INTERNATIONAL CORPUS OF ENGLISH

Markup Manual for Written Texts

Gerald Nelson

2002

Contents

1 INTRODUCTION

This document presents the **complete set** of markup symbols for ICE Written Texts, as used, for example, in the ICE-GB corpus.

> ***Note***
> **The application of *all* the markup symbols can be time-consuming and labour-intensive. Some of the markup symbols are only required for later stages of annotation, including POS-tagging and syntactic parsing. Therefore some ICE teams have chosen to apply a reduced set of markup symbols. Details of essential, recommended, and optional markup can be found in Appendix 3.**

This manual presents and explains the markup symbols to be applied to written texts in the ICE corpus. The term `written' includes both printed and manuscript material. Markup ensures that we will preserve as many of the features of the original text as possible, such as boldface, italics, etc, which are lost when the text is computerised. Some of the markup, such as the normalization symbols, is inserted to facilitate the Nijmegen tagger and parser.

In this document the term **corpus text** refers to a 2,000-word text in the corpus. There are 500 corpus texts in each national or regional corpus. Many of these will consist of two or more shorter texts combined to create one corpus text. These shorter texts are referred to as **subtexts**.

1.1 The Markup Symbols

In its simplest form a markup symbol consists of an opening tag and a closing tag. The opening tag is <XXX> and the closing tag is </XXX>. For example, to mark a series of words in boldface, place the tag <bold> immediately before them and </bold> immediately after them.

The second type of markup symbol is the entity tag, which has the form <XXX> and requires no closing symbol. Use this to mark a *location* in the text; it has no scope beyond this location. For example, if you wish to mark a sentence boundary, insert <sent> at the boundary (2.2). Similarly, insert <space> to mark an orthographic space (4.7) and <l> to mark the boundary between two lines (4.6).

A complete list of markup symbols may be found in Appendix 1.

2 GENERAL MARKUP

This markup indicates features of the entire text or subtext.

2.1 Subtext Markers <I> </I>

Indicate the boundaries of *every* subtext by putting <I> before all text and markup in the subtext and </I> after all text and markup. Use this symbol even if the corpus text consists of a single text. If there is extra-corpus text at the end of the subtext, insert the </I> *after* the extra-corpus text.

2.2 Text Unit Markers <#>

Text unit markers are location codes inserted in the text to facilitate referencing and automatic searching. They will eventually be numbered and will be assigned the filename, corresponding to the text category of the text. They will have the form <ICE-GB:W1A-001#12:3>, for example. This refers to the twelfth text unit, which occurs in the third subtext of W1A-001, in the British (ICE-GB) corpus. The numbering of the text units can be done automatically, and need not be done during markup. Simply insert the symbol <#> at the start of every sentence in the text.

3 TYPOGRAPHIC MARKUP

3.1 Boldface <bold> </bold>

Example: Readers **must** return all books to the library
Markup: Readers <bold>must</bold> return all books to the library

3.2 Italics <it> </it>

Example: You must attend *every* day during term
Markup: You must attend <it>every</it> day during term

3.3 Underlining <ul> </ul>

Example: It was simply unbelievable!
Markup: It was simply <ul>unbelievable</ul>!

3.4 Small capitals <smallcaps> </smallcaps>

Example: THE STYLE of this quotation indicates its antiquity
Markup: T<smallcaps>he style</smallcaps> of this quotation indicates its antiquity

3.5 Subscript <sb> </sb>

Example: This is the composition of H_2O
Markup: This is the composition of H<sb>2</sb>O

3.6 Superscript <sp> </sp>

Example: For further discussion, see Levin^6
Markup: For further discussion, see Levin<sp>6</sp>

3.7 Roman type <roman> </roman>

Roman is complementary to italics and hence is optional in all cases. If a text is predominantly in italics mark words in roman type.
Example: If you require more information ring 071-123-4567 *between 9am and5pm.*
Markup: If you require more information ring <roman>071-123-4567</roman> between 9am and 5pm.

3.8 Typeface <typeface> </typeface>

Mark a *significant* change in typeface. Give each typeface an identifying number or name and add this to the symbol.

