PERGAMON

Accident Analysis and Prevention 31 (1999) 31–43

ACCIDENT ANALYSIS & PREVENTION

Junction road accidents during cross-flow turns: a sequence analysis of police case files

David D. Clarke [a,*,1], Richard Forsyth [b], Richard Wright [c]

[a] *Department of Psychology, University of Nottingham, University Park, Nottingham, NG7 2RD, UK*
[b] *University of West England, UK*
[c] *Unilever Research, UK*

Received 4 November 1997; received in revised form 27 June 1998

Abstract

In-depth studies of behavioural factors in road accidents using conventional methods are often inconclusive and costly. In a series of studies exploring alternative approaches, 200 cross-flow junction road accidents were sampled from the files of Nottinghamshire Constabulary, UK, coded for computer analysis using a specially devised Traffic Related Action Analysis Language, and then examined using different computational and statistical techniques. For comparison, the same analyses were carried out on 100 descriptions of safe turns, and 100 descriptions of hypothetical accidents provided by experienced drivers. The present study employed a range of sequence analysis techniques to examine the patterns of events preceding accidents of different types. Differences were found between real accidents, hypothetical ones and safe turns; between accidents turning onto and off a road with the right of way; between the accidents of younger and older drivers; between accidents on minor roads and major roads; and between the accident expectations (but not the real accidents) of male and female drivers. Pairs of successive events often provided particularly good cues for discriminating accident types. © 1998 Elsevier Science Ltd. All rights reserved.

Keywords: Accident causation; Road junction; Police records; Sequence analysis

1. Introduction

Behavioural factors in road accidents are difficult to study by traditional research methods for a number of reasons (Clarke et al., 1998a). Accidents are relatively rare and unpredictable, so direct observation is often impossible. Statistical comparisons of accident rates for different kinds of driver and circumstance leave out the various stages in the causal sequence (Wagenaar and Reason, 1990). Multidisciplinary Accident Investigation (MDAI) teams have a number of disadvantages. As Grayson and Hakkert (1987) point out, 'in spite of the tremendous amount of information collected in this type of study, the definitive conclusions reached on the crash process are very limited'.

Simple self-report studies using interviews and questionnaires with accident-involved road users are limited by the difficulties of reporting over-learned behaviours; problems of post hoc interference and forgetting; and deliberate concealment of critical information. Most in-depth accident studies use mixed samples of accident types, which make it difficult to find meaningful general patterns, even though case-study approaches are known to work best with tightly defined samples of similar accidents (England, 1981; Midland, 1992). On-the-spot accident investigations also tend to be biased towards injury accidents and certain times of day.

Finally, most behavioural in-depth studies tend to sacrifice the glue of a rich relational network in the data (Sheehy and Chapman, 1988) because they only look for complex patterns once the data have been aggregated over cases, rather than finding patterns in individual cases, and then aggregating those. This is a problem since most data coding methods can only preserve the elements of a case and not its structure, and the structural information, once lost, cannot be recovered by later analysis.

* Corresponding author. Tel.: +44-115-9515284; Fax: +44-115-9515324; e-mail: ddc@psychology.nottingham.ac.uk

[1] All three authors were working at the University of Nottingham when this study was conducted.

PII: S0001-4575(98)00042-6

This study is one of a series aimed at developing and demonstrating alternative approaches to accident causation research which could overcome many of these problems. Each study explores a different way of coding and analysing police case files describing right-turn junction accidents, sampled from the records of Nottinghamshire Constabulary (the regional police force for the county of Nottinghamshire, UK) for the year of 1988. (Note that in Britain, drivers use the left side of the road, so a right turn involves crossing the oncoming stream of traffic when turning off a major road, and two converging streams of traffic when turning onto a major road. In general the implications and hazards of this manoeuvre are the same as a left turn in countries such as the USA and mainland Europe, where drivers use the right side of the road).

Police files were chosen, partly because they were convenient and suitable for the purpose; partly because they cover all locations in the region, all seasons and all times of day; and partly to show how these methods enable us to use this large body of relatively neglected data. Police reports have sometimes been used for similar purposes before, for instance by Fell (1976), who recommended their use in the development of accident causal schemata. Massie et al. (1993) produced classifications of collision avoidance strategies using approximately 200 police files, concluding 'The review of selected police accident reports is an essential element of the analytic process'.

Following the recommendations of Grayson and Hakkert (1987), a specific type of accident was chosen in order to cut down the diversity of the sample, and thus improve the chances of getting meaningful results which were consistent across cases. Right-turning accidents, either onto or off a larger road, were chosen for the following reasons:

1. They occurred often enough in Nottinghamshire to give an adequate pool of accidents to sample.
2. There was a reasonably long chain of separable events leading up to each accident, which is a requirement for sequence analysis.
3. Junction accidents are of special interest to researchers in their own right, partly because they involve disproportionate numbers of older drivers (Moore et al., 1982; Viano et al., 1990; Verhaegen, 1995). The ageing population has a greater susceptibility from side impacts (Viano et al., 1990), so the human and financial cost of this type of accident is likely to increase with time, in relative terms.

In addition to the accident case reports, a sample of experienced drivers were each asked to write an account of an imaginary right-turning accident, and an accident-free right turn. These two kinds of hypothetical accounts were compared with each other and with the real case data, using the same methods throughout. By comparing hypothetical with real accidents, we hoped to see how the accident-expectations of drivers differ from the real dangers, and thus where their precautions against accidents are likely to be misconceived. By comparing real accidents with the safe manoeuvres, we hoped to find the distinctive features of the accident sequences.

The first study based on these cases (Clarke et al., 1998a) had used a genetic algorithm to create predictive rules which could discriminate injury from damage-only accidents, for example. This is a rule-finding computer program which allows efficient rules to evolve by a process which mimics mutation, breeding and natural selection. A number of interesting rules were found, such as:

(PULLOVER|(VEHICLE2 < 2.88))
(FAILSTN > (TYPE > WEATHER)).

meaning

The right-turner changes lane from left to right before making the turn OR the colliding vehicle has less than four wheels

and

The Turner fails to notice a vehicle or pedestrian
AND
the Turner is turning Off a larger road
OR
the Turner is turning Onto a larger road in poor weather.

