Paper presented at the ACH/ALLC '95 Conference, Santa Barbara 11-15/7-1995

Figures and tables at the end of the article.

A Program for Aligning English and Norwegian Sentences

Knut Hofland
Norwegian Computing Centre for the Humanities (NCCH)
Harald Hårfagres gt. 31
N-5007 Bergen
Norway

Telephone +47 5521 2954/5/6
Fax +47 5532 2656

E-mail: Knut.Hofland@hd.uib.no

URL for the project: http://www.hd.uib.no/enpc.html

Knut Hofland is senior computing officer and systems manager at NCCH.

Abstract

A program for aligning sentences in the English-Norwegian Parallel Corpus is presented. The program makes use of a simple bilingual lexicon (anchor words), but in addition uses information like proper nouns, special characters and tags, cognates and sentence length in characters. The program has been tested on a part of the corpus (51 text pairs of totally 1.3 million words) with an error rate of 1.9 per cent. The program is an invaluable aid in preparing the parallel corpus.

1. Introduction

The English-Norwegian Parallel Corpus (ENPC) project was started in the beginning of 1994, after a pilot study had been carried out in 1993, and it is expected to end in early 1996. The parallel corpus is meant to be a research tool for linguists and students interested in contrastive linguistics. It will consist of two sub-corpora: a translation corpus and a comparative corpus. The translation corpus will contain English and Norwegian originals and their translations (including both English-to-Norwegian and Norwegian-to-English translations), while the comparative corpus will consist of original texts, in both languages which are matched with respect to text type.

The corpus will only contain written material, but both fiction and non-fictional texts will be included. To make it possible to include as many different writers and translators as possible, text extracts of 10,000 - 15,000 words will be selected and not complete books. Each text extract will start at the beginning of the book, and if possible, end at a chapter boundary. We hope to be able to compile as many as 160 text extracts during the first year (40 English and 40 Norwegian texts, with their translations), which will give us a corpus of approximately 2 million words. By July 1995 150 texts had been scanned and 51 pairs had

been aligned. As regards the structure of the corpus and the projected uses of the material, see (Johansson and Hofland, 1994) and (Johansson and Ebeling, forthcoming).

2. Processing and coding of the texts

Before alignment can be attempted, a text must be available in electronic form and contain some basic markup. The coding of the texts is in broad agreement with the recommendations of the Text Encoding Initiative (TEI) (cf. Sperberg-McQueen & Burnard 1994).

All texts are scanned using 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 these are also inserted automatically. However, the program is not foolproof, and it 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, we also insert markers of page breaks manually at this point in the process.

Contractions in the texts are split and orthographic sentences are recognised and tagged by using two small programs. An empty TEI header is placed at the beginning of the file and is filled out during the proofreading.

The texts are then proofread and additional markup is added. As regards the coding of the texts, see (Johansson, Ebeling and Hofland forthcoming). An example of a text extract before and after markup is given in fig 1.

After the markup has been completed and checked, each pair of texts is run through the alignment program. The alignment is checked manually, sometimes the markup for sentence division or other input data is changed, and the program is then run once more.

3. Alignment

By the time the ENPC pilot project started in spring 1993, a good deal of work had already been done on alignment, (Brown et al.,1991), (Gale and Church,1991) and (Simard et. al, 1992). Since then several other approaches have been used, both for sentence, word and character alignment, (Kay and Röscheisen 1993), (Church 1993), (McEnery et. al 1994), (Papageorgiou 1994).

All of these methods are mainly based on statistics. In our project we wanted to focus on lanuage specific information, so-called anchor words or a simple bilingual lexicon. The considerations in putting together the list were: 1) the words had to be reasonably frequent; 2) the words had to have fairly straightforward equivalents in the two languages. Some items were included even though the words were not very frequent, provided that a good correspondence could be established, e.g. *photo*/foto**, (the asterisk indicates truncation). Conversely, some frequent words were included even though the correspondence pattern was not straightforward. For example, *open* /åpen*, åpn** caters for a number of different forms in the two languages: *open, opens, opened, openly, openness,* etc. vs. *åpen, åpner, åpnet, åpent, åpenhet,* etc. The list was partially based on intuition, and partially on (manual) matching of original and translated texts. An extract of the anchor list is given below:

all / all*, alt, hel*
almost / nesten
alone, single / alene
already / allerede

also / også
always / alltid, bestandig
among / blant
and / og
and so on / osv, og så videre

A slash separates English and Norwegian forms, a comma separates alternatives in one of the languages. Sequences of up to three words can be given. The anchor list has currently about 900 lines. The list is read into the program and stored as two text trees, one for each language, and with one character in each node. A node in the tree may contain a list of anchor line numbers and also a flag if the node is the end of a truncated form.

