Markup Manual for Spoken Texts

Gerald Nelson

2002

CONTENTS

1. Introduction

This document presents the **complete set** of markup symbols for ICE Spoken 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 spoken texts in the ICE corpus. The term `spoken' includes both unscripted and scripted ("written to be spoken") material.

The term *corpus text* refers to a 2,000-word text in the corpus. There are 300 spoken texts in each national or regional corpus. Many of these will consist of two or more shorter texts combined to create a single corpus text. These shorter texts are referred to as *subtexts.*

Markup symbols are of two types:

1. Markup which has scope over one or more words consists of an opening symbol and closing symbol. The opening symbol is <*symbol*> and the closing symbol is </*symbol*>. To mark a series of quoted words, for instance, place <quote> immediately before them and </quote> immediately after them: <quote>To be or not to be</quote>.

2. Markup which indicates a position in the transcription has the form <symbol>. This has no scope over any words and includes text unit markers and pauses.

A list of all markup symbols for spoken texts may be found in Appendix 1.

2. General Markup

This markup indicates features of the entire corpus text or subtext.

2.1 Subtext Markers <I> </I>

Since many corpus texts will be composite, we must distinguish clearly between the subtexts which they contain. To do this, place the symbol <I> before all text and markup in the subtext and </I> after all text and markup. The subtext markers *must* be the first and last items in all subtexts. *Use these markers even if the corpus text consists of a single 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:S1A-001#12:3:A>, for example. This refers to the twelfth text unit, which occurs in the third subtext of S1A-001, in the British (ICE-GB) corpus. The "A" indicates that it is uttered by speaker "A" (see 2.3). 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 text unit.

As a general rule, text units should correspond loosely to orthographic sentences, though this will clearly be impossible in some spoken texts. An incomplete response to a question, for example, will be a text unit.

Note: The beginning of a speaker turn must coincide with a new text unit. This applies also to turns by extra-corpus speakers.

2.3 Speaker IDs <$>

Mark the beginning of each speaker turn with a speaker ID symbol. The first speaker will be <$A>, the second <$B>, and so on. Always identify the same speaker using the same letter.
If the speaker's identity is uncertain, use uncertain transcription symbols (3.12) to query your identification, eg. <?><$A></?>. In monologues, refer to the single speaker as <$A>. If a corpus text contains more than one subtext, the sequencing of speaker turns must begin anew for each subtext.
If you transcribe the words of an extra-corpus speaker, identify him or her as <$Z>.

3. Content Markup

3.1 Pauses <,> and <,,>

Mark short and long pauses using <,> and <,,> respectively. A short pause is defined as any perceptible break in phonation equal in length to a single syllable, taking into account the tempo of the speaker's delivery. A long pause is defined as any break in phonation which is longer than a single syllable.

The following points should be noted:

1. If a pause occurs within a word, insert the pause marker *after* the word.

2. When inserting pauses, do not be influenced by the probable punctuation in a written text.

3. If a pause occurs between speaker turns, insert the pause at the end of the first speaker turn.

3.2 Overlapping speech <[> </[> and <{> </{>

The simplest form of overlap involves just two overlapping speech strings. Mark each string individually with <[> and </[>. Then enclose both strings within <{> and </{>. In the following example $A's utterance "Nothing stands out" overlaps completely with $B's "Yeah I suppose".

<$A> <#><{><[>Nothing stands out</[>
<$B> <#><[>Yeah I suppose</[></{>

It is more usual for $B to interrupt $A, causing $A to bring his utterance to an end:

<$A> <#>Is that because you've got lots or <{><[>nothing stands out</[>

<$B> <#><[>Yeah I suppose</[></{> <#>I can't say anything really stands out

If there are two or more overlaps within a speaker turn, mark them as in the example above and number them 1, 2, 3 etc. *At the end of the first speaker's turn*, list the second speaker's overlapping strings and number them in sequence:

<$A> <#>I keep watching the film over and over <{1><[1>again</[1> <#>I just <{2><[2>think</[2> its really good

<$B> <#><[1>Right</[1></{1>

<$B> <#><[2>Uhm</[2></{2>

In this example "again" overlaps with "right" and "think" overlaps with "uhm", as the numbering indicates. No matter how many overlaps occur within a speaker turn, they must be placed *after* the complete turn. A speaker turn must not be broken, since a complete turn is necessary for parsing.

Frequently an overlap will cause one string to become unclear. In this case use unclear words markup (3.11) within overlap markup:

<$A> <#>I'd really have to pick and choose <{><[><unclear>one or two words</unclear></[>

<$B> <#><[>Yes you would</[></{>

In conversation it is very common for one speaker to pause just long enough to allow the other to interject a single word, usually a monosyllable. This should be treated as a string which overlaps with a pause, and not as a change of speaker turn. In the following example $A's short pause after "out" allows $B to say "uhm". $A then continues the turn.