However the technique suffered from two main disadvantages. Firstly, the algorithm (like the evolutionary mechanism on which it is modelled) is non-deterministic, so different runs of the same procedure with the same data give somewhat different outcomes. Up to a point this is an advantage, in that it allows the robustness of the findings to be assessed. Secondly, the rules were quite hard to interpret in some cases.

These problems were largely overcome in a second study (Clarke et al., 1998b), using a different machine-learning program, based on Quinlan's (1986) ID3 algorithm, which is deterministic, and produces outputs which are easier to understand.

However, neither of these studies addressed the key issue of how the action and event *sequences* leading up to an accident affect its outcome. As with most previous accident research, they dealt with cases merely in terms of the presence or absence (or frequency) of relevant events and features in accidents of different types. The main aim of the present study was to see what else can be discovered by looking at sequential patterns in the events preceding the accidents.

Sequence analysis methods preserve the patterning of the data over time, so that the chains of events which are most likely to lead to an accident can be picked out. They also deal with events and actions in their context, so they translate more directly into safety advice. To take a very simplistic example, a conventional study may find that event X is more characteristic of safe

manoeuvres than accidents. But what recommendation then follows from this? Should drivers do X? When? All the time? For most driving actions this would be inappropriate, and sometimes dangerous. A sequential study, on the other hand, may find that the sequence WX is commoner in safe manoeuvres, and WY is commoner in accidents. From this a much more specific and practical recommendation follows: 'After W, be sure to do X rather than Y'. This allows for the possibility that in other contexts Y might be a safer action than X.

2. Method

2.1. Data collection

2.1.1. Case selection

A total of 200 police case-files on right-turning accidents were randomly selected from the records held at police headquarters for Nottinghamshire, UK, (Nottingham Constabulary) for 1988, to include 100 right turns off a main road and 100 right turns onto a main road. Accidents at T-junctions and cross-roads were included. The year was fixed by the start date of a larger project which included this study, and the sample size was chosen to be adequate but manageable for the different kinds of analysis to be used in the project as a whole. Of these initial 200 cases, 16 met one or more of the following exclusion criteria:

1. the accident occurred at a roundabout (these were excluded to reduce the diversity of the sample);
2. there was not enough data on the police report for the accident to be reliably coded, or;
3. the accident had obviously been mis-classified as a right turn.

This left 184 accident cases, 90 turning onto, and 94 turning off, a major road.

2.1.2. Coding

Each case file contained a wealth of detailed information about the accident, typically including the four page summary form completed by the attending police officers, witness statements, vehicle examiners' reports, breath test reports, scale drawings of the accident site, and photographs from several distances and positions. Some files also contained further information and letters about court proceedings and insurance claims. (There is no national police force in the UK, and procedures vary in detail between the forces operating in the different regions and counties. In Nottinghamshire, a four page G126 summary form is required for all accidents attended by officers. This contains mainly categorical information like road type, speed limit, time of day, signalling, map reference, vehicle type, drivers details, and so on. However, the degree of detail, and the amount of supporting information given, tend to vary with the severity of the accident). All these sources of information were considered when making the coded summaries for analysis here.

Static features were recorded (such as weather and road conditions, time of day, carriageway type, and so on) together with the sequence of events making up the accident. The static features were straightforward, but a special coding scheme was needed to deal with the sequential information. This was called TRAAL—Traffic-Related Action Analysis Language—(see Clarke et al. (1995), for full details). Most road accidents involve at least two participants, who act independently for part of the time. This creates problems for the usual methods of sequential coding, because without direct observation it is not possible to interweave the actions of the different participants reliably into a single sequence of events. Therefore, TRAAL was only used to code the actions of the right-turner, and any other relevant actions and events which occurred in known relation to that. The codings are rather like simplified sentences, whose grammar represents patterns in the data, as opposed to lists of features which can merely be present or absent. An example coding of a single accident sequence is shown below.

```
UNEVENTFUL DRIVING
APPROACHES JUNCTION
INDICATES/RIGHT
SLOWS
STOPS
VIEW OBSTRUCTED BY VEHICLE/RIGHT/
  STATIC
FAILS TO NOTICE VEHICLE/RIGHT/MOVING
STARTS RIGHT TURN
IMPACT/AT Y10/ON FRONT/BY NEARSIDE
STOPS
```

Successive events are listed down the page. Modifiers for each event type are listed across the page, and separated by slashes. INDICATES means the same thing as uses turn signal. Y10 is a location code describing where in the intersection the accident occurred. This form of coding is human and machine readable, reasonably transparent, and yet reasonably detailed.

Each case was coded by one of two coders who had jointly devised and piloted the coding scheme, and agreed the definitions of all the terms. A further check of inter-coder reliability was carried out by setting a rule-finder program (BEAGLE—Bionic Evolutionary Algorithm Generating Logical Expressions) to try to predict which accidents had been coded by which coder. Where the rule-finder was successful, it not only showed the presence of a discrepancy between the coders, but also described its nature. On the first pass, some such rules were found, so all accidents were checked by both coders jointly, and differences in coding were resolved. After this the rule-finder was used again, and failed to

find any effective rules to discriminate one coder's work from the other's (which is as it should be).

Missing information was handled by contacting the people involved in the accident, or the attending police officers, or else by visiting the accident site to gather further information. The research was carried out some time after the accidents had occurred, so there was no question of visiting the sites while relevant physical evidence, such as debris or tyre marks, was still present.

2.2. Hypothetical accidents and safe transits

Some 100 drivers, 52 males and 48 females, with a mean age of 40.7 years were paid a small honorarium to take part in the study. They were recruited through advertisements in the local press and shop windows, and by word of mouth. Each driver provided a written description of a hypothetical right-turning accident and a safe right turn, in a format which was like the police accident reports.

The hypothetical accident and safe-transit data did not pose the same problems of missing detail as the real accident data. Otherwise, however, the same checks were carried out on the hypothetical accidents and safe transits as on the real-accident data.