In the first version of the program a fixed number of paragraphs were read from the two texts and the sentences were aligned within paragraph boundaries. But it turned out that very few pairs of texts had an equal number of paragraphs (only 6 out of 51 tested pairs). In the current version, therefore, a window of 15 sentences is moved through the two texts to be aligned, with an overlap of five sentences. An overview of the input and output files is given in fig. 2.

Words in each of the 15 sentences are first looked up in the anchor text trees. Each part of a hyphenated word is also looked up. The words are then stored in a list. The anchor numbers found are kept in another list.

In addition to the anchor numbers, other information from the sentences is recorded. Words with an initial capital are kept in a list as candidates for proper nouns. Some characters like colon, question mark, exclamation mark and per cent are kept in a character list. The program also makes a list of selected tags in the sentence like tags for start or end of division, heading or paragraph, and tags for boldface or italics. The number of characters in the sentence (excluding markup) is counted. An example of these lists are found in fig. 3.

When the 15 sentences have been read in the two languages, each of the 15 English sentences is checked against each of the 15 Norwegian sentences and a 15 times 15 matrix is built. If the same anchor number is found in the list for an English sentence *i* and in the list for a Norwegian sentence *j,* then the value in cell *i,j* in the matrix is increased by one. In the same way, the lists for special characters and tags are checked. The lists of capitalized words are also checked, either for equality between two words or to see if a word in one of the languages is the first part of a word in the other language. This is because in Norwegian (and German) a proper noun can be the first part of a compound word. Proper nouns which are different in the two languages like *Norway / Norge*, cannot be found in this way and are entered in the anchor list. All the combinations of the three sentences in fig. 3 are given in tab. 1 and the resulting 3x3 matrix in table 2.

A recent addition to the program is an automatic extraction of cognates in the texts, based on principles used in the CRATER project, (McEnery et al. 1994). Truncation and Dice's similarity coefficient are the methods used to find cognates. If the first *n* (*n* is read as a parameter to the program) characters are equal for a word in an English and a Norwegian sentence, the two words are assumed to be cognates. For Norwegian/English a value of *n* =6 or 7 gives good results. Dice's similarity coefficient is the number of matching bigrams in the two words divided by the mean of the number of bigrams for the two words (2a /(b+c), where *a* is the matching number of bigrams, *b* and *c* are the number of bigrams in the two words). For Norwegian and English a value of more than 0.7 or 0.8 gives reasonable results. For other languages the acceptable value for the coefficient can be less. This parameter is also read by the program. For a novel, the cognates, however, only account for 0.5 per cent of the values in the matrix for an English-Norwegian or French-Norwegian pair, and for 3-4 per cent for an

English-French pair of texts. For non-fiction texts or technical texts this figure will be much higher. Examples of cognates found in an English-Norwegian text are given in table 1.

A sample matrix is given in table 4. Rows correspond to the English original, while columns correspond to the Norwegian translation. The figures above and beside the rows and columns give the sentence number in the window and the number of characters in each sentence.

The next step is to find a match between the English and Norwegian sentences so that the sum of the values from the matrix is maximized. For a given position *i,j* in the matrix the following combinations are tested:

combine E_i with N_j (1:1)
combine E_i with N_{j+1} (skip a Norwegian sentence, 1:0)
combine E_{i+1} with N_j (skip an English sentence, 0:1)
combine E_i and E_{i+1} with N_j (2:1)
combine E_i with N_j and N_{j+1} (1:2)

For each of these combinations an entry is placed on a queue list with the following information:

the next position to go to
the sum so far
the combinations so far

Combinations which have been tried before are not put on the queue list. The queue is served until it is empty. The current sum is tested against the current maximum sum and the combination giving a new maximum value is remembered. When testing a combination, the value from the matrix is adjusted according to the correspondence between the length of the English and the Norwegian sentences. The correspondence index is measured as the absolute value of the difference in length divided by the mean of the two lengths. A good correspondence (low index) increases the value in the matrix and a bad correspondence (high index) decreases the value. When two sentences in one language are combined with one in the other language, the combination is discarded if the correspondence index is too high. The different thresholds for the index are read by the program. An example of the first entries on the queue list for the matrix in table 4 is given in fig. 4.

