English-Norwegian Parallel Corpus: Manual

Stig Johansson
Jarle Ebeling
Signe Oksefjell

Department of British and American Studies
University of Oslo, 1999/2002

Contents

---

1 Introduction

The main purpose of this manual is to describe the structure and explain the coding of the English-Norwegian Parallel Corpus. In addition, the manual briefly introduces the programs developed within the project and provides documentation on the corpus texts.

The project was made possible through the generosity of the authors, translators, and publishers who have given us permission to include their texts, and through financial support from the Faculty of Arts and the Department of British and American Studies, University of Oslo. The project also benefited greatly from the Nordic network ‘ Languages in Contrast’ , financed by the Nordic Academy of Advanced Study (1994-1996), and the project on corpora

and cross-linguistic research at the Centre for Advanced Study at the Norwegian Academy of Science and Letters (1996-1997).

The principal members of the research team have been: Stig Johansson, University of Oslo, and Knut Hofland, Norwegian Computing Centre for the Humanities, Bergen (project leaders), and Jarle Ebeling and Signe Oksefjell, University of Oslo (research assistants).

1.1 Aim

The aim of the English-Norwegian Parallel Corpus (ENPC) project is to produce a computer corpus for use in contrastive analysis and translation studies. The original plan was to compile a core corpus consisting of original texts and translations (Norwegian to English and English to Norwegian), and a larger supplementary corpus consisting of English and Norwegian texts matched by genre. In the course of the project, the focus changed and, rather than compiling a supplementary corpus of this kind, we decided to focus on a more detailed analysis of the core corpus (see Section 3) and on a multilingual expansion of the corpus (see Section 5).

1.2 Structure and uses of the corpus

The core corpus contains original texts and their translations (English to Norwegian and Norwegian to English). In order to make it possible to include material by a range of authors and translators, the texts of the core corpus are limited to extracts of some 10,000 - 15,000 words each. The core corpus contains both fictional and non-fictional texts, distributed as follows:

	Original texts		Translated texts	
	English	Norwegian	English	Norwegian
Fiction	30	30	30	30
Non-fiction	20	20	20	20
Total texts	50	50	50	50
Total number of words	671,700	629,900	699,400	661,500

All in all, there are thus 100 original texts and 100 translated texts, amounting to some 2.6 million words in all.

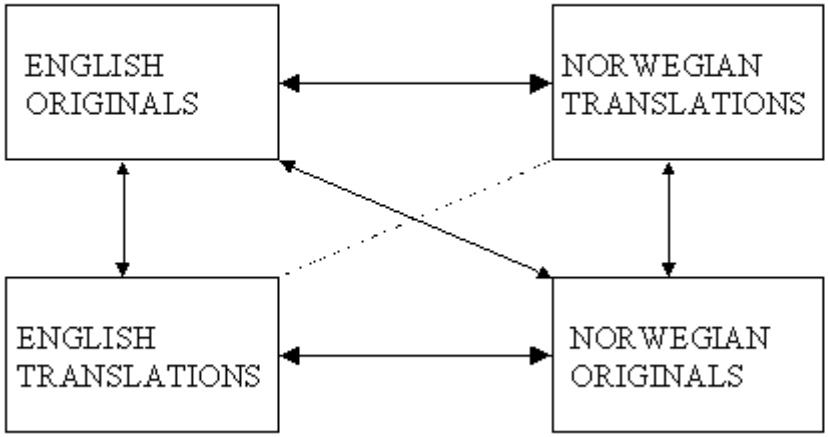

Figure 1 The structure of the English-Norwegian Parallel Corpus

Figure 1 shows the schematic structure of the core corpus. The main parts of the corpus are indicated by the four boxes, and the lines between them show the main types of studies which are made possible by this structure.

- contrastive studies based on parallel original texts (see the solid diagonal line in Figure 1);

- contrastive studies based on original texts and their translations, going from source text to translation and/or from translation to source text (see the solid horizontal lines in Figure 1);

- various types of translation studies, e.g. focusing on (a) translation problems viewed from either language (see the solid horizontal lines in Figure 1), (b) deviations of translated texts as compared with original texts in the same language (see the vertical lines in Figure 1), and (c) general features of translated texts (see the broken diagonal line in Figure 1).

The size of the boxes in Figure 1 is equal in terms of the number of texts. By balancing the corpus in this way, we hope to simplify the different types of comparison made possible by the structure.

1.3 Text selection

In selecting texts for the corpus, we decided to choose as recent texts as possible, in order to have a corpus that was as homogeneous as possible as far as the time dimension was concerned. A list of the texts, with bibliographical information, is given in Appendix 1.

Although it would have been preferable to sample systematically from different types of English (British, American, etc.) and Norwegian (*bokmål* and *nynorsk*), we could not do this within the scope of our project. The majority of the English texts were written by English or American authors, but the corpus also contains texts from other parts of the English-speaking world, e.g. Canada, Australia, and South Africa. Information on the origin of each text is

given in the header (see 2.3 below). With a couple of exceptions, the Norwegian original texts are in *bokmål*, and all the translations are in *bokmål*, no doubt reflecting the realities as far as Norwegian translations are concerned.

The decision to build up the corpus using original texts and translations means of course that we have only been able to choose from among texts that have been translated. It is a problem that far fewer texts are translated from Norwegian into English than the other way around. As regards Norwegian original texts, we made a survey of what had been translated in the last 10-15 years and collected as many of these as possible until we had filled our quota. For English we made an attempt to include a variety of recent texts of the types that are generally translated, including texts by major authors of fiction and non-fiction. We cannot claim, however, that the corpus is representative in a statistical sense.

There were special problems in finding non-fictional texts. Norwegian writers of non-fiction who want to reach an international audience often prefer to publish in English, rather than having their texts published first in Norwegian and then translated. English non-fictional texts are often translated only into one Scandinavian language, so if there is a Swedish translation, for example, we may not find a translation into Norwegian. Moreover, non-fiction texts are often adapted, making them unsuitable for the sort of linguistic studies that we had in mind in compiling the corpus. For these reasons, the non-fiction texts constitute a smaller proportion of the corpus than the fiction texts.

The choice of original texts and their translations has some advantages, however. Texts that are chosen for translation are often those which are especially valued or which have attracted a wide readership. The decision to use published texts also means that the texts have presumably gone through an editing process, which should provide some kind of guarantee of quality as regards the language of the texts.

