16

Towards Cross-Linguistic Standards or Guidelines for the Annotation of Corpora

PETER KAHREL, RUTHANNA BARNETT
and GEOFFREY LEECH

16.1 Introduction

The production of an annotated corpus is without doubt an expensive task, in terms of both time and effort, and therefore the reusability and shareability of such a resource is of great importance. Standardization of annotation practices can ensure that an annotated corpus can be used to its greatest potential. For an annotated corpus, standardization can be seen as important on two levels:

1. Standard encoding of corpora and annotations
2. Standard annotation of corpora

This first level has been addressed on a world-wide basis by the Text Encoding Initiative (TEI), using recommendations for the use of SGML for mark-up in corpora (see Sperberg-McQueen and Burnard 1994: Chapter 26).[1] The second level, which is being addressed in Europe by the EAGLES initiative (Expert Advisory Group for Language Engineering Standards), is the one with which we will be concerned in this chapter. As EAGLES represents the world's first major attempt at cross-linguistic annotation ground rules, in this chapter we will look in some detail at this initiative, and discuss the various problems encountered, and the possible solutions proposed.

We will first explain in more detail why standards are necessary (Section 16.2). In Section 16.3 we discuss a number of problems in connection with formulating standards and how these problems can be dealt with. Section 16.4 illustrates the EAGLES proposals that have been made recently for adopting standards. An important but often overlooked aspect of any annotation scheme is its documentation; this is taken up in the final Section 16.5.

The issues discussed in this chapter may apply to any level or type of annotation (syntactic, semantic, morphosyntactic, phonological, pragmatic, etc.). However, the most widely applied annotations to date are morphosyntactic (tagging individual words; see Chapter 2) and, slightly less commonly, syntactic annotation (marking constituent structure and syntactic relations such as Subject and Object; see Chapter 3). Since these types of annotation produce similar and interrelated problems, we will concentrate mainly on these two levels in order to illustrate the issues involved in any attempt at standardization.

16.2 Why are Standards Considered Necessary?

There are a number of reasons why standards are helpful and, in many cases, necessary (see also Section 1.3):

1. Although a great deal of annotation work, both morphosyntactic and syntactic, has been done on English, many projects have been undertaking (or at least starting) work on other languages. Standardization of annotation practices will ensure to a degree that text corpora annotated by different groups in different countries are comparable, which is vital for research done, for example, on aligning parallel corpora of different languages (see Section 15.1).
2. Annotation work on only one language has been carried out by various research teams in various countries. Without standards, none of this work is easily comparable, and much unnecessary extra work would be required to allow further research on the same corpus.
3. Annotating a large corpus can be an extremely expensive activity, and in the current situation tools developed by one group cannot (or can hardly) be re-used by other groups. Setting annotation standards means that tools developed for the annotation of one corpus have a greater chance of being interchangeable and reusable, thus saving time, effort, and funds.
4. On a completely different level, standardization could also facilitate the exploitation of corpus research. Annotated corpora may be used for a variety of applications: many industries or research groups would benefit from the use of a corpus, but would not want to formulate their own annotation scheme from scratch. A standard scheme can act as an 'off-the-shelf product' which they can (relatively) easily implement.

16.3 Problems with Standardization

There are a number of problems associated with standardization of annotation practices, which are reflected in the use of the term 'guidelines' alongside 'standards' in the title of this chapter.

1. *Relevance of standards to existing and parallel research* Any standards to be produced must take account of work previously carried out. The standards should be sufficiently flexible so that any already existing annotated corpus that has proved useful for its intended purpose will conform to the standards with little effort. The standards should also be compatible with parallel areas of research, e.g. lexicon building. Lexicons are not covered in detail in this book. However, since the structure of sentences is determined to a large extent by the lexical properties of the words that make up the sentences, the codings in the lexicon should be compatible with both the morphosyntactic and the syntactic annotation schemes. After all, it should be possible to annotate a corpus morphosyntactically for a particular task, but it should also be possible to use the tagged corpus later for other tasks such as syntactic annotation and lexicon enrichment.
2. *Acceptability of standards* Standardization of annotation seems to presuppose that there is agreement about the linguistic analysis of the corpus. But this is hardly the case. In the case of morphosyntactic annotation, it is virtually impossible to lay down rules for absolute consistency in the application of tags to text (see Section 17.1). Ideally, an annotation scheme should be so precise that when two annotators apply that scheme to a corpus, both annotations will be the same. But there are many fuzzy boundaries. For example, in English it is not clear whether one should analyse *gold* in *a gold watch* as a noun or an adjective. In syntactic annotation, not only do we have to determine which labels to apply to segments of the text, but the segments themselves have to be chosen from among many possibilities. The way these segments relate to one another also has to be determined. Fortunately, there is considerable consensus about some of the syntactic segments which have to be recognized in syntactic annotation – e.g. noun phrases and prepositional phrases. On the other hand, there is less consensus about how other syntactic segments should be defined, as illustrated by the following anecdote (Sampson 1995: 4). During the annual conference of the Association of Computational Linguistics in 1991, NLP researchers from nine institutions were asked to specify the bracketing of a number of example sentences. One of the examples was:

   He said this constituted a [very serious] misuse [of the [Criminal Court] processes].