The program gives output in several forms. The form in figure 5 is used for proofreading and can also be used to detect new words which could be added to the anchor word list.

The windows are moved in the texts with an overlap of at least five sentences. The two windows will not always be moved forward the same number of sentences. This is due to the translator's tendency to skip and join sentences.

While aligning the sentences, the headings, paragraphs and sentences are given a unique identifier based on the identifier given in the last division id. Another output from the program is the texts with these identifiers and an attribute giving the alignment for headings and sentences. A sample of these files is given in figure 6. When there is no correspondence, the *corresp* attribute links to the nearest previous paragraph.

The program will still run even if a sequence of 5-7 sentences is missing in one of the texts. But in a few texts too many sentences are not translated and the rest of the output will be out of phase. It is then possible manually to give a skip attribute to a sequence of sentences.

4. Results

The program has so far been run on 51 pairs of texts (a total of 89,000 sentences and 1.3 million words) and the results have been checked manually. The error rate is 1.9 per cent. One problematic text has an error rate of 19.5 per cent (it is a very free translation) and has consequently been excluded from the following statistics. Except for this text the error rate varies from 0 per cent to 7.2 per cent. The distribution of errors is shown in figure 7.

In table 5 the results have been broken down into languages and text types. The total number of sentences in each group is also given.

Table 6 gives statistics about how many sentences in the other language a sentence is linked to (the contents of the *corresp* attribute).

Table 7 gives an overview of the kind of errors the program makes when aligning sentences. From this table we see that the most common error is where one sentence has been linked to a wrong sentence. The next most common errors are where a zero or double link should have been a single link. Links involving three or more sentences only account for about ten per cent of the errors. So even if the program was improved to handle this kind of alignment, the error rate would only drop from 1.9 to 1.7-1.8 per cent (the net result would be less than 10 per cent because of the introduction of new errors).

5. Usage of the program on other language pairs.

A French-Norwegian parallel corpus project has been started based on the ENPC. Currently 40 texts have been scanned and a few have also been aligned. A French-Norwegian version of the anchor list has been made. The results are so far promising. The two anchor lists have also been combined to make an English-French list. Work is in progress on a German, Spanish and Italian version of the English-Norwegian list. The program has been tested on texts from the European Corpus Initiative Multi lingual CD-ROM. The English-Swedish parallel corpus project in Lund will also make use of the program and has made an English-Swedish anchor list.

A word frequency program has been written to make a rank list with information on the number of sentences in which a word is used and whether the word is used in several sentences consecutively. In this list it is also marked (with a #) if a word is not found in the anchor list. An example of the first part of such a list for a small Polish text is given in table 8. By translating 100 of these words into English, a success rate of 96.6 per cent was achieved when aligning an English-Polish text pair (a total of 267 sentences).

6. Browsing tool

At the University of Oslo's department of British and American studies, a browsing tool has been written by Jarle Ebeling for the ENPC. The aligned and proofread texts are indexed, making it possible to search the text database for words in one of the languages and retrieve the sentences together with their translations in the other language. Words in the two languages can be combined with the *and* or the *not* operator so that only pairs of sentences with a specific word in the first language together with (or not together with) another word in the second language are found. Words can also be truncated and a distance (in number of words) between particular words can be given.

7. Conclusion

Because of the nature of translation, automatic alignment will never be perfect. However, judging by our experience 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.

Figures and tables

'Because — because I *know* we're never going to get to the bottom of that case. And there's nothing that irritates a policeman more than to feel that a case has gone unsolved.'

16

'But...'

