAna
          PUBLIC "-//CES//DTD cesAna//EN">
<cesAna version="1.5"
         type="SENT TOK LEX DISAMB"
         doc=MyText1>
      <cesHeader version="2.3">
          ...
      </cesHeader>
        <chunkList>
           <chunk doc="MyText1" from='1.2\1'>
             <s >
                <tok class='tok' from='1.2\1'>
```

```
        <orth>Les</orth>
        <disamb>
            <ctag>DMP</ctag>
        </disamb>
        <lex>
            <base>le</base>
            <msd>Da-fp--d</msd>
            <ctag>DFP</ctag>
        </lex>
        <lex>
            <base>le</base>
            <msd>Da-mp--d</msd>
            <ctag>DMP</ctag>
        </lex>
        <lex>
            <base>le</base>
            <msd>Pp3fpj-</msd>
            <ctag>PPJ</ctag>
        </lex>
        <lex>
            <base>le</base>
            <msd>Pp3mpj-</msd>
            <ctag>PPJ</ctag>
        </lex>
    </tok>. . .
```

5.3 The cesAlign DTD

The cesAlign DTD defines the annotation document containing alignment information for parallel texts. It consists entirely of links between the documents that have been aligned.

Alignment may be between primary data documents or between annotation documents containing segmentation information for the aligned units (paragraphs, sentences, tokens etc.). Alignment may be between two or more such documents, which are identified in the header of the alignment document.

The most common situation in aligning parallel translations is to align data that comprises the content of an entire SGML element, such as an `<s>`, `<par>`, or `<tok>` element. Especially when the aligned data is not in the SGML original document, it is likely that the elements to be associated will have id attributes by which they can be referenced in the alignment document, in order to specify the elements to be aligned or "linked".

Note that when the SGML ID and IDref mechanism is used to point from one element to another in the same SGML document, the SGML parser will validate the references to ensure that every IDREF points to a valid ID. In the CES, all alignment documents are separate from the documents that are being aligned, and therefore this validation of IDrefs by the SGML parser is lost. However, other software may be used to validate cross-document references, if necessary.

The CES provides a simple means to point to SGML elements in other SGML documents by referring to IDs or any other unique identifying attribute on those elements, using the xtargets attribute on the `<link>` element. Here is a simple example:

```
DOC1:      <s id=p1s1>According to our
     survey, 1988 sales of mineral water
     and soft drinks were much higher than
     in 1987, reflecting the growing
     popularity of these products.</s>
     <s id=p1s2>Cola drink manufacturers in
     particular achieved above-average
     growth rates.</s>
<!-- ... -->

DOC2:      <s id=p1s1>Quant aux eaux
     minérales et aux limonades, elles
     rencontrent toujours plus
     d'adeptes.</s>
  <s id=p1s2>En effet, notre sondage fait
     ressortir des ventes nettement
     supérieures à celles de 1987, pour les
     boissons à base de cola notamment.</s>

ALIGN DOC:
  <linkGrp targType="s">
    <link xtargets="p1s1 ; p1s1">
    <link xtargets="p1s2 ; p1s2">
</linkGrp>s
```

When the data to be linked does not include IDs on relevant elements (or for some reason it is not desired to use IDrefs for alignment), or when the data to be linked is not the entire content of an SGML element, it is necessary to reference locations in the documents using the CES notation, which consists of a combination of ESIS tree location and character offset.

Conclusion

By far the greatest need for the development of linguistic corpora is to ensure their usability and reusability in integrated platforms. This demands (at least):

- the development and use of consistent and coherent encoding formats for data representation, as well as standardized schemes for annotation of linguistic information;
- the development of reusable, integrated systems and tool architectures for language processing and analysis, including the corresponding development of a data architecture to best suit research needs.

It is imperative that these activities be undertaken in collaboration. For example, an encoding format that maximizes processability and

retrievability must be devised in view of the capabilities and architecture of the tools that will handle them; similarly, reusable tool design must be informed by full knowledge of the nature and representation of linguistic information, desired processes, etc.
The development of the CES is an attempt to achieve this kind of integration between the development of encoding schemes and corpus processing and use. Very little study has been made to date of the relation between encoding conventions and the demands of processing and retrieval, despite the fact that with the development of digital libraries and web-based document delivery, consideration of these relationships is critical. The CES is in some sense an experiment to develop a principled basis for further work on this topic; it is in no way intended to be the complete and final answer to the problem. Rather, the CES is being developed from the bottom-up, by starting with a relatively minimal set of encoding conventions and successively incorporating feedback to enlarge the standard as needed by the language processing community, and as processing and retrieval needs become better understood. Testing of the current CES specifications and feedback are both invited and encouraged, as well input and suggestions concerning the treatment of other areas of corpus encoding.

Acknowledgements

The present research has been partially funded by US NSF RUI grant IRI-9413451, and EU funding through the EAGLES and MULTEXT projects. The author would like to acknowledge the contribution of Greg Priest-Dorman to the preparation of the web version of the documentation and DTDs. I would also like to acknowledge the input of the many people who have so far contributed to the CES, and in particular the partners of the Multext-East project.

References

Hovy, E., and Ide, N. (eds.) (1998). *Proceedings of the Workshop on Translingual Information Management: Current Levels and Future Abilities*. Held in conjunction with the *First International Language Resources and Evaluation Conference (LREC),* May 31–June 1, 1998, Granada, Spain.

Ide, N., and Véronis, J. (1993). Background and context for the development of a Corpus Encoding Standard, EAGLES Working Paper, 30p. Available at <http://www.cs.vassar.edu/CES/CES3.ps.gz>.

Ide, N. (1998). *Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora.* Proceedings of the First International Language Resources and Evaluation Conference (LREC), Granada, Spain, pp. 463-470.

Ide, N., Priest-Dorman, G., and Véronis, J. (1996). *Corpus Encoding Standard.* Available at <http://www.cs.vassar.edu/CES/>.

International Organization For Standards (1986) ISO 8879: *Information Processing--Text and Office Systems--Standard Generalized Markup Language (SGML),* ISO, Geneva.

Maler, E. & DeRose S. (1998). XML Pointer Language (Xpointer), WWW Consortium Working Draft, Working Draft, 3 March 1998, http://www.w3c.org/TR/WD-xptr.

Sperberg-McQueen, C.M., Burnard, L. (eds.) (1994) *Guidelines for Electronic Text Encoding and Interchange*, Text Encoding Initiative, Chicago and Oxford.