<$A> <#>The maintenance for the children that you've worked out <{><[><,></[> uh is something that you and I have discussed on the phone

<$B> <#><[>Uhm</[></{>

If the overlap involves more than two speakers, treat it in exactly the same way. In the following example $A's "happen very often" overlaps with $B's "I just thought" and with $C's "shouldn't happen". The position of the opening symbol <{> and the closing symbol </{> indicates clearly that all three strings overlap with each other.

<$A> <#>It really shouldn't <{><[>happen very often</[>

<$B> <#><[>I just thought</[> it was worth mentioning

<$C> <#><[>Shouldn't happen</[></{> at all really

3.3 Anthropophonics
Mark utterances such as coughs, sneezes and laughs using untranscribed text markup.
Examples: <O>cough</O> <O>sneeze</O> <O>laugh</O>

There are two standard orthographic representations for the voiced pause. These are *uh* and *uhm*. For background noises see 4.2.

3.4 Abbreviations, numerals and dates
Standard abbreviations such as Mr, Mrs and Dr may be used. Do not use a full stop (period) after these abbreviations. Acronyms must be transcribed as they are spoken, so NATO is simply transcribed as NATO. If a word or an abbreviation is spelled out by the speaker, transcribe the letters with a space between them, eg, 'the U S A'.

Transcribe numerals and dates exactly as they are spoken, eg, 'one point five', *not* 1.5. Similarly, `nineteen sixty-two', *not* 1962.

3.6 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 transcriber.

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.

3.6 Mentions <mention> </mention>
Mark all words which are cited *as* words.
Example: Dog is a noun
Markup: <mention>Dog</mention> is a noun

3.7 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 descriptions (3.10) in addition to foreign word markup.

Example: This is what Derrida called the logic of the pharmakon
Markup: This is what Derrida called the logic of the <foreign>pharmakon</foreign>

3.8 Indigenous words <indig> </indig>
Mark words which are non-English but are indigenous in the country or region being sampled. For example, in the Hong Kong corpus, Cantonese words are marked in this way:

Example: Little bit more for the <indig> gwailo </indig>

The indigenous word may be followed by an explanatory editorial comment (4.3):

Example: Little bit more for the <indig> gwailo </indig> <&> Cantonese=foreigner </&>

3.9 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 speaker a demands it. If you do use it, be sure to preserve the word-class of the original. Make sure your replacements are consistent within a corpus text or subtext.

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

3.10 Unusable characters &XXX;
If a speaker utters a word whose orthographic spelling includes an unusable character, use the appropriate SGML character description. If 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

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
- '

A list of standard SGML character descriptions may be found in Appendix 2.

3.11 Unclear words <unclear> </unclear>
Mark words which cannot be deciphered using unclear word markup. Do not use untranscribed text markup for this. If a guess can be made use uncertain transcription (3.12).

Example: This is my ?
Markup: This is my <unclear>word</unclear>

3.12 Uncertain transcription <?> </?>

If you are in any doubt about the accuracy of your transcription, mark your it uncertain transcription.

Example: This is a typical example of <?>radiation</?> filtering

Use uncertain transcription only when you can make a good guess about the original text. If you cannot make any guess, use unclear text markup (3.11).

4. Non-corpus material

4.1 Extra-corpus text <X> </X> and quotations <quote> </quote>

Transcribe utterances by non-corpus speakers - Americans in the British corpus, for example - if they are necessary for context, and mark them as extra-corpus text. Extended extra-corpus utterances which are not contextually necessary need not be transcribed. Simply mark them as untranscribed text (4.2). For example, <O>speech by George Bush</O>.

Mark very brief quotations *which the speaker has incorporated into the structure of his own utterance* using <quote> and </quote>, as in the following example:

> I asked him if it was better <quote>to be or not to be</quote>

NOTE: It is important to remember that extra-corpus material will be completely ignored by the tagger and parser, whereas material marked as a quotation will not. This means that a single text unit cannot contain both corpus and extra-corpus material. *Either the whole text unit is extra-corpus or none of it is.*

Example:

> And in the words of my predecessor Britain must not relinquish sovereignty at any cost. Nothing could be worse for the nation or for this parliament.

Here the quoted part begins at `Britain must not' and continues to the end of the utterance. It would be incorrect to mark the whole quotation as extra-corpus, since the first part of it has been integrated into the first sentence. From `Britain' to the end of the first sentence is marked as a quotation, and the second sentence is additionally marked as extra corpus.

Markup:

> <#>And in the words of my predecessor <quote>Britain must not relinquish sovereignty at any cost. <X><#> Nothing could be worse for the nation or for this parliament.</quote></X>

4.5 Untranscribed text <O> </O>

Use this markup to describe text which has not been transcribed because it is not contextually relevant, eg. <O> commercial break </O>

4.6 Editorial comments <&> </&>

Mark *essential* comments by the transcriber as editorial comments. Use sparingly, and only when no other markup symbol is applicable. In particular, check whether extra-corpus text, unclear text or uncertain transcription are more appropriate.

