1. Background

Parallel corpora have proved extremely useful resources for cross-linguistic research and translation studies in recent years. They provide an empirical basis for contrastive and typological research and they give new insights into the languages compared — insights that are likely to be unnoticed in studies of monolingual corpora. They are also important for practical applications in various fields, such as lexicography, language teaching and computer-aided translation.

The compilation of the *English-Swedish Parallel Corpus* (ESPC) began at Lund University in 1993 with financial support from the *Swedish Council for Research in the Humanities and Social Sciences* (HSFR). From 1997 the corpus has been developed as a cooperative project by the Departments of English at the Universities of Lund and Göteborg. The principal members of the project have been Karin Aijmer, University of Göteborg, and Bengt Altenberg, Mats Johansson and Mikael Svensson, University of Lund. From the start the design and compilation of the corpus have been carried out in close cooperation with sister projects in Norway and Finland (see section 5).

This manual describes the second, enlarged version of the corpus completed in the spring of 2001.

2. The corpus

2.1 Structure and uses of the corpus

Parallel corpora can consist of comparable original texts in two or more languages ('comparable corpora') or of original texts and their translations into another language ('translation corpora'). The English-Swedish Parallel Corpus (ESPC) combines the advantages of these two types. The structure of the corpus is shown in Figure 1.
Figure 1. Structure of the English-Swedish Parallel Corpus

The original texts from both languages have been matched as far as possible in terms of text type, subject matter, purpose and register (see 2.4). This means that the corpus can be used:

- as a comparable corpus (vertical line joining left-hand boxes)
- as a translation corpus (horizontal lines)
- for comparing original and translated texts in the same language (diagonal lines)
- for comparing translated texts in the two languages (vertical line joining right-hand boxes)

2.2 Size and composition of the corpus

The current (expanded) version of the ESPC consists of 64 English text samples and their translations into Swedish and 72 Swedish text samples and their translations into English (see Table 1). The samples from each language have been drawn from two main text categories, fiction and non-fiction. Most of the samples are extracts from larger works and consist of 10,000-15,000 words, but in the non-fiction component there are also some shorter complete texts as well as some composite texts consisting of several shorter complete texts (see section 2.5). As a result, there are more samples representing non-fiction than fiction, especially among the Swedish original texts, but the size and proportion of the two text categories (in terms of running words) are roughly the same in the two languages. The total size of the corpus is 2.8 million words.

Table 1. Size and composition of the corpus

<table>
<thead>
<tr>
<th></th>
<th>English originals</th>
<th>Swedish transl.</th>
<th>Swedish originals</th>
<th>English transl.</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fiction</td>
<td>Text samples</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>No. of words</td>
<td>340,745</td>
<td>346,649</td>
<td>308,160</td>
<td>333,375</td>
</tr>
<tr>
<td>Non-fiction</td>
<td>Text samples</td>
<td>39</td>
<td>39</td>
<td>47</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>No. of words</td>
<td>364,648</td>
<td>344,131</td>
<td>353,303</td>
<td>413,500</td>
</tr>
<tr>
<td>Total</td>
<td>Text samples</td>
<td>64</td>
<td>64</td>
<td>72</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>No. of words</td>
<td>705,393</td>
<td>690,780</td>
<td>661,463</td>
<td>746,875</td>
</tr>
</tbody>
</table>
2.3 Text selection

The selection of texts for the corpus has been guided by the following principles:

- the texts should reflect contemporary language and not be restricted to any specific regional variety;
- the corpus should include a wide range of text types, authors and translators;
- the text types from both languages should be as comparable as possible, i.e. well matched in terms of genre, subject matter, type of audience and register;
- where possible, the text samples should consist of fairly large and coherent extracts of 10,000 - 15,000 words (30 - 40 pages) taken from the beginning of a text and ending at a natural breaking point (chapter or section);
- as far as possible, the English source texts should be available in Finnish and Norwegian as well as Swedish translations (see section 5).

In reality, the selection has been constrained by various practical circumstances, especially by what is translated between the two languages. The following features should be noted.

With few exceptions, the samples have been taken from texts published since 1980. Most major regional varieties of English are represented (British, American, Canadian, Irish, South African) but no attempt has been made to achieve a systematic or 'representative' distribution of these. Only written texts are represented. A number of prepared speeches have been included but they have their origin in writing and do not reflect genuine speech. Other categories that are missing in the corpus are, for example, newspaper text, private letters and business correspondence.

