Universiteit Leiden
The Netherlands

Archaeology and the application of artificial intelligence : case-studies on use-wear analysis of prehistoric flint tools

Dries, M.H. van den

Citation

Dries, M. H. van den. (1998, January 21). *Archaeology and the application of artificial intelligence : case-studies on use-wear analysis of prehistoric flint tools*. Retrieved from https://hdl.handle.net/1887/13148

Version:	Corrected Publisher’s Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/13148

Note: To cite this publication please use the final published version (if applicable).

7 Test results[1]

7.1 Introduction

A knowledge-based system should not merely be evaluated on its design and on the degree to which requirements have been met (chapter 5), but also on its practical functionality. The latter is decisive for the final acceptance by the end users. Especially when the interpretation of an expert system application is crucial for safety purposes, all aspects need to be tested under all possible circumstances. Also the abilities of WAVES could only be experienced by testing them in practice.

Expert system testing consists of two components: a *theoretical validation* and a *practical evaluation*. The former mainly concerns the correctness and the reliability of the interpretations. The underlying goal of this validation is to determine the extent to which the knowledge base reflects the knowledge of the expert. The confidence of the future users can only be gained if the application performs as accurate as the expert. Since this component validates the applications accuracy, consistency and completeness, it can be seen as a quality control. These aspects can, for example, be tested by measuring the number of correct interpretations, by verifying the repeatability of interpretations in case of a repeated submittance of data and by mapping the application's sensitivity to incorrect or incomplete input data. In practice, this can be done by comparing the performance of an application with that of the expert by confronting them both with the same cases. Usually, this is carried out by or in dialogue with the human expert who guided the development process. The meaning of this test is not to detect which of the two parties performs better: it is assumed that in case of differences, the expert's interpretation is correct. Nevertheless, it is advisable to define in advance the kind of interpretations that will be rewarded and the minimal rate of success that is still acceptable. Since such a test may be an extremely time consuming task when a knowledge base consists of several thousand rules, automated validation tools, so-called *rule checkers* (*cf*. Perkins *et al*. 1989) have been designed. They not only quicken the test procedure, but also guarantee a complete and thorough validation. In comparison with human beings, they are better in checking all possible interactions of these huge amounts of rules.

A practical evaluation, on the other hand, concerns the usability of the application. The underlying goal is to verify whether the content addresses the basic functionality that it intends to cover and whether it provides the expected results when it is employed by independent users. Usually, this evaluation is performed by the end users in order to obtain independent results and to discover how they experience the use of the application. It validates the validness of interpretations in a given context, the application's user-friendliness and comprehensibility, the transparency of the knowledge, the flexibility of the explanation facilities, etc. Whereas there are various means to validate the reliability of an application, there are no standard methods for practical validations. This implies that it may be difficult to compose realistic and adequate criteria. Therefore, the results of such a test must always be interpreted with care, especially since they do not represent an objective measurement. They are highly influenced by human factors, such as concentration during use, prejudices (both positive and negative), level of intelligence. An additional complicating factor is that the interpretations must also be assessed in relation to the amount and quality of the information that the application has to deal with and in relation to the limitations of the underlying method of analysis. If the method cannot handle particular cases it is to be expected that the application cannot either. There are, however, some general applicable criteria for practical evaluations. One of the most important is that garbage input should result in no output rather than in garbage output (Hollnagel 1989: 394). Another criterion is that the information of the application consists of a good and useful advice which may improve the quality of the final decision or interpretation of the user.

In reference to the consequences that may be drawn from the results of both theoretical and practical validations, we must be aware that all evaluation methods bear methodological problems and that none of them provides an exhaustive validation of all aspects of an application (*ibid*.: 410). Except maybe in case of small and closed knowledge domains, hardly any test covers all possible situations. It seems to be beyond human cognition to evaluate complex reasoning processes, let alone to design absolute infallible validation means. According to Hollnagel we are therefore trapped in an impasse, because "…*reasoning mechanisms are introduced to compensate for the shortcomings of human reasoning, but these very shortcomings make it extremely difficult to determine whether the reasoning mechanisms work correctly*." (*ibid.*: 399).

Since the reasoning mechanism of WAVES is not very complex and the amount of rules not extreme, our test will probably raise less insurmountable problems. Nevertheless, this evaluation should be validated on its reliability, validity and usability for there are all kinds of aspects that may have influenced the test results. Moreover, it must be tried to make the test as representative as possible for the application as a whole. It should anticipate to situations that have not been explicitly taken into account in the composition of the knowledge base but which are likely to be encountered in practice. It should also be realised, however, that this evaluation will merely be a cursory check that offers nothing more than an impression of the application's functionality.

Despite its limited meaning, the above described dual approach of a theoretical and practical validation has been applied for the evaluation of WAVES. This implies that the application was subjected to two tests. A purely theoretical test was carried out after the basic knowledge had been implemented. The aim was to experience the abilities of the applied reasoning approach in order to measure the completeness of the conceptual knowledge and to discover what additional expert knowledge would be required. There was no need to employ an automated 'rule-checker', because the amount of rules was not so large that only a computer program could check them. Moreover, the syntax of the rules and their interaction with each other had constantly been checked during the application building process.

Usually, there is no reason to publish the results of a preliminary test which only gives an impression of the state of affairs. Moreover, there is always a chance that somebody uses these figures out of its context. However, these preliminary results were published because this test had an additional aim. I wanted to compare the achievements of WAVES (chapter 5) with those of WARP (chapter 6). Since both applications were based on the same experimental reference collection, they comprised the same knowledge, although in a different format. For WAVES the information had been analyzed, modelled and edited, while WARP had been fed with the original, unmodified data. I was curious whether they would perform comparably. At that time the technique of neural networks was depicted as being superior to that of experts systems in terms of functional abilities and social acceptability (see chapter 6), something which I wanted to verify. I therefore intended to judge the functional superiority by subjecting both applications to the same test. On the basis of the results of this first test, the knowledge base of WAVES was adapted and supplemented with expert knowledge. It was only after the development process had been finished that a second test was executed. This test focused on the practical evaluation of the application. While in the first test the information had been provided by myself, in the second this was done by four independent analysts. The reason to opt for four analysts instead of for one was that the influence of 'the human factor' on the test results was expected to be considerable. This would only be recognizable, however, if several analysts with various levels of experience would participate. In both tests only the analysis procedure of WAVES (see paragraph 5.5) has been involved. One of the reasons for this is that at the time of the first test the hypothesis validation procedure had not yet been developed. The most important reason, however, was that only the analysis procedure could be tested in a fashion that resembles traditional blind tests. Naturally, the rules of the knowledge base of the hypothesis validation module have also been submitted to a theoretical test.