   The brackets here represent the only constituents the nine researchers could agree on: viz.: the adjective phrase *very serious*, the prepositional phrase *of the Criminal Court processes* and the nominal constituent *Criminal Court*.

While standards must be explicit and usable, they cannot be too stringent or limiting. As shown above, there may be disagreement about the definitions or applicability of particular kinds of linguistic analysis. At the same time, while much of the work in this area is still at an early stage, a scheme chosen for annotation may have a marked effect on the success of the automation of this annotation. With syntactic annotation this can easily be illustrated – many syntactically annotated corpora are used as a training tool for automatic parsers. Since no completely successful parser has as yet been developed, the imposition of any particular scheme for syntactic annotation could be detrimental to future research.

3. *Task dependence of corpora* Strict standards pose a problem in an unrelated way as well. Annotation schemes may be produced for a wide variety of uses. An annotated corpus may be intended for use purely in linguistic research (e.g. studying variation in language use across genres, or across time; studying the frequency of particular vocabulary, or structures); or for natural language processing (as a testbed for an automatic parser; as example text for example-based translation, etc.). However, since the production of corpora is expensive in terms of time and effort, the most desirable corpus would be one that is suited to both theoretical and applied ends of the research spectrum. This is not so easy in practice – the aims of these two approaches could be very different, and this would be reflected in the corpus produced and the annotation applied to it. If a corpus is to be processed automatically, certain phenomena which may be problematic for automatic annotation may be left out for reasons of practical expediency, although these phenomena may be more interesting from a linguistic point of view. With reference more specifically to syntactic annotation, from an NLP perspective, the most important (and difficult) task may be the simple grouping together of certain parts of sentences into constituents, while from the linguist's perspective, this is the simplest (and mainly intuitive) task.
4. *Relevance of standards to a wide range of languages* Because much of the work on annotated corpora has been carried out on English, projects now underway on other languages may tend to be unduly influenced by that previous work. Cross-linguistic standards should be flexible enough to comprehend a wide range of languages. In the case of EAGLES, for the moment the aim is to move towards standardization in the treatment of European languages (initially those of the European Community), but optimally the emerging standards should be applicable to as many other languages or language families as possible.

For all these reasons standardization in the commonly-understood sense of 'seeking uniformity' is too confining in the current state of research: setting such standards would seriously constrain annotation practices without leaving any room for development. Furthermore, setting rigid standards does not acknowledge that annotated corpora can be used for different uses and, indeed, prohibits task dependent annotation.

So the dilemma the research community is faced with is: standards of annotation would be a good thing, but they are very difficult to implement in practice. Recently, in spite of the problems associated with standardization, EAGLES produced two documents on the standardization of morphosyntactic and syntactic annotation.[2] While recognizing all the pitfalls, these documents take the form of provisional guidelines (essentially, a tentative move towards standards) and sets of recommendations. They leave enough scope for researchers to vary.[3]

Now let us take a closer look at how the EAGLES recommendations have been formulated.

16.4 EAGLES Guidelines for Annotation

EAGLES has so far undertaken to propose sets of provisional guidelines for the morphosyntactic and syntactic annotation of corpora. To counter the danger of overrigidity mentioned above, three levels of constraint on annotation practices have been suggested. These three levels, **obligatory**, **recommended**, and **optional** annotations, are naturally different for morphosyntactic and syntactic annotation, but in both types of annotation the three levels are distinguished. We will discuss and illustrate these three levels in separate sections. The guidelines proposed for morphosyntax are discussed in Section 16.4.1; and for syntactic annotation, in Section 16.4.2.

16.4.1 Morphosyntactic annotation (or part-of-speech tagging)

The three levels of constraint proposed for morphosyntactic annotation are the following (Leech and Wilson 1994: 8):

1. **Obligatory attributes or values.**[4] These are characteristics that have to be included in any POS tagset. They include the major parts of speech, such as Noun and Verb.
2. **Recommended attributes or values.** These are widely recognized grammatical categories which occur in conventional grammatical

descriptions, such as Person, Number, Gender, Case and Tense, as well as major subcatgories such as 'Common' and 'Proper' for Nouns.

