An Evaluation of Anaphor Generation in Chinese

Ching-Long Yeh *and Chris Mellish
Department of Artificial Intelligence
University of Edinburgh
80 South Bridge
Edinburgh EH1 1HN, UK
{chingyeh, chrism}@aisb.ed.ac.uk

Subject: Evaluation
Type of submission: Paper

Abstract

In this paper, we present an evaluation of anaphors generated by a Chinese natural language generation system. In the evaluation work, the anaphors in five test texts generated by three test systems employing generation rules with different complexities were compared with the ones in the same texts created by twelve native speakers of Chinese. We took the average number of anaphors matching between the machine and human texts as a measure of the quality of anaphors generated by the test systems. The results show that the one we have chosen and which has the most complex rule is better than the other two.

1 Introduction

We have established several rules for the generation of anaphors in Chinese, including the decision between zero, pronominal and nominal anaphors, and a rule for the choice of a description if a nominal anaphor is decided upon. These rules were implemented in our Chinese natural language generation system and a number of texts for describing entities in a national park were generated (Yeh 95). As shown in our previous studies (Yeh & Mellish 94; Yeh & Mellish 95; Yeh 95), these rules were obtained from empirical studies. The experimental results show that the anaphors generated by using these rules largely match the ones in the test texts we used , assuming the same semantic structures and contextual information. This shows the performance of the rules. However, in this previous work the same data served as both training and test data. Furthermore, the assumed contextual information, for example, discourse structures, may be difficult to implement in a real system. Thus, the performance of a real anaphor generation algorithm based on the previous rules may be different to the experimental results. In this paper, we attempt a post-evaluation by asking some native speakers of Chinese to judge the result of the anaphors generated by our system.

2 Previous Work and Our Approach

Though the field of natural language generation has progressed towards composing complex texts, the evaluation of natural language generation systems has remained at the discussion stage (Maybury 90; Meteer & McDonald 91). Two broad methods have been identified for evaluating natural language generation systems: *glass box* and *black box* evaluation (Meteer & McDonald 91). The *glass box* method is concerned with examining the internal working of individual components in a system, while the latter looks at the behaviour of the input and output to the generation systems. The difficulty of the *glass box* method is the lack of a clear division between components in generation systems. The problem of the *black box* method, on the one hand, is determining the appropriate input for generation. On the other hand, it is very difficult to be objective in evaluating the output text.

In this paper, we aim to investigate the quality of anaphors generated by the referring expression component in our Chinese natural language generation system. The referring expression component lies between the text planner and the linguistic realisation component in the system, as shown in Fig. 1. On accepting an input goal from the user, the system invokes the text planner which uses the operators in the plan library to build up a plan which is a hierarchical discourse structure to satisfy the input goal. After the text planning is finished, the decision of anaphoric forms and descriptions is then made by traversing the plan tree. As is discussed in (Yeh & Mellish 95; Yeh 95), the algorithm of the referring expression component first determines an appropriate form for an anaphor to be generated. This includes zero anaphors which are omissions of noun phrases in surface sentences, pronominal anaphors, `ta` (s/he, it), like *s/he* and *it* in English,

*To whom correspondence should be addressed. Also with the Department of Computer Science and Engineering at Tatung Institute of Technology, 40, Section 3, Chungshang Road, 10457, Taipei, Taiwan.

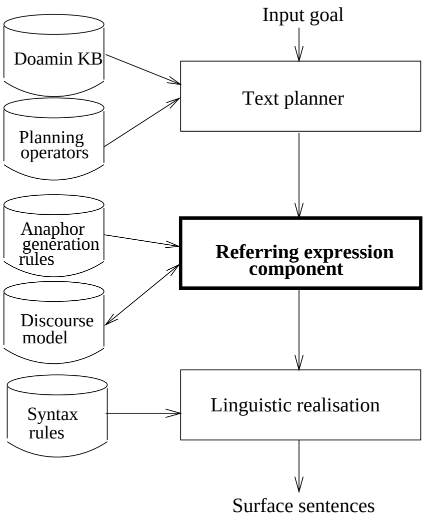

Figure 1: Referring expression component in the Chinese natural language system.