In outline this chapter first illuminates the meaning of blind tests for use-wear analysts and the guidelines that have been proposed for composing and evaluating such tests (paragraph 7.2). Subsequently the two tests introduced in the above will be presented in section 7.3 and 7.4. Both the compositions and the performances will be described. Additionally, the results of the second test, the practical evaluation, will be compared with the achievements of other analysts in order to put them in perspective (section 7.4.5). In this comparison, all blind tests on use-wear analysis that I have knowledge of have been incorporated. Finally, in paragraph 7.5 the findings will be discussed and some conclusions will be drawn about the applicability of WAVES.

7.2 Blind tests in use-wear analysis

Use-wear analysis differs from other specialistic methods in archaeology in that it has been subjected to blind tests almost from the moment of its introduction in the western world. Blind tests are considered to be an important means to evaluate the method and the results obtained with it. The first test was carried out by Keeley and Newcomer (1977) in order to demonstrate the abilities of the high-power analysis method (see also chapter 4, section 4.3). As a reaction, Odell and Odell-Vereecken (1980) initiated a test in which they focused on the possibilities of the 'low-power technique'. Obviously, this evoked other tests, from european (Gendel and Pirnay 1982; Newcomer *et al.* 1986) and american analysts (Bamforth *et al.* 1990), and even an international one (Unrath *et al.* 1986), that alternately confirmed or contradicted previous findings.

By employing this approach, use-wear analysts have maneuvered themselves in a vulnerable position. They had the nerve not only to show the possibilities, but also the limitations of the method and of themselves. The fact that this has not always been in the best interest of this discipline has clearly been illustrated by the sour discussions that were provoked by the test of Newcomer *et al*. (1986). Whereas the participants in the other above mentioned tests had been fairly positive on the assumed validity of the use-wear analysis technique, Newcomer *et al*. uttered severe doubts on the usefulness of the technique. As a consequence, numerous archaeologists became highly reserved towards the achievements that wear-trace analysts accomplished, notwithstanding the bulk of positive results that had already been obtained and were again achieved afterwards and despite the arguments that had put the concerning test in perspective (*e.g*. Moss 1987; Bamforth 1988; Hurcombe 1988).

In spite of the damage that this discussion has done to the method, the tradition of blind testing has also yielded valuable information that contributed to the improvement of the method. An additional advantage of this tradition and the discussions that it evoked, is that it made analysts aware of the influence of the composition of such tests and of the test-conditions on the results. Several analysts argued that bad achievements can, to a certain degree, be ascribed to poor test compositions (Moss 1987; Bamforth 1988; Hurcombe 1988). For instance, the Newcomer test was said to have relied too much on implements that hardly showed diagnostic traces due to short durations of use. It also became clear that a test which consists of only unusual contact materials will yield results completely different from one that consists of general categories of contact materials, and that test results are not only dependent on the applied method, but also on the person performing it.

Consequently, there are now some recommendations for carrying out blind tests, even though generally accepted standards for conducting and evaluating blind tests are still lacking. Due to the fact that the composition of the test set influences the results, one of the recommendations is to publish not only the details of the test composition, but also the complete interpretation of the analysts. Moreover, prior to the rewarding of the interpretations explicit statements are required of what constitutes a 'correct' answer (Bamforth *et al*. 1990: 424). Since some polishes look identical and cannot be interpreted at a high level of specificity, it must be specified how exact the answers must be in order to be accepted. Furthermore, one ought to define in advance the rate of error that is maximally accepted and provide information on the microscopic equipment and chemical cleaning procedures that have been involved. Additional guidelines are that the test tools should only be employed for task-oriented activities, not merely to obtain traces; that they should be cleaned prior to the analyses; that they should be used for more than five minutes (*ibid*.: 414), in order to enlarge the change of interpretable traces. Hurcombe also stressed the need to isolate the interpretations from the observations, for "*Evaluating why correct and incorrect interpretations were made would have enabled us to learn from them*." (Hurcombe 1988: 3).

Apart from the valuable information and the recommendations for carrying out blind tests, another advantage of this tradition is that it has yielded data for comparisons. These may for instance be used for monitoring the progress of students or for validating whether adjustments of the method lead to improved results. Moreover, they could be helpful in putting the results that would be obtained by WAVES in the right perspective. For the sake of comparability, it has been tried to comply with the above mentioned recommendations in testing WAVES.

7.3 The first test

7.3.1 TEST-SET COMPOSITION

WAVES was subjected to the first test after the experimentally obtained knowledge had been implemented. This test was meant to make an intermediate validation of this knowledge and to trace its deficiencies. It was also seen as a perfect opportunity to compare the achievements of an expert system application with those of a neural network application. The timing was considered optimal because both systems were in the same stage of development since they had both been implemented with knowledge that was derived from the experimental programme only. For this reason, these two applications were subjected to exactly the same test.

The test case consisted of 16 replicated flint artefacts that had been used for experimental purposes and of 10 archaeological artefacts of the Dutch Linear Bandceramic site of Elsloo, Limburg. The tools were selected in consultancy with the expert and it was made sure that the wear patterns on these implements were entirely new to both systems: none of them had been used for the composition of the knowledge base of WAVES or for the training of the network. Furthermore, it was decided to apply a test of a rather high complexity. It is not difficult to compose a test that yields optimal results, but that is absolutely not informative. My main interest was to reveal the limitations of the applications. For that reason, implements were selected that displayed traces which the two applications were familiar with as well as tools with polishes that would be difficult to interpret. For instance three experimental tools (346, 378 and 385) were included that had been used on materials of which it was known beforehand that both systems would not be able to identify them. The purpose of the latter was to study the difference in performance between the two applications when they are confronted with unknown situations. A number of the other experimental tools were selected because they displayed slightly different wear patterns although they had been in contact with similar materials, or because they showed less diagnostic traces.

Table 2. The actually worked materials compared with the interpretation of the expert system and the neural network.
* Both systems have not been provided with knowledge about these materials. With reference to the interpretation of the expert system, only those that generated the highest diagnostic value have been included in the comparison.

exp.	worked material	expert system interpretation	neural network interpretation
344	soaked antler	–	dry antler/fresh bone
345	medium hard wood*	hard wood/soft wood	soft wood
346	shell*	–	soft plants
350	soft wood	soft wood	soft wood
351	soaked antler	–	hard wood
352	soft wood	soft wood	dry hide
360	soft wood	–	fresh bone
363	soft wood	–	fresh bone
367	fresh hide	fresh hide	fresh hide
370	fresh hide	fresh hide	fresh hide
371	fresh hide	fresh hide	fresh hide
378	hide with ochre*	–	soft wood/dry antler
383	soft wood	soft wood	soft wood
385	dry clay*	soaked antler	soaked antler/soft wood
386	fresh hide	fresh hide	fresh hide
388	dry bone	butchering	fresh bone/dry antler

The procedure of the test consisted of two steps. First, the characteristics of the wear-traces were described by experienced analysts.[2] The reason for this is that the test was intended to validate the knowledge rather than the applications practical usability. By using experienced analysts, the possibility could be excluded that bad achievements could be caused by a user's lack of experience, something which was very well possible. Subsequently, the descriptions of these analysts were presented to both systems by myself. This also prevented that the achievements would be influenced by a user's lack of experience in working with the two knowledge-based applications.