Example: Warhol is `alive` and well
Markup: Warhol is <typeface:courier>alive</typeface:courier> and well

3.9 Unusable characters &XXX;

Replace any character that is not included in the following list by the corresponding SGML character. Appendix 2 contains a standard list of SGML characters. Note that the mathematical symbol `<' (`less than') is reserved for markup and must be treated as unusable. The *usable* characters are:

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z a b c
d e f g h i j k l m n o p q r s t u v w x y z
0 1 2 3 4 5 6 7 8 9 / - , ; : . ! ? ' " ( ) [ ] { }
% & + = @ * ` ^ _ ~ | \

Where the unusable character forms part of a word no spaces are allowed either surrounding or within the markup. Here is an example of the correct markup:

coup de grâce ? coup de gr&acircumflex;ce.

Note:

- Fractions such as ½ must be changed into the form 1/2, so 3½ will be 3 1/2, *not* 31/2.

Appendix 2 contains a list of standard SGML characters.

4. CONTENT MARKUP

4.1 Headings <h> </h>

Mark the start and end of each heading. If there is a subheading, mark it as a *separate* heading and as a *separate* text unit. There is no markup specifically for subheadings. Do not include by-lines in heading markup, but mark them as separate text units. Newspaper headlines should be treated in the same way as headings in other printed material. The following example shows the markup for a newspaper headline, subheading and by-line:

Government in legal battle
Cabinet to meet today
by John Smith

Markup:
<#><h>Government in legal battle</h>
<#><h>Cabinet to meet today</h>
<#>by John Smith

Note that headings and subheadings are separate text units.

4.2 Paragraphs <p> </p>

Example:

> These days we hear and read a great deal about the threat to the climate system posed by the increasing population, atmospheric pollution, and deforestation. A key realisation is that the impacts are not localised, but affect the whole planet. In spite of much scientific activity, predictions about what will happen in the future remain uncertain.

Markup:

> <p>These days we hear and read a great deal about the threat to the climate system posed by the increasing population, atmospheric pollution, and deforestation. A key realisation is that the impacts are not localised, but affect the whole planet. In spite of much scientific activity, predictions about what will happen in the future remain uncertain.</p>

4.3 Footnotes <footnote> </footnote> and References to footnotes <fnr> </fnr>

Mark a footnote with <footnote> and </footnote> and insert it *immediately after the sentence which refers to it.* The footnote reference, usually in the form of a number, is marked <fnr> </fnr>. Footnote references are often printed in superscript (See 3.6).

Example:

> Mascot 2 has been in use since 1976 and a number of toolsets have been developed to support it. In 1987 the Mascot 3 handbook was published (1), and toolsets are now being developed to support Mascot 3 with Ada. The transputer has many benefits for embedded systems.
>
> (1) The Official Handbook of Mascot, RSRE, June 1987

Markup:

> Mascot 2 has been in use since 1976 and a number of toolsets have been developed to support it. In 1987 the Mascot 3 handbook was published <fnr>(1)</fnr>, and toolsets are now being developed to support Mascot 3 with Ada. <footnote>(1)

The Official Handbook of Mascot, RSRE, June 1987</footnote>
The transputer has many benefits for embedded systems.

4.4 Marginalia <marginalia> </marginalia>

Marginalia are rarely found in modern printed material, though you may find them in handwritten letters or examination scripts. Remember that they must have been written by the author of the main text, and not by an editor or examiner, and that they must refer to the body of the text. Captions should not be treated as marginalia, since they refer to a drawing, chart or photograph, and not to the text itself (See 5.6). Insert marginalia immediately *after* the sentence(s) to which they refer.

Example:

Work is going really well. I really
enjoy it, though there's still so
much that I don't know how to do
yet. The people are really friendly,
especially Jeanne, the lady that I
work with. Surprisingly, I managed
to find my way to work with no
problems even though it involved a
change of buses etc.

Or is it Jean?

Markup:

Work is going really well. I really
enjoy it, though there's still so
much that I don't know how to do
yet. The people are really friendly,
especially Jeanne, the lady that I
work with. <marginalia>Or is it Jean?
</marginalia> Surprisingly, I managed
to find my way to work with no
problems even though it involved a
change of buses etc.

4.5 Deleted text <del> </del>

Deleted text refers to deletions made by the author, and will be found only in unpublished material, such as letters, essays and examination scripts. Only use this markup when you can actually decipher what the deleted word is. For example, if the word `was' is deleted in `He was is a good man', mark this: `He <del>was</del> is a good man'. If the deletion renders the word illegible, use unclear word markup (See 5.5). Mark one deleted and illegible word as follows: <del><unclear>one word</unclear></del>

This markup should not be confused with normative deletions, which are made by the person applying markup (See 6: Normalizing the text).

