To appear in proceedings from symposium «Languages in Contrast, March 1994», Department of English, Lund University, Sweden.

Coding and Aligning the English-Norwegian Parallel Corpus

Stig Johansson (Stig.Johansson@iba.uio.no),
Jarle Ebeling (Jarle.Ebeling@iba.uio.no), University of Oslo, and
Knut Hofland (Knut.Hofland@hd.uib.no), Norwegian Computing Centre for The Humanities, Bergen

1 Introduction

The aim of the English-Norwegian Parallel Corpus (ENPC) project is to produce a computer corpus for use in contrastive analysis and translation studies. There will be a core corpus consisting of original texts and translations (Norwegian to English and English to Norwegian), and a supplementary corpus consisting of English and Norwegian texts matched by genre. As regards the structure of the corpus and the projected uses of the material, see Johansson and Hofland (1994).

At present, we are at the stage of building up the core corpus. This paper will focus on the encoding of texts and on the development of our method of automatically aligning original and translated texts. These are both critical steps in building a corpus of this kind. The purpose of text encoding is to prepare the texts in such a way that they adequately reflect the source texts and are maximally useful for computational studies. The aim of the alignment program is to specify equivalent points in the original and the translation, so that corresponding sections of the texts can be easily retrieved. Alignment and encoding are closely connected, in that decisions at the encoding stage have important consequences for alignment.

2 Text encoding

Before alignment can be attempted, a text must be available in electronic form and contain some basic markup. The pre-alignment text encoding consists of three phases:

I Scanning and insertion of initial mark-up
II Insertion of additional mark-up
III Proofreading

The markup is inserted into the text as tags; see Appendix. Angle brackets are used to separate tags from the text: <...> for 'start-tags' and </...> for 'end-tags'. Tags may have attributes, e.g. <div1 type=part id=ST1.1> indicating the start of a part of a book with the identifier ST1.1. The use of tags and attributes is defined formally by a so-called 'document type definition' (DTD). The coding of the texts is in broad agreement with the recommendations of the Text Encoding Initiative (TEI). See further Sperberg-McQueen & Burnard (1994).

2.1 Scanning and insertion of initial mark-up

All the texts are scanned by the use of an OCR reader. The scanner outputs word processing files. This means that paragraphs are uniquely identifiable, and can be tagged as such when we convert a text processing file to ASCII format, which is the format in which we store our files. The scanner can also recognise highlighted text, e.g. italic or bold, and tags for this are also inserted automatically. However, this is not foolproof, and the scanner often misses out on highlighted text, or places tags for start and end of highlighted text wrongly.

In addition to paragraph mark-up ( ) and highlighted text (<hi rend=italic> </hi>), we also insert markers of page breaks (<pb n=3>) at this point in the process. Example (1) shows a section from a text sample after scanning.

(1) <pb n=28> Night and day melted into one. They slept on deck scaffolds, cramped into the space between barrels that smelled sourly of rotten fish and sea water. There were many prisoners. An English girl cried out to her mother in the dark, and men wailed in foreign languages. From a dark corner came the voice of a monk reciting long prayers in Latin: <hi rend=italic>In nomine Patris et Filii et Spiritus Sancti</hi>...

2.2 Insertion of additional mark-up

The next phase comprises three steps. First, contractions are split to simplify the alignment process, then s-units (orthographic sentences) are recognised and tagged, and finally an empty header file is placed before the text.

Each file (text extract) has its own header. The header consists of four main parts: a file description, an encoding description, a profile description, and a revision description. These are tagged as follows:

(2)
```
<header>
    <fileDesc></fileDesc>
    <encodingDesc></encodingDesc>
    <profileDesc></profileDesc>
    <revisionDesc></revisionDesc>
</header>
```

At the outset, we had our own header, but the TEI header (tagged <teiHeader>) can now also be used (they differ only slightly). See Appendix 1.

The splitting of contractions is done with the help of a small computer program which separates the constituent words (e.g. *isn't --> is n't, I'll --> I 'll*). This does, however, result in some funny words, e.g. wo (*won't --> wo n't*), and ca (*can't --> ca n't*).

Insertion of s-unit mark-up is also done by a computer program. The program searches for sequences of end punctuation (.?!) followed by a capital letter, and inserts an s-unit end tag (</s>) after the end punctuation and an s-unit start tag (<s>) before the capital letter. It also utilises the paragraph tags (,) introduced above since an s-unit start tag (<s>) always occurs after a paragraph start tag ( ) and an s-unit end tag (</s>) before a paragraph end tag ( ). The program also has to take into account a number of exceptions to the rules above, for example in connection with abbreviations (*Mrs. Brown, St. Patrick*). Example (3) shows a section of tagged text:

(3) <pb n=3> <s>Time is not a line but a dimension, like the dimensions of space.</s> <s>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.</s>

2.3 Proofreading

Apart from correction of all errors produced by the scanner and of possible errors produced by the programs splitting contractions and inserting s-unit markup, the proofreading stage also involves the insertion of tags for the start and end of text, and the start and end of parts and chapters. Highlighted text which contains a book title, a foreign word, name of a ship, building, etc. is given the appropriate tag (<title>, <foreign lang=xx>, or <name>), and tags for omitted text, pictures, drawings, etc. are inserted. The header is also completed with the appropriate information at this stage (see Appendix 1 for illustration of a complete header).