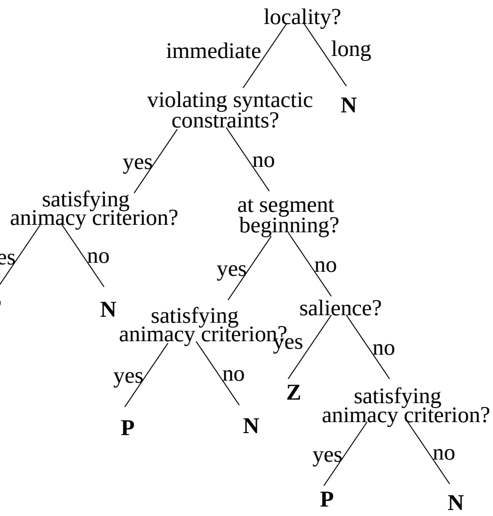

Figure 2: A Chinese anaphor generation rule.

and nominal anaphors like fang zuozi (square table). It then chooses a description, for example, either fang zuozi (square table), or zuozi (table) if a nominal form is decided upon.

Suppose that the referring expression components we wish to compare all adopt the above basic algorithm. Then the essential characteristic to distinguish them from each other becomes the rules used in the components and how these rules are implemented. If all of these referring expression components are embedded in the same Chinese natural language generation system, as in Fig. 1, for example, then, given an input to the system, anaphors in the resulting texts can be characterised by the rules used in the referring expression component and their implementation.

By adopting this approach, we need not worry about the problems of either of the evaluation methods stated above, except the objective evaluation of output text. Since there is no machine that can read the generated texts and give an impartial judgement about them, we rely on the opinions of human readers who are native speakers of Chinese to investigate the quality of the generated anaphors. This is an easier task than assessing the quality of whole texts. To compensate for possible bias among the individual readers, we sent the output texts to a group of readers for viewing and took the average of their outcomes as the measurement.

In brief, each object system in our evaluation work is thought of as having the same individual components, including control and knowledge bases, except that the anaphor generation rules used in the referring expression components are different to each other. In the existing literature, we cannot find other work on the generation of Chinese referring expressions, which means that we have no real working systems to compare with. In practice, we employ our Chinese natural language generation system described in (Yeh 95) as the backbone of the evaluation work because it is easy for us to control and maintain. What we have to do for each generation system is simply to insert the corresponding generation rule.

3 Systems to Compare and the Test Task

Having described the framework of the evaluation, in this section, we give details about the object systems to be compared in the evaluation work and the tasks to be performed in the evaluation work.

3.1 Systems to compare

The anaphor generation rule we obtained in our previous studies (Yeh & Mellish 94; Yeh & Mellish 95; Yeh 95) is shown in Fig. 2, where the internal nodes represent constraints and the terminal nodes are the decisions of using a zero (Z), pronominal (P), or nominal (N) form. The *locality* constraint checks whether the anaphor in question occurs either in the immediately previous utterance or at a long distance. The second constraint determines whether an anaphor occurs in a position violating syntactic constraints on zero anaphors. We adopted the concept of discourse segment structure in (Grosz & Sidner 86) to build up the constraint *at segment beginning*. It checks whether an anaphor is at the beginning of a discourse segment. The *salience* constraint says that both the positions of an anaphor and its antecedent are the topics of their respective utterances. The *animacy* constraint checks whether the anaphor in question is animate. Then the following rule is used if a nominal form is decided on.

> If a nominal anaphor, n, is at the beginning of a "sentence" [1], or is the first