Due to the fact that WARP and WAVES had only been trained to interpret polishes, this test exclusively focused on the interpretation of this wear category. For the same reason neither the applied motions or the relative hardness of the worked materials were included. With regard to the rewarding of the obtained interpretations two different methods were followed. Since the contact materials of the experimentally used tools were known, the interpretations concerning these tools could be evaluated as a 'blind test'. However, the interpretations concerning the prehistoric polishes were more difficult to evaluate because the worked material could, of course, not be known with certainty. Therefore, these results were compared with interpretations that a professional human use-wear analyst had given prior to the test.[3] Hereby, the assumption was that in case of dissimilar interpretations, those of the human expert would be considered correct.

7.3.2 THE EXPERT SYSTEM'S ACHIEVEMENTS

In table 2 the results regarding the experimentally used artefacts are presented. WAVES could not identify the traces of 6 tools (344, 346, 351, 360, 363 and 378) and, therefore, refrained from giving an interpretation (see section 5.7.3) However, of the 10 interpretations that it could give, only one was incorrect (tool 385). In two other instances (tool 345 and 388), the system's suggestion of the applied material was acceptable because it approached the right answer sufficiently. In some cases this can be justified because it is known that different activities can cause similar traces.

The number of missing interpretations (six) concerns a rather large part of the test-set, but is not very surprising regarding the composition of the test. There are several reasons responsible for this. Firstly, the traces that were analyzed deviated from the traces that the system had knowledge about. For instance, tool 378 had been used on hide with ochre while this combination had not been included in the experimental programme. Moreover, some of the other artefacts showed combinations of wear-characteristics that had not been experienced with the artefacts of the experimental programme either. This is inherent to the fact that this knowledge is derived from experimentally obtained traces. An experimental programme cannot contain the entire range of traces that may occur archaeologically. It has often been experienced that some traces cannot be replicated with experimental tools even though they occur frequently on archaeological tools. An example of this is the so-called polish '23' (Van Gijn 1989: 85). This type of polish (bright, plant-like on one side, hide-like on the other) has been observed by several other analysts (Keeley 1977; Cahen *et al*. 1986; Juel Jensen 1989, 1994), though its origin has not yet been discovered by means of experiments. Since only the human experts have knowledge about the variability of the traces that the archaeological record exhibits, this kind of expert knowledge had to be incorporated in WAVES as well.

Tool nr.	analyst' interpretation	expert system interpretation	neural network interpretation
1	dry hide	–	fresh hide
3a	dry hide	–	fresh hide
3b	bone	butchering	butchering
5	hide?	fresh hide	fresh hide
6	bone	–	butchering
10	fresh hide	–	fresh hide
19	wood	hard wood/soft wood	hard wood/soft wood
20	fresh hide	fresh hide	fresh hide
31	hide	–	fresh hide
34	antler	soaked antler	soaked antler

Table 3. Interpretations of polish on 10 Linear-band-ceramic artefacts, given by a human analyst, the expert system and the neural network.

A second reason for missing interpretations was thought to be due to the subjective nature of the variables that are used to describe the wear-traces. Most of the descriptions are based on relative 'measurements'. It is, for instance, difficult to decide whether a polish looks 'bright' or 'very bright'. This implies that, even though experienced analysts were involved, the descriptions of the wear characteristics given by the analysts do not always match those given by the expert and on which the system is based. Therefore, this may cause discrepancies between the descriptions, yielding information the system cannot interpret correctly.

The incorrect interpretation of tool 385 can also be ascribed to a discrepancy in the knowledge base. This implement had been used for an experiment (carving dried clay) that had not been included in the experimental programme. The fact that the system did come up with an interpretation means that, according to the system, the observed traces showed a resemblance with those caused by working soaked antler. For a use-wear analyst this may be a strange misinterpretation, but it can be explained by the fact that the observed traces coincidently resembled those experienced on another implement of the experimental programme. This artefact had been used on soaked antler, but showed non-diagnostic wear-attributes that resembled those on the implement that had been used on the dried clay.

The results concerning the analysis of the prehistoric polishes (table 3) were less easy to validate than those relating to the experiments, because the correct interpretations were unknown. However, 50% of the application's suggestions turned out to be in accordance with the answers that the human analyst had given. This included, however, the answer that WAVES gave with respect to the traces on tool 3b. Since bone working and butchering may cause similar traces, this answer was accepted. Although no misinterpretations were given, again a large percentage did not lead to any suggestion at all. Despite these lacunae the results were considered promising. The failures were ascribed to the insufficient amount of knowledge of the application.

7.3.3 The neural network's achievements

A major difference between an expert system and a neural network is that the latter will always generate an answer, even if it is an unsure one.[4] In case it cannot find an exact match, a neural network simply searches for examples of contact materials from which the traces come closest. This explains why the network made more mistakes in interpreting the experimentally obtained polishes (table 2). Most of these mistakes concern exactly those tools (344, 346, 351, 360, 363 and 378) that WAVES could not identify either, but since WARP tried anyhow, it failed more often. In some instances such an 'educated guess' gives a correct indication of the relative hardness category of the worked contact material, but in other cases it not always yields correct answers. The problem with these guesses is, however, that you will never know which answers are reliable. Moreover, the reason for misinterpretations cannot be traced and explained, because the reasoning process of neural networks is invisible.