It should be stressed that, although the corpus is based on original texts and their translations, it can also be used in a comparison of original texts in the two languages (see Figure 1). In other words, it can be used both as a *translation corpus* and as a *comparable corpus*. This is the unique feature of the ENPC, which makes possible a variety of studies, focusing on translation and/or language comparison.

Like any corpus, the ENPC is, however, not suitable for all kinds of studies. For lexical studies it will often be too small. As pointed out above, the selection of texts is skewed towards the types that are generally translated. Depending upon the focus of study, it will often be necessary to go beyond the corpus and study further material. Moreover, the wise corpus user realises that the corpus is only one tool in the exploration of language. The ultimate aim is to study the language, not the corpus.

1.4 Text classification

The header of each text (see 2.2.3) contains a classification code. The fiction texts are grouped into three types: children’ s fiction (FC), detective fiction (FD), and general fiction (FG). They are distributed as follows:

	Number of texts	
Genre	English original	Norwegian original

Children’ s fiction	3	7
Detective fiction	8	4
General fiction	19	19
Total	**30**	**30**

For the non-fiction texts, a sub-classification scheme was developed for the purpose of the project (see Aijmer et al. 1996:81), but it proved to be too difficult to apply, mainly because of the low degree of specification. The non-fiction texts have now been reclassified using a modified version of the Dewey decimal classification system.

	Number of texts	
Category	English original	Norwegian original
Religion	1	
Social sciences (incl. Economics)	4	7 (2)
Law	3	1
Natural sciences	5	
Medicine (incl. Psychology)	1	
Arts (sports, literature, rhetoric, philosophy, biography, music, etc.)	2	4
Geography & History (incl. travel)	4	10 (2)
Total	**20**	**22**

As the table shows, the non-fiction material is quite heterogeneous, and it was not possible to achieve a balance between the subcategories of the texts in the two languages. The non-fiction material must therefore be used with caution.

1.5 Definition of text

As already pointed out, the texts of the core corpus are mostly extracts from books and contain 10,000 to15,000 words (about 30 - 40 pages). All extracts are taken from the beginning of the books. The front matter - prefaces, forewords, list of contents, etc. - is not included in the extracts. In some cases, introductions have been left out as well, e.g. introductions by scholars to works of fiction. The length of each text is specified in the file description of the header (see 2.2).

The reasoning here was that we wanted to include fairly long and coherent text extracts, rather than short text samples taken from different parts of the texts. Although the extracts have all been taken from the beginning of books, they should be sufficiently long to get into the body of the text. In order to get a coherent piece of discourse, we chose to end each extract at a natural breaking point, preferably the end of a chapter (see also 2.3.1). For

this reason, there is some variation in the length of the extracts, but the total material for the main components of the corpus does not vary very much.

1.6 Text preparation

Initially, we made a survey of recent texts translated from English into Norwegian and from Norwegian into English. For the texts selected, we wrote to the copyright holders - authors, translators, and publishers - asking for permission to include the texts in our corpus and use them for the purposes of linguistic research. Only texts for which we received permission were included in the corpus.

After permission had been secured, we selected and scanned the text extracts and inserted - manually - the relevant information in the header of each text (see 2.2). Markup for major divisions of the text (see 2.3.2), such as chapters or parts of books, were also inserted manually. Paragraph and s-unit tags (see 2.3.3-4) were handled by a special program making use of paragraph markers, punctuation, capital initials, etc. in the scanned text. The file was proofread and corrected.

In the next stage, the original and the translation for each text were aligned, using a program for automatic sentence alignment. As a result, each sentence in the original and the translation received an identifier and one or more pointers to the relevant sentence(s) in the parallel text. After proofreading and correction, the aligned texts were put into a database for use with the browser developed within the project. As regards the programs, see further 4.1 and 4.2.

1.7 Availability

To be allowed to store and use the corpus, we have been subjected to strict conditions, formulated in agreement with the Norwegian authors' and translators' associations. The corpus can only be used for research. No commercial use is permitted. Use of the corpus is also limited to the institutions mentioned in our letters of permission, signed by the copyright holders. These are the Department of British and American Studies at the University of Oslo and the Norwegian Computing Centre for the Humanities at the University of Bergen. Scholars and students outside these institutions can gain access to the corpus by visiting, or co-operating with, one of these institutions.

2 Coding

2.1 General principles

The coding of the texts is in broad agreement with the TEI guidelines for electronic texts, as presented in Sperberg-McQueen and Burnard (1994). Textual features are marked by tags enclosed within angle brackets. For example, a heading is marked by a start-tag <head> and an end-tag </head>. Tags may have attributes, to provide an identifier of the element or characterize it in some other way, e.g. <p id=p1> to identify a particular paragraph or to mark a chapter. Some tags do not enclose text, e.g. <pb n=2> marking a page break at a particular point in the text. So-called entity references (bounded by & ;) can be used for a variety of purposes, e.g. to represent characters which are not available or to carry a grammatical tag. The occurrence of tags, attributes, and entity references in a particular type of document is called a document type definition.

The document type definition for the texts in the corpus differs in some respects from the TEI model; see Appendix 3. The differences are, however, mainly additions to the TEI model; a few new tags and entities have been introduced. These tags and entities can be found in the files ENPC.DTD and ENCP.ENT respectively. Together with ENPC.TXT, which invokes the appropriate TEI tag sets, they constitute the complete ENPC tag set.

The overall structure of an ENPC text is shown by this example:

<tei.2 id=AT1>
<teiHeader type=text>
</teiHeader>
<text>
</text>
</tei.2>

In other words, there are two main parts: a header and the main text. Every text has a unique identifier, in this case AT1 (indicating text 1 by Anne Tyler). The corresponding coding for the translation would be: <tei.2 id=AT1T>

The value of the identifier of the translated text is identical to that of the original, with the addition of a letter (T) marking it as a translation. Each text in the corpus thus has a unique identifier.