The beginning and end of a text, with the header unspecified, will look like example 4 (from Margaret Atwood's *Cat's Eye*).

(4)
```
<partext lang=en type=novel orig=yes id=MA1>
<header>
</header>
<text>
<front>
</front>
<body>
<div1 type=part id=MA1.1>
<head>I Iron Lung</head>
<div2 type=chapter id=MA1.1.1>
<head>1</head>
<pb n=3><p><s>Time is not a line but a dimension, like the dimensions of space.</s>
<s>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.</s></p>
...
<p><s>I know he is n't the real Santa Claus, just someone dressed up like him.</s>
<s>Still, my idea of Santa Claus has altered, has acquired a new dimension.</s>
<s>After this it becomes hard for me to think of<pb n=38> him without thinking also
of the snakes and the turtles and the pickled eyes, and the lizards floating in their
yellow jars, and of the vast, echoing, spicy, ancient and forlorn but also comforting
smell of old wood, furniture polish, formaldehyde and distant mice.</s></p>
</div2>
</div1>
<omit extent="rest of text">
</body>
<back>
</back>
</text>
</partext>
```

Front and back matter (<front>, <back>) are left empty, while <div1> - <div5> are used to tag the different partitions of a text. The <head> tag is used to tag part, chapter and section headings. The outermost tag, <partext>, is really superfluous in that it is a synopsis of the information contained in the header, and can be replaced by <tei.2>, which is the tag recommended by the TEI guidelines. The unique identifier attributes in the tags are used in the alignment process.

Finally, the text extract is run through an SGML parser to ensure that the encoding is consistent, i.e. that the text conforms to the DTD. The DTD lists all possible elements, attributes, and entities which the text can contain, and describes the structure of the text, e.g. specifies that <s> tags can only occur within tags. Appendix 3 gives an example of a DTD.

3 Coding problems

Lack of standardisation of layout and lack of agreement between original and translation create problems at the encoding stage. So far we have tried to be as true to the printed text as possible, even if this causes problems for the alignment program. This means, for instance, that differences in paragraph layout between original and translation are kept:

(5) <s>Robert M. Turner stanset og betraktet sin bordfelle, slik en tryllekunstner kan nøle et<pb n=13> øyeblikk for å gi publikum en sjanse til å oppdage hemmeligheten i det han foretar seg.</s> <s>Så fortsatte han:</s>
<s>Hvilken rørende naivitet!</s> ...

<s>Robert Turner stopped and observed his table companion as a magician may hesitate a moment to give his audience a chance to discover how he is doing his trick.</s> <s>Then he continued: What touching naïvety!</s> ...

Different ways of rendering direct speech also cause problems. Both single and double quotation marks are used, and also, especially among Norwegian writers, the dash (—).

(6) <s>'Oh, you do have electricity then?' she asked.</s>
<s>'No.</s> <s>A comrade in the next street fills it for me every morning,' he said.</s>

<s>The Queen Mother looked quite pleased to hear this.</s> <s>"So I shan't starve?"</s>
<s>"Under the People's Republican Party's rule, nobody in Britain will starve," said Jack.</s>

<s>‐ Tordensønnen er et godt navn, sa han.</s>
<s>‐ Tordensønnen er et godt navn, gjentok Raske Hjort.</s>

We have kept the dash and double quotation marks, and converted all the single quotation marks to double in direct speech, in quotations, and elsewhere where words or expressions are set out with the use of quotation marks. The dash is represented by an entity reference (see Appendix 2) to distinguish it clearly from hyphens. Single quotation marks are avoided to prevent confusion with the apostrophe. All kinds of quotes within quotes are tagged using the <q> tag.

(7) <s>"Lately he 's discovered <q>breakfast meetings</q>."</s>

<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, <q>Stop picking your nose</q>, and he could n't do anything about it.</s> <s>He looked an awful fool."</s>

It would have been preferable to mark dialogue consistently using the <q> tag. As there is a variety of devices marking dialogue (quotation marks, dash, indentation), this is a major task and will be deferred until a later stage of the project.

Even though the program that inserts <s> and </s> tags is fairly sophisticated, proofreading is necessary and errors occur, especially in connection with proper names, the pronoun *I*, and ellipsis marked by full stops (...).

Problems further arise in connection with lists and embedded texts like poems and songs. Lists are normally treated as paragraphs or sequences of paragraphs. Lists which contain very little ordinary text are replaced by an <omit> tag. Poems, songs, etc. that are embedded in prose text are tagged <poem>. The internal structure is not specified.

In addition to certain types of lists, all drawings, pictures, tables, etc., including captions, are replaced by an <omit> tag. In some cases a note specifying the type and amount of omitted material is included in the <notesStmt> in the header.

Apart from the traditional divisions into parts and chapter, writers quite often use one or more blank lines to signal some kind of change in the line of events of the story. These «scene changes» are not given a <div1>-<div5> tag; the blank lines are simply replaced by one <blankline> tag. The <blankline> tag is also used whenever a poem, list, quotation, etc. is separated from the text by one or several blank lines.