Apart from such obvious gaps in the corpus, what has turned out to be especially problematic has been to achieve a good genre match between the text samples from the two languages. Many text types (e.g. popular fiction, newspaper reporting) are generally only translated in one direction, typically from English into Swedish. Moreover, in areas where English is used as an international language (such as scientific writing) there is no need for translation. As a result, the selection of texts has been limited to text types that are translated in both directions (see Table 2).

2.4 Text classification

The header of each text contains a text classification code indicating its text type (see 3.2.3). The classification is shown in Table 2. The fiction texts are divided into three types: children's fiction (FC), crime and mystery (FD), and general fiction (FG). For the non-fiction texts the original plan was to make a division of the material into 'popular' and 'specialised' texts reflecting the purpose and intended audience of the texts (see Aijmer et al 1996), but this idea proved to be difficult and had to be abandoned. Instead the non-fiction texts have been classified into eight broad subject areas, partly based on the Dewey Decimal Classification System.
Table 2. Classification of texts in the corpus

<table>
<thead>
<tr>
<th>Text types</th>
<th>Number of texts</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>English originals</td>
<td>Swedish originals</td>
<td></td>
</tr>
<tr>
<td>Fiction</td>
<td>25</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Children’s fiction</td>
<td>1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Crime and mystery</td>
<td>7</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>General fiction</td>
<td>17</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Non-fiction</td>
<td>39</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>Memoirs and biography</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Geography (travel, leisure)</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Humanities (history, religion, dance, folklore)</td>
<td>5</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Natural sciences (astronomy, evolution, zoology)</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Social sciences (economics, politics, welfare)</td>
<td>7</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Applied sciences (medicine, environment)</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Legal documents (acts, EU documents, treaties)</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Prepared speech (Nobel lectures, political speeches)</td>
<td>15</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>64</strong></td>
<td><strong>72</strong></td>
<td></td>
</tr>
</tbody>
</table>

As the table shows, it has not been possible to achieve a complete balance between all the subcategories of the two languages. Stylistic comparisons across the two languages must therefore be made with caution.

**2.5 Text samples**

As mentioned, most of the texts are extracts from books and contain 10,000 - 15,000 words (about 30-40 pages). The extracts have been taken from the beginning of the books. Front matter - prefaces, forewords, list of contents and (in some cases) introductions - has not been included in the extracts. The principle has been that each extract should represent a fairly long and coherent piece of text rather than short samples from different parts of the texts. For the same reason, the extract generally ends at a natural breaking point (chapter or section) of the text. As a result, 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.

Some of the samples consist of smaller complete texts. This applies to most of the legal documents (EU directives, Swedish acts) and to the prepared speeches (Nobel lectures,
political speeches). In some of these cases several short texts have been conflated into 'composite' texts (the EU directives and all speeches delivered by the same speaker in the EU Parliament).

2.6 Text preparation

Some of the texts in the corpus were obtained in electronic form but most of them have been scanned from an original published source. They were then converted to a common text format and proofread. Despite the proofreading, there are some typographical errors in the corpus. These will be corrected in future versions of the corpus.

Each text is supplied with a ‘header’ with information about the text (see 3.2). In addition, certain features of the text have been annotated with TEI-conformant 'markup'. The features coded in this way are of two main types:

(a) major divisions of the text, such as chapters, paragraphs and sentences (s-units);

(b) other typographical features of the text, such as headings, highlighted elements, titles, etc.

A description of this coding is given in section 3.

2.7 Availability and note for users

The ESPC project has been allowed to store and use the texts under certain strict conditions stated in the permissions from the copyright holders. The corpus can only be used for research. No commercial use is permitted. Moreover, the corpus is only available for research at the Department of English at the Universities of Lund and Göteborg. Scholars and students outside these departments can gain access to the corpus by visiting, or cooperating with, one of these departments.

Many copyright holders have stressed that references should be given to the printed texts. In publications based on 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 corpus webpage.

3. Coding

3.1 General principles

The coding of the texts is in broad agreement with with the TEI guidelines for electronic texts (see Sperberg-McQueen & Burnard 1994). The system used in the ESPC is identical with that used for the English-Norwegian Parallel Corpus, which is described in detail in Johansson, Ebeling & Øksefjell (1999). The following survey is an abbreviated version of that description.

Textual features are marked by 'tags' enclosed within angle brackets. For example, headings are enclosed by a start-tag <head> and an end-tag </head>. Some tags have 'attributes' to identify or characterise an element, e.g. <p id=p1> which identifies the beginning of a particular paragraph or <div type=chapter> which marks the beginning of a chapter. Some
tags do not enclose text, e.g. `<pb n=2>` which marks a new page (and its number) in the text. So-called 'entity references' (bounded by & and ;) are used for a variety of purposes, e.g. to represent characters which are not available.