2.2 The header

Each text is described by a header which has four main parts, in accordance with the TEI guidelines: a file description, an encoding description, a profile description, and a revision description. These are tagged as follows (see also the example below):

<teiHeader>
<fileDesc></fileDesc>
<encodingDesc></encodingDesc>
<profileDesc></profileDesc>
<revisionDesc></revisionDesc>
</teiHeader>

Header and main text structure

<tei.2 id=AT1>
<teiHeader type=text>
<fileDesc>
<titleStmt>
<title>The Accidental Tourist: Extract in machine-readable
form</title>
<author>Anne Tyler</author>
<respStmt>
<resp>tagger</resp>
<name>BHL</name>
</respStmt>
</titleStmt>
<extent>12,000 words from beginning of text</extent>
<publicationStmt><distributor>English-Norwegian
Parallel Corpus (ENPC) Project</distributor></publicationStmt>

```
<notesStmt><note resp=tag></note></notesStmt>
<sourceDesc>
<biblStruct>
<monogr>
<author>Anne Tyler</author>
<respStmt>
<resp></resp>
<name></name>
</respStmt>
<title>The Accidental Tourist</title>
<imprint>
<pubPlace>New York</pubPlace>
<publisher>Alfred A. Knopf</publisher>
<date>1985</date>
</imprint>
</monogr>
</biblStruct>
</sourceDesc>
</fileDesc>
<encodingDesc>
<p>Modified TEI P3. See the ENPC project manual.</p>
</encodingDesc>
<profileDesc>
<langUsage><language>AmE</language></langUsage>
<textClass><classCode>FG</classCode></textClass>
</profileDesc>
<revisionDesc>
        <change>
                <date></date>
<name></name>
<what></what>
        </change>
</revisionDesc>
</teiHeader>
<text>
<body>
<div1 type= id= >
<div2 type= id= >
<p id= >
<s id= corresp= ></s>
</p>
</div2>
</div1>
</body>
</text>
</tei.2>
```

2.2.1 File description

The file description gives bibliographical information on the machine-readable file and the source text. Note that the <titleStmt> describes the machine-readable file, while the source text is specified in the <sourceDesc>. The title in the <titleStmt> should indicate that this is a machine-readable version and should not be identical to the title of the source text. The file description also specifies author, tagger, translator, publication information and the extent of the text extract.

Irregularities, e.g. omissions and non-standard spellings, of the text are noted in the <notesStmt> (see also 2.12.1 and 2.12.2).

2.2.2 Encoding description

The TEI encoding description may include a project description, editorial declarations (on correction, normalization, etc.), information on sampling, reference systems, and any classification schemes. In our case the encoding description can be very brief; it chiefly consists of a reference to the manual for the corpus and any additional comments on special features of encoding applying to the individual text.

2.2.3 Profile description

The profile description is of particular interest in the encoding of corpora, in that it makes it possible to describe each text in a very detailed manner. The present project will chiefly use the following main parts of the TEI profile description:

<langUsage><language> where the language/dialect of the text is described;

<textClass><classCode> where the text is classified in terms of a classification scheme;

The description under <langUsage><language> is in terms of labels like: American English (AmE), Australian English (AuE), British English (BrE), Canadian English (CaE), New Zealand English (NZE), etc. This section may also include observations on special linguistic features of the text (cf. 2.7 below). As regards the classification under <textClass><classCode>, see 1.4 above.

2.2.4 Revision description

The revision description takes the form of a series of changes. It is structured as

follows:

```
<revisionDesc>
<change>
<date></date>
<name></name>
<what></what>
</change>
</revisionDesc>
```

In other words, this is a list of changes specifying the date of the change, the person responsible for the change, and the nature of the change.

2.3 Text units

The corpus texts are segmented into the following main units: text, division (where applicable), paragraph, s-unit, and word. Words are simply marked by spacing as in ordinary written text. The other units are explicitly tagged.

2.3.1 Text

Where complete texts are encoded, these have the structure recommended by the TEI guidelines:

<text>
<body>
</body>
</text>

In the case of text extracts from books, [part of] the body only is included. The encoded text starts with the body of the main text, including headings, and ends with the nearest chapter or section division after the required number of words for the text extract has been reached. If the nearest chapter or section division extends considerably beyond the required number of words, the encoded text ends with the nearest paragraph.

The end of a text extract is marked by an <omit> tag; see 2.12.2.

2.3.2 Divisions

Most written texts include some sort of segmentation in terms of parts, chapters, sections, etc. According to the TEI guidelines, these units are tagged as numbered or unnumbered divisions. This corpus uses numbered divisions, where a lower number indicates a higher level. The type of division is described by an attribute. Example structure:

```
<body>

   <div1 type=part id=NN1.1>

      <div2 type=chapter id=NN1.1.1>

         <div3 type=section id=NN1.1.1.1></div3>

      </div2>

   </div1>

</body>
```

Each unit has an identifier which is built up by successively adding to the identifier of the text (in this case text NN1: cf. the example in 2.1 above).

Low-level divisions in the text which are only marked by a blank line, asterisks, or the like, are not tagged as divisions. The tag <blankline> is inserted at the appropriate point in the text. This may be taken to signal a major paragraph break.

2.3.3 Paragraphs

Divisions primarily contain a sequence of paragraphs (in addition, there may be

headings, notes, etc.). Continuing our example above, these are marked as follows:

```
<div3 type3=section id=NN1.1.1.1>

<p id=NN1.1.1.1.p1></p>
```

</div3>

Each paragraph has an identifier which adds yet another layer to the immediately superordinate identifier.

Paragraphs are identified as sections of texts marked by indentation, a blank line, or a combination of the two. Lists are marked as paragraphs or sequences of paragraphs; see 2.9.

2.3.4 S-units

Paragraphs are divided into orthographic sentences, here called s-units to underline that they are not necessarily sentences in a grammatical sense. They are tagged as follows:

<s id=NN1.1.1.1.s1 corresp=NN1T.1.1.1.s1></s>
<s id=NN1.1.1.1.s2 corresp=NN1T.1.1.1.s2></s>

S-units are numbered within the nearest division, as shown above. In this way, each s-unit is given a unique identifier. After alignment, each s-unit in the core corpus has a 'corresp' attribute containing a reference to the corresponding unit(s) in the parallel text.

An s-unit always opens after a paragraph start and ends before an end-of-paragraph marker. S-units are split within paragraphs where a mark of end punctuation (.?! or ... marking ellipsis) is followed by a word beginning with a capital initial (ignoring intervening parentheses, dashes, and quotation marks). No split is made between a colon or semi-colon followed by a word beginning with a capital initial (unless there is an end-of-paragraph marker).