[1] A "sentence" is in general a meaning-complete

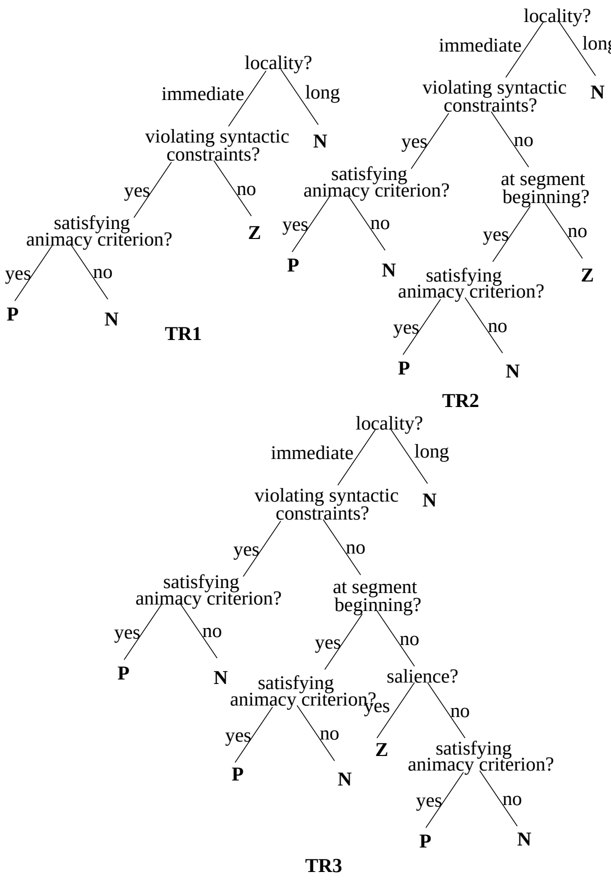

Figure 3: Rules used in the comparison systems.

> mention of the referent in a "sentence," then a full description is preferred; otherwise, if n is within a "sentence" or has been mentioned previously in the same "sentence" without distracting elements, then a reduced description is preferred; otherwise a full description is preferred.

The constraints in the anaphor generation rule were established by consulting relevant linguistic studies (Yeh & Mellish 94; Yeh & Mellish 95; Yeh 95). Consequently, subsets of constraints in the above rule can be thought of as possible rules, if not complete, for the generation of anaphors in Chinese. As described previously, the systems to compare in this evaluation work are assumed to share the same individual components, except the anaphor generation rules. In this paper, we equipped each system with such a possible anaphor generation rule.

We chose three rules, termed TR*1* TR*2* and TR*3*, with different complexities among the possible candidates as the targets of the test. The rules are shown in Fig. 3. The first one uses locality, syntactic constraints and animacy. The second and the third rules have one additional constraint, namely, discourse segment boundaries and salience, respectively, added to their predecessors. In the following, we use the above rule names to represent the systems.

unit (Liu 84). A sentential mark is used to indicate the full stop of a "sentence"; a comma within a "sentence" indicates a temporary stop.

3.2 The test task

The task can be divided into an annotation and a comparison stage. Each of twelve native speakers of Chinese was given a number of test sheets to finish. On each sheet is a text generated by our generation system. Each anaphor position in a generated text was left empty and all candidate forms of the anaphor, including zero, pronominal, and full, or reduced descriptions were put under the empty space. The task for a speaker to perform was to annotate which form he or she preferred for each anaphor position on the sheets.

We selected five texts generated by our system for the test. The numbers of clauses in the texts are 5, 12, 12, 21 and 34; the numbers of anaphors in the texts are 4, 11, 11, 20 and 34. For convenience, we summarise the occurrence of anaphors in the test texts in a graphical form as shown in Fig. 4. In the figure, each box represents a clause and at the right end is the accompanying punctuation mark. Each box is divided into three parts which represent the topic, the subject and the direct object positions of the clause. The numbers in a box, except for the first occurrences in the text, are the indices of anaphors in the corresponding clauses. Initial references are indicated by bold italics. For example, in Text 2, the numbers 1, 2, 3 and 4 occurring in the first, 5*th*, 8*th* and 10*th* clauses, respectively, are initial references; others are anaphors.

After the annotations were collected, we carried out comparisons between the speakers' results and the generated texts to investigate the performance of the test rules. In each comparison, we noted down the number of matches between the computer generated text and the human result. In the following, we use C_{ij} to denote the text indexed j generated by the system equipped with Rule TRi, where i is 1 to 3 and j is 1 to 5; and H_{kl} to denote the resulting text indexed l of speaker k, where k is 1 to 12 and l is 1 to 5. The comparison work is summarised procedurally as below.

```
for each rule TRi
    for each speaker j
        for each text k
            compare C_ik with H_jk and
            note down the number of matches
            of anaphors between them
```

4 Results

In this section, we investigate the result of the comparisons made in the last section. The comparison result is shown in Table 1. The average matching rates for all test texts are 72, 74 and 76%.