Despite some unfortunate guesses, WARP performed rather well. It interpreted the traces of six tools exactly correct (350, 367, 370, 371, 383, 386). Since the system has no output neuron for medium hard wood, only for hard wood and for soft wood, I also rewarded the interpretation of tool 345. In two other cases (tool 344 and 388) the interpretation was almost correct, but rejected anyway. This decision may be doubted, especially because in the case of the expert system application 'butchering' was rewarded when it concerned 'dry bone'. Whatever the decisions on these instances should have been, they demonstrate that the network had some difficulties in separating the traces of similar materials, like those of bone and antler working. It must also be stressed, however, that this is not surprising since professional analysts may have difficulties with this as well. With respect to tool 385, it is remarkable to notice that, like the expert system, the network interpreted the traces that were caused by carving dried clay (385) as originating from soaked antler. This implies that the observed traces must

	expert system		neural network	
experimental replica's (N=16)				
correct	9	(56.3%)	7	43.8%
incorrect	1	(6.3%)	9	(56.3%)
no interpretation	6	(37.5%)	0	
archaeological artefacts (N=10)				
correct	5	(50.0%)	8	(80.0%)
incorrect	0		2	(20.0%)
no interpretation	5	(50.0%)	0	
total (N=26)				
correct	14	(53.8%)	15	(57.7%)
incorrect	1	(3.8%)	11	(42.3%)
no interpretation	11	(42.3%)	0	

Table 4. Final comparison of the results of the first test.

indeed have been comparable with those caused by working the soaked antler.

Regarding the archaeological artefacts (table 3), the network's interpretation was similar to that of the human analyst in no less than eight cases (3b, 5, 6, 10, 19, 20, 31 and 34). Once more this includes a case in which 'butchering' was judged positively whilst it (presumably) concerned traces of bone working. The suggestions concerning two fresh hide working tools (1 and 3a) were not rewarded, but again this is disputable because they were not absolutely false.

7.3.4 CONCLUSION

From a comparison of the achievements (table 4), it can be concluded that in reference to the experimental tools WAVES performed slightly better than WARP, whereas the opposite is true for the interpretations of the archaeological implements. The reason for this is not very clear. It may pertain to the composition of the test-set, because the replicated tools displayed relatively more wear-patterns that are not very diagnostic, whereas the archaeological tools contained relatively more diagnostic patterns.[5] Expert system applications, assuming that they have been provided with the appropriate knowledge, may be better in interpreting exceptions, *i.e.* in extrapolating, than neural networks. When interpreting data, the latter focus on recognizing similarities with the examples that they have learned. They try to relate new data and thus also exceptions to their generalized knowledge. Therefore, they can only interpret exceptions correctly if they have been provided with enough 'learn examples'. Unfortunately, the difficulty with exceptions is that the examples are not abundant. However, when it comes to real exceptions that occurred never before, the expert system will not be able to give an interpretation. It will simply lack the appropriate knowledge. A neural network, on the other hand, might be able to give an interpretation that is in the right direction (for example the right hardness category).

From the results it can also be concluded that both systems can be useful if a human analyst wants a second opinion on his interpretation. For example the analyst was uncertain about the traces on tool number five, but both WAVES and WARP confirmed the interpretation. An argument, however, that favours the first is that, in contrast to the neural network, its achievement on the replicated tools was not different than on the archaeological ones. It performed consistently.

The final conclusion of this comparison was that both applications performed already quite well, especially considering their stage of development and the fact that they were based on a rather small and unbalanced set of examples. The expert system interpreted 54 percent (14 out of 26 tools) correctly and the neural network 58 percent (15 out of 26 tools). This seemed to favour the latter, but if the total number of false interpretations is taken into consideration, the opposite is true: 3.8 percent in case of the expert system versus 42.3 percent in case of the neural network. From this it can be concluded that none of the techniques performed absolutely better than the other. Therefore, I disagree with Gibson on the supposed functional superiority of neural networks (Gibson 1992: 265). At most it can be concluded that the one approach serves particular purposes better than the other (Van den Dries 1993). But, this does not seem to be determined by its achievements but rather by the principle of the approach.

One other thing that the misinterpretations illuminate is the problem of identifying non-diagnostic wear patterns. Such a problem certainly shows one of the limitations of expert

experiment*		tool type	activity	duration in minutes
1	(2)	blade	cutting roots (turnip)	20
2	(53)	flake	butchering meat (roe deer)	15
3	(96)	scraper	scraping soaked antler (reindeer)	15
4	(110)	blade	carving fresh bone	26
5	(120)	blade	reaping cereals (emmer)	30
6	(186)	flake	cutting dry grass	30
7	(197)	scraper	scraping fresh hide (hare)	60
8	(226)	scraper	scraping fresh hide (elk)	60
9	(297)	blade	butchering fish (rudd)	35
10	(352)	waste (block)	splitting soft wood (willow)	20
11	(367)	scraper	scraping hide (swine) with flower	115
12	(385)	point	carving dry clay	20
13	(383)	quartier d'orange	scraping soft wood (birch bark)	20
14	(388)	point	carving fresh bone	45
15	(346)	retouched blade	sawing shell	10

Table 5. Composition of the blind test set. The numbers between parenthesis refer to the numbers that the experiments have in the reference collection.

systems. If a situation or problem differs too much from those from which the knowledge was derived, a system might be unable to deal with it. Even though some similar problems may be prevented by expanding the application with expert knowledge and by enlarging the experimental programme, no system will ever have sufficient knowledge to exclude all such misinterpretations. Non-diagnostic wear and especially generic weak polish may simply be hard or impossible to interpret.

7.4 Second test

7.4.1 INTRODUCTION

On the basis of the results of the first test, the knowledge base of WAVES was refined and supplemented with expert knowledge. This broadened the range of the wear patterns that it is able to recognize. It was only after the entire development process had finished that a second test was carried out. In this test WARP was not included, since this test was meant to be the final evaluation of the application before it would become operational. The network has not been adapted on the basis of the results of the first test, because this prototype had merely been made for a comparison of both techniques. Moreover, within the scope of this study it was not intended to develop an operational neural network as well.

Since the second test would be the practical evaluation of WAVES, it had to be carried out by independent analysts and students rather than by myself or the involved expert. The aim was to find answers to questions like:

* What success rates can be obtained when the application is employed by students or by analysts who were trained by other experts or originate from different methodical schools?

* To what degree can WAVES substitute the support of the human expert in training students?

* Are the results that are obtained by WAVES comparable to those of other human analysts that participated in blind tests?

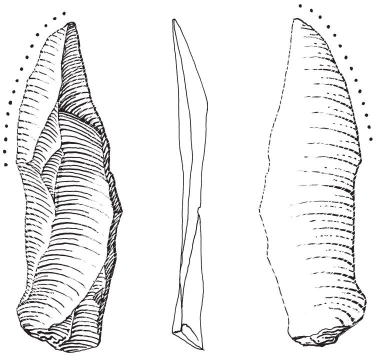
* How do students appreciate the application: will they accept it as an initial tutoring system?

* To what degree is it acceptable to have students working with the system without the help of the expert and without a basic introduction into use-wear analysis?