The set of 184 coded real cases, together with the hundred hypothetical accidents and the hundred safe right turns, then made up the Nottingham Accident Database for Right Turns (NAD/RT).

3. Results

The following results relate mainly to the sequential information in the NAD/RT database, while the static features of the cases were kept for use in other studies, using different methods of analysis.

3.1. Stochastic analysis

One of the first stages in analysing sequential data is to estimate their degree or order. This is normally done by treating the sequences as Markov chains and using information theory to calculate the unpredictability of each symbol given a knowledge of varying numbers of preceding symbols (Attneave, 1959).

The issue is whether each event is influenced by the preceding one alone (which of course will be influenced by the one before that) or by particular combinations of two, three or more events. The complexity of a sequence, in this sense, is called its order. Sequences where each event depends on just one before are called second order sequences, because the basic unit is the event pair. (Some authors use a different convention in which this kind of sequence is called first order, because each event depends on one previous event.)

For this study, the 48 commonest actions were each given a single character code and a 49th character (@) was used to mark the place of a rare event. (The rare events were those which occurred less than three times in the NAD/RT database.) This gave an alphabet of 49 symbols.

In this way, the event descriptions were transformed into strings of characters, which were analysed as if they were texts. For example, the accident sequence:

UNEVENTFUL DRIVING
SLOWS
APPROACHES JUNCTION
CHECKS BEHIND
FAILS TO NOTICE VEHICLE/BEHIND
INDICATES/RIGHT
STARTS RIGHT TURN
NOTICES VEHICLE/BEHIND
STRAIGHTENS UP
IMPACT/AT Z9/ON BACK/BY FRONT
STOPS
END
*

was coded as:

cgFJSWja@VkP*

for the present analysis. (In this case the rare event is STRAIGHTENS UP, but in other sequences the rare-event symbol might stand for something else).

Most estimates of the complexity or order of event sequences treat the data as a single sample and quote significance values based on that sample alone (Thomas et al., 1983). For the present study though, a new program was written which allowed the data to be split into training and test sets, so that predictive information from one set of cases could be evaluated on a different set of unseen cases. This is a standard strategy in machine learning, to reduce the problem of overfitting, or accidentally tailoring a description to the peculiarities of a given sample of cases, producing pseudo-findings which do not generalise (Clarke et al., 1998a). It uses essentially the same idea as the cross-validation procedure used in many factor analytic and regression studies.

The real accidents, the hypothetical accidents and the safe transits were all analysed in this way—both backwards and forwards. The main finding was that all six modes are best described as second-order stochastic processes. It seems that symbol *triplets* (or longer fragments) do not significantly add to the information provided by *pairs*.

The results obtained in forward prediction of the real-accident dataset are typical of the other five modes, and are given below. Table 1 shows the results when the data are treated in the conventional way, as a single set. Table 2 shows the result of using findings from a training subset of cases to predict the events in an unseen test subset.

Table 1
Predictions based on the data as a single set of cases: percentage correct and mean entropy per symbol at different orders of approximation

Order	Percentage correct	Mean entropy per symbol
0	2.04	5.61 (bits/symbol)
1	6.49	3.37
2	54.10	1.34
3	65.59	1.04
4	72.58	0.85
5	79.86	0.65
6	85.75	0.48
7	92.44	0.37
8	95.23 (max)	0.31 (min)

In Tables 1 and 2, the percentage-correct values are calculated by working through each case, N events at a time, assuming at each step that the next event to occur will be the one which most commonly follows the previous N-1 events, in the data as a whole. The proportion of times that this makes the right prediction is the percentage-correct value.

The mean entropy per symbol was calculated using the formula described in Attneave (1959). This measures the improbability, or unpredictability, of each symbol under various conditions of prior knowledge. For the so called zero-order model, all 49 symbols are assumed to be equally likely. This gives the baseline unpredictability of the data when no empirical information is being used. In the first-order model the relative frequency of each symbol on its own is used, but no information about which events tend to follow which (their transitional probabilities). When N is 2 or more, the entropy value is based on the unpredictability of the last symbol in each group of N, given a knowledge of the previous N-1.

When the data are treated as a single set, the apparent unpredictability of each symbol continues to drop almost indefinitely as N grows larger, though the biggest fall occurs between 1st and 2nd order. When the data are split into training and test sets, however, the entropy estimate rises after order 2, while the

Table 2
Predictions made from a training subset of cases evaluated on a test subset: percentage correct and mean entropy per symbol at different orders of approximation

Order	Percentage correct	Mean entropy per symbol
0	2.17	5.61 (bits/symbol)
1	6.15	3.56
2	55.80 (max)	1.38 (min)
3	52.70	1.80
4	49.83	2.11
5	44.15	2.41

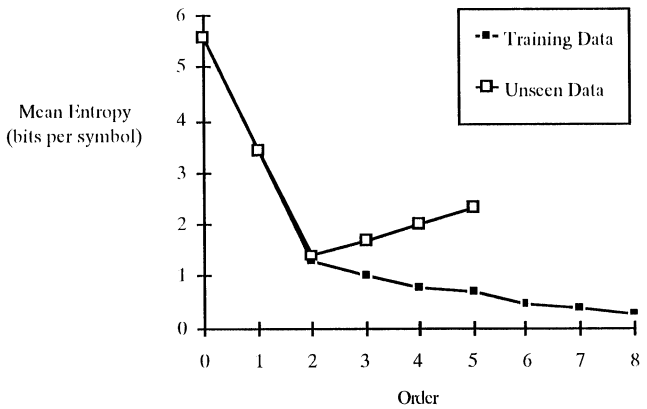

Fig. 1. Entropy of N-grams. (The unpredictability of the last event in each group of N, given the previous N-1, as a function of N).

percentage of correct predictions starts to fall. This shows that knowledge of two previous events is no better than knowledge of one, for predicting the next event. The analysis performed on the whole data set had indeed succumbed to overfitting, producing a false impression of greater predictability when longer histories were considered. This relationship is shown in Fig. 1.

As most of the sequential information in the data could be captured by the second-order model, it seemed appropriate to use event-pairs as critical markers in some of the later analyses.