S-units are not allowed to nest, i.e. they cannot be contained within each other. If there is an included sentence, e.g. within parentheses or between dashes, it is not coded separately, but is part of the s-unit it is included in. S-units may contain embedded poems, quotations, etc.

The division into s-units is complicated in some cases involving abbreviations and direct speech. Examples:

<s>Dr. Smith, St. George</s>
<s>"Hurry up!" Wolfram interrupted.</s>
<s>"Why didn't you come straight to me?" I asked her.</s>

No split is made in such cases, where the capital does not mark the beginning of an s-unit, but rather the nature of the word.

Headings, epigraphs, notes, and poems embedded in the text are not split into s-units.

2.3.5 Words

As pointed out above, words are simply marked by spacing as in ordinary written text. The exception is that contractions are split into two words (in order to facilitate alignment). Examples:

can't	ca n't
I'll	I 'll
it's	it 's
d'you	d' you

In the early stages of the project words were not grammatically annotated, with a couple of exceptions, such as:

let's let 's&pron;
soon's soon 's&subord;

The -s is here disambiguated by the following entity reference, which may be regarded as a grammatical tag. As regards word-class tagging, see 3.2.

2.4 Headings and other openers

Headings occur at the beginning of a division. They are marked by the tag <head>. Examples:

<div1 type=part id=NN1.1>
<head id=NN1.1.h1>Part 1</head>
<div2 type=chapter id=NN1.1.1>
<head id=NN1.1.1.h1>1 Mind in myth</head>

The "enumerator" is encoded as part of the head, as in these examples. Headings carry an 'id' attribute which is built up according to the same principle as the 'id' of paragraphs and s-units, i.e. they are numbered within the nearest but using 'h1, h2, etc.' rather than 'p1, p2, etc.' and 's1, s2, etc'. See 2.3.3-4.

Where there is more than one heading at a particular point, the tag <head> may be repeated. The typographical rendition of the heading is regularly left unmarked, but it can be specified by a 'rend' attribute; see 2.6.1.

Running heads at the top of pages are not encoded.

Epigraphs at the beginning of divisions have the following structure:

<epigraph>
<quote></quote>
<bibl></bibl>
</epigraph>

As regards the encoding of other opening elements, see the TEI guidelines.

2.5 Punctuation

The punctuation is regularly left as in the original text. Some problems of detail are taken up below.

2.5.1 Full stop

The full stop is retained both as a marker of abbreviation and when marking the end of an orthographic sentence. The two uses are disambiguated by the tagging of s-units (see 2.3.4).

The marking of ellipsis by successive full stops is regularized; any spaces before or between the dots are removed.

2.5.2 Hyphen

Line-end (soft) hyphens are removed where they are not part of the regular spelling of the word. In cases of doubt, guidance was sought elsewhere in the same text or in dictionaries. If doubt still remained, a hyphen was retained rather than removed.

2.5.3 Dash

Dashes are marked by an entity reference (—). No distinction is made between different types of dashes.

2.5.4 Quotation marks

All single quotation marks (') are converted to double quotation marks in direct speech and other contexts. Examples:

<s>"I do n't know how he stays so thin."</s>
<s>She used her "meeting voice".</s>

The single quotation mark, ('), is only used in contractions (*She 's, y' enjoy*) and to mark the genitive (*next week's Sunday newpapers' review section*). Quotations within quotation are tagged <qq>. Examples:

<s>"The finger got stuck inside his nose," Matilda said, "and he had to go around like that for a week.</s> <s>People kept saying to him, <qq>Stop picking your nose</qq>, and he could n't do anything about it.</s> <s>He looked an awful fool."</s>

<s>"Lately he 's discovered <qq>breakfast meetings</qq>.</s> <s>Now he gorges and guzzles all day.</s> <s>I do n't know how he stays so thin."</s>

As regards the treatment of quotations, see further 2.6.5.

2.5.5 Apostrophe

The apostrophe is left as it is. It is distinguished from single quotation marks (cf. 2.5.4).

2.6 Highlighting and quotation

No attempt is made to capture the full typography of the original text. Variation between upper and lower case is reproduced as in the original text. Use of typographical highlighting is marked where it is judged to be significant for the interpretation of the text.

2.6.1 Typographical highlighting

Typographical highlighting is marked by a "rend" (=rendition) attribute, if it applies to a whole element: a paragraph or an s-unit, as in:

<p rend=italic>
<s rend=bold>

Where there is no applicable element, the tag <hi> is used:

I <hi rend=italic>hate</hi> it.

The TEI guidelines propose the tag <emph> for linguistically emphatic or stressed sections of the text. The TEI tag <hi> is preferred in the present corpus, to avoid problems in identifying the purpose of typographical highlighting.

Where part of a text is highlighted typographically because it is identified as foreign, the tagging presented in the next section is preferred (though the 'rend' attribute may be used in addition).

2.6.2 Foreign words and expressions

Foreign words and expressions are marked by a 'lang' attribute. This is simple if the foreign element carries a tag:

<head lang=fr>
<s lang=la>

Where there is no applicable element, the tag <foreign> is used:

He was tried <foreign lang=la>in absentia</foreign>

Some possible values of the 'lang' attribute are:

de	German
en	English
es	Spanish
fr	French
gr	Greek
la	Latin
no	Norwegian
sv	Swedish

Foreign words and expressions are only marked where they are clearly recognizable as foreign (by being reproduced as typographically distinct from the surrounding text). The 'lang' attribute may also be used in the cases taken up next. Long passages in a foreign language are replaced by an <omit> tag; see 2.12.2.

2.6.3 Titles

Titles of books, newpapers, magazines, films, songs, paintings, etc. are tagged <title>, as in:

Have you read <title>Paradise Lost</title>?

Titles are only tagged if they are typographically highlighted in some way, eg by italic, bold or underscore.

2.6.4 Names

Names of persons, ships, boats, buildings, etc. are tagged <name>, as in:

I went on board <name>Tumble</name> and set sail.

Names are only tagged if they are typographically highlighted in some way, eg by italic, bold or underscore. The 'type' attribute is optional, and is usually not inserted.

Names of places, organizations, etc. are usually not tagged.

2.6.5 Quotations

Quotations from extraneous sources are tagged <quote>, as in:

The Apostle Paul said concerning some that <quote>"By good words and fair speeches they deceived the heart of the simple."</quote>

Foreign quotations are marked by a 'lang' attribute. Long foreign quotations are omitted and replaced by an <omit> tag; see 2.12.2.