'And because,' he went on, ignoring me, 'because it was so easy to see the *pattern* in it all.'

```
<p><s>"Because &dash; because I <hi rend=italic>know</hi> we 're never going
to get to the bottom of that case.</s> <s>And there 's nothing that irritates
a policeman more than to feel that a case has gone unsolved."</s></p>
<pb n=17><p><s>"But..."</s></p>
<p><s>"And because," he went on, ignoring me, "because it was so easy to see
the <hi rend=italic>pattern</hi> in it all."</s>
```

Fig. 1. Original Text and Text with Markup

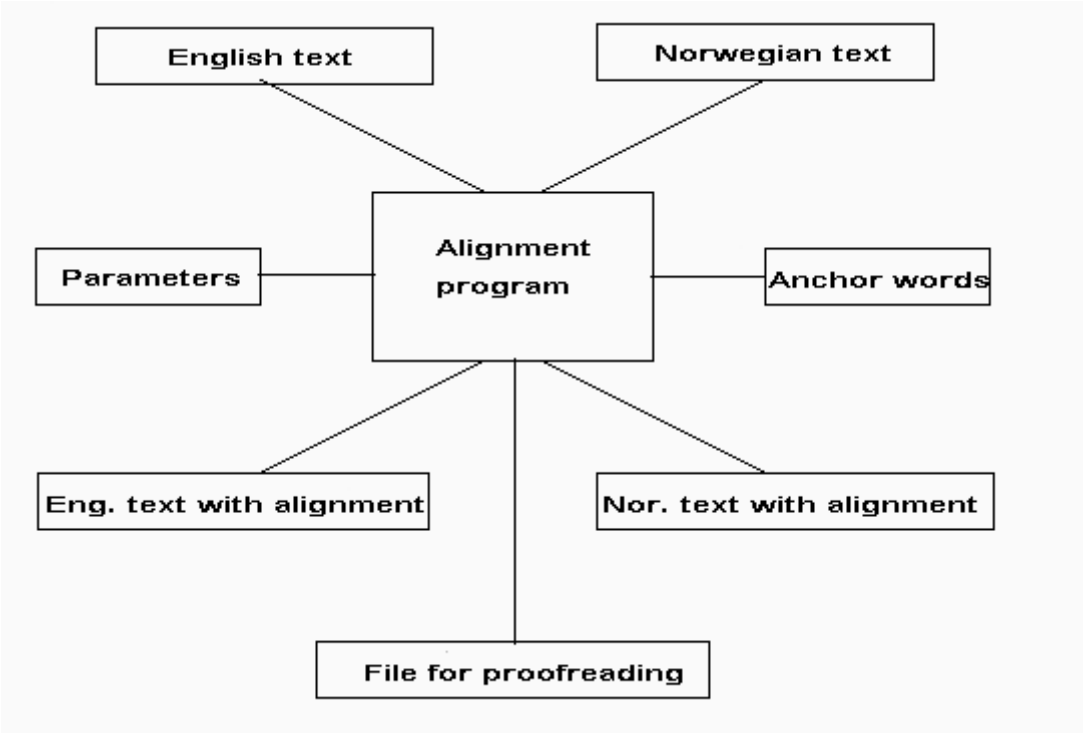

Fig. 2. An Overview of the Alignment Program

E1: *Oh, the wages of sin!*
Anchor numbers: 545 (oh), 839 (wages), sin (693)
Special characters: !
Capitalized words: Oh

N1: *Å, syndens lønn!*
Anchor numbers: 545 (å), 693 (syndens), 839 (lønn)
Special characters: !
Capitalized words: Å

E2: *Natalie Harris sinned, and her husband Harry left for work one fine morning and did n't come back.*
Anchor numbers: 693 (sinned), 91 (and), 383 (her), 406 (husband), 454 (left), 880 (work), 502 (morning), 536 (n't), 196 (come), 111 (back)
Special characters:
Capitalized words: Natalie, Harris, Harry

N2: *Natalie Harris syndet og en vakker dag dro Harry, mannen hennes på jobb og kom ikke tilbake.*
Anchor numbers: 693 (syndet), 91 (og), 226 (dag), 472 (mannen), 406 (mannen), 383 (hennes), 880 (jobb), 159 (kom), 536 (ikke), 532 (ikke), 530 (ikke), 111 (tilbake)
Special characters:
Capitalized words: Natalie, Harris, Harry

E3: *The morning was fine only temporarily.*
Anchor numbers: 502 (morning), 845 (was), 548 (only)
Special characters:
Capitalized words: The