(8) <s>"Nor do I, Bev.</s> <s>I mean, the Queen next door.</s> <s>We 'll put in for a transfer, eh?"</s>
<s>Slightly comforted, Beverley went to sleep.</s>
<blankline>
<s>It was Tony Threadgold who had prised the boards from the front door, but it was Prince Philip who had taken the key from his wife, turned it in the lock and entered the house.</s> <s>It was ludicrously small, of course.</s>

Footnotes are embedded in the text where they occur; the internal structure of the note is not encoded. If required, the number of the note is specified by an attribute: <note n=xx>.

(9) <s>"I want you to take Son-of-Thunder and travel north to the Chiricahua Mountains.</s> <s>My brother Nachi is there, he is chief of all the Chokonen.</s> <note resp=auth place=foot>The Chokonen was the name of one of the Diné bands. It was also known as the Central Chiricahua Band.</note> <s>Tell him my last great wish was that Nachi should bring up Son-of-Thunder as his own child."</s>

The coding scheme is quite elaborate, and may be too detailed for other projects. The tags required for the alignment program to work are only: <div1>-<div5>, <head>, , and <s>. For a complete list of the tags currently used in the project, see Appendix 2.

4 Alignment method

Many of the decisions at the coding stage have been taken with a view to simplifying the process of aligning the texts. This is why we have been careful to identify the major units of the texts, in particular chapters and similar divisions, paragraphs, and s-units (orthographic sentences). The splitting up of contractions simplifies the matching of words in the source and target texts. This is crucial, as our alignment method relies heavily on word correspondences.

The basic features of our alignment method have been described in some detail in Johansson and Hofland (1994). The matching is done at the level of s-units. The object is to

provide each s-unit with a link to the corresponding s-unit(s) in the parallel text. The alignment program attempts to find the best match using a combination of:

- shared anchor words as defined in a simple bilingual lexicon (an asterisk indicates truncated forms; a slash is used to separate English and Norwegian forms when there are more than one in each language)
  e.g. morning* / morgen*, formiddag*
  most / mest, meste, flest*
  mother* mor*
  mouth* mun*
  mov* / rør*, flytte*
  much mye
  music* musik*
  must må
  my, mine / min, mine, mitt
  myself / selv, sjøl*, meg;

- sentence length in terms of characters.

Since the method was first presented in Johansson and Hofland (1994), it has been refined in the following ways:

- the list of anchor words has been revised;

- in addition to the list of anchor words, the program takes into account correspondences as regards proper names;

- the program also takes into account certain correspondences as regards punctuation marks (chiefly question marks, exclamation marks, and colons).

The result of these modifications is a more delicate correspondence measure, resulting in more accurate alignment.

A further important modification of the alignment program is that it no longer matches s-units paragraph by paragraph. Although texts usually match very well at the level of the paragraph, there are enough deviations to invalidate this procedure. S-units are now matched by moving a fixed window of 15 s-units successively through the text.

4.1 Procedure

The alignment program was originally written for the UNIX operating system, but has now been transferred to DOS, which will be the only supported platform from now on.

The input to the program consists of the two texts that are to be aligned and the anchor word list. The program tries to find the best possible match between the 2 x 15 s-units it has read, based on the criteria outlined above. The program outputs several files; the most important are the two aligned texts with s-units specified with attributes for 'id' (identifier) and 'link' (the identifier of the corresponding s-unit in the parallel text). Example (10) shows samples of aligned texts (from Sue Townsend's *The Queen and I/Dronninga og jeg*):

(10a) Original:

<div1 type=part id=ST1.1>

<head>April</head>
<omit desc=drawing>
<div2 type=chapter id=ST1.1.1>
<head>1 Uneasy Lies the Head</head>
<pb n=13>

<s id=ST1.1.1.s1 link=ST1T.1.1.s1>The Queen was in bed watching television with Harris.</s>
<s id=ST1.1.1.s2 link=ST1T.1.1.s2>It was election night, 11.20 pm, Thursday 9 April 1992.</s>
<s id=ST1.1.1.s3 link=ST1T.1.1.s3>Harris yawned, displaying his sharp teeth and liver-coloured tongue.</s>

<s id=ST1.1.1.s4 link=ST1T.1.1.s4>"Are you bored with the election, my darling?" asked the Queen, stroking Harris's back.</s>
...
<s id=ST1.1.1.s20 link='ST1T.1.1.s20 ST1T.1.1.s21'>The computer men vanished from the screen to be replaced yet again by anxious politicians being interviewed and Harris lost interest and jumped back onto the bed.</s>

(10b) Translation:

<div1 type=part id=ST1T.1>
<head>APRIL</head>
<div2 type=chapter id=ST1T.1.1>
<head>1. UTRYGT HVILER HODET</head>
<pb n=11>

<s id=ST1T.1.1.s1 link=ST1.1.1.s1>Dronningen lå i sengen sin og så på TV sammen med Harris.</s>
<s id=ST1T.1.1.s2 link=ST1.1.1.s2>Det var valgnatt, klokken var 23.20 torsdag den 9 april 1992.</s>
<s id=ST1T.1.1.s3 link=ST1.1.1.s3>Harris gjespet og blottet de skarpe tennene og den leverfargede tungen.</s>