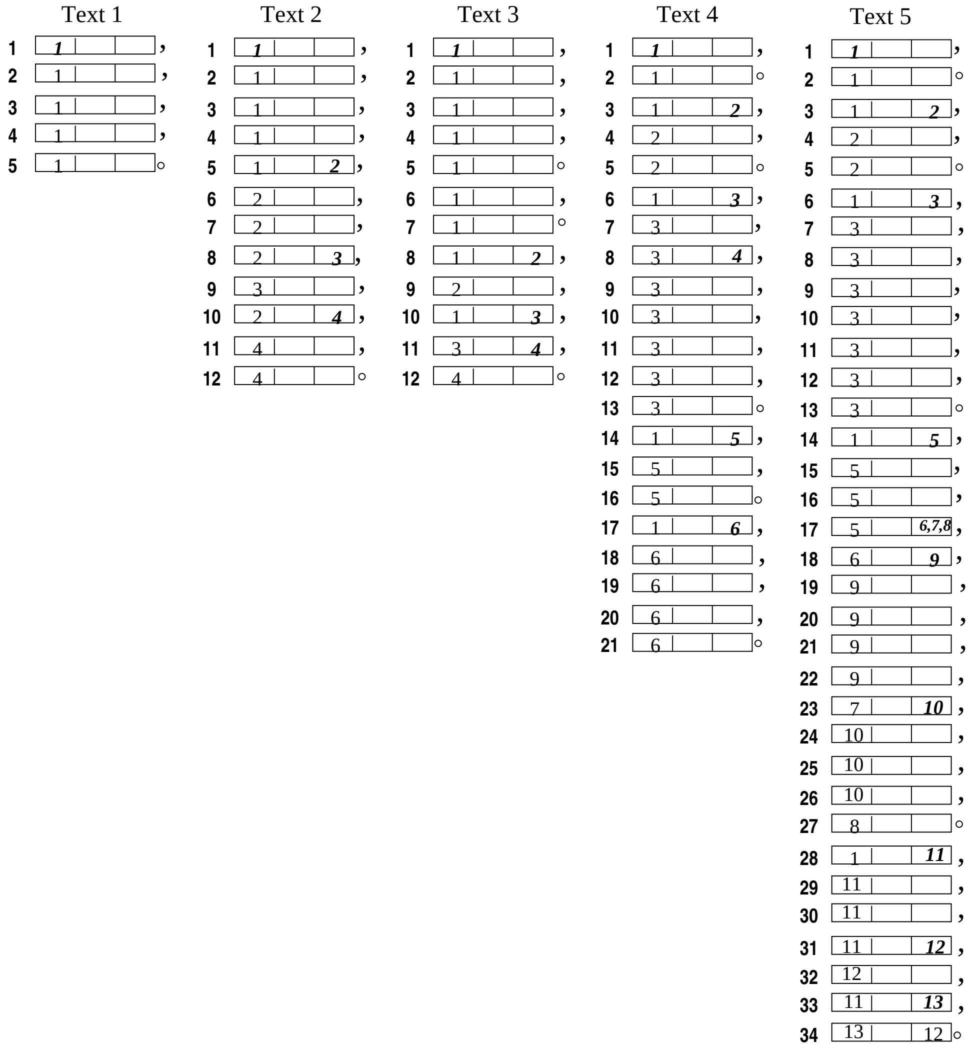

Figure 4: Occurrence of anaphors in the test texts.

Table 1: Match between the results of test systems and native speakers.

System	Speaker	Text 1	Text 2	Text 3	Text 4	Text 5
TR*1*	1	4	10	9	16	27
	2	4	8	6	16	24
	3	4	7	5	15	24
	4	4	8	5	13	23
	5	3	7	8	14	23
	6	4	8	7	15	28
	7	4	7	6	16	25
	8	4	10	9	17	32
	9	2	6	7	9	14
	10	4	8	9	14	23
	11	2	5	6	10	20
	12	4	9	5	13	23
	Average	3.6	7.8	6.8	14	23.8
	Total anaphors	4	11	11	21	34
		90%	70%	62%	70%	70%
TR*2*	1	4	10	8	17	26
	2	4	8	6	17	25
	3	4	7	5	16	23
	4	4	8	5	14	24
	5	3	7	9	15	24
	6	4	8	7	16	29
	7	4	7	7	17	26
	8	4	10	9	18	33
	9	2	6	8	9	14
	10	4	8	10	15	24
	11	2	5	7	11	19
	12	4	9	6	14	24
	Average	3.6	7.8	7.3	14.9	24.3
	Total anaphors	4	11	11	21	34
		90%	70%	66%	75%	71%
TR*3*	1	4	7	5	13	18
	2	4	11	9	16	25
	3	4	10	8	19	29
	4	4	9	4	14	24
	5	3	9	9	16	26
	6	4	11	6	14	24
	7	4	8	10	13	23
	8	4	7	6	14	25
	9	2	7	7	12	20
	10	4	9	7	16	23
	11	2	6	6	12	21
	12	4	10	9	16	30
	Average	3.6	8.7	7.1	14.6	24
	Total anaphors	4	11	11	21	34
		90%	79%	65%	73%	71%