In the original version of the corpus, direct speech in fiction is left unmarked and is simply shown by quotation marks. Note that direct speech may not be identifiable, as it is not always indicated by quotation marks.

As regards the marking of direct speech and thought in some parts of the corpus, see 3.1 below.

2.7 Linguistically distinct material

The marking of foreign elements has already been dealt with (see 2.6.2). Other cases of linguistically distinct material, such as dialect words or idiosyncratic spellings are often tagged <distinct>, with an attribute indicating the type of deviance. Examples:

<distinct type=nonstand>Mister Carlyle sure give it to yuh, he finds out!</distinct>

Why do we not treat <distinct type=nonceword>bunkraptcy</distinct> precisely as we treat bankruptcy?

The main value used for the 'type' attribute in the present project is "nonstand", indicating deviance of different kinds: dialect, slang, idiosyncratic spelling, etc. If such features are

pervasive in the text, this is noted in the header (under <notesStmt>), and each individual case is not marked.

2.8 Notes

Notes in the source text are tagged <note> and are inserted at the place in the text marked by the reference to the note. Attributes include 'resp' and 'place'. Example:

<note resp=auth place=foot>Unless otherwise specified, all remarks about bilingualism apply as well to multilingualism, the practice of using alternately three or more languages.</note>

Values of the 'resp' attribute used in the project are: auth (author), ed (editor), tr (translator), tag (tagger). References to notes are omitted.

Notes are not counted as included in the text proper, and are not split into s-units.

In special cases notes were omitted and replaced by an <omit> tag. See 2.12.2.

2.9 Lists

Lists which contain very little ordinary language text (e.g. lists of references) are omitted and replaced by an <omit> tag; see 2.12.2. Other lists are treated as paragraphs or sequences of paragraphs (the latter in case each list item is set out typographically as a paragraph). S-units are used for subdivision, as for ordinary paragraphs.

2.10 Figures, diagrams, and tables

Figures, diagrams, and tables are left out and replaced by an <omit> tag. See 2.12.2.

2.11 Embedded texts

Poems, songs, etc. that are embedded in a prose text are tagged <poem>. The internal structure is not specified. Verse lines are reproduced with a line break between each. There is a blank line between stanzas. Poems are included in the nearest s-unit. There is no internal division into s-units.

In some cases a poem is left out and replaced by an <omit> tag. See 2.12.2.

Embedded texts in prose are simply reproduced as part of the main text. Ordinary paragraph and s-unit marking is used. Frequently they are tagged as quotations; see 2.6.5.

2.12 Editorial comment

The mechanisms for editorial comment are those recommended by the TEI guidelines for simple editorial changes.

2.12.1 Correction and regularization

Correction is marked as shown by this example:

... to render that service to poor <corr sic="poele" resp=tag>people</corr>

Where it is apparent that there is a typographical error, the main text is corrected and the original reading is given as a value of a 'sic' attribute. A 'resp' attribute specifies the person responsible for the correction (normally "tag" for "tagger"; cf. 2.8).

The tag <sic> is used where there is no straightforward correction, but it is apparent that the text is inaccurate. A suggested correction may be given as a value of a 'corr' attribute. A 'resp' attribute specifies the person responsible for the correction. Repeated wrong spelling of words throughout a text is noted in the <notesStmt>, and is not tagged using the <corr> tag on each occasion.

Beyond correction of obvious typographical errors, the language of the corpus texts is not normalized or regularized.

2.12.2 Addition, deletion, and omission

Omission of passages in the text may be marked by an <omit> tag; see 2.3.1, 2.6.2, 2.6.5, 2.8, 2.9, 2.10, 2.11. The tag has the following attributes:

desc: describing the omitted text
reason: giving the reason for the omission
extent: indicating the extent of the omission
resp: specifying the person responsible for the omission

The 'desc' and 'resp' attributes are normally used. Sample 'desc' values include: table, figure, foreign text.

Addition and deletion in the main text are avoided. Where they occur, they are indicated by <add> and <del> tags.

2.13 Special characters

Special characters are encoded as entity references, for example:

š	š
£	£
—	—

Entity references specific to the project are listed in the project entity file (ENPC.ENT); see Appendix 3. All others are found in one of the public entity sets that comes with TEI P3, e.g. ISOpub.ENT.

Accented and special characters used in Western European languages (de, en, fr, no) are **not** encoded as entity references. They are, therefore, system dependent.

2.14 Page breaks

Page breaks in the source text are kept to make it easier to refer back to the source. They are tagged <pb n= >, i.e. with the number as the value of an attribute. The placement of <pb> is normalized and is always given at the beginning of the relevant page. If there is a page break in the middle of a hyphenized word in the original text, <pb> is placed after the relevant word in the encoded text.

2.15 Reference system

A reference system is built up using the identifiers of the text units. See 2.1 (text), 2.3.2 (division), 2.3.3 (paragraph), 2.3.4 (s-unit), 2.4 (heading).

2.16 Links

Links between parallel texts are indicated by attributes of s-units, as shown in 2.3.4. Example:

<s id=DL2.1.s18 corresp='DL2T.1.s18 DL2T.1.s19'>At once, feeling her advantage, she said, "Do n't forget you 've been living soft for four years."</s>

<s id=DL2T.1.s18 corresp=DL2.1.s18>Hun hadde fått et lite overtak og fulgte det opp.</s>

<s id=DL2T.1.s19 corresp=DL2.1.s18>"Ikke glem at du har levd godt i fire år nå."</s>

3 Analysis

To increase the usefulness of the corpus, we have added marking of direct speech and thought to all the original fiction texts in the corpus (3.1), and all the English original texts have been provided with part-of-speech tags (3.2).

3.1 Marking of direct speech and thought

All the original fiction texts in the corpus, English and Norwegian, have been marked for direct speech and thought. The entity &qb; (‘ quote begin’) has been used for start of direct speech/thought and &qe; (‘ quote end’) for end of direct speech/thought. This will facilitate and encourage research in the area of direct speech as opposed to straightforward narrative.