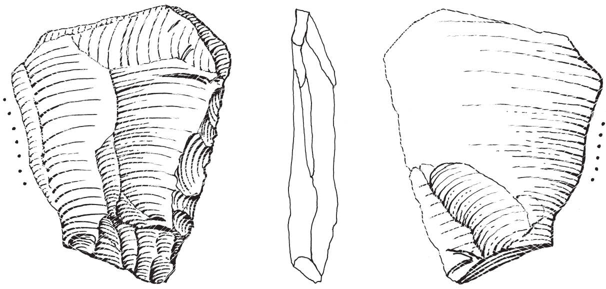
7.4.2 TEST-SET COMPOSITION

In order to be acceptable as a final practical validation of the system's abilities and usefulness, the test had to comply with various criteria. For instance, it had to gear to the situations and circumstances that may be encountered in educational environments. This implied that a broad range of activities and traces had to be involved: not only tools with diagnostic traces, but also slightly developed wear. Moreover, the traces had to be different from those the knowledge was deduced from and the test had to be carried out by analysts that had not been involved in the development process, *i.e.* they had to answer the profile of a future user.

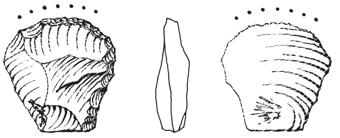
It was experienced in the first test that if different interpretations are encountered, it is impossible to validate them and to decide whether that of the user or of the application is most likely. Therefore, this second test contained exclusively experimentally obtained use traces. It consisted of 15 tools (table 5, fig. 44), but in order to avoid the association of particular tool forms with specific activities it was tried not