N3: *Morgenen var fin bare en stund.*
Anchor numbers: 502 (morgenen), 845 (var), 346 (fin), 548 (bare), 860 (stund)
Special characters:
Capitalized words: Morgenen

Fig. 3. Lists with Information from the sentences

English sentence	Norwegian sentence	Common items from lists	Total
1	**1**	545, 839, 633, !	4
1	**2**	693	1
1	**3**		0
2	**1**	693	1
2	**2**	693, 91, 383, 406, 880, 91, 536, 111. Natalie, Harris, Harry	11
2	**3**	502	1
3	**1**		0
3	**2**		0
3	**3**	502, 845, 548	3

Table 1. Common Anchor Numbers, Characters and Proper Nouns

	1	**2**	**3**
1	4	1	0
2	1	11	1
3	0	0	3

Table 2. Resulting Matrix

English	Norwegian	Dice coefficient	Truncation length
allegory	allegori	0.86	7
appetite	appetitt	0.86	7
argument	argumentet	0.88	8
arrogant	arrogante	0.93	8
assortment	sortiment	0.71	0
banality	banaliteten	0.71	7
barrier	barrieren	0.86	7
candidate	kandidat	0.80	0
capital	kapitalen	0.71	0

Table 3. Example of Cognates in an English-Norwegian Text Pair

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

Table 4: Sample Matrix

From Queue list	To queue list	New queue list
	(1 1 0)	(1 1 0)
(1 1 0)	(3 2 11 1+2,1) (2 1 0 1,0) (1 2 0 0,1) (2 2 9 1,1)	(3 2 11 1+2,1) (2 1 0 1,0) (1 2 0 0,1)(2 2 9 1,1)
(2 2 9 1,1)	(4 3 15 1,1 2+3,2) (2 3 9 1,1 0,2) (3 3 10 1,1 2,2)	(3 2 11 1+2,1) (2 1 0 1,0) (1 2 0 0,1)(4 3 15 1,1 2+3,2) (2 3 9 1,1 0,2)(3 3 10 1,1 2,2)
(3 3 10 1,1 2,2)	(5 4 21 1,1 2,2 3+4,3) (3 4 10 1,1 2,2 0,3) (4 4 12 1,1 2,2 3,3)	(3 2 11 1+2,1) (2 1 0 1,0) (1 2 0 0,1)(4 3 15 1,1 2+3,2) (2 3 9 1,1 0,2)(5 4 21 1,1 2,2 3+4,3) (3 4 10 1,1 2,2 0,3) (4 4 12 1,1 2,2 3,3)

Fig. 4. Queue List

```
 Result= 1+2,1 3,2 4,3 5,4 6,5+6 7,7 8,8 9,9 10,10

--------------------------------------------------------------------------------
 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)
```

Fig. 5. File Format for Proofreading

Original text file

<div1 type=part id=ST1.1>
<head id=ST1.1.h1 corresp=ST1T.1.h1>April</head>
<omit desc=drawing>
<div2 type=chapter id=ST1.1.1>
<head id=ST1.1.1.h1 corresp=ST1T.1.1.h1>1 Uneasy Lies the Head</head>
<pb n=13>

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

.....

<s id=ST1.1.1.s20 corresp='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>

.....

<s id=ST1.1.2.s157 corresp=ST1T.1.2.s170>"She 's been well paid for it," snapped Jack.</s>
<s id=ST1.1.2.s158 corresp=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 corresp=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>

Translation text file

<div1 type=part id=ST1T.1>
<head id=ST1T.1.h1 corresp=ST1.1.h1>APRIL</head>
<div2 type=chapter id=ST1T.1.1>
<head id=ST1T.1.1.h1 corresp=ST1.1.1.h1>1. UTRYGT HVILER HODET</head>
<pb n=11>

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

.....

<s id=ST1T.1.1.s20 corresp=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 corresp=ST1.1.1.s20>Harris mistet all interesse og hoppet opp i sengen igjen.</s>

.....