All texts do not have clearly marked boundaries between direct speech and the rest of the text. At the one end we find texts with quotation marks at the beginning and the end of each utterance:

"Are you bored with the election, my darling?" asked the Queen, stroking Harris's back. (ST1.1.1.s4)

At the other extreme, however, we find no overt marking at all:

Jeg vet noe, sier Rut, noe fælt noe. (BV2.1.1.s1)
Lit.: I know something, says Rut, something terrible.

The marking has been done partly automatically and partly manually, calling for some interpretation on the part of the person responsible for the dialogue marking. All the manual checking was done by Berit Løken, who also worked part time as a tagger on the project.

3.2 Word-class tagging

The English part of the ENPC has been tagged for part-of-speech (P-O-S). The tagging was done automatically using the English Constraint Grammar parser (http://www.ling.Helsinki.FI/~tapanain/cg/cgdemo.html) developed by Atro Voutilainen, Juha Heikkilä, Arto Anttila and Pasi Tapanainen according to the Constraint Grammar framework originally proposed by Fred Karlsson. We are grateful to Atro Voutilainen, Helsinki, for doing the actual tagging.

Before the tagger could be applied, the tagger' s lexicon was updated – the texts in the corpus were checked for words not already in the lexicon, and these were manually given P-O-S tags. Additionally, all tags and entities had to be removed from the texts. After the tagger had been run on the texts, the original text, i.e. the text with the alignment information, had to be merged with the P-O-S tagged text (see the table on the next page).

The tagging information is contained in the <w> (word) element, which contains three attributes: 'lemma', 'pos', and 'feature'. The merging of the original and P-O-S tagged texts was carried out by Diana Santos and Helge Hauglin at the Text Laboratory, University of Oslo.

Finally, the intermediate tag set used in the tagging process was converted to the compact representation of the EngCG-2 (http://www.ling.Helsinki.FI/~avoutila/cg/index.html). Appendix 2 contains an overview of all the tags. So, what the user of the corpus is presented with is this format:

<div1 type=part id=PM1.1>
<head id=PM1.1.h1 corresp=PM1T.1.h1><w p="Nadv">JANUARY</head><pb n=1>

<s id=PM1.1.s1 corresp=PM1T.1.s1><w p="DET">The <w p="Nadv">year <w l="begin" p="Vpast">began <w p="PREP">with <w p="N">lunch.</s>

The <w> element is now moved in front of the word it specifies (following the layout of the British National Corpus), and only the 'pos' (reduced to 'p') and the 'lemma' (reduced to 'l') attributes remain. The original 'feature' attribute is partly included in the 'p' attribute of the compact tag set.

There are some points which the user of the P-O-S tagged English texts should keep in mind. The English Constraint Grammar parser has a high level of accuracy, but errors are made, and often a word is left with more than one P-O-S tag or lemma tag. After the automatic tagging, there was a proofreading and correction stage. We examined particularly the cases where one word had been assigned more than one tag, and all tags that were obviously incorrect were eliminated. We did not, however, have the time and the resources to check all the material for errors and consistency of tagging.

As regards the application of the individual tags, note in particular that participle forms such as *walked* and *walking* are tagged EN and ING, respectively, and do not receive a V tag. Elements of some sequences which behave as units are given identical tags, e.g. *as well as* is tagged Cc Cc Cc and *in order to* is tagged TO TO TO.

The Norwegian part of the corpus has not been tagged as of January 1999, but a Norwegian tagger using the Constraint Grammar framework is under development, and, given sufficient resources, we hope to tag the Norwegian original texts as well.

Before merging (P-O-S tagged text):	After merging (of aligned and P-O-S tagged text):
"<JANUARY>" "January" <ADV-N> <Proper> N NOM SG "<$>" "<$>"	<div1 type=part id=PM1.1> <head id=PM1.1.h1 corresp=PM1T.1.h1> JANUARY <w lemma="January" pos="N NOM SG" feature="ADV-N Proper"> </head> <pb n=1>
"<The>" "the" DET SG/PL	<p id=PM1.1.p1> <s id=PM1.1.s1 corresp=PM1T.1.s1> The <w lemma="the" pos="DET SG/PL">
"<year>" "year" <ADV-N> N NOM SG	Year <w lemma="year" pos="N NOM SG" feature="ADV-N">
"<began>" "begin" V PAST	Began <w lemma="begin" pos="V PAST">
"<with>" "with" PREP	With <w lemma="with" pos="PREP">
"<lunch>"	Lunch <w lemma="lunch" pos="N NOM SG">

"lunch" N NOM SG "<$.>" "<$<s>>"	. </s></p>

4 Programs

A number of programs were written in connection with the project, e.g. a program for splitting the texts into s-units. The main programs are the Translation Corpus Aligner, which aligns texts automatically at the sentence level (4.1), and the Translation Corpus Explorer, which is a browser for parallel texts (4.2).

4.1 The Translation Corpus Aligner

The Translation Corpus Aligner was crucial in the building of the corpus. It takes as input machine-readable versions of the original and the translation, and produces versions of the texts where each s-unit in the original and the translation is provided with a unique identifier ('id' attribute) and a 'corresp' attribute pointing to the corresponding s-unit(s) in the parallel text (cf. 2.3.4 above). The program is described in Hofland (1996) and Hofland & Johansson (1998).

The program was originally written for English and Norwegian, but has later been developed to handle a number of other language pairs.

4.2 The Translation Corpus Explorer

The Translation Corpus Explorer (TCE) was developed to allow the user to search and browse the corpus texts. The program makes use of the 'id' and 'corresp' attributes of s-units (cf. 2.3.4 above) and produces pairs of matching text extracts from original and translation. The program is described in some detail in Ebeling (1998b), and the current version (WebTCE) has a Web interface which makes it accessible over the Internet. Since the use of the corpus texts is subject to strict conditions, access to the corpus is also restricted, and a password is needed to browse the texts.

A detailed account of the search options, the versions of corpus that are available, etc. can be found in the Help menu to the program. To access the corpus using WebTCE, follow the link from the project home page (see under References).

5 Expansion of the corpus

In order to facilitate more general cross-linguistic and translation studies, we have created two additional corpora: a corpus of multiple translations (5.1) and a multilingual corpus (5.2).

5.1 Multiple translations

In the English-Norwegian Parallel Corpus, each original text is linked with its published translation. A related project focuses on the degree of variation among professional translators translating the same original text.