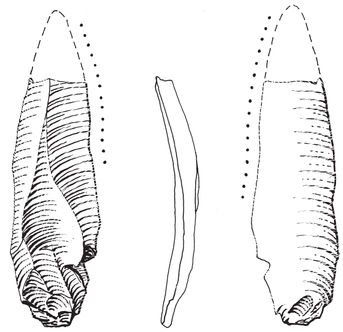
1



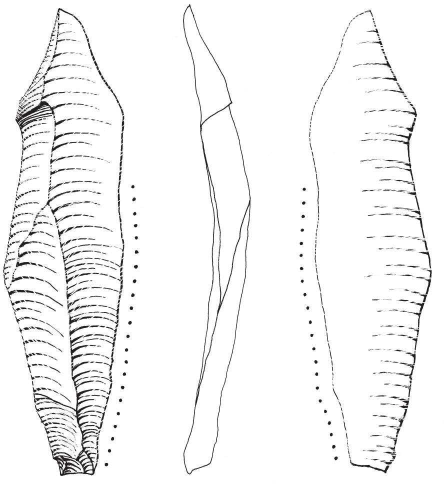
2



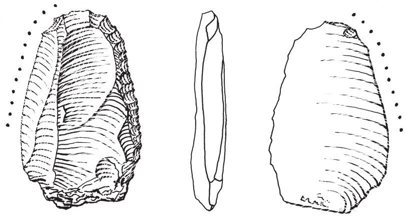
3



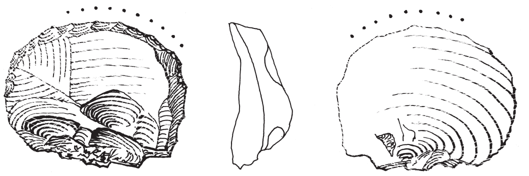
4



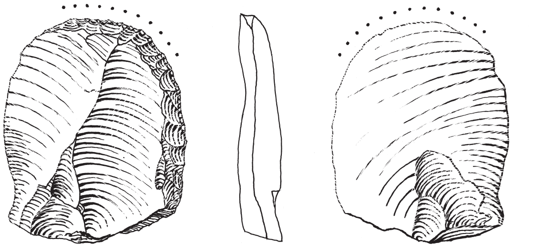
5



6



7



8

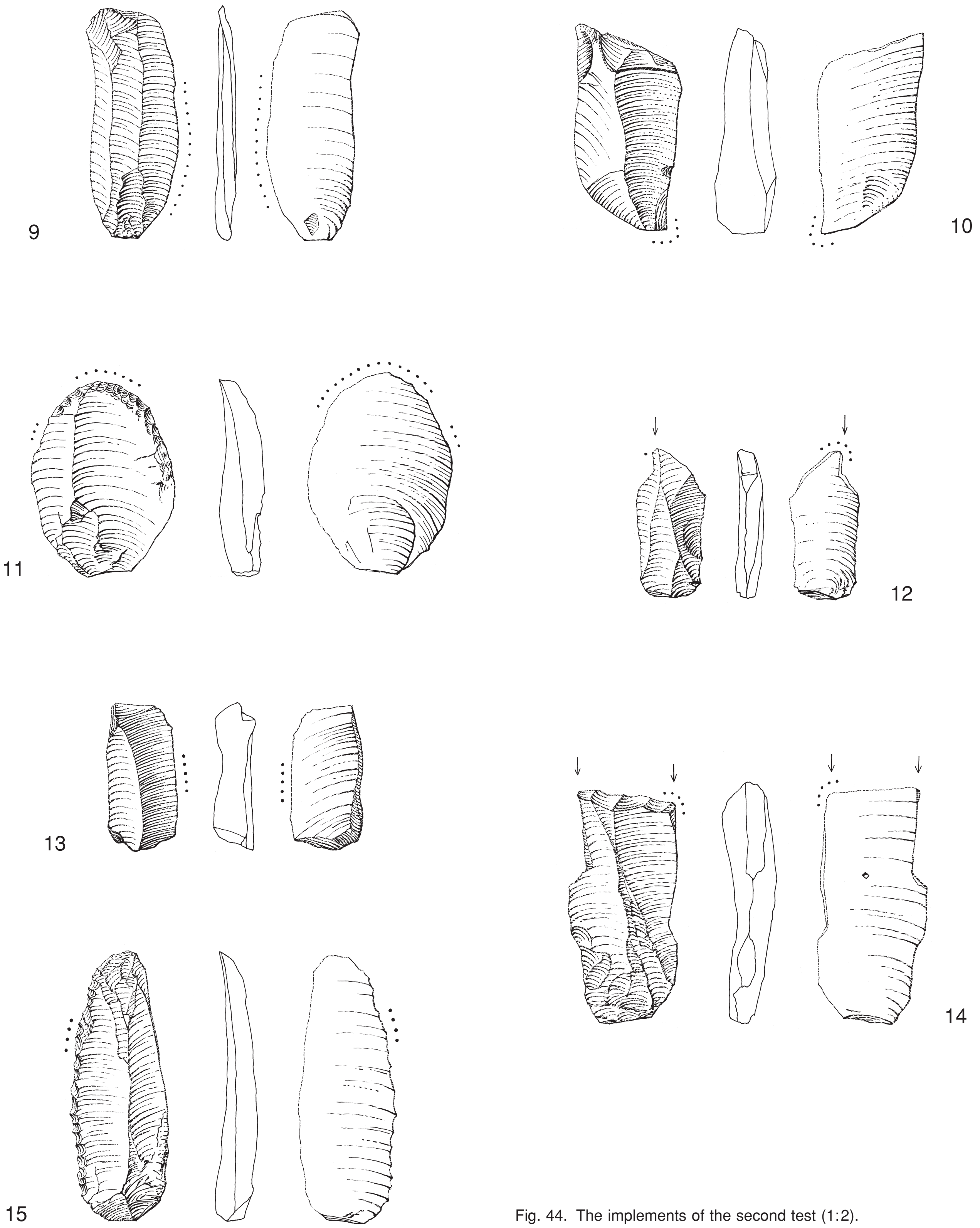
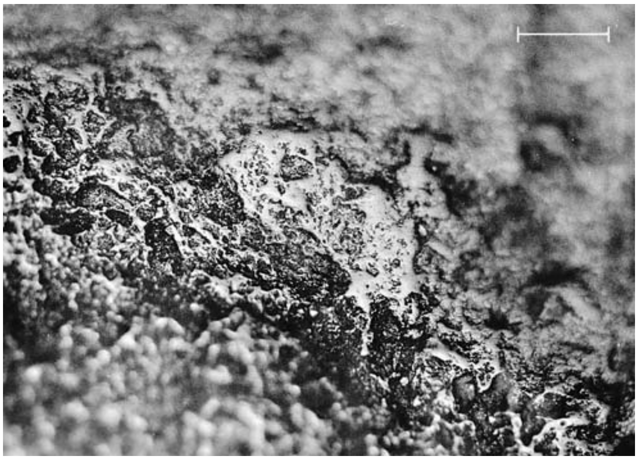
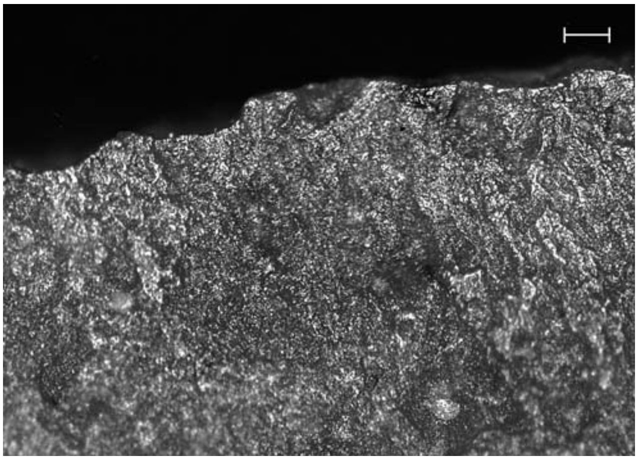


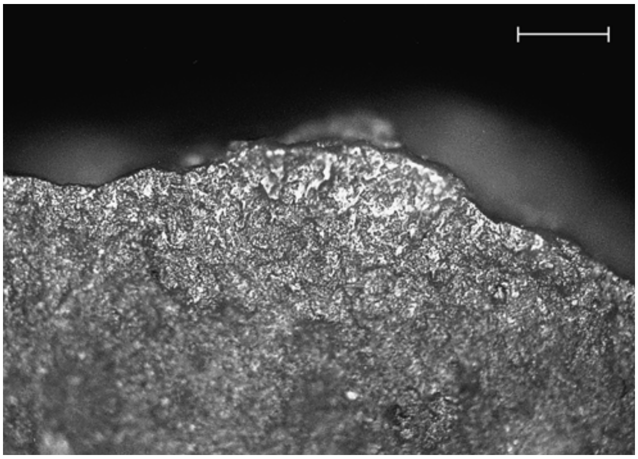
Fig. 44. The implements of the second test (1:2).