<s id=ST1T.1.1.s4 link=ST1.1.1.s4>"Kjeder valgsendingen deg, lille venn?" spurte dronningen og strøk Harris over ryggen.</s>
...
<s id=ST1T.1.1.s20 link=ST1.1.1.s20>Dataherrene i floss forsvant fra skjermen og ble nok en gang avløst av intervjuer med bekymrede politikere.</s>
<s id=ST1T.1.1.s21 link=ST1.1.1.s20>Harris mistet all interesse og hoppet opp i sengen igjen.</s>

Each s-unit is given a unique identifier, the value of the 'id' attribute. As shown by the underlined portions of (10), the identifier consists of the text id, ST1 (Sue Townsend, text 1) for the original and ST1T for the translation, the number of the current division ST1.1.1, and the sequential number of the s-unit within the current division ST1.1.1.s1. The value of the 'link' attribute is the value of the 'id' attribute of the corresponding s-unit in the translation, ST1T.1.1.s1.

The 'id' and 'link' attribute values are checked with the help of the SGML parser, which will report any empty links and invalid 'link' attribute values. Links which point to

nothing will either be deleted, or, if an s-unit has not been translated, will be set to point to the nearest previous paragraph ‘id’ of the corresponding text (see example (11)). Invalid ‘link’ attribute values are corrected.

(11a) Original:

<s id=ST1.1.2.s157 link=ST1T.1.2.s170>"She 's been well paid for it," snapped Jack.</s>
<s id=ST1.1.2.s158 link=ST1T.1.2.s171>"And I could give you the names of a dozen people I know personally who have worked twice as hard for their country as your mother and have been paid nowt."</s>
<s id=ST1.1.2.s159 link=**ST1T.1.2.p61**>Jack's use of the word "nowt" came from his childhood, a time of poverty and humiliation, when his political philosophy was formed.</s>

(11b) Translation:

** **
<s id=ST1T.1.2.s170 link=ST1.1.2.s157>"Hun har fått godt betalt," bet Jack ham av.</s>
<s id=ST1T.1.2.s171 link=ST1.1.2.s158>"Og jeg kunne gitt deg navnene på et dusin hederskvinner, personlig bekjente av meg, som har arbeidet dobbelt så iherdig for landet sitt som mora di uten å få fem flate pence for jobben."</s>

The examples above illustrate the final output of the alignment process and the format in which the texts will be stored.

4.2 Results

Another file which the alignment program outputs shows how the program has «reasoned» when matching s-units. This output also displays the aligned s-units in a more reader-friendly format. The output is, first and foremost, used to proofread the aligned texts. It can also be used to detect new words which could be added to the anchor word list and to fine-tune the alignment algorithm.

Example (12), again from Sue Townsend’s *The Queen and I/Dronninga og jeg*, shows the alignment matrix of one text window and the aligned s-units. Rows correspond to the English original, while columns correspond to the Norwegian translation. The values of the matrix are calculated on the basis of anchor words, proper nouns, punctuation marks, and sentence length in characters. The figures above and beside the rows and columns give the s-unit numbers within each chunk and the number of characters in each s-unit. The bold figures inside the matrix show how the alignment program has moved through the matrix when matching s-units. The line below the matrix shows how the s-units have been matched.

(12)