3.2. Overlay analyses

A further program was written that built complete flow diagrams of all event sequences in the data, backwards and forwards, for contrasting subsets of data (such as injury and non-injury accidents; or accidents involving male and female Turners). It also pinpointed hot-spots, or choice points where the frequency of alternative pathways differed significantly between the two data-sets. (The person or vehicle making the right turn is always called the Turner here, and the person or vehicle they collide with is called the Collider).

This analysis did not provide the best discrimination between accident categories, but it did enable us to find the commonest sequence in various sorts of incident. Three examples are shown below.

Real accidents (seven instances)

UNEVENTFUL DRIVING	171
APPROACHES JUNCTION	162
SLOWS	125
FAILS TO NOTICE VEHICLE	10
STARTS RIGHT TURN	10
IMPACT	8
STOPS	7
END	7

Hypothetical accidents (four instances)	
UNEVENTFUL DRIVING	99
APPROACHES JUNCTION	95
INDICATES	64
SLOWS	59
STOPS	40
VIEW OBSTRUCTED	11
FAILS TO NOTICE VEHICLE	7
STARTS RIGHT TURN	6
IMPACT	4
STOPS	4
END	4

Safe transits (five instances)	
UNEVENTFUL DRIVING	90
APPROACHES JUNCTION	86
INDICATES	61
SLOWS	59
STOPS	44
WAITS	27
STARTS RIGHT TURN	7
CONTINUES	6
UNEVENTFUL DRIVING	5
END	5

Here the numbers on the right give the frequency of the sequence in the data-set, up to that point.

The sudden drop in frequency, from 125 to 10, between SLOWS and FAILS TO NOTICE VEHICLE in the real accident sequences indicates that this is a stage at which there are many options open, hence a point of particular difficulty for the Turner, and a point where the sequence is likely to go wrong. This was not reflected in the hypothetical accident descriptions, where our informants imagined STOPS to follow SLOWS much more often than it does in real life.

Real and hypothetical sequences tend to unfold in different ways. The commonest beginning for both is UNEVENTFUL DRIVING + APPROACHES JUNCTION, but then real accidents tend to continue with SLOWS, followed by either STOPS or FAILS TO NOTICE VEHICLE. Hypothetical accidents however typically go next to INDICATES then SLOWS.

These differences probably occur for the following reasons: (a) the real accident records fail to capture some of the detail which is written into the hypothetical ones; and (b) the hypothetical accidents describe unrealistically well-behaved Turners, because the informants usually cast themselves in that role, and imagine accidents in which they are victims of other people's actions.

3.3. Typical sequences

The next step was to identify typical and distinctive sequences, using a method of sequence comparison known as the Gestalt algorithm. This is a procedure which takes two symbol-strings and calculates a value between 0 and 1 that represents their similarity. This measure takes account, not only of the elements the two strings have in common, but also their relative locations. There are many ways of computing string-similarity (Sankoff and Kruskal, 1983), of which we tested three: (a) the proximity scan procedure (Rosenthal, 1984); (b) Levenshtein's algorithm (Levenshtein, 1965); and (c) the Gestalt algorithm (Ratcliff and Metzener, 1988). The latter performed best on the trial data, so it was used for all the analyses reported here.

One program (called TYPICAL) finds the best single prototype for a given group of cases. It takes a set of strings and calculates the average similarity between each one and all the rest. The string with the highest score is the typical member of the set. This sort of similarity analysis has been used previously for comparing bird songs (Bradley and Bradley, 1983); for estimating the evolutionary distance between segments of DNA (Sellers, 1980); for correcting misspellings in text files (Alberga, 1967); and for other purposes. However, we believe this is the first application of this technique in road-safety research. Malaterre (1990) provided the general principle of representing a complex set of accident cases by reporting an automatically identified prototype for each of the main categories.

A second program (called CONTRAST) finds the accident which best captures the distinctive features of a certain type of case. It takes two sets of strings, which are instances and non-instances of the type in question, for instance accidents where the Turner was under 25, contrasted with accidents where the Turner was not under 25 (called positive and negative instances of the type), and computes the average similarity of each string to the positive and the negative examples. Then the string with the largest difference between these two averages is the best one for capturing what the positive strings are like, in contrast to the negative ones. The advantage of using the TYPICAL and CONTRAST programs in this way is that they consider global patterns and similarities between sequences, whereas most methods of sequence analysis just rely on transitional frequencies, which only capture local regularities.

This type of analysis was applied in six areas: (a) sex differences; (b) age effects; (c) road type; (d) turn type; (e) real versus hypothetical accidents; and (f) severity.

3.3.1. Sex differences

To look at sex differences, we separated the accidents involving a male Turner from those involving a female Turner. According to the TYPICAL program, the most typical male accident was an Off accident which went as follows.

UNEVENTFUL DRIVING
APPROACHES JUNCTION
SLOWS
FAILS TO NOTICE VEHICLE/BEHIND
STARTS RIGHT TURN
IMPACT
STOPS
END

This sequence is the same as the most typical accident overall, which is not surprising as the majority of the Turners in the sample (75.3%) were male. More remarkably, however, the typical sequence for female Turners (also an Off accident) was just the same. This suggests that male and female drivers follow the same sequence of behaviours—at least at the level of detail captured by the TRAAL coding scheme—when turning right at junctions.

The hypothetical accidents invented by male and female informants are also quite similar. The typical sequences given by both sexes start in the same way:

UNEVENTFUL DRIVING
APPROACHES JUNCTION
INDICATES /RIGHT
SLOWS
STOPS

then they diverge slightly:

Male informants	Female informants
CONTINUES/THROUGH GREEN LIGHT	WAITS/FOR GAP IN TRAFFIC
VIEW OBSTRUCTED/BY VEHICLE	FAILS TO NOTICE VEHICLE
FAILS TO NOTICE VEHICLE	

and then continue again in the same way:

STARTS RIGHT TURN
IMPACT
STOPS
END

As it happens, both examples involved accidents turning Onto a larger road which resulted in damage only. So males and females do not imagine accidents which are markedly different in sequence, nor do they seem to have accidents with different sequential patterns. The differences which do occur may be a reflection of different patterns of exposure, with male informants doing relatively more driving on major roads, and therefore imagining signalised junctions, and female informants being relatively more used to driving on local roads, and so imagining unsignalised junctions. The project did not set out to contrast signalised and unsignalised junctions a priori, which on reflection may have been a mistake, given the different driving procedures and risks they involve. However, as with other relevant contrasts, the approach was to code as many features of the accidents as possible without pre-judging their significance, and then to see which ones turned out to be associated with different outcomes, and different types of drivers.