<s id=ST1T.1.2.s170 corresp=ST1.1.2.s157>"Hun har fått godt betalt," bet Jack ham av.</s>
<s id=ST1T.1.2.s171 corresp=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>

<s id=ST1T.1.2.s172 corresp=ST1.1.2.s160>Prins Charles gned seg over nesevingen med en manikyrert pekefinger.</s>

Fig. 6. Text Files with Alignment Information

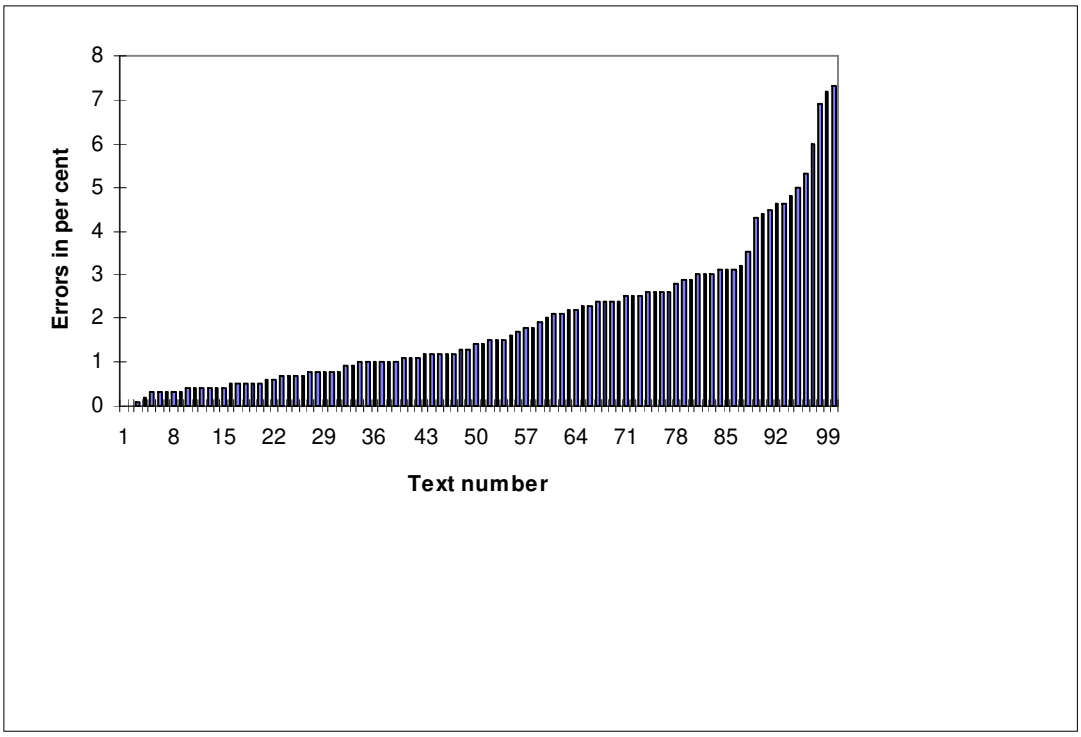

Fig. 7. Distribution of Errors

Original language	Fiction	Non-fiction	Total
Norwegian	1.53/40120	2.36/20558	1.81/60678
English	2.15/31651	5.13/1539 (*)	2.29/33190
Total			1.98/93868

(*) There is only one text in this group at the moment; consequently this figure is not representative.

Table 5. Error Rates in per cent with Number of Sentences in each Group

Number of sentences in one link	Number of links	Percentage
0	551	0.56
1	94304	96.40
2	2791	2.85
3	167	0.17
4 or more	17	0.01
Total	97830	100.00

Table 6. Number of Different Kind of Links between Sentences

Number of links suggested by program	Correct number of links						
	0	**1**	**2**	**3**	**4**	**5**	**6**
0	34	544	12	5	0	1	0
1	179	629	219	13	4	1	1
2	2	459	37	148	8	2	0

Table 7. Type of Errors

Number of		Percentage of words in		Word
words	**sentences**	**2 consecutive sentences**	**3 consecutive sentences**	
120	94	28	13	w#
102	88	19	3	sie#
100	87	30	8	nie#
89	76	21	1	i#
80	69	32	9	na#
77	67	24	10	ze#
74	65	23	6	do#
68	47	32	15	—
58	53	11	2	z#
44	40	25	5	to#
33	31	6	0	a#
28	25	28	8	krwi#
27	24	13	0	po#

Table 8. Rank List with Statistics on Word Distribution