3. **Optional extensions to the list of attributes or values.** This category is subdivided into two:
   (a) Generic attributes or values. These are not usually encoded, but may be included by anyone tagging a corpus for specific purposes. For example, it may be desirable for some purposes to mark semantic classes such as temporal nouns, manner adverbs, place names, etc.
   (b) Language-specific attributes or values. These may be important characteristics of particular languages (e.g. honorifics in Japanese and other East Asian languages; cases in Finnish), and indeed might be recommended by someone annotating texts in those languages.

Below we illustrate a number of the levels.

Obligatory attributes/values

Only the major word categories, or parts of speech, are assigned to the obligatory level. These are the following (note that it is the categories, not the labels, that are obligatory):

1.	N	Noun	8.	C	Conjunction
2.	V	Verb	9.	NU	Numeral
3.	AJ	Adjective	10.	I	Interjection
4.	PD	Pronoun/Determiner	11.	U	Unique/unassigned
5.	AT	Article	12.	R	Residual
6.	AV	Adverb	13.	PU	Punctuation
7.	AP	Adposition			

Most of these are familiar, and need no comment. The Adposition category subsumes both prepositions and postpositions. (Prepositions are, of course, dominant in the wider-known European languages; but arguably postpositions can be exemplified in the *'s* genitive morpheme and the temporal particle *ago* in English.) The Unique value (U) is applied to categories with a unique or very small membership, such as the infinitive marker (*to* in English, *zu* in German) and the existential particle (*there [is/are]*) in English, *er* in Dutch.

The residual value (R) is assigned to classes of word token which lie outside the range of 'canonical' grammatical classes, although they do occur quite commonly in many texts. For example, foreign words, or mathematical formulae. It can be argued that these are on the fringes of the grammar or lexicon of a language; nevertheless, they need to be tagged – for automated corpus analysis no part of the text can be ignored.

Punctuation marks (PU) are (perhaps surprisingly) treated as a part of morphosyntactic annotation, as it is very common for punctuation marks to be tagged and to be treated as equivalent to words for the purpose of automatic tagging and corpus parsing.

Recommended attributes/values

Of the recommended attributes/values, we illustrate just one (Nouns):

Nouns
(i) Type 1. Common 2. Proper
(ii) Gender 1. Masculine 2. Feminine 3. Neuter
(iii) Number 1. Singular 2. Plural
(iv) Case 1. Nominative 2. Genitive 3. Dative 4. Accusative 5. Vocative

Here, Type, Gender, Number and Case are attributes; what follows the arabic numerals are values. For specific languages, both the attributes and the values can be easily extended as the need arises. For example, some languages have not only Singular and Plural, but also Dual number. Similarly for Case, many languages have fewer or more than the five cases listed under (iv). The number of attributes can be extended as well, to handle languages with different alignment systems than nominative-accusative, e.g. ergative languages like Basque.

16.4.2 Syntactic annotation ('treebanks')

Guidelines for syntactic annotation need to take account of a number of 'flexibility' issues, as discussed in Section 16.3 (3–4). Among the reasons for flexibility are: (a) annotated corpora can be used for a wide variety of uses (we called this 'task dependence' above); (b) since annotation practices are still developing, it would be inadvisable to impose a straitjacket on such an immature research area. This caveat is even more true of syntactic annotation than of morphosyntactic annotation. It follows that EAGLES should not propose one standard in this area, but, rather, a set of preliminary recommendations. To handle the first problem, the EAGLES documentation specifies a number of different layers of annotation. Roughly in order of increasing complexity or abstraction, these layers are as follows (cf. Table 3.1, p. 49):

(a) Bracketing of segments;
(b) Labelling of segments;
(c) Marking of dependency relations (see Section 3.3.5);
(d) Indicating functional labels, such as Subject and Object;

(e) Marking subclassification of syntactic segments;
(f) Deep or 'logical' information;
(g) Information about the rank of a syntactic unit (e.g. Clause, Phrase, Word);
(h) Special syntactic characteristics of spoken language.

By allowing these different layers of annotation, without making any of them obligatory, the guidelines meet the requirement that a corpus can be annotated appropriately for a specific purpose. We briefly illustrate some of these layers below.