4.6 Line-breaks and hyphens <l>

End-of-line hyphens require special attention as it is sometimes unclear whether the author intended a hyphenated word or the word was simply split because of the end of the line. Three types of hyphenation should be noted during markup:

- Words which are *definitely hyphenated* occurring at a line-break. In this case, do not use any markup. Simply move the entire word (including the hyphen) onto the same line.

Example: Take a left-
hand turn here.

Markup: Take a
left-hand turn here.

- Words which are *definitely unhyphenated*. If a definitely unhyphenated word is split at a line-break, join the parts of the word and insert a line-break marker <l> to indicate where it was split in the original.

Example: Britain entered the Exchange Rate Mech-
anism after several months of debate.

Markup: Britain entered the Exchange Rate Mech<l>anism after several months of debate.

- Ambiguous hyphens. If a word can be used with or without a hyphen, and it is split at a line-break, keep the hyphen and mark it with <?> and </?> symbols to indicate that it is ambiguous.

Example: The Chancellor visited centres in the north-
east during a two-day tour

Markup: The Chancellor visited centres in the north<?>-</?>east during a two-day tour

To determine if a word requires a hyphen or is definitely unhyphenated, refer to instances of the word in the same subtext or to a reputable dictionary.

4.7 Orthographic spaces <space>

Mark all instances of unusual spacing in the text, for example a space before a punctuation mark.

Example: I never , ever did that
Markup: I never<space>, ever did that

4.8 Orthographic words <w> </w>

All words containing internal apostrophes, such as *you'll, I'd, d'you, we're, John's,* etc, will automatically be given orthographic word markup prior to tagging, so this need not be applied by the person marking up the text. However if the apostrophe is *not* internal, ie, if it is either preceded or followed by a space, then the word must be marked to distinguish the apostrophe from a closing quotation mark.

Example: The boys' books.
Markup: The <w>boys'</w> books.

Similarly a word with an initial apostrophe must be marked as an orthographic word.

Example: In the late '60s....
Markup: In the late <w>'60s</w>....

4.9 Mentions <mention> </mention>
Mark words which are cited *as* words.

Example: Dog is a noun
Markup: <mention>Dog</mention> is a noun

4.10 Changed names and words <@> </@>
Mark names and words that have been changed to preserve anonymity. You will probably use this very rarely, but you may need it if the author of a social or business letter, for example, demands it. If you do use it, be sure to preserve the case (upper/lower) of the original, together with its word-class. Make sure your replacements are consistent within a corpus text or subtext.

Example: Dr Brown of London University
Markup: Dr <@>Payne</@> of <@>Camden</@> University

4.11 Quotations <quote> </quote>
Very brief quotations which the author of the text has incorporated into the structure of his own sentence should be marked as quotes, as in the following example:

Example:

> The starting point for such a method was to be "simple sensuous perception", from which the mind would be led step-by-step, "as if by machinery", to inductive proof.

Markup:

> The starting point for such a method was to be <quote>"simple sensuous perception"</quote>, from which the mind would be led step-by-step, <quote>"as if by machinery"</quote>, to inductive proof.

If quotations comprise one or more complete sentences, mark them *additionally* as extra-corpus text. This always applies to indented quotations.

Example:

> In his 1987 Mansion House address he spoke as follows:

> I would like to see the medieval street plan of pre-war Paternoster reconstructed, not out of mere nostalgia but to give meaning to surviving fragments like Amen Court and the Chapter House.

Markup:

In his 1987 Mansion House address he spoke as follows:

> <X><quote>I would like to see the medieval street plan of pre-war Paternoster reconstructed, not out of mere nostalgia but to give meaning to surviving fragments like Amen Court and the Chapter House.</quote></X>

4.12 Foreign words <foreign> </foreign>
Mark a word or sequence of words that is foreign and non-naturalised. If the word contains unusable characters, use unusable character markup (see 3.9) in addition to foreign word markup.

4.13 Indigenous words <indig> </indig>
Mark words which are non-English but are indigenous to the country in which the corpus is being compiled as indigenous rather than foreign words. For example, in the Indian corpus, *Lok Sabha* (Parliament) should be marked as indigenous, though it should be marked as foreign in every other ICE corpus.