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 corpus differs in some respects from the TEI model. These differences are, however, mainly additions to the TEI model (see Johansson, Ebeling & Oksefjell 1999: Appendix 3).

The overall structure of an ESPC text is shown by the following example:

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

This coding indicates that 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 identifier of the translated version of this text is identical to that of the original, except that a letter T is added to the identifier: `<tei.2 id=AT1T>`. Hence, each text in the corpus has a unique identifier.

### 3.2 The header

Each text is described by a header which has four main parts: a file description, an encoding description, a profile description, and a revision description. These are tagged as follows (see also Figure 2 below):

```xml
<teiHeader>
  <fileDesc></fileDesc>
  <encodingDesc></encodingDesc>
  <profileDesc></profileDesc>
  <revisionDesc></revisionDesc>
</teiHeader>
```
3.2.1 File description

The file description contains bibliographical information about the machine-readable file and the source text. The <titleStmt> describes the machine-readable file, while the <sourceDesc> identifies the source text. These must be differentiated as they are not identical. The file description also specifies author, tagger, translator, publication information and the extent of the text extract. Irregularities in the electronic text (e.g. omissions) are noted in the <notesStmt> (see 3.12.2).
3.2.2 Encoding description

The encoding description simply contains a reference to the manual of the English-Norwegian Parallel Corpus (Johansson, Ebeling & Oksefjell 1999).

3.2.3 Profile description

The profile description specifies two features of the text:

- `<langUsage><language>` indicating the language/dialect of the text
- `<textClass><classCode>` indicating the genre or text type of the text

The `<langUsage><language>` description specifies the regional variety of language used in the English texts in terms of labels like: AmE (American English), AuE (Australian English), BrE (British English), CaE (Canadian English), IrE (Irish English), etc. This section may also include comments on other pervasive linguistic features of the text (see 3.7 below). As regards the text classification under `<textClass><classCode>`, see 2.4. above.

3.2.4 Revision description

The revision description specifies any changes made in the text, including the date of the change, the person responsible for the change, and the nature of the change.

3.3 Text units

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

3.3.1 Text

The text samples in the corpus either represent complete texts or extracts from books. In the former case the text is enclosed by the following tags:

```
<text>
<body>
...
</body>
</text>
```

In the case of text extracts, part of the body only is included. The encoded text starts with the body of the main text, including headings, and ends at 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 to indicate that the rest of the source text has been omitted (see 3.12.2).
3.3.2 Divisions

The major divisions of the text (parts, chapters, sections, etc) are tagged as numbered divisions, where a lower number indicates a higher level. The type of division is described by an attribute as illustrated in the following example:

```
<body>
  <div1 type=part id=NN1.1>
    <div2 type=chapter id=NN1.1.1>
      <div3 type=section id=NN1.1.1.1> ...
    </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).

Low-level divisions in the text which are only marked by a blank line, an asterisk, etc. 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.

3.3.3 Paragraphs

Paragraphs are identified as sections of text marked by indentation, a blank line, or a combination of these. Lists are also marked as paragraphs or sequences of paragraphs (see 3.9). The beginning and end of each paragraph are marked by a tag; the beginning tag has an identifier which adds yet another numbered layer to the immediately superordinate identifier. Continuing the example above, the first paragraph in subdivision 3 (a section) of text NN1 will be marked as follows:

```
<div3 type=section id=NN1.1.1.1>
  <p id=NN1.1.1.1.p1> ...
</p>
</div3>
```

3.3.4 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 numbered within the nearest division (typically a paragraph), as shown below:

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

In this way, each s-unit is given a unique identifier. After alignment, each s-unit in the corpus has a 'corresp' attribute containing a reference to the corresponding unit(s) in the parallel text (see 3.15).
The demarcation of s-units is sometimes problematic. Headings, epigraphs, notes, and poems embedded in the text are not split into s-units. For a more detailed description of the principles used, see Johansson, Ebeling & Oksefjell (1999).

3.3.5 Words

Words are simply marked by spacing. The exception is that contractions are split into two words (in order to facilitate alignment). Examples:

can't > can't
I'll > I'll
it's > it's
d'you > d'you

Words have not been grammatically tagged, but there are some exceptions:

let's > let's
soon's > soon's

The -s is here disambiguated by the following entity reference, which may be regarded as a grammatical tag.

3.4 Headings and epigraphs

Headings occur at the beginning of a division and are marked <head>:

<div 1 type=part id=NN1.1>
<head id=NN1.1.h1>Part 1</head>
<div 2 type=chapter id=NN1.1.1>
<head id=NN1.1.1.h1>1 Mind in myth</head>

As this example shows, the <head> tag contains an 'id' attribute which is built up according to the same principle as the 'id' of paragraphs and s-units. The typographical form of the heading is normally left unmarked, but it can be specified by means of a 'rend' attribute (see 3.6.1). 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>

3.5 Punctuation

The punctuation is largely left as in the original text. Some deviations and special features are described below. It should be noted, however, that these have not been applied consistently in the material and mainly concern the English original texts.
The marking of ellipsis by successive stops is regularized: any spaces before or between the dots have been removed. Dashes are marked by an entity reference (—). No distinction has been made between different types of dashes. All single quotation marks (') have been converted into double quotation marks in direct speech and other contexts. Single quotation marks are only used in contractions and to mark the genitive. Quotations within quotations are tagged <qq>.

3.6 Highlighting and quotation

No attempt has been made to capture the full typography of the original text. Variation between upper and lower case is reproduced as in the original text. Typographical highlighting (italics, bold face, etc) is marked where it is judged to be significant for the interpretation of the text.

3.6.1 Highlighting

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

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

In other cases the tag <hi> is used:

```xml
I <hi rend=italic>hate</hi> it
```

Where part of a text is highlighted to indicate a foreign expression, the tagging presented below has been preferred.

3.6.2 Foreign words and expressions

Foreign words and expressions are marked by the tag <foreign> and a 'lang' attribute:

```xml
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
sv Swedish

If the foreign element already carries a tag, the attribute is inserted in the tag:

```xml
<head lang=fr> 
<s lang=la>
```
Foreign words and expressions are only marked when they are clearly recognizable as foreign (e.g. by highlighting). Long passages in a foreign language have been omitted and replaced by an <omit> tag (see 3.12.2).

3.6.3 Titles

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

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

3.6.4 Names

Names of ships, boats, buildings etc are also tagged if they are typographically highlighted in some way. The 'type' attribute has not always been used in these cases:

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

3.6.5 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 3.12.2).

3.7 Linguistically distinct material

Apart from foreign expressions, 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:

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

If nonstandard features (dialect, slang, idiosyncratic spellings, etc) are pervasive in a text, this is noted in the header (under <noteStmt>) and each individual case is not marked. On the treatment of typographical errors, see 3.12.1 below.

3.8 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 corpus 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 (especially when merely containing a bibliographical reference) and replaced by an <omit> tag.

3.9 Lists

Lists which contain very little ordinary language (e.g. lists of references) are omitted and replaced by an <omit> tag. Other lists are treated as paragraphs or sequences of paragraphs. S-units are used for subdivision, as for ordinary paragraphs.

3.10 Figures, diagrams, and tables

Figures, diagrams, and tables are left out and replaced by an <omit> tag. Pictures and other illustrations have been silently omitted.

3.11 Embedded texts

Poems, songs, etc. that are embedded in a prose text are tagged <poem>. They have been included in the nearest s-unit without an internal division into s-units. In some cases a poem is left out and replaced by an <omit> tag. Embedded texts in prose are simply reproduced as part of the main text, with ordinary paragraph and s-unit marking. Frequently they are tagged as quotations (see 3.6.5)

3.12. Editorial comment

3.12.1 Correction and regularization

Where it is apparent that there is a typographical error, the error has been 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"):

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

The tag <sic> is used where there is no straightforward correction but it is apparent that the text is inaccurate. Beyond correction of obvious typographical errors, the language of the corpus texts is not normalized or regularized.

3.12.2 Omission

Omission of passages in the text may be marked by an <omit> tag, normally with a 'desc' and 'resp' attribute describing the omitted text and the person responsible for the omission. Typical values of the 'desc' attribute are: table, figure, foreign text, note.

3.13 Special characters

Special characters are encoded as entity references, for example:

\[ &\text{scaron}; \\
\£ &\text{pound}; \\
\–– &\text{mdash}; \]
Accented and special characters used in Western European languages (English, French, German, Swedish) are not encoded as entity references. They are, therefore, system dependent.

3.14 Page breaks

Page breaks in the source text are kept to make it easier to refer back to the source. They are tagged \(<pb n= >\), with the page number as the value of an attribute. The \(<pb>\) tag is always placed 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 generally placed after the relevant word in the encoded text, but in some texts it occurs before the word.