(a) **Bracketing of segments** consists in the delimitation by some annotative device (for our purposes, square brackets) of sentence segments (normally hierarchically organized) which are recognized as having a syntactic integrity (e.g. sentences, clauses, phrases, words). For example:

[[He] [walked [into [the garden]]]]

(b) **Labelling of segments** amounts to specifying the formal category of the non-terminal syntactic units or constituents identified by bracketing, such as Noun Phrase, Verb Phrase, Relative Clause. Thus adding labels to the above string might yield the following labelled analysis:

[S [NP He NP] [VP walked [PP into [NP the garden NP] PP] VP] S]

(c) **Marking subclassification of syntactic segments**. This means assigning attribute values to constituents such as clauses or phrases, e.g. marking a Noun Phrase as singular, or a Verb Phrase as past tense. A feature-based syntax has been modelled by the Text Encoding Initiative (Sperberg-McQueen and Burnard 1994), in which syntactic information may be represented by means of attribute-value pairs. If necessary, this kind of information can be added in a compact way by adding various subscripts to syntactic segments. Figure 16.1 is an example from the SUSANNE Corpus (showing only part of a sentence), with, in the right hand column, a singular (proper) noun phrase (Nns), and past tense verb phrase (Vd).

A01:0010b	AT	The	the	[O[S[Nns:s.
A01:0010c	NP1s	Fulton	Fulton	[Nns.
A01:0010d	NNL1cb	County	county	.Nns]
A01:0010e	JJ	Grand	grand	.
A01:0010f	NN1c	Jury	jury	.Nns:s]
A01:0010g	VVDv	said	say	[Vd.Vd]

Figure 16.1 Marking subclassification in SUSANNE

(f) **Marking logical (or deep structure) relations of various kinds**. This includes a variety of syntactic phenomena, such as co-referentiality (for example in control structures), cross-reference (or substitution), ellipsis, traces, and syntactic discontinuity. Such information is found, for example, in the SUSANNE Corpus and in the second phase of the Penn Treebank (see Sections 3.3.2 and 3.3.4).

(h) **Information about spoken language non-fluency phenomena**. Spoken language corpora show a range of phenomena that do not normally occur in written language corpora, such as blends, false starts, reiterations, and filled pauses. In syntactic annotation, it has to be decided whether to include such phenomena in a parse tree, and if so, how. There is now increasing interest in this layer of annotation. For example, the British National Corpus contains a small syntactically-annotated subcorpus with inclusion of non-fluency phenomena in the skeleton parsing of spoken data (Eyes 1996). Sampson (personal communication) is now beginning a new project extending the SUSANNE Corpus to spoken data, as discussed in Sampson (1995: Ch. 6). There is also a proposal to include an analysis of such phenomena in the parsing of the British component of the International Corpus of English (Aarts 1992, Greenbaum 1992).

A second issue mentioned at the beginning of this section is that it is not advisable to set standards in a research field that is still developing. The EAGLES guidelines cater for this in two unrelated ways. First, no standard is set for the formalization of syntactic structures or relations. It is recognized that there are two main methods of representing syntactic relations in terms of tree-like structures: phrase structure and dependency structure (see Section 3.3.5). However, there are no good reasons for accepting one of these as a standard and rejecting the other. Second, as with the guidelines for morphosyntactic annotation, the information types given in the guidelines for syntactic annotation are specified on three levels of constraint:

1. Obligatory annotations
2. Recommended annotations
3. Optional annotations.

Obligatory

Because of the variable nature of syntactic annotation, and the many combinatorial possibilities, it is suggested that no part of the syntactic annotation be treated as obligatory. The first layer (a) in the 'hierarchy' of annotation, bracketing, could be seen to be a possibly obligatory level, and indeed for a constituent structure analysis, it would be. However, as we have seen, there are dependency-based schemes that do not actually group together the words making up constituents (e.g. ENGCG – see Section 3.3.5), and these must still undoubtedly be regarded as a useful form of syntactic annotation.

Recommended

On the recommended level, certain non-terminal categories are proposed as annotations within a phrase structure model. They comprise the widely recognized major constituents:

Sentence/Clause	Adjective Phrase
Noun Phrase	Adverb Phrase
Verb Phrase	Prepositional Phrase

as well as coordination phenomena. Although these non-terminal categories are widely recognized, it is not easy to agree on precisely how they are instantiated in texts. The documentation accompanying a corpus should therefore give a clear account of how these constituents are defined, with sufficient attention to problem cases.