3.3.2. Age effects

To look at age effects, all the real accidents where the Turner was under 25 (young drivers) were put into one data-set, and all the accidents where the Turner was 60 or over (older drivers) were put into another. The typical older driver's accident, according to the TYPICAL program, is:

UNEVENTFUL DRIVING
APPROACHES JUNCTION
SLOWS
INDICATES/RIGHT
STOPS/AT TRAFFIC LIGHTS
WAITS/FOR GREEN LIGHT
FAILS TO NOTICE VEHICLE
STARTS RIGHT TURN
IMPACT
STOPS
END

The typical young person's sequence, on the other hand, comes from an Onto accident, and is:

UNEVENTFUL DRIVING
APPROACHES JUNCTION
SLOWS
FAILS TO NOTICE VEHICLE
STARTS RIGHT TURN
IMPACT
STOPS
END

Both sequences start and finish in the same way. In the middle, the older Turner (unlike the young Turner) INDICATES, STOPS and WAITS but still crashes. This is consistent with the idea that young people go wrong through risk acceptance or not taking precautions, while older drivers get into trouble more often through failure of observation or situation awareness. (Note that in Britain, a driver turning right on a green light still has to give way to oncoming traffic, so a collision is likely to reflect the poor observation of the Turner, rather than a signal violation by the oncoming driver, who will be crossing the junction on green, with right of way.) It also seems that once again the difference between signalised and unsignalised junctions is affecting different groups of drivers in different ways.

3.3.3. Road type

The differences were also examined between accident sequences on major roads (A or B) and minor roads (C or unclassified). The typical major-road sequence comes from an Off accident and goes as follows.

UNEVENTFUL DRIVING
APPROACHES JUNCTION
SLOWS
INDICATES/RIGHT
FAILS TO NOTICE VEHICLE/AHEAD
STARTS RIGHT TURN
IMPACT
STOPS
END

The typical minor-road accident is exactly the same except that the INDICATES/RIGHT is omitted. Thus the typical major and minor road event sequences are much the same, but an application of the CONTRAST program indicates that the most distinctive sequences from each category are rather different.

For major-roads, the most distinctive sequence came from an Off accident which resulted in slight injury:

UNEVENTFUL DRIVING
APPROACHES JUNCTION
SLOWS
INDICATES/RIGHT
STOPS
FAILS TO NOTICE VEHICLE/BEHIND
STARTS RIGHT TURN
IMPACT
STOPS
END

For minor-roads, the most distinctive sequence came from an Onto accident that resulted in damage only:

UNEVENTFUL DRIVING
APPROACHES JUNCTION
STARTS RIGHT TURN
LOSES CONTROL OF VEHICLE
LEAVES CARRIAGEWAY
IMPACT [hit a road-sign]
REJOINS CARRIAGEWAY
UNEVENTFUL DRIVING [drove off at high speed]
END

Although the latter is not typical of what happens in right-turning accidents at junctions on minor-roads, it is a kind of episode which is only found on minor roads, in our data. Presumably, even careless drivers are not this careless on a major road. (It happened at 22:45, and no other vehicle was involved—two features which are typical of accidents involving alcohol. It may be preferable in future to exclude such accidents from the more general categories being considered here, and to examine them in a separate study).

3.3.4. Turn type

Road-type is confounded with turn-type, since in the NAD/RT database only 42% of the major-road accidents are Onto cases, whereas 58% of minor-road accidents are Onto. So we also checked for sequences typical of Onto and Off accidents.

For accidents turning Off a larger road (not always an A or B road), the typical sequence is the same as the typical major-road sequence shown above. For Onto accidents, the most typical sequence is still the same, except that (a) the stage INDICATES/RIGHT is missing and (b) the event FAILS TO NOTICE VEHICLE is qualified by /LEFT rather than /AHEAD—the direction of danger is from the left not the front.

Applying the CONTRAST program tells a slightly different story. The most distinctive sequences from Off and Onto accidents are given below.

Distinctive Off sequence	Distinctive Onto sequence
UNEVENTFUL DRIVING	STARTS FROM PARKED
SLOWS	APPROACHES JUNCTION
INDICATES/RIGHT	SLOWS
APPROACHES JUNCTION	VIEW OBSTRUCTED/LEFT
CONTINUES/THROUGH GREEN LIGHT	FAILS TO NOTICE VEHICLE/LEFT
FAILS TO NOTICE PEDESTRIAN/RIGHT	STARTS RIGHT TURN
STARTS RIGHT TURN	IMPACT
IMPACT	CONTINUES/TURNING RIGHT
STOPS	UNEVENTFUL DRIVING
END	END
*SLIGHT INJURY	*DAMAGE ONLY

The distinctive Off accident is a rather ordinary right-turning accident, except that the Turner goes through a green light (more likely when turning Off than Onto a bigger road, for obvious reasons). The turner then fails to notice, and hits, a pedestrian rather than a vehicle, causing injury rather than just damage (which is also commoner among Off than Onto accidents).

The distinctive Onto accident sequence is very distinctive. It is not very like the typical Onto sequence, but it is very unlike the typical Off sequence. Starting from a parking place, having an obstructed view, and continuing after an impact (the second UNEVENTFUL DRIVING event) are all rare features of Off sequences.

We also contrasted Off and Onto accidents on fast roads only (roads with a speed limit over 35 mph). The typical sequences are almost identical but the most distinctive sequences are not, as shown below.