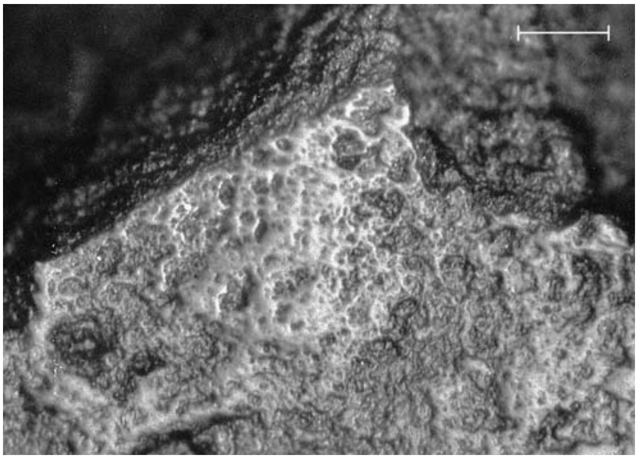
experiment 1



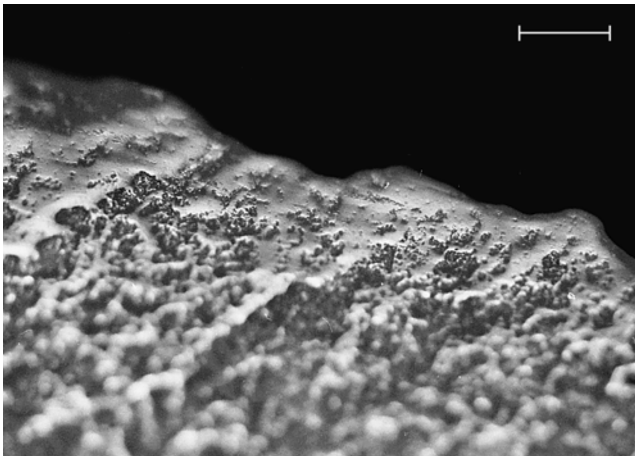
experiment 2



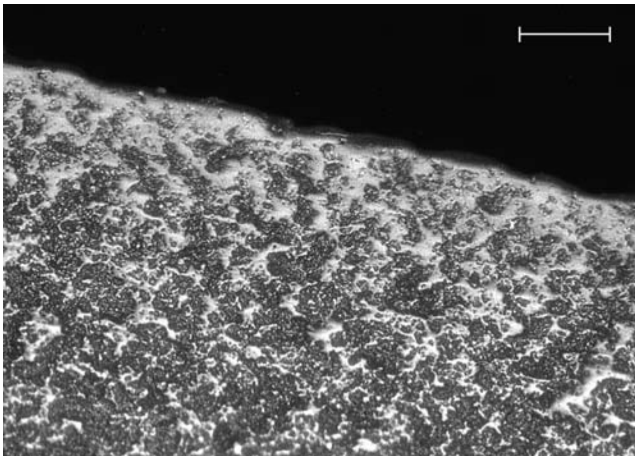
experiment 3



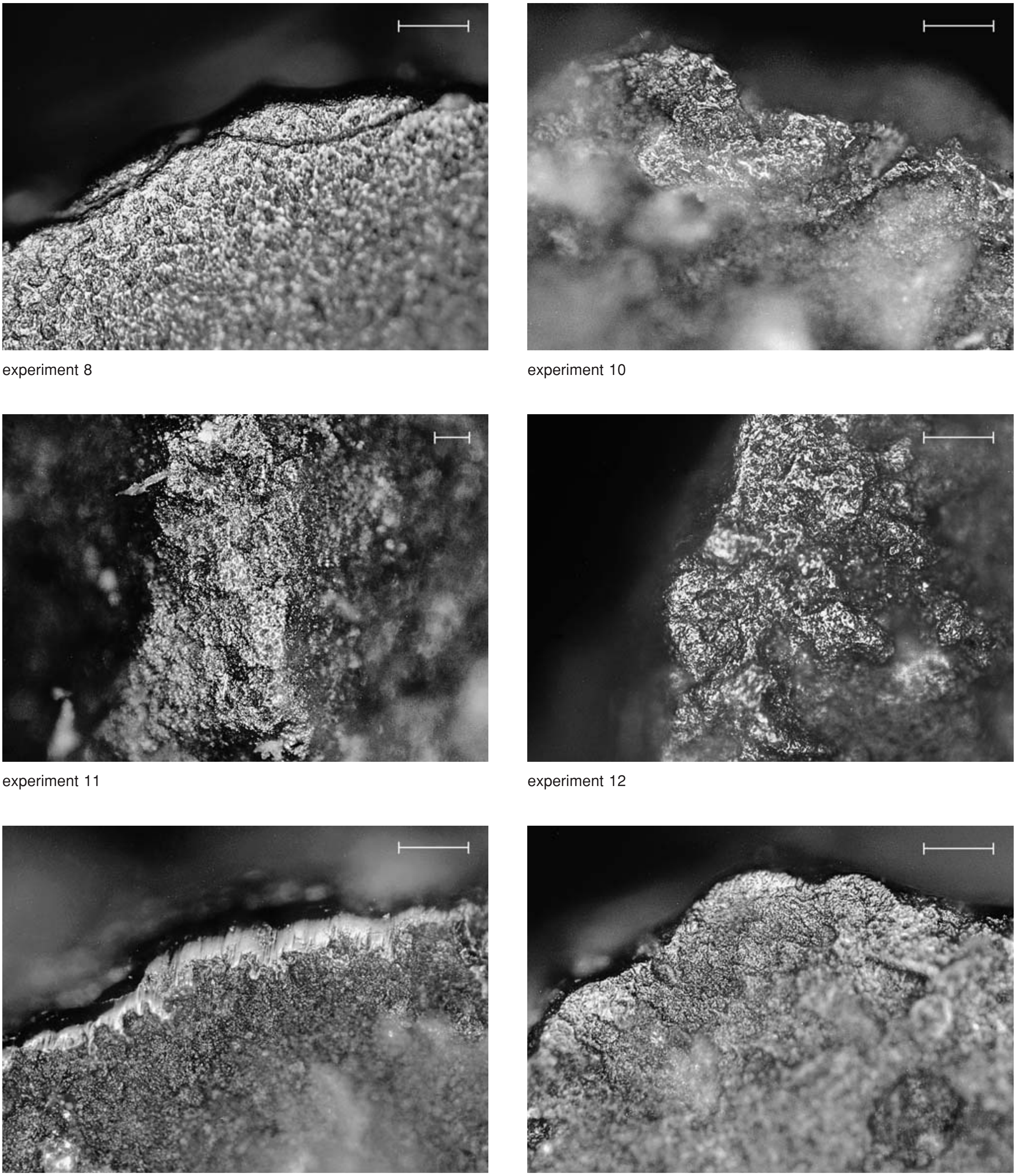
experiment 4



experiment 5



experiment 6



experiment 8

experiment 10

experiment 11

experiment 12

experiment 14

experiment 15

Fig. 45. Wear traces observed on the implements of the second test. (All scale bars equal 50 micron)

to select artefact types that are characteristic for a certain prehistoric period. With reference to the activities that had been involved in the experiments, the worked materials were not limited to a particular prehistoric period either, only to temperate Europe. This corresponds with the extent of the knowledge in WAVES.

Moreover, all tools had been employed in a realistic and task-oriented fashion and the inclusion of problematic traces, like traces caused by multiple use, hafting, trampling or obliterated by curation or post-depositional surface modifications, had been avoided. Nevertheless a wide range of contact materials was involved, including some that the analysts may have less experience with such as dry clay and shell. All tools had been used, and none had been used for less than five minutes. It had been made sure that all artefacts showed sufficient and interpretable traces (fig. 45). The numbers 1 till 9 had been part of the reference collection from which the basic knowledge for WAVES was derived, while the others (10 till 15) had been carried out after the knowledge base had been composed. It was no coincidence that the latter had also been part of the first test. As half of them could not be interpreted then, it would be interesting to see how the application would react to them now.

After the experiments were finished all tools were cleaned according to a standard procedure: in order to remove tissues from the contact materials they were first put in an ultrasonic tank for 5 minutes in a HCL (3.6%) solution, then they were washed with water and subsequently soaked in a weak KOH solution.[6] During the test the analysts were only allowed to clean the pieces with alcohol in order to remove grease left by handling and the remains of plasticine by which the tools are placed in position under the microscope.

The test-team was rather heterogeneous. Two of the analysts came from different methodical schools. One of them was a French student with quite some experience, although not with the method of description that is employed at Leiden University. Another volunteer came from Australia and was more experienced in residue analysis than in the kind of wear trace analysis WAVES focuses on. The two other analysts were both students from Leiden University: one had already been instructed by our expert and had some experience with wear analysis, the other had never observed wear traces before and started the test completely ‘blank’. The latter was added to the team in order to get an impression of the support WAVES would be able to give a fresh student and to validate the educational value of WAVES. No member of the team had ever worked with an expert system application.