In collaboration with Stig Johansson, Linn Øverås has commissioned multiple translations of the same two English original texts into Norwegian: "A Lamia in the Cevennes", a short story by A. S. Byatt, and the scientific article "Communication and cooperation in early infancy: a description of primary intersubjectivity" by Colwyn Trevarthen. Norsk Oversetterforening (The Norwegian Association of Literary Translators) and Norsk faglitterær forfatter- og oversetterforening (The Norwegian Non-fiction Writers and Translators Association) have contributed funds for the project.

The translators are among the best and most experienced translators in Norway. Each has provided a draft version and an edited version of the text. The material will be used in the first instance as the basis for a doctoral thesis on general aspects of translation.

The multiple-translation project is an extension of the English-Norwegian Parallel Corpus in the sense that it uses the techniques and the tools developed for the purpose of the original project. The same mark-up and the same software are used as in the main project.

5.2 A multilingual corpus

If we want to gain insight into language and translation generally, and at the same time highlight the characteristics of each language, it is desirable to extend the comparison beyond pairs. For these reasons we have collected translations of many of our English original texts into three other languages: German, Dutch, and Portuguese. Together with our Norwegian translations and the translations into Swedish and Finnish from our sister projects, we can then compare across six languages using English as a starting-point (see Figure 2).

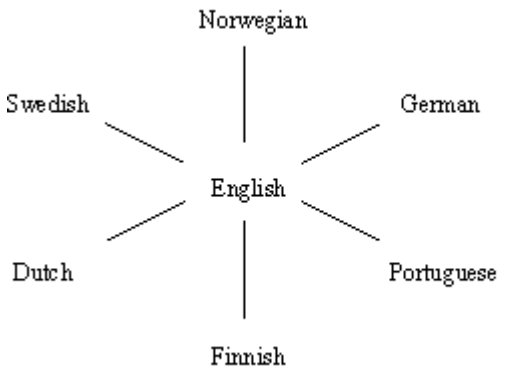

Figure 2 Multilingual expansion of the ENPC

The multilingual project uses the techniques and the tools developed for the purpose of the original project. The same mark-up and the same software are used as in the main project. A number of databases for different language combinations have been built up, e.g. English-German-Norwegian and English-German-Dutch. For information, see the list of databases in the Translation Corpus Explorer (cf. 4.2).

For economic reasons, we have not been able to build up the texts in other languages to the same extent as for English and Norwegian. A list of the texts available is given on our web page: http://www.hf.uio.no/iba/prosjekt/

In the future, we plan to extend the multilingual corpus focusing on English/German and Norwegian/German, in particular, in collaboration with the Department of Germanic Studies and the Section for Applied Linguistics, University of Oslo. A trilingual database which matches the structure of the original ENPC will then be built up for use in a cross-linguistic study of English, German, and Norwegian.

6 Important note for the user

Copyright holders have given us permission to include their texts in the corpus on the conditions specified in 1.7 above. Please be careful to observe the conditions of use. Many copyright holders have stressed that references should be given to the printed texts. In publications making use of the corpus, remember to give full bibliographical information on the texts (author, translator, title, publisher, etc.). In shorter publications, it may be sufficient to give a reference to the web page of the project: http://www.hf.uio.no/iba/prosjekt/

References

http://www.hf.uio.no/ilos/english/services/omc/enpc/ENPCbib.html

---

Appendix 1: List of corpus texts

- Fiction http://www.hf.uio.no/ilos/english/services/omc/enpc/ENPCfiction.html
- Non-fiction http://www.hf.uio.no/ilos/english/services/omc/enpc/ENPCnon.html

For information on texts in other languages, see our web page:
http://www.hf.uio.no/ilos/english/services/omc/enpc/languages.html

---

Appendix 2: List of word-class tags

The tag set used for the English texts is that of the EngCG-2 (see http://www.ling.helsinki.fi/~avoutila/cg/doc/engcg2-tags/engcg2-tags.html), with two alterations. The tag **EX** is introduced to separate existential *there* from adverb *there*, and the negative particles *not* and *n't* are tagged **NEG** and not **ADV** (other adverb) as in the original EngCG-2.

To keep the number of tags to a minimum, the tag set used for the Norwegian texts overlap to a great extent with the English one with some exceptions (see the list below).

Several of the tags are exclusive to one of the languages. In most cases this is transparent, as with the tags **ADVwh** and **Infmerke**. Tags exclusive to the Norwegian texts have Norwegian

examples only. If in doubt about a particular tag, consult the lists of tag frequencies which accompany the TaggedTCE browser.

Tag	***Explanation***	***Examples*** (Comment)
A	adjective	*hilarious, Brown*
Acmp	comparative adj.	*better*
Apresp	present participle	*lengtende* (Nor. texts only)
Asup	superlative adj.	*best*
ADV	other adverb	*very, enough, more, up*
ADVwh	**WH**-adverb	*while, when, how*
Cc	coordinator	*and, as+well+as*
Cs	subordinator	*if, though, for*
DET	other determiner	*the, both, those*
DETdem	demonstrative determiner	*de, denne, slik* (Nor. texts only)
DETintens	intensive determiner	*egen, selv* (Nor. texts only)
DETkvant	quantitative determiner	*én, noe, syv* (Nor. texts only)
DETposs	possessive determiner	*hans, sin* (Nor. texts only)
DETwh	**WH**-determiner	*whose, which, whatever*
EN	past participle	*walked, been*
EX	existential *there*	*there*
I	interjection	*oh, hi*
Infmerke	infinitive marker	*å* (Nor. texts only)
ING	**ING**-form	*walking, interesting*
N	noun, fraction	*Joe, dog, one-third*
Nabbr	abbreviation	*Dr., Ms, A*
Nadv	adverbial **N**	*time, today, Monday*
Ncard	cardinal number	*one, 123, 9%*
Ngen	genitive **N**	*Dr.'s, boy's, boys', Paul's*