Distinctive Off sequence (fast road)	Distinctive Onto sequence (fast road)
UNEVENTFUL DRIVING	STARTS FROM PARKED
APPROACHES JUNCTION	APPROACHES JUNCTION
SEES TURN-OFF SIGN LATE	STOPS
FAILS TO NOTICE VEHICLE/BEHIND	WAITS/FOR GAP IN TRAFFIC
INDICATES/RIGHT	FAILS TO NOTICE VEHICLE/RIGHT
SLOWS	STARTS RIGHT TURN
PULLS OVER/TO OUTSIDELANE	IMPACT
IMPACT	STOPS
FORCED INTO ONCOMING TRAFFIC	END
IMPACT	*SLIGHT INJURY
STOPS	
END	
*SERIOUS INJURY	

Here the program has picked up some genuine distinctive features of accidents turning Off high-speed roads:

1. late recognition of turn-off;
2. consequent late indication and slowing;
3. failure to appreciate danger from behind;
4. changing lanes from left to right;
5. double impact.

All these features, while not common in right-turn accidents Off a fast road, are extremely rare in Onto accidents. On the other hand, starting from a parking place is found only in Onto accidents. (The fact that this particular Off sequence leads to a serious injury, while the Onto sequence leads to a slight injury, is also realistic).

3.3.5. Real versus hypothetical accidents

We looked at typical sequences among hypothetical accidents, real accidents, and safe transits. The typical hypothetical sequence is rather like the typical safe transit.

Typical hypothetical accident	Typical safe transit
UNEVENTFUL DRIVING	UNEVENTFUL DRIVING
APPROACHES JUNCTION	APPROACHES JUNCTION
INDICATES /RIGHT	INDICATES /RIGHT
SLOWS	SLOWS
STOPS	STOPS
WAITS /FOR GAP IN TRAFFIC	WAITS /FOR GAP IN TRAFFIC
FAILS TO NOTICE VEHICLE /AHEAD	STARTS RIGHT TURN
STARTS RIGHT TURN	CONTINUES /TURNING RIGHT
IMPACT	UNEVENTFUL DRIVING
STOPS	END
END	

Thus our informants tend to imagine a prototypical accident as one where the Turner does everything right, just as in an idealised safe turn, until reaching the junction, and then goes wrong by failing to observe a hazard. This may reflect something about our informants' hypotheses on accident causation. If so, they are right to focus on failure of observation as playing a decisive part in the process, but wrong to assume that indicating, stopping and waiting are normal parts of the sequence leading to a right-turning accident.

In real accidents, the typical sequence is shorter and less like the ideal safe turn.

Typical real accident
UNEVENTFUL DRIVING
APPROACHES JUNCTION
SLOWS
FAILS TO NOTICE VEHICLE /LEFT
STARTS RIGHT TURN
IMPACT
STOPS
END

Taking serious-injury accidents on their own, some systematic differences emerged between real and hypothetical cases which could have implications for driver education. The most distinctive serious-injury accident, that is the one which was most like other real serious accidents and least like the hypothetical ones, is given below, including its static features.

TYPE	OFF	
WHEN	WEDNESDAY, 14-12-88, 22:35	
ROADTYPE	UNCLASSIFIED	
SPEEDLIM	30	
URBANITY	URBAN	
CARRIAGEWAY-TYPE	SINGLE	
JUNCTION-TYPE	PRIVATE DRIVE	
JUNCTION-CONTROL	NONE	
WEATHER	FINE	
SURFACE	DRY	
LIGHTING	DARK	
AGE-OF-DRIVER	19	11

SEX-OF-DRIVER	M	M
SEATBELTS	IN USE	N/A
BREATH-TEST	NOT REQUIRED	N/A
VEHICLE-TYPE	CAR	PEDESTRIAN

UNEVENTFUL DRIVING
APPROACHES JUNCTION
SLOWS
FAILS TO NOTICE PEDESTRIAN/RIGHT/MOVING
STARTS RIGHT TURN
IMPACT /AT Y12/ON FRONT-OFFSIDE/BY BODY
STOPS
END
*SERIOUS INJURY

This case illustrates most of the features that distinguish real and imaginary serious-injury accidents.

More likely in real serious-injury accidents
Collider on 2-wheeler or on foot
Fine weather, urban road
Time-of-day, pm
Danger from behind

More likely in hypothetical serious-injury accidents
Turner INDICATES then SLOWS then STOPS
Accident on Major road with fast speed limit
Turner EDGES FORWARD to gain better view
Collider fails breath-test

3.3.6. Severity

Finally, we used this method to look at the important question of severity, by contrasting injury with non-injury accident sequences. The most typical injury accident sequence is given below. (It happens also to be the most typical serious-injury accident sequence).

UNEVENTFUL DRIVING
APPROACHES JUNCTION
SLOWS
FAILS TO NOTICE VEHICLE/RIGHT
STARTS RIGHT TURN
IMPACT
STOPS
END

The most typical damage-only accident sequence was exactly the same, except that the Turner failed to notice a vehicle approaching from behind, not from the right. Both, in fact, are typical accident sequences, so the injury and damage-only accidents do not appear to differ in sequential structure.

The CONTRAST program found that the distinctive serious-injury sequence (as opposed to damage-only sequence) was the same as the typical injury accident sequence given above, except that SLOWS was replaced by MISJUDGES SPEED and DISTANCE OF VEHICLE/AHEAD. However the measure of discrimination between the two kinds of sequence was very low.

Our overall conclusion from this phase of the study was that the sequential structure of serious-injury accidents, slight-injury accidents and damage-only accidents is not markedly different—at least in so far as it can be captured by TRAAL. The three kinds of incidents unfold in a similar pattern, and the factors that determine how serious the result will be are things like the size of the vehicle(s) involved and the speed at which they are travelling.

3.4. Sequential pattern-detection

A further program, based on the ID3 algorithm (Quinlan, 1986; Clarke et al., 1998b) was written to make use of the earlier finding that event pairs are the main information-carrying units. It builds discrimination trees (rather like decision trees) for distinguishing different kinds of accident sequences, represented as text strings as above. A discrimination tree can be used to classify unseen cases, or to read off rules for separating two types of sequences.