Optional

On the optional level, such annotation types as the following are suggested, being commonly useful in parsing and in providing syntactic information in the lexicon:

- sentence subcategorization: different sentence types (declarative, imperative, interrogative)
- syntactic clause subcategorization: clauses annotated as to their formal or functional characteristics (nominal, adverbial, relative, etc.)
- syntactic phrase subcategorization: further subcategorization of phrases to include values such as Person, Number, Case, Tense, Voice and Aspect.
- grammatical function: inclusion of syntactic functions such as Subject, Object, Indirect Object.
- semantic phrase subcategorization: specification of semantic functions of constituents, such as Locative, Temporal adverbials.
- deep/logical information: annotation of various phenomena indicated in (f) above.

16.5 Documentation: a standard after all

There is one area which deserves obligatory standards, namely the documentation of the annotation scheme (see Section 1.3). Without adequate documentation provided by its originators, an annotated corpus can be extremely difficult for other users to apply to their own research tasks. Decisions taken in the development of an annotation scheme, as well as in its application, should be well documented in order to ensure that future users will be able to apply the scheme in a manner consistent with that of the originators of the scheme, and which will then be consistent in the new application. It would be unrealistic to expect optimal documentation practices; but at least the documentation should include some reference to each of the following classes of information:

(a) *What level or layers of annotation have been undertaken?* The documentation should include information as to what particular phenomena are marked in the annotation scheme.

(b) *What is the set of annotation devices used (e.g. brackets, labels) and what are the meanings of these devices?* Each symbol should be described, defined and illustrated with one or more examples.

(c) *What are the conventions for the application of the annotation devices to texts?* An annotation scheme[5] (i.e. a tagging scheme or parsing scheme) is more than (a) and (b) above. It includes the set of guidelines or conventions whereby the annotation symbols are to be applied to text sentences, such that (ideally) two different annotators, implementing the scheme manually to the same sentence, would agree on the analysis to be applied (see further Sections 1.3, 2.5, 3.3.4). To increase its coverage, an annotation scheme may include reference to a lexicon, to a grammar or to a reference corpus of annotated sentences.

(d) *What is the measurable quality of the annotation?* Answers to this should include: (i) to what extent the corpus has been manually checked; (ii) accuracy rate; (iii) consistency rate. These different measures of quality of annotation will depend mainly on how the corpus is annotated. An automatic annotation will require figures of accuracy – often given in terms of a percentage success rate, or in terms of recall and/or precision (see Chapter 7, n. 1; also Voutilainen 1995). A recall of less than 100 per cent indicates that some correct readings have been discarded, while a precision of less than 100 per cent indicates that superfluous readings remain in the output in the form of system ambiguities.

(e) *How detailed/shallow is the analysis?* To a certain extent, the specificity of the analysis may be shown by the levels/layers of annotation that have been applied. However, more detailed documentation may be necessary in order to make clear the granularity or level of detail to which an annotation is undertaken – for example some aspects of a deep or logical grammar may be included in an annotation, while others are not marked (e.g. marking of discontinuity, but no marking of 'traces').

(f) *To what extent and in what respects has disambiguation (of machine-generated ambiguities) been carried out?* During the annotation of a corpus, ambiguous structures may be left in the mark-up (see Section 9.3). Resolution of problematic ambiguities should be documented, as should any ambiguities that are left in the corpus.

(g) *To what extent and in what respects is the annotation at any particular level or layer incomplete?* At any particular level of annotation, certain markings may be ignored by the annotation scheme, for ease of automated annotation, or because of the intended purpose of the resource. Information of this sort should also be included in the documentation.

Standardization is difficult and, especially in the case of syntactic annotation, controversial to the extent that it will be impossible to formulate one agreed 'consensus' standard. Therefore EAGLES proposes tentative standards on different levels to accommodate different theoretical approaches to language.[6]

Notes

1. Examples of TEI conformant encoding of corpus annotation are given in Section 2.4 and Chapter 3, n. 8.
2. Morphosyntactic guidelines are presented in Leech and Wilson (1994), and syntactic guidelines in Leech, Barnett and Kahrel (1995).
3. It should be emphasized that the guidelines are preliminary and subject to later modification. A critique of the syntactic annotation guidelines (Leech, Barnett and Kahrel 1995) is provided by Atwell (forthcoming).
4. The use of 'attribute' and 'value' is illustrated as follows: *Feminine* is a value of the attribute *Gender*; *Singular* is a value of the attribute *Number*.
5. Also termed a 'grammatical representation' by Voutilainen (1994).
6. The provisional EAGLES recommendations on morphosyntactic and syntactic annotation of corpora were the result of teamwork. We gratefully acknowledge the contributions of the following committee members: Gerardo Arrarte, Nicoletta Calzolari, Paula Guerreiro, Jean-Marc Langé, Monica Monachini, Simonetta Montemagni, Anne Schiller, Hans van Halteren, and Atro Voutilainen.