Trans(No)		165	37	56	37	61	85	73	159	69	45	117	62	57	29	73
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Orig	(En)															
1	107	**9**	1	1	0	1	1	1	6	1	1	3	3	1	0	0
2	29	**1**	0	0	0	1	0	1	0	1	1	0	1	1	1	0
3	32	0	**7**	2	1	0	0	0	0	0	0	0	0	0	1	0
4	46	0	3	**10**	0	1	0	1	0	0	1	1	0	0	1	0
5	23	0	0	0	**2**	0	0	0	0	0	0	0	0	0	0	0
6	118	2	0	1	0	**3**	**2**	3	2	1	1	2	4	3	0	0
7	64	1	0	1	0	2	0	**3**	1	2	1	2	3	4	0	2
8	150	4	0	0	0	1	1	2	**13**	2	0	1	1	1	0	0
9	51	0	0	1	0	2	0	1	0	**3**	3	1	3	3	0	0
10	57	1	0	1	0	2	0	1	1	2	**3**	1	3	3	0	0

```
 11 116      3   0   0   0   1   1   1   3   1   1   5   2   1   0   0
 12  72      3   0   0   0   2   2   3   3   2   1   2   9   3   1   2
 13  58      1   0   1   0   2   0   1   1   2   1   1   4   7   0   1
 14  28      0   0   0   0   1   1   1   1   1   0   0   1   0   4   0
 15  76      0   0   0   0   0   1   1   1   1   0   0   1   1   0   4

 Result= 1+2,1 3,2 4,3 5,4 6,5+6 7,7 8,8 9,9 10,10 11,11 12,12 13,13 14,14 15,15

-------------------------------------------------------------------------------
 1: <s>He sometimes put on a silly voice and said things about the Goons and
     stuff, but mostly he was dead serious.</s>  (ST1.1.4.38)
 2: <s>Frowning and giving lectures.</s></p>  (ST1.1.4.39)

 1: <s>Iblant la han an en fjollet stemme og bablet om busemenner og sånt,
     men som oftest sto han med rynket panne og lekset opp for dem om plikt og
     ansvar, dødsens alvorlig.</s></p>  (ST1T.1.4.40)
-------------------------------------------------------------------------------
 3: <p><s>Mama said, "This is our new house."</s></p>  (ST1.1.4.40)

 2: <p><s>"Dette er det nye huset vårt," sa mammi.</s></p>  (ST1T.1.4.41)
-------------------------------------------------------------------------------
 4: <p><s>William said, "How can it be new when it 's old?"</s></p>  (ST1.1.4.41)

 3: <p><s>"Hvordan kan det være nytt når det er gammelt?" sa William.</s></p>  (ST1T.1.4.42)
-------------------------------------------------------------------------------
 5: <p><s>Again the boys laughed.</s>  (ST1.1.4.42)

 4: <p><s>Nok en gang brølte guttene av latter.</s>  (ST1T.1.4.43)
-------------------------------------------------------------------------------
 6: <s>William lost control and needed support, he leaned against the
     creosoted fence dividing their house from its neighbour.</s>  (ST1.1.4.43)

 5: <s>William mistet helt <pb n=26>besinnelsen og måtte støtte seg til
     noe.</s>  (ST1T.1.4.44)
 6: <s>Han lente seg mot det kreosotinnsatte stakittet som delte den vesle
     hageflekken i to.</s>  (ST1T.1.4.45)
-------------------------------------------------------------------------------
 7: <s>The tired fence gave way under his fragile weight and collapsed.</s>  (ST1.1.4.44)

 7: <s>Det slitne gjerdet ga seg under ham, og William tumlet inn i naboens
     hage.</s>  (ST1T.1.4.46)
-------------------------------------------------------------------------------
 8: <s>Seeing him there, fallen and shrieking amongst splintered wood, Diana
     automatically looked for Nanny, who always knew what to do, but Nanny was
     n't there.</s>  (ST1.1.4.45)

 8: <s>Han satte i et skingrende hyl der han lå og kavet i plankehaugen, og
     Diana så seg om etter Dadda, som alltid visste hva som skulle gjøres, men
     Dadda var ikke der.</s>  (ST1T.1.4.47)
-------------------------------------------------------------------------------
 9: <s>She bent down and lifted her son from the wreckage.</s>  (ST1.1.4.46)

 9: <s>Diana bøyde seg ned og løftet sønnen opp av de sørgelige
     vrakrestene.</s>  (ST1T.1.4.48)
-------------------------------------------------------------------------------
10: <s>Harry whimpered and clung to the hem of her denim jacket.</s>  (ST1.1.4.47)

10: <s>Harry kvinket og trakk henne i dongerijakken.</s>  (ST1T.1.4.49)
-------------------------------------------------------------------------------
11: <s>Charles booted furiously at the front door, which opened, releasing a
     stench of neglect and damp and ghostly chip fat.</s>  (ST1.1.4.48)

11: <s>Charles ga døren et rasende spark, og dermed fløy den opp og slapp ut
     en stank av vanstell og fukt og gammelt matfett.</s>  (ST1T.1.4.50)
-------------------------------------------------------------------------------
12: <s>He switched on the hall light and beckoned his wife and children
     inside.</s></p>  (ST1.1.4.49)

12: <s>Han slo på lyset i gangen og vinket sin kone og sine barn inn.</s></p>  (ST1T.1.4.51)
-------------------------------------------------------------------------------
13: <p><s>Tony Threadgold lit a cigarette and passed it to his wife.</s>  (ST1.1.4.50)

13: <p><s>Tony Threadgold tente en sigarett og ga den til sin kone.</s>  (ST1T.1.4.52)
-------------------------------------------------------------------------------
14: <s>Then he lit one for himself.</s>  (ST1.1.4.51)

14: <s>Så tente han en til seg selv.</s>  (ST1T.1.4.53)
-------------------------------------------------------------------------------
15: <s>His good manners were often mocked at the Flowers Estate Working Men's
     Club.</s>  (ST1.1.4.52)

15: <s>De pene manerene hans ble ofte gjort narr av nede på
     arbeidsmannsklubben.</s>  (ST1T.1.4.54)
-------------------------------------------------------------------------------
```

Besides matching 1:1 correspondences, the program can also match 1:0, 0:1, 1:2, and 2:1 correspondences (see s-units 1+2,1 and 6,5+6). It cannot at present make 3:1 or 1:3 matches.

5 Alignment problems

The level of success in correct alignment is high. The error rate varies somewhat depending upon the text, but is generally as low as 2-3 per cent, i.e. only 2-3 sentences per hundred are incorrectly aligned. Where errors occur, the alignment tends to get on track again very quickly. Two recurrent problems are:

A. Short s-units

As the matching is heavily dependent upon word correspondences, there are problems particularly where the number of words is very low (the dotted lines in the examples below show how the s-units where grouped by the alignment program):

(13)

```
--------------------------------------------------------------------------------

10: <s>"Vær så inderlig snill!"</s></p>  (ST1T.1.2.11)
--------------------------------------------------------------------------------
10: <s>Please!"</s></p>  (ST1.1.2.11)
11: <p><s>"May we open the windows, Mr Barker?" asked the Queen.</s>  (ST1.1.2.12)

11: <p><s>"Kan vi få åpne vinduene, Mr Barker?" spurte dronningen.</s>  (ST1T.1.2.12)
--------------------------------------------------------------------------------
```

The alignment program has not been able to match the English one-word sentence *Please!* with the Norwegian *Vær så inderlig snill!*. The reason is that Norwegian has a multi-word expression, 'vær så snill', for 'please', and that the intensifying adverb 'inderlig' has been added in the translation.