For this purpose, the cases were randomly divided into a training set for forming the rule-trees, and a test set for assessing the rules on unseen data. The significance levels given are for the results on unseen data. The program was unable to find a robust rule for discriminating injury from non-injury accidents on the basis of distinctive event pairs. However, it was easy to find a rule-tree that distinguished hypothetical from real accidents:

IF APPROACHES JUNCTION + INDICATES is present
THEN case type is Hypothetical
ELSE IF APPROACHES JUNCTION + SEES TURN-OFF SIGN LATE is present
THEN case type is Hypothetical
ELSE IF VIEW OBSTRUCTED + STARTS RIGHT TURN is present
THEN case type is Hypothetical
ELSE IF FAILS TO NOTICE VEHICLE + EDGES FORWARD is present
THEN case type is Hypothetical
ELSE case type is Real

This rule was developed on 174 cases, and tested on a random selection of 110 other cases. Of the 43 unseen cases classed as Hypothetical by this rule, 32 really were hypothetical, and 11 were not. Of the remaining 67 cases classed as Real by this rule, 57 were real, and ten were not ($\chi^2 = 39.27$, d.f. $= 1$, $P < 0.001$). The program also estimates the probability of each case being of each type. These probability estimates were correlated with

the true categories (point-biserial correlation = 0.59). This shows the rule to account for 35% of the variance. In other words, the program discovered four event-pairs that can reliably distinguish hypothetical from real accident sequences.

The same procedure was used to look for consistent differences between the accident sequences of male and female Turners, and between young and older drivers, but the rule-trees which emerged did not generalise to unseen cases.

The program was also set to look for differences between hypothetical accidents and safe-transits. This turns out to be too easy (since pairs involving IMPACT, for instance, were never found in safe turns). Having removed all such cheating pairs, we obtained a rule-tree that was still successful at categorising unseen cases, where it gave 32 true positives, three false positives, 39 true negatives, and 12 false negatives ($\chi^2 = 38.30$, d.f. = 1, $P < 0.001$). The rule is rather complex, but can be summed up by saying that safe-transits were marked by:

STARTS RIGHT TURN immediately followed by
AVOIDED BY VEHICLE, *or*
CONTINUES/TURNING RIGHT, *or*
ACCELERATES, *or*
BRAKES + CONTINUES/TURNING RIGHT.

This may seem unsurprising, but it probably reflects a real tendency on the part of informants to have the following in mind as features of safe right turns: good behaviour by other road users (AVOIDED BY VEHICLE); pure luck (CONTINUES/TURNING RIGHT); or skill at last-minute avoiding action (BRAKES, ACCELERATES). To put it the other way round, they seem to attribute right-turning accidents to (a) lack of skill or care by other road users; (b) lack of skill in emergency action by the Turner; or (c) bad luck, rather than to failures of observation by the Turner.

4. Discussion

4.1. Main findings

The accident sequences in the NAD/RT database turn out to have a second-order sequential structure. In other words, it is unhelpful to consider transitional probabilities linking more than immediately successive events, as this leads to overfitting and to misleading predictive information that does not generalise to unseen cases. Event pairs, however, can be appropriate units of analysis in various ways, including the induction of classification rules which discriminate hypothetical accidents from real ones, and from safe turns.

These rules suggest that drivers exaggerate the importance of good behaviour by other road users, luck, and skill at last-minute avoiding action. They also tend to imagine right-turn accidents in which the Turner does everything correctly, just as in an idealised safe turn, until reaching the junction, and then goes wrong by failing to observe a hazard. Hypothetical accidents also tend to be over-dramatic, with a large number of risk factors added together. In part this can be explained by the expectations of the experimental situation, but it also reflects the general tendency for people to perceive safe situations as safer than they really are, and risky ones as more dangerous (Howarth, 1987).

Real accident sequences, on the other hand, are shorter and different in character, with a crucial danger point occurring just after the Turner stops. Now the network of possible events suddenly divides in many more ways, and the possibilities of error are more numerous. Of course, the next few events are then the decisive ones in determining whether a collision will occur.

Accidents turning Onto and Off a major road also show different patterns. The distinctive Onto sequence starts with the vehicle parked, there is obstruction of view, and the driver continues after the impact, all of which is seldom found in Off sequences. Off accidents are distinctive in involving late recognition of turn-off; consequent late indication and slowing; failure to appreciate danger from behind; changing lanes from left to right; and double impact.

Older drivers are more likely to have accidents with the events INDICATES, STOPS, WAITS (in that order) before the turn, suggesting perhaps that young people go wrong through risk acceptance or not taking precautions, while older drivers get into trouble more often through failure of observation or situation awareness.

Minor-road accidents sometimes involve degrees of carelessness which never occur in the major-road accidents, in this database.

Male and female drivers are not very different in their real accidents, according to this study, although the hypothetical right-turn accidents they envisage are somewhat different.

Serious-injury accidents, slight-injury accidents, and damage-only accidents are not very different in sequential structure, although there are some particularly interesting differences between real serious-injury accidents and hypothetical serious-injury accidents.

4.2. Methodology

The analysis described here has two main limitations. Police files on road accidents, elaborate though they are, describe events at a level of detail which does not firmly establish the ordering of certain high-speed events.

Secondly, it would appear that right-turning accidents do not have as much detailed sequential structure as one might expect. Although driving is a temporal process, giving rise to long chains of behaviour, no significant dependencies were found between acts separated by more than a few others. In so far as accidents have a grammar it is a very simple and rather inflexible one, in which the omission of a key element at a fixed point in the sequence (such as failing to notice an approaching vehicle just before starting a turn) renders an otherwise valid sentence completely invalid. To be more concrete: the boundary between low-risk and high-risk situations is sharp and can be crossed in a moment by a single act (or omission), not by deviating progressively from an ideal path, with opportunities to recover along the way.

It could be argued that some of the findings of this study are already familiar from more conventional research, but it is reassuring to have further confirmation from this rather different source, and it seems likely that this method would show its real diagnostic power when applied to more tightly defined accident categories.