Nord	ordinal number	*fifth, 7th*
Npl	plural **N**	*dogs*
Nprop	proper **N**	*Perugia, Vatikanet* (Nor. texts only)
NEG	negation	*not, n't*
P	pronoun	*all, this, everything, none*
Pgen	genitive **P**	*my, his, its, whose*
Pint	interrogative **P**	*which, who*
Pinterr	interrogative **P**	*hva, hvem* (Nor. texts only)
Pnom	nominative **P**	*I, he, she, it*
Pobl	oblique (accusative, dative) **P**	*him, her, it, us*
Ppers	personal **P**	*den, hun, jeg* (Nor. texts only)
Prefl	reflexive **P**	*seg* (Nor. texts only)
Prel	relative **P**	*who, that, which*
Presipr	reciprocal **P**	*hverandre, kvarandre* (Nor. texts only)
Pubest	indefinite **P**	*man* (Nor. texts only)
Pwh	other wh-**P**	*what, whatever*
PREP	preposition	*to, for, in+spite+of*
TO	infinitive marker	*to, in+order+to*
U	unknown word	(Nor. texts only)
V	other verb	(Nor. texts only)
Vimp	verb, imperative	*be, walk*
Vimpaux	auxiliary verb, imperative	*bli, få, vær* (Nor. texts only)
Vinf	verb, infinitive	*be, walk*
Vinfaux	auxiliary verb, infinitive	*bli, burde, være* (Nor. texts only)
Vmod	modal auxiliary	*can, will, may*
Vpast	verb, past tense	*was, walked*

Vperfp	verb, past participle	*advart, halset, sendt* (Nor. texts only)
Vperfpaux	auxiliary verb, past participle	*blitt, fått, kunnet* (Nor. texts only)
Vpres	verb, present tense	*is, walks, walk*
Vpresaux	auxiliary verb, present tense	*blir, bør, er* (Nor. texts only)
Vpret	verb, past tense	*kverna, la, rettet* (Nor. texts only)
Vpretaux	auxiliary verb, past tense	*ble, skulle, var* (Nor. texts only)
Vsbj	verb, subjunctive	*be, walk*

Appendix 3: Extensions to the TEI Guidelines

To understand how the files below fit into the TEI scheme, the *TEI Guidelines* have to be consulted, especially chapter 29.

The first file listed below (ENPC.TXT) invokes the TEI tag set, and must be included whenever an ENPC text is parsed, that is checked, against the TEI DTD (see 2.1). All the foreign language attributes must be listed in this file.

ENPC.ENT, the next file below, contains the additional entities introduced and used in the ENPC texts. The file also holds the so-called class extensions, that is the new elements to be introduced into the TEI DTD.

The last file, ENPC.DTD, contains the descriptions of the new elements, and the attribute 'skip' which is an additional feature of the <s> element of the TEI. The skip attribute tells the alignment program not to include the current s-unit in the alignment process.

ENPC.TXT

```
<!DOCTYPE teiCorpus.2 system 'tei2.dtd' [

<!ENTITY % TEI.corpus 'INCLUDE'>

<!ENTITY % TEI.prose 'INCLUDE'>

<!ENTITY % TEI.analysis 'INCLUDE'>

<!ENTITY % TEI.linking 'INCLUDE'>

<!ENTITY % ISOLat1 SYSTEM "ISOLat1.ENT">

<!ENTITY % ISOLat2 SYSTEM "ISOLat2.ENT">

<!ENTITY % ISOpub SYSTEM "ISOpub.ENT">
```

```
<!ENTITY % ISOnum SYSTEM "ISOnum.ENT">

%ISOLat1;

%ISOLat2;

%ISOpub;

%ISOnum;

]>

<teiCorpus.2 id=ENPC>

<teiHeader type=corpus>

<fileDesc>

<titleStmt>

<title></title>

</titleStmt>

<publicationStmt><distributor>English-Norwegian Parallel (ENPC)
Project</distributor>

</publicationStmt>

<sourceDesc>

<biblStruct>

<monogr>

<author></author>

<title></title>

<imprint>

<pubPlace></pubPlace>

<publisher></publisher>

<date></date>

</imprint>

</monogr>

</biblStruct>

</sourceDesc>

</fileDesc>

<profileDesc>

<langUsage>
```

```
<language id=af>

<language id=da>

<language id=de>

<language id=en>

<language id=es>

<language id=fr>

<language id=gr>

<language id=inuit>

<language id=it>

<language id=iw>

<language id=la>

<language id=nl>

<language id=no>

<language id=sv>

<language id=sw>

</langUsage>

<textClass><classCode><sic id=tag></sic></classCode></textClass>

</profileDesc>

</teiHeader>
```

File: ENPC.ENT

```
<!-- English-Norwegian Parallel Corpus (ENPC) project

     Entities added to TEI P3

     Last modified: 1995-02-17

     File: enpc.ent

     Jarle.Ebeling@iba.uio.no    -->


<!-- GENERAL ENTITIES -->

<!--      ENTITIES CONTENT  -->

<!ENTITY  pron     "_pron"   > <!-- Pronoun contraction "'s" in "let's" -->

<!ENTITY  subord   "_subord" > <!-- Subordinator contraction "'n" in
"More'n" -->
```

```
<!ENTITY  pm       "_pm" > <!-- Plus sign above a minus sign -->

<!-- FOR THE NEW CONTENT MODELS, SEE FILE ENPC.DTD -->
<!ENTITY % s 'IGNORE' >

<!-- ADDITION TO TEI P3: Class extension
     Cf file ENPC.DTD and chapter 29 in TEI P3 -->
<!ENTITY % x.chunk    'blankline |' >
<!ENTITY % x.globincl 'omit |'      >
<!ENTITY % x.hqphrase 'poem | qq |' >
```

File: ENPC.DTD

```
<!-- English-Norwegian Parallel Corpus (ENPC) project
     Elements changed and added to TEI P3
     Last modified: 1995-02-17
     File: enpc.dtd
     Jarle.Ebeling@iba.uio.no    -->

<!-- Cf file ENPC.ENT and chapter 29 in TEI P3 -->
<!ELEMENT s          - -        (%phrase.seq;) -(s) >
<!ATTLIST s            %a.global;
                       %a.seg;
                       skip     NUMBER              #IMPLIED >

<!ELEMENT blankline - o        EMPTY            >
<!ELEMENT omit      - o        EMPTY            >
<!ATTLIST omit         desc     CDATA               #IMPLIED
                       reason   CDATA               #IMPLIED
                       extent   CDATA               #IMPLIED
                       resp     (auth | ed | tr | tag) #IMPLIED >
<!ELEMENT poem      - -        (#PCDATA | corr | qq)+   -(s) >
```

```
<!ELEMENT qq        - -        (#PCDATA | hi | distinct)+ -(s) >
```