3.15 Links

Links between parallel texts are indicated by attributes of s-units, as shown in 3.3.4. The following example shows a sentence from an English source text that is rendered as two sentences in the Swedish translation. The link between the sentences is indicated by the values of the 'id' and 'corresp' attributes which point from s73 in the source text to s73 and s74 in the translation and vice versa:

\(<s id=AT1.1.s73 corresp='AT1T.1.s73 AT1T.1.s74'>He passed a line of cars that had parked at the side of the road, their windows opaque, their gleaming surfaces bouncing back the rain in shallow explosions.</s>\)

\(<s id=AT1T.1.s73 corresp=AT1.1.s73>Han passerade en lång rad bilar som hade parkerat vid sidan av vägen.</s>\)

\(<s id=AT1T.1.s74 corresp=AT1.1.s73>Deras rutor var ogenomskinliga, och regndropparnas studsar mot de glänsande ytorna var som små explosioner.</s>\)

4. Alignment program and search tool

The use of the aligned version of the ESPC is dependent on a number of computer programs developed by the English-Norwegian Parallel Corpus team at Bergen and Oslo. Apart from 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, developed by Knut Hofland, Bergen, and Stig Johansson, Oslo, 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 text and the translation are linked by means of a unique identifier ('id' attribute) and a 'corresp' attribute pointing to the corresponding s-unit(s) in the parallel text (cf 3.3.4 above). The program was originally written for English and Norwegian, but has later been adapted for English and Swedish and a number of other language pairs. The program is described in Hofland (1996) and Hofland & Johansson (1998).
4.2 The Translation Corpus Explorer

The Translation Corpus Explorer (TCE), created by Jarle Ebeling, Oslo, was developed to allow the user to search and browse the aligned corpus texts. The program makes use of a database version of the aligned corpus and produces pairs of matching text extracts from original and translation. An internet version of the program (WebTCE) is available for researchers permitted to use the corpus (see 9 and the ENPC project home page). A detailed account of the search options can be found in the Help menu of the program. The program is described in some detail in Ebeling (1998).

4.3 Paragraph-aligned version

The ESPC also exists in a paragraph-aligned version where individual source texts and translations are are linked paragraph by paragraph. This version makes use of the table facility in MS Word. It is especially useful when a full view of the running text is needed and in cases where a linguistic feature cannot be spotted by the browser but has to be identified by a close reading of the text. The following extract from the opening paragraphs of an English novel illustrates this possibility:

```
<pb n=3>1
Time is not a line but a dimension, like the dimensions of space. If you can bend space you can bend time also, and if you knew enough and could move faster than light you could travel backwards in time and exist in two places at once.

It was my brother Stephen who told me that, when he wore his ravelling maroon sweater to study in and spent a lot of time standing on his head so that the blood would run down into his brain and nourish it. I didn't understand what he meant, but maybe he didn't explain it very well. He was already moving away from the imprecision of words.
```

```
<pb n=11>1
Tiden är inte en linje utan en dimension, precis som rummets dimensioner. Kan man kröka rummet sa kan man också kröka tiden, och om man bara visste tillräckligt och kunde röra sig fortare än ljuset så kunde man färdas baklänges i tiden och finnas till på två platser samtidigt.

Det var min bror Stephen som berättade det för mig, då när han hade på sig sin utslitna rödbruna tröja medan han pluggade, och tillbringade en massa tid med att stå på huvudet så att blodet skulle strömma ner i hjärnan på honom och ge den nåring. Jag begrep inte vad han menade, men kanske förklarade han det inte så bra. Han var redan på väg bort från ordens brist på exakthet.
```

5. Cooperation

From the start the design and compilation of the ESPC have been carried out in close cooperation with sister projects in Norway (Stig Johansson, Oslo) and Finland (Kari Sajavaara, Jyväskylä). One important aim of the three Nordic projects has been to select English original texts that are available in Norwegian, Finnish and Swedish translations (see 2.3). As a result, many of the English source texts are shared by the three national corpora: the English-Norwegian Parallel Corpus (ENPC), the Finnish-English Contrastive Corpus, and the English-Swedish Parallel Corpus (ESPC). This has important methodological and theoretical
advantages: not only are the design and structure of the three corpora similar, they also permit interesting linguistic and typological comparisons of two closely related languages (Norwegian and Swedish) and one language belonging to a totally different language family (Finnish).

The Norwegian team has also collected translations of many of the English original texts into three other languages: German, Dutch, and Portuguese. This means that the English-Swedish Parallel Corpus is a branch of a larger multilingual corpus which permits comparisons across six languages using English as a starting-point. For information on this multilingual corpus, see the Norwegian web page: https://www.hf.uio.no/ilos/english/services/omc/