Errors in examples of this kind led to the inclusion of punctuation in the calculation of alignment (cf. Section 4). Short s-units are very often questions or exclamations, where punctuation marks may be an important clue for alignment (but there is by no means a complete match between the use of question marks and exclamation marks in the original and the translated text).

B. Complex correspondences

Where the translation is very free, there will be few word correspondences, and errors in alignment easily occur:

(14)

```
--------------------------------------------------------------------------------
13: <s><title>The Times</title> had described Pa as a brilliant tactical
     officer with a courageous war record.</s>  (JA1.1.2.104)

14: <s><title>The Times</title> hadde omtalt Pa som en fremragende offiser
     med et talent for taktisk krigføring.</s>  (JA1T.1.2.109)
--------------------------------------------------------------------------------


 1: <s>Hans rulleblad fra krigen viste mot og utholdenhet.</s>  (JA1T.1.2.110)
--------------------------------------------------------------------------------
```

Here the Norwegian sentence 1 should also have been matched with the English sentence 13. The program has discarded this possibility even if it matched two word pairs: 'courageous'/'mot' and 'war'/'krigen' (the latter is pure chance since the translation of 'war record' is 'rulleblad'). The reason is that it has given more weight to sentence length in

characters, and, put together, the two Norwegian sentences contain far more characters than the single English one.[1]

In spite of such problems, the overall results are encouraging. The alignment program has been successfully tested with texts and anchor word lists for other language pairs (Swedish-English, Norwegian-French), but it may be less satisfactory for more distantly related languages.

Our aim is to develop the alignment program further. The anchor word lists will be expanded and refined. We will fine-tune the combination of measures taken into account (anchor words, proper names, punctuation, s-unit length). Eventually alignment may be based on texts which have been grammatically tagged. Chances will then increase of successfully handling more complex correspondences.

Because of the nature of translation, automatic alignment will never be perfect. However, judging by our experiences so far, it will be an invaluable aid in preparing parallel corpora for contrastive studies. Even where it fails, automatic alignment may be significant in highlighting cases where a less direct translation has been chosen, thus pinpointing differences between languages and perhaps even providing a window into the mind of the translator.

[1] The problem in matching these sentences is compounded by the fact that they occur on the border between the 15-unit groups used in the matching (cf. Section 4). Currently, the last s-unit in each group is transferred to the next group as the «window» moves forward in the text. In further work with the alignment program we will study the effect of increasing the overlap.

References

Johansson, Stig and Knut Hofland. 1994. Towards an English-Norwegian parallel corpus. In *Creating and using English language corpora*, ed. by Udo Fries, Gunnel Tottie and Peter Schneider. 25-37. Amsterdam: Rodopi.

Sperberg-McQueen, Michael C.M. & Lou Burnard (eds.). 1994. *Guidelines for electronic text encoding and interchange*. *TEI P3*. Chicago & Oxford: Association for Computers and the Humanities/Association for Computational Linguistics/ Association for Literary and Linguistic Computing.

Appendix 1: The header

A. Example of ENPC header

```
<header>
	<fileDesc>
		<titleStmt>
			<title>Cat's Eye: Extract in machine-readable form</title>
			<author>Margaret Atwood</author>
			<resp>
				<role>tagger</role>
				<name>JE</name>
			</resp>
		</titleStmt>
		<extent>11,000 words from beginning of text</extent>
		<publicationStmt></publicationStmt>
		<notesStmt></notesStmt>
		<sourceDesc>
			<biblStruct>
				<monogr>
					<author>Margaret Atwood</author>
					<resp>
						<role></role>
						<name></name>
					</resp>
					<title>Cat's Eye</title>
					<imprint>
						<publisher>McClelland and Stewart</publisher>
						<place>Toronto</place>
						<date>1988</date>
					</imprint>
				</monogr>
			</biblStruct>
		</sourceDesc>
	</fileDesc>
	<encodingDesc></encodingDesc>
	<profileDesc>
		<langUsage>CaE</langUsage>
		<textClass>FG</textClass>
		<textDesc></textDesc>
	</profileDesc>
	<revisionDesc></revisionDesc>
</header>
```

Note:
Some of the fields in the header contain little information. The <fileDesc> is always fully specified. The <encodingDesc> will indicate the level of markup provided in the text (as defined in the project manual). The <profileDesc> only specifies <langUsage> and <textClass>, using a simple taxonomy (here: CaE = Canadian English, FG = fiction, general). At a later stage of the project, more use will probably be made of the <profileDesc>. At the

present stage, no use is made of the <revisionDesc>, but this will no doubt be needed later as errors are discovered and corrected.

B. The ENPC vs. the TEI header

The only differences with respect to a header conforming to Sperberg-McQueen and Burnard (1994) is that the top tag in the latter is: <teiHeader type=text>. The TEI guidelines also provide a header for the whole corpus: <teiHeader type=corpus>. At present, there is no tag for the whole corpus in ENPC, and hence no header for the whole corpus.