Table 2: Agreement of annotations among speakers.

Speaker	Text 1	Text 2	Text 3	Text 4	Text 5
1	3.9	8	7.5	14.3	24
2	3.9	9.5	7.8	16.1	26.5
3	3.9	9.1	7.8	15.8	26.3
4	3.9	8.9	6.6	15.4	23.9
5	3.3	8.5	8.3	15.2	25.4
6	3.9	9.5	8.3	14.1	26.5
7	3.9	8.3	7.1	15	26.2
8	3.9	8.1	7.9	15.8	26.4
9	2.4	6.8	7	12.1	20.5
10	3.9	8.6	8.1	14.5	25
11	2.3	5.7	7	12.7	21.2
12	3.9	9.4	7.8	15.3	26.3
Average	3.6	8.4	7.6	14.7	24.9
Total anaphors	4	11	11	21	34
	90%	76%	69%	73%	73%

This average matching rate, however, is lower than the matching rates, about 92%, we obtained in the empirical studies described previously (Yeh & Mellish 94; Yeh 95). The problem is mainly because the test texts used in the former comparison are human-created, while the test texts used here are machine-generated. The grammatical structures of the machine-created texts are simplified; they are not so sophisticated as human texts. In the evaluation work, when the speakers were asked to decide their preferences for anaphors in the machine-generated texts, they may find less complete information shown in the test texts than what they are used to in creating their own texts and hence it may be difficult for them to make their own decisions. In the empirical study, the human-created texts perhaps provided more sufficient information for the hypothetical machine to decide on an appropriate anaphoric form.

Another reason why the matching rates are lower than before is that in some circumstances there may be more than one acceptable solution and the speakers may not always choose the same one as the machine. This hypothesis can be investigated by looking at the extent to which the speakers agree among themselves. To see how the speakers agree among themselves, we further made a comparison between the speakers' annotations, which is summarised as below.

for each speaker i
  for each text j
    compare i's with the rest of speakers' annotations and note down the average number of the matches

The comparison result is shown in Table 2. For each speaker, the number for each test text is the average number of matches with the other eleven speakers. For example, Speaker 1 receives, on average, 8 matches for Text 2. At the end of the table are the average numbers for the speakers' agreement among themselves. The figures in the table show that the speakers do not achieve an agreement among themselves for the use of anaphors in this test.

As shown in Fig. 4, the anaphors in Text 1 form a "topic chain" [2] within a single "sentence". These anaphors are all zeroed according to the conditions of locality and syntactic constraints in the three test rules. All three systems produce the same result for Text 1 and, hence, unsurprisingly all three systems have the same matching rate, 90%, as shown in Table 1.

Text 2 similarly contains a single "sentence" but has three topic shifts in addition to "topic chains" within the "sentence" as shown in Fig. 4. Since no discourse segment boundaries occur within the "sentence", the discourse segment boundary constraint in TR*2* has no effect on this test text, which means that both TR*1* and TR*2* produce the same output. However, there are three topic shifts within the "sentence", namely, clauses 5 and 6, 8 and 9, and 10 and 11, as shown in Fig. 4. The shifts would make the rule containing the salience constraint, TR*3*, obtain different output from those without this constraint,. TR*1* and TR*2* obtain the same matching rate, 70%. TR*3* obtains higher matching rates than the other two, 79%, which shows the effectiveness of the salience constraint in it.