4.14 Captions and graphics
All graphics in a text must be marked as untranscribed text using <O> and </O/>. Graphics include drawings, graphs, photographs, diagrams, tables and charts. Also mark the captions which accompany these as untranscribed text. Captions must not be marked as marginalia (4.4). Mark graphics as <O>chart</O>, <O>diagram</O>, <O>photograph</O>, etc.

4.15 Uncertain transcription <?> </?>
In handwritten texts, you will frequently have to guess what a particular word is. Mark your guess as an uncertain transcription, as in the following:

Example: the ? car
Markup: the <?>red</?> car

You should also use uncertain transcription to mark ambiguous hyphens (See 4.6).

Note: Only use uncertain transcription when you can make a reasonable guess about the original text. If you cannot make any guess, use unclear words markup (4.16).

4.16 Unclear words <unclear> </unclear>

Mark words which are illegible in hand-written texts with <unclear> and </unclear>.

Example: This is ? more difficult than before
Markup: This is <unclear>one word</unclear> more difficult than before

5 NON-CORPUS MATERIAL

In this section we are concerned with material which is excluded from the corpus but is included with the text to provide context or clarification. The corpus text itself may also contain some extra-corpus material, such as quotations and foreign words.

5.1 Extra-corpus text <X> </X>

If a complete chapter has been scanned onto the computer, the selected 2,000-word text will be the corpus text, and the remainder will be extra-corpus. Mark this material with <X> at the start and </X> at the end. Remember to put the closing subtext marker after this material (see 2.1). Extra-corpus text need not be marked up. Give priority to material which is actually in the corpus, though you should eventually mark up all text.

5.2 Untranscribed text <O> </O>

Use this markup to mark items which exist in the original text but have not been reproduced in the computerized version. This applies to tables, mathematical formulae and graphics (See 4.14). It must not be used to mark unclear words (4.16).

5.3 Editorial comments <&> </&>

This is used to mark essential comments made by the person applying the markup. Use it sparingly, and only when no other markup symbol is applicable. In particular, check whether extra-corpus text, untranscribed text or uncertain transcription are more appropriate. An editorial comment can be used, for example, to explain an unusual abbreviation:

Example: He went out and bought a vid.
Markup: He went out and bought a vid.<&>vid=video</&>

6. NORMALIZING THE TEXT

In this section we are concerned with normalization of the original text using deletions and insertions. Normalization is carried out when a word or sequence of words deviates from the norm of written English, and it is intended to facilitate the parser. As a general rule, normalize as little as possible, and only when it is absolutely essential to do so. Normalization is chiefly used to mark misspellings, obvious grammatical errors, repetitions and incomplete words.

There are three types of normalization:

Normative deletion <-> and </->
Normative insertion <+> and </+>
Original normalization <=> and </=>

Both normative deletion and normative insertion can be used on their own. A misplaced punctuation mark, for example, can simply be marked for deletion.

Example: I didn't; know that!
Markup: I didn't<->;</-> know that!

Similarly a word or sequence of words can be inserted if it is essential to do so for the parser.

Example: It rains here all time
Markup: It rains here all <+>the</+> time

When making insertions of this kind, make sure that they are *absolutely essential*.

It is most common to use the normalization symbols in combination with each other, in which case the entire combination must be enclosed within the opening symbol <}> and the closing symbol </}>.

6.1 Misspellings

Mark all misspellings in the original text using <-> and </->. Insert the correct form within the symbols <+> and </+>, and enclose the entire sequence within <}> and </}>.

Example: But that's rank hypocrasy!
Markup: But that's rank <}><->hypocrasy</-><+>hypocrisy</+></}>!

An important point to note about deletion is that it does not actually delete or remove the word from the text. Every feature of the original is retained. The misspelling is normalized so that the word will be recognised by the tagger, and to ensure accurate word frequency statistics. The normalization also ensures that we can retrieve all words from the corpus, whether they have been misspelled or not.

6.2 Grammatical errors

Do not correct grammatical errors, or what may be perceived as grammatical or stylistic deviations.

6.3 Repetition

Use this when the repetition of a word or a sequence of words causes deviation from the norm of written English.

Example: I I know that.
Markup: <}><->I<-><=>I</=></}> know that.