Appendix 2: List of tags and other coding conventions

With the exception of <partext> and <blankline>, these tags are all recommended in the TEI guidelines (Sperberg-McQueen & Burnard 1994):

author	
back	back matter of text (e.g. index)
bibl	bibliographical information
biblStruct	structured bibliographical entry
blankline	one or more blank lines in source text
body	main body of text
corr	corrected text
date	
distinct	linguistically distinct material (e.g. dialect words)
div1 - div5	major divisions of text
encodingDesc	description of the coding of the electronic text
epigraph	
extent	extent of the electronic text
fileDesc	description of the electronic text
foreign	foreign words and expressions
front	front matter of text (e.g. title page)
gloss	gloss which is singled out in some way in the text (e.g. by quotation marks)
head	headings in text
header	header of electronic text
hi	highlighted text (e.g. italic)
imprint	
langUsage	description of language or dialect of text
list	
mentioned	words and expressions mentioned rather than used
monogr	
name	names of ships, buildings, etc
note	any kind of note (e.g. footnote)
notesStmt	notes on the electronic text
omit	omitted text and illustrations
p	paragraph
partext	top tag (encloses header plus text)
pb	page break
place	
poem	poem, song, etc
profileDesc	description of text (classification, situational parameters, etc.)
publicationStmt	publication of the electronic text
publisher	
q	quoted matter or direct speech (used particularly for quotes within quotes)
quote	matter quoted from extraneous source
resp	responsibility (e.g. for correction, omission)
revisionDesc	summary of changes with respect to previous versions of the text
role	role of person responsible for the text (e.g. editor)
s	s-unit (orthographic sentence)
sic	text exactly reproduced (though possibly erroneous)
sourceDesc	description of the source of the electronic text

term	term which is singled out in some way in the text (e.g. by italics or quotation marks)
text	the tag enclosing the actual text (as opposed to the header)
textClass	classification of the text
textDesc	description of the text in terms of situational parameters
title	title of book, play, etc.
titleStmt	title statement for the electronic text
what	change in revision of the electronic text

Some of the most important attributes are:

desc	description of omission
id	unique identifier
lang	language
link	pointer to identifier of equivalent unit
n	number of note or page
resp	responsibility for editorial intervention
type	type of text, division, etc.

With the exception of 'link' (equivalent to the TEI attribute 'corresp'), these attributes are recommended in the TEI guidelines (Sperberg-McQueen and Burnard 1994). See further the formal description in Appendix 3.

Entity references (identified by & ;) are used very sparingly, chiefly to mark special characters. See the entity list in Appendix 3.