We then examine another middle-sized test text, Text 3, which is broken into three "sentences," as shown in Fig. 4. The beginning of a "sentence" is the beginning of a discourse segment in our implementation (Yeh 95). Furthermore, there are three topic shifts occurring in Text 3, i.e., clauses 8 and 9, 10 and 11, and 11 and 12. The constraint of discourse segment begin-

[2] A "topic chain" is a situation where a referent is referred to in the first clause, and then several more clauses follow talking about the same referent, namely, the topic.

nings in TR*2* and TR*3* and the salience constraint in TR*3* would therefore have some effects on the output texts. The matching rates, as shown in Table 1, increase from 62 to 66% for TR*2*, which shows that the constraint on discourse segment beginnings in TR*2* is effective. TR*3* obtains 65% matching rate, on average, which is 1% lower than its predecessor TR*2*. However, this decrease of average matching rate does not deny the effectiveness of the salience constraint in TR*3*. TR*2*'s text differs from TR*3*'s in the three topic shifts: TR*2* generates zero anaphors for these shifts, while TR*3* generates full descriptions. The speakers varied greatly in choosing anaphoric forms for these topic shifts: among twelve speakers, four chose all full descriptions, three used all zero anaphors, and the other five chose zero, pronominal and nominal anaphors. Thus, four among the twelve speakers completely agree with TR*3*, while three agree with TR*2*. This shows that the salience constraint in TR*3* is still effective.

Then we examine the more complicated texts, Texts 4 and 5. As shown in Table 1, the increases of matching rates show the effectiveness of the constraint of discourse segment beginning in TR*2*. Again, the average matching rates of TR3 are sightly lower than TR*2* for these two texts. However, similar to the situation in Text 3, the speakers have varied agreement on the choice of anaphors for the topic shiftings in these two texts. For Text 4, three and one speaker completely agree with TR*2* and TR*3*, respectively. As for Text 5, two speakers completely agree with TR*2*, while the others partly agree with TR*2* and TR*3*.

The above discussions show that the salience constraint in TR*3* is sometimes effective in getting small improvements in the output texts. This shows the difference of concepts of salience used between the speakers and TR*3*. In brief, the more sophisticated constraints a rule contains, the better it performs. Both TR*2* and TR*3* perform better than TR*1*. TR*3* performs better than TR*2* for texts with simple discourse segment structure. For the texts having complicated discourse segment structures, TR*2* is slightly better than TR*3* on average matching rates.

5 Conclusion

In this paper, we evaluated the quality of anaphors in the texts generated by using various rules. As shown in the results of comparisons between the anaphors created by computers and native speakers of Chinese, the individual constraints we collected in our previous studies (Yeh & Mellish 94; Yeh & Mellish 95; Yeh 95), are effective to a large extent in the generation of anaphors in Chinese. Also they can be implemented successfully. The comparison results reveal the fact that a Chinese natural language generation system employing the combination of these constraints would produce more effective anaphors than one using individual constraints. Although the average matching rates between the different rules and the speakers are lower than those from our previous experiments based on human-generated texts, this at least in part reflects considerable disagreement among the speakers themselves.

References

(Grosz & Sidner 86) B. J. Grosz and C. L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

(Liu 84) Y. C. Liu. *Zuowen de fangfa (Approaches to Composition) (in Chinese)*. Xuesheng Chubanshe, Taipei, Taiwan, 1984.

(Maybury 90) M. T Maybury. *Planning Multisentential English Text Using Communicative Acts*. Unpublished PhD thesis, Cambridge University, 1990.

(Meteer & McDonald 91) M. Meteer and D. McDonald. Evaluation for generation. In J. G Neal and S. M. Wlater, editors, *Natural Language Processing Systems Evaluation Workshop*, pages 127–131, NY, 1991. Rome Laboratory.

(Yeh & Mellish 94) C. L. Yeh and C. Mellish. An empirical study on the generation of zero anaphors in Chinese. In *Proc. of the 15th International Conference on Computational Linguistics*, pages 732–736, Kyoto, Japan, 1994.

(Yeh & Mellish 95) C. L. Yeh and C. Mellish. An empirical study on the generation of descriptions for nominal anaphors in Chinese. In *Prod. of Recent Advances in Natural Language Processing*, Tzigov Chark, Bulgaria, 1995.

(Yeh 95) C. L. Yeh. *Generation of Anaphors in Chinese*. Unpublished PhD thesis, University of Edinburgh, Edinburgh, Scotland, 1995.