In this example the second `I' is marked for retention using <=> and </=> while the first is deleted using <-> and </->.

Not all repetitions require normalization. Some, like *very, very* and *many, many* are used for emphasis and need no markup.

6.4 Incomplete words <.> </.>

Incomplete words can sometimes appear in handwritten material. They are usually deleted by the author and replaced by the complete word.

Example: I thought it was dil delicious
Markup: I thought it was <}><-><.>dil</.></-><=>delicious</=></}>

The incomplete word is first marked with <.> and </.> and then normatively deleted.

6.5 Discontinuous words <(> </(> and <)> </)>

Discontinuous words typically occur in cases of tmesis, when one word is inserted within another, and are quite rare in written texts. To mark them, enclose the original form between the symbols <(> and </(>. Insert the normalized form and enclose it within <)> and </)>.

Example: fan-bloody-tastic
Markup: <(>fan-bloody-tastic</(><)>bloody fantastic</)>

Discontinuous words should not be confused with *incomplete* words (6.4).

APPENDIX 1
MARKUP SYMBOLS FOR WRITTEN TEXTS

<#>	Text unit marker
<I>...</I>	Subtext marker
<l>	Linebreak marker
<p>...</p>	Paragraph marker
<h>...</h>	Heading
<w>...</w>	Orthographic word
<X>...</X>	Extra-corpus text
<?>...</?>	Uncertain transcription
<O>...</O>	Untranscribed text
<.>...</.>	Incomplete word
<->...</->	Normative deletion
<+>...</+>	Normative insertion
<=>...</=>	Original normalization
<}>...</}>	Normative replacement
<&>...</&>	Editorial comment
<(>...</(>	Discontinuous word
<)>...</)>	Normalized disc. word
<@>...</@>	Changed name or word
<sb>...</sb>	Subscript
<sp>...</sp>	Superscript
<ul>...</ul>	Underline
<it>...</it>	Italics
<bold>...</bold>	Boldface
<typeface>...</typeface>	Change of typeface
<roman>...</roman>	Roman type
<smallcaps>...</smallcaps>	Small capitals

Markup symbols for written texts
(continued)

<footnote>...</footnote>	Footnote
<fnr>...</fnr>	Reference to footnote
<space>	Orthographic space
<quote>...</quote>	Quotation
<del>...</del>	Deleted text
<marginalia>...</marginalia>	Marginalia
<mention>...</mention>	Mention
<indig>...</indig>	Indigenous word(s)
<foreign>...</foreign>	Foreign word(s)

APPENDIX 2
SGML Character Descriptions

à	&agrave;
À	&Agrave;
á	&aacute;
Á	&Aacute;
â	&acircumflex;
Â	&Acircumflex;
ä	&auml;
Ä	&Auml;
æ	&aeligature;
Æ	&AEligature;
ç	&ccedille;
Ç	&Ccedille;
è	&egrave;
È	&Egrave;
é	&eacute;
É	&Eacute;
ë	&euml;
Ë	&Euml;
ö	&ouml;
Ö	&Ouml;
œ	&oeligature;
Œ	&OEligature;
ü	&uuml;
Ü	&Uuml;
ñ	&ntilde;
Ñ	&Ntilde;
£	&pound;
$	&dollar;
α	&alpha;
β	&beta;
γ	&gamma;
Γ	&GAMMA;
δ	&delta;
Δ	&DELTA;
θ	&theta;
Θ	&THETA;
µ	&mu;
#	&hash;
□	&square;
o	&circle;
•	&bullet;
...	&dotted-line;
·	&dot;
v	&square-root;
8	&infinity;
°	&degree;
~	&approximate-sign;
§	&sect;
< ("less than")	&lt;
> ("greater than")	&gt;

APPENDIX 3
Essential, Recommended, and Optional Markup in Written Texts

Essential	Recommended	Optional
Text units	Incomplete words	Normalization
Subtexts	Deleted text	Boldface
Extra-corpus	Footnotes	Italics
Editorial comments	Footnote references	Typeface
Untranscribed text	Marginalia	Roman
Unclear words	Mentions	Underline
Unusable characters	Orthographic words	Smallcaps
Uncertain transcription	Changed names	Subscript
	Orthographic space	Superscript
	Foreign words	Line-breaks
	Indigenous words	Discontinuous words
	Quotations	
	Headings	
	Paragraphs	