Appendix 3: Document type definition (DTD)

```
<!DOCTYPE partext [
<!-- DTD for the English-Norwegian Parallel Corpus Project
     Last modified: 1994-09-12
     Jarle.Ebeling@iba.uio.no    -->

<!-- GENERAL ENTITIES -->
<!--      ENTITIES CONTENT  -->
<!ENTITY  shy      "-"  > <!-- Soft hyphen -->
<!ENTITY  dash     "--" > <!-- Dash -->
<!ENTITY  pron     ""   > <!-- Pronoun 'us' ('s) let 's&pron; -->
<!ENTITY  subord   ""   > <!-- Subordinator 'than'('n) more 'n&subord; -->
<!ENTITY  pound    ""   > <!-- Pound sign -->
<!ENTITY  dollar   ""   > <!-- Dollar sign -->
<!ENTITY  Ccaron   ""   > <!-- European characters -->
<!ENTITY  Ecaron   ""   >
<!ENTITY  ecaron   ""   >
<!ENTITY  rcaron   ""   >
<!ENTITY  Scaron   ""   >
<!ENTITY  Zcaron   ""   >
<!ENTITY  zcaron   ""   >

<!-- PARAMETER ENTITIES -->
<!--      ENTITIES CONTENT  -->
<!ENTITY  % las    "lang (la | gr | fr | de | it | es | no | en | sw |
                          nl | ar | ru | af | da | bo | inuit) #IMPLIED"  >
<!ENTITY  % las2   "lang (la | gr | fr | de | it | es | no | en | sw |
                          nl | ar | ru | af | da | bo | inuit) #REQUIRED" >
<!ENTITY  % style  "rend (italic | bold | normal)  #IMPLIED" >
<!ENTITY  % role   "resp (auth | ed | tr | tag) #REQUIRED"> <!-- Role -->
<!ENTITY  % divh   "head*, epigraph?" > <!-- div1-5 epigraph/head -->


<!--      ELEMENTS    MIN  CONTENT (EXCEPTIONS)                      -->
<!ELEMENT partext     - -  (header, text)                              >

<!-- HEADER -->
<!ELEMENT header      - -  (filedesc, encodingdesc, profiledesc,
                            revisiondesc)                             >

<!ELEMENT filedesc    - -  (titlestmt, extent, publicationstmt,
                            notesstmt, sourcedesc)                    >
<!ELEMENT titlestmt   - -  (title, author, resp)                       >
<!ELEMENT title       - -  (#PCDATA)                                   >
<!ELEMENT author      - -  (#PCDATA)                                   >
<!ELEMENT resp        - -  (role, name)                                >
<!ELEMENT role        - -  (#PCDATA)                                   >
<!ELEMENT name        - -  (#PCDATA)                                   >
<!ELEMENT extent      - -  (#PCDATA)                                   >
<!ELEMENT publicationstmt - -  (#PCDATA)                               >
<!ELEMENT notesstmt   - -  (note*)                                     >
<!ELEMENT sourcedesc  - -  (biblstruct)                                >
<!ELEMENT biblstruct  - -  (monogr)                                    >
<!ELEMENT monogr      - -  (author, resp, title, imprint)              >
<!ELEMENT imprint     - -  (publisher, place, date)                    >
<!ELEMENT publisher   - -  (#PCDATA)                                   >
<!ELEMENT place       - -  (#PCDATA)                                   >
<!ELEMENT date        - -  (#PCDATA)                                   >

<!ELEMENT encodingdesc - - (#PCDATA)                                   >

<!ELEMENT profiledesc - -  (langusage, textclass, textdesc)            >
<!ELEMENT langusage   - -  (#PCDATA)                                   >
<!ELEMENT textclass   - -  (#PCDATA)                                   >
<!ELEMENT textdesc    - -  (#PCDATA)                                   >

<!ELEMENT revisiondesc - - (date, name, what)?                         >
<!ELEMENT what         - - (#PCDATA)                                   >
```

```
<!-- TEXT -->
<!ELEMENT text         - - (front?, body, back?)
                             +(omit | pb | blankline)  >
<!ELEMENT front        - - (#PCDATA)                     >

<!ELEMENT body         - - (div1+)                       >
<!ELEMENT div1         - - (%divh, (div2 | p)+) +(note)  >
<!ELEMENT div2         - - (%divh, (div3 | p)+)          >
<!ELEMENT div3         - - (%divh, (div4 | p)+)          >
<!ELEMENT div4         - - (%divh, (div5 | p)+)          >
<!ELEMENT div5         - - (%divh, p+)                   >
<!ELEMENT epigraph     - - (quote, bibl)                 >
<!ELEMENT quote        - - (#PCDATA | s)+                >
<!ELEMENT bibl         - - (#PCDATA)                     >
<!ELEMENT head         - - (#PCDATA)      +(hi)          >
<!ELEMENT p            - - (#PCDATA | quote | s)+
                                  +(hi | mentioned | term |
                                  gloss | title | distinct |
                                  list | sic | corr |
                                  poem | foreign | name |
                                  quote | q)              >
<!ELEMENT s            - - (#PCDATA)                     >

<!ELEMENT back         - - (#PCDATA)                     >


<!--      FLOATING
          ELEMENTS    MIN  CONTENT (EXCEPTIONS)          -->
<!ELEMENT omit          - O EMPTY                        >
<!ELEMENT pb            - O EMPTY                        >
<!ELEMENT blankline     - O EMPTY                        >
<!ELEMENT corr         - - (#PCDATA)                     >
<!ELEMENT distinct     - - (#PCDATA)                     >
<!ELEMENT gloss        - - (#PCDATA)                     >
<!ELEMENT hi           - - (#PCDATA)                     >
<!ELEMENT list         - - (#PCDATA)                     >
<!ELEMENT mentioned    - - (#PCDATA)                     >
<!ELEMENT note         - - (#PCDATA)                     >
<!ELEMENT sic          - - (#PCDATA)                     >
<!ELEMENT term         - - (#PCDATA)                     >
<!ELEMENT poem         - - (#PCDATA)                     >
<!ELEMENT foreign      - - (#PCDATA)                     >
<!ELEMENT q            - - (#PCDATA)                     >

<!-- ATTRIBUTES -->
<!--      ELEMENTS    NAME     VALUE                DEFAULT        -->
<!ATTLIST partext     lang     (en | nor)           #REQUIRED
                      type     (novel | act | shortstory | textbook |
                                biography | documentary) #REQUIRED
                      orig     (yes | no)           #REQUIRED
                      id       CDATA                #IMPLIED       >
<!ATTLIST (div1 | div2 | div3 | div4 | div5)
                      type     (part | chapter | section) #REQUIRED
                      id       CDATA                #IMPLIED       >
<!ATTLIST head        %style
                      %las                                         >
<!ATTLIST p           %style
                      %las
                      id       CDATA                #IMPLIED       >
<!ATTLIST s           %style
                      %las
                      id       ID                   #IMPLIED
                      link     IDREFS               #IMPLIED       >
<!ATTLIST omit        desc     CDATA                #IMPLIED
                      reason   CDATA                #IMPLIED
                      extent   CDATA                #IMPLIED
                      resp     (auth | ed | tr | tag) #IMPLIED     >
<!ATTLIST pb          n        NUMBER               #REQUIRED      >
<!ATTLIST corr        sic      CDATA                #REQUIRED
                      resp     CDATA                #REQUIRED      >
```

```
<!ATTLIST distinct    type     (nonstand)          nonstand          >
<!ATTLIST (hi | term) %style                                        >
<!ATTLIST mentioned   %style
                      %las                                          >
<!ATTLIST note        n        NUMBER              #IMPLIED
                      %role
                      place    CDATA               #IMPLIED          >
<!ATTLIST sic         corr     CDATA               #IMPLIED
                      resp     CDATA               #REQUIRED         >
<!ATTLIST title       %las                                          >
<!ATTLIST foreign     %las2                                         >
]>
```