Example: <&>break in recording</&>

Note: `Noises off', such as telephones ringing, should be marked <&>telephone rings</&> only if they are relevant to the context. Non-relevant background noises, such as traffic noises, may be ignored.

5. Normalizing the text

In this section we are concerned with normalization of the original text using deletions and insertions. As a general rule, normalize as little as possible, and only when it is absolutely essential to do so for parsing purposes.

There are three types of normalization:

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

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 </}>. Normalization is chiefly used to deal with repetitions, hesitations, self-corrections, and incomplete words.

5.1 Repetitions and hesitations

Hesitation and stammering causes the speaker to repeat words, often several times. To mark this, delete all instances of the word except the final one, which is marked for preservation using original normalization symbols. Then enclose the entire sequence in <}> and </}>.

Example: I I don't want to
Markup: <}><->I</-><=>I</=></}> don't want to

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 repetition is normalized primarily to facilitate the parser, and of course to record the hesitation phenomenon. The deletion symbols have the effect of suppressing the words they enclose while the text is being automatically processed.

Not all repetitions require normalization. They may sometimes be used for emphasis, as in *very very* and *many many.*

5.2 Self-corrections

Mark a self-correction in exactly the same way as a repetition. Mark the `incorrect' version for deletion, and the `correct' one for preservation. Then enclose the entire sequence in <}> and </}>.

Example: I will not discuss this part of the case this part of the summons
Markup: I will not discuss <}><->this part of the case</-><=>this part of the summons</=></}>

The `incorrect' and `correct' versions in a self-correction are not always adjacent to each other. When they are not, they are often separated by a voiced pause.

Example: I won't see him until the end of uh the beginning of August
Markup: I won't see him until <}><->the end of</-> uh <=>the beginning of</=></}> August

5.3 Grammatical errors

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

5.4 Incomplete words <.> </.>

Incomplete words occur when the speaker stammers, tails off or is interrupted, and they are very common in cases of hesitation and self-correction. If the `word' is so incomplete that you cannot guess what complete word was intended, use incomplete word markup:

Example: I have always been interested in ant
Markup: I have always been interested in <.>ant</.>

If you can guess the complete word, then use incomplete word markup in conjunction with deletion and insertion:

Example: I have always been interested in anthropol
Markup: I have always been interested in <}><-><.>anthropol</.></-><+>anthropology</+></}>

Here the complete word is inserted by the transcriber using <+> and </+>. This should only be done if you are *absolutely certain* what complete word was intended by the speaker.

Frequently, a speaker will use an incomplete word followed by its complete version. In this case, use incomplete word markup in conjunction with original normalization:

Example: I have always been interested in anthro anthropology
Markup: I have always been interested in <}><-><.>anthro</.></-><=>anthropology</=></}>

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

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

Example: fanbloodytastic
Markup: <(>fanbloodytastic</(><)>bloody fantastic</)>

APPENDIX 1
MARKUP SYMBOLS FOR SPOKEN TEXTS

<$A>, <$B>, *etc*	Speaker identification
<I>...</I>	Subtext marker
<#>	Text unit marker
<O>...</O>	Untranscribed text
<?>...<?>	Uncertain transcription
<->...</->	Normative deletion
<+>...</+>	Normative insertion
<=>...</=>	Original normalization
<.>...</.>	Incomplete word
<}>...</}>	Normative replacement
<[>...</[>	Overlapping string
<{>...</{>	Overlapping string set
<,>	Short pause
<,,>	Long pause
<(>...</(>	Discontinuous word
<)>...</)>	Normalized disc. word
<X>...</X>	Extra-corpus text
<&>...</&>	Editorial comment
<@>...</@>	Changed name or word
<w>...</w>	Orthographic word
<quote>...</quote>	Quotation
<mention>...</mention>	Mention
<foreign>...</foreign>	Foreign word(s)
<indig>...</indig>	Indigenous word(s)
<unclear>...</unclear>	Unclear word(s)

APPENDIX 2
SGML Character Descriptions

à	à
À	À
á	á
Á	Á
â	&acircumflex;
Â	&Acircumflex;
ä	ä
Ä	Ä
æ	&aeligature;
Æ	&AEligature;
ç	&ccedille;
Ç	&Ccedille;
è	è
È	È
é	é
É	É
ë	ë
Ë	Ë
ö	ö
Ö	Ö
œ	&oeligature;
Œ	&OEligature;
ü	ü
Ü	Ü
ñ	ñ
Ñ	Ñ

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

Essential	Recommended	Optional
Text units	Incomplete words	Normalization
Subtexts	Mentions	Discontinuous words
Extra-corpus	Orthographic words	
Editorial comments	Changed names	
Untranscribed text	Foreign words	
Unclear words	Indigenous words	
Unusable characters	Quotations	
Uncertain transcription	Pauses	
Speaker IDs	Overlapping speech	