The TRAAL coding-scheme has aroused some interest in the USA and in France already, but it is clearly far from perfect. A major improvement could be made by checklisting the main action terms. This would mean dividing the episode in question into two or three phases which are reliably identifiable from the police records, and recording the presence or absence of acts or short sequences of acts, rather than attempting to specify the detailed sequence of events *de novo*. Such an approach would fit the coding scheme better to the data on which it is based, and would have several additional advantages. Each case would take less time to code. It would be easier to cross-check between coders. Above all, it would allow the acts of the Collider to be recorded on the same basis as those of the Turner. It could also be used on other kinds of manoeuvre besides right turns.

Our experience with Nottinghamshire Constabulary accident records leads us to believe that police files, although unsuitable for some kinds of detailed moment-by-moment coding, do contain useful information that is not usually exploited for the purposes of road accident research. Such records give valuable behavioural information which is lost by the time the national statistics are compiled. TRAAL can be seen as a first step in the evolution of coding practices that will allow a more productive use to be made of this neglected data source.

Overall, we hope this series of studies will help to open up new modes of road-accident research which provide a viable compromise between traditional in-depth studies and nomothetic methods. They have the potential of being cheaper than traditional in-depth studies, but at the same time, more informative about behavioural details than conventional statistical analyses of large aggregated data sets.

Acknowledgements

This study was sponsored by the United Kingdom Department of Transport (since incorporated into The Department of the Environment, Transport and the Regions) and carried out under contract to the Transport Research Laboratory, Crowthorne, UK. This paper is abridged with permission from the TRL final project report CR305 The Analysis of Pre-Accident Sequences. We are most grateful to Nottinghamshire Constabulary for their patient assistance in locating suitable cases for analysis; to other members of the Accident Research Unit, Department of Psychology, University of Nottingham, for their helpful comments and suggestions; to Chris Ashton of the Nottinghamshire County Council Accident Investigation Unit for assistance with the selection of the sample; to Graham Grayson and Geoff Maycock of the Transport Research Laboratory, for their expert guidance and advice; and to two anonymous referees for their helpful comments and suggestions.

References

Alberga, C.N., 1967. String similarity and misspellings. Communication of the Association for Computing Machinery. 10, 302–313.

Attneave, F., 1959. Applications of Information Theory to Psychology. Holt, Rinehart and Winston, New York.

Bradley, J., Bradley, B., 1983. Distance measures for comparing birdsong among different populations. In: Sankoff, D., Kruskal, J.B. (Eds.), Time Warps, String Edits and Macro-Modules: The Theory and Practice of Sequence Comparison. Addison-Wesley, Reading, MA.

Clarke, D.D., Forsyth, R.S., Wright, R.L., 1995. The Analysis of Pre-Accident Sequences. Contractor Report no. 305. Department of Transport/Transport Research Laboratory, Crowthorne, Berkshire, UK.

Clarke, D.D., Forsyth, R.S., Wright, R.L., 1998. Behavioural factors in accidents at road junctions: the use of a genetic algorithm to extract descriptive rules from police case files. Accident Analysis and Prevention. 30, 223–234.

Clarke, D.D., Forsyth, R.S., Wright, R.L. Machine learning in road accident research: decision trees describing road-accidents during cross-flow turns, Ergonomics (in press).

England, L., 1981. The role of accident investigation in road safety. Ergonomics 24, 409–422.

Fell, J.C., 1976. A motor vehicle accident causal system: The human element. Human Factors 18, 85–94.

Grayson, G.B., Hakkert, A.S., 1987. Accident analysis and conflict behaviour. In: Rothengatter, J., de Bruin, R. (Eds.), Road Traffic Safety. Van Gorcum, The Netherlands.

Howarth, C.I., 1987. Perceived risk and behavioural feedback: strategies for reducing accidents and increasing efficiency. Work and Stress 1, 61–65.

Levenshtein, V.I., 1965. Binary codes capable of correcting spurious insertions and deletions of ones. Problems of information transmission 1, 8–17.

Malaterre, G., 1990. Error analysis and in-depth accident studies. Ergonomics 33, 1403–1421.

Massie, D.L., Campbell, K.L., Blower, D.F., 1993. Development of a collision typology for evaluation of collision avoidance strategies. Accident Analysis and Prevention. 25, 241–257.

Midland, K., 1992. In-Depth Accident Investigation Teams as a Tool for Traffic Safety. Report 135/1992. Institute of Transport Economics [TØI], Oslo, Norway.

Moore, R.L., Sedgeley, I.P., Sabey, B.E., 1982. Ages of Car Drivers Involved in Accidents, with Special Reference to Junctions. Supplementary Report 718. Transport and Road Research Laboratory, Crowthorne, Berkshire, UK.

Quinlan, J.R., 1986. Induction of decision trees. Machine Learning 1, 81–106.

Ratcliff, J.W., Metzener, D.E., 1988. Pattern matching-the Gestalt approach. EXE Magazine 3, 24–38.

Rosenthal, S., November, 1984. The PF474: A Co-Processor for String Comparison. Byte Magazine.

Sankoff, D., Kruskal, J.B., 1983. Time Warps, String Edits and Macro-Modules: The Theory and Practice of Sequence Comparison. Addison-Wesley, Reading, MA.

Sellers, P.H., 1980. The theory and computation of evolutionary distances: pattern recognition. Journal of Algorithms 1, 359–373.

Sheehy, N., Chapman, A., 1988. Reconciling witness accounts of accidents. In: Rothengatter, J.A., de Bruin, R.A. (Eds.), Road User Behaviour: Theory and Practice. Van Gorcum, The Netherlands.

Thomas, A.P., Roger, D., Bull, P., 1983. Conversation as a Markov process. British Journal of Social Psychology 22, 177–188.

Verhaegen, P., 1995. Liability of older drivers in collisions. Ergonomics 38, 499–507.

Viano, D.C., Culver, C.C., Evans, L., Frick, M., 1990. Involvement of older drivers in multivehicle side-impact crashes. Accident Analysis and Prevention 22, 177–188.

Wagenaar, W.A., Reason, J.T., 1990. Types and tokens in road accident causation. Ergonomics 33, 1365–1375.