Logistically it was impossible to have all analysts perform the test simultaneously, so they carried it out one after the other: three at the laboratory of Leiden university, one at his ‘own’ laboratory. Although the tools had been packed individually to avoid damage during transport, this could not prevent one of the tools (exp. 4) from breaking. It was decided that it could stay in the test set because the remaining part showed sufficient interpretable traces. Once the test had begun, there was no communication between the analysts, because they either did not know each other or were unaware of each others participation. Nevertheless, by way of precaution the original numbers of the experiments had been replaced by new ones. Since half of the test tools were derived from the experimental programme which had already been published (Van Gijn 1989) it had to be prevented that the analysts would be tempted to verify some details in the publication.

Before the test began, the specifications concerning the parameters of the experiments were communicated to the analysts. Therefore, they knew that the experiments had not been based upon a particular cultural framework and that they could not rely on form-function relationships. Since it was not a speed contest and it was considered more important to obtain a qualitative good interpretation which takes a long time than a fast answer that is incorrect, the analysts did not get a restriction on the amount of time that was available for one analysis. On the average, they needed approximately half an hour to examine each stone.[7]

The analysts were asked to give a personal interpretation before that of the application was known.[8] They were allowed to give more than one answer or none at all. The purpose of this was not to compare the achievements of one analyst with another, but to study whether the interpretation of the more experienced analysts would be influenced by the analysis method followed by WAVES, or whether they would come up with totally different interpretations.

Despite the fact that three nationalities were involved, at the end of the test the descriptions and interpretations were fully comparable (see appendix IV). Terminological difficulties were avoided by making sure that the analysts described their observations according to the approach of WAVES. Each analyst was provided with paper copies of the interpretation screens which enabled them to record their descriptions and the subsequent answers of the application in a standard manner. This did not prevent all translational difficulties, however. It was experienced that it took some time, especially for the French analyst, to become acquainted with the method of description WAVES employs. Therefore, the French analyst was initially supported in using the application. Especially the explanation of the meaning of some of the variables and attributes needed a little extra attention.

Since this test was not meant as a validation of the knowledge base anymore, this aspect had to be excluded from the test. Therefore, it was considered important that wear patterns were selected of which it was known that WAVES would be able to interpret them. For this reason the test has been composed of implements from the experimental programme and from the first test. Only these circumstances would make it possible to attribute misinterpretations to biases in the practical functionality rather than to deficiencies in the knowledge.

It was also for this reason that I carried out a control test before the tools were handed over to the analysts. This control test means that a description was gathered of the traces on all tools and that these were presented to the application. The descriptions were given by the expert and by two experienced analysts that were trained by her.[9] It was made sure that they described the traces according to the method that is used by WAVES. Subsequently, both the descriptions and the interpretations that were obtained from WAVES could be used as a standard against which the recordings and the inferences of the participants would be compared. Additionally, this standard description could serve as a means to trace the reason for misinterpretations that would be obtained from WAVES by the analysts. Apart from this control test an additional check of the results has been carried out. Since both the descriptions and the interpretations were reported by the analysts, their analyses could be repeated. This has indeed been done in order to rule out that interpretational mistakes were due to input mistakes. It was only in one or two cases that minor discrepancies were detected and the interpretations were almost 100% consistent.

The above described test may seem rather ordinary, but it must be stressed that it differs at various points in comparison with traditional blind tests. First of all it was not intended to compare the achievements of expert analysts or to assess the methodical aspects of the analysis. On the contrary, it was meant to validate the system's achievements when it is employed by (inexperienced) human analysts. Therefore, the test team did not only consist of experienced analysts but predominantly of students from different levels. Another difference with previous tests is that the analysts were told which part of the artefact had been used for the experiment. Normally, locating the traces is an integral aspect of a blind test. In this case, however, it was preferred to gather results that would be optimally comparable rather than polluted with wrong descriptions and therefore wrong interpretations. Since there was a rather large chance of wrong description due to the lack of experience of the users, this chance was reduced as much as possible.

A third difference is that the analysts had to describe their observations using the terminology that WAVES provides.

A final difference concerns the composition of the interpretations. With other blind tests wear analysts usually base their interpretations on the entire pattern of the wear traces. This has also been the case with the personal interpretations that the analysts gave in our test (appendix V). WAVES, however, has explicitly been designed to analyze the polish features independent of the use retouch and edge rounding (see chapter 5). This separation also underlies the results of our test. The deduction of the exact contact material is based on the polish features and the relative hardness on the edge rounding and the use retouch. The reconstruction of the applied motion consists of two components: one is based on the characteristics of the polish, the other on the retouch and rounding.

7.4.3 ACHIEVEMENTS

The interpretations that the participants obtained from WAVES have been rewarded on the basis of a comparison with the responses of WAVES to the standard description.[10] Prior to this evaluation, however, it was decided that the interpretations would be rigorously judged. One reason is that the artefacts displayed enough characteristic traces to enable accurate answers. The other reason, however, was that the test did not intend to assess the achievements of the human analysts, but of the computer application.

With respect to the exact contact material, an interpretation was only considered correct if the applied material actually received a diagnostic value, whether this was the highest score or not. In chapter 5 it has already been explained that the actually worked material not always receives the highest diagnostic value, because not all wear traces are very diagnostic. It is shown in table 11, 12, 13 and 14 how many of the rewarded answers received the highest diagnostic value, the second best, the third best or less. Furthermore, an interpretation would be rewarded if it resembled the conclusions on the standard description. However, inferences that seemed to be in the right direction but did not mention the exact material were not rewarded. For instance if an antler working tool was reconstructed as a bone working tool, and 'antler' was excluded from the interpretation, then this was not accepted as a correct answer.

The criteria for rewarding the interpretations on the relative hardness of the worked material slightly deviated from those on the exact contact material. For instance, it was decided that only the hardness category with the highest diagnostic value would be taken into account rather than the whole of the interpretation. The reason for this is that the interpretation can consist of only three possibilities. By handling the criterion that an interpretation is correct if the right hardness category is part of it, it would be too easy to achieve perfect results. Moreover, by taking only the hardness category with the highest value into account, the results of this test would be comparable with other blind tests.

This did not mean, however, that the criteria for rewarding the interpretations were less difficult to establish. In fact, it was even more complex than with the evaluation of the interpretations of the exact contact material. First of all, the materials that had been involved in the experiments could not easily be categorised into the three hardness classes (soft, medium, hard) that were distinguished, because it was impossible to apply objective, measurable means. Consequently, the dividing lines between the hardness classes were rather diffuse. Secondly, some materials turned out to cause other wear traces than was expected on the basis of their resistivity. For instance, materials that seemed to be rather resistant, still caused edge damage that was thought to be typical for medium hard materials.

Unlike those of the exact contact material and the relative hardness, the interpretations of the applied motion were more easy to assess. Since the differences between the motions are distinct, dubious decisions did not occur. Again, for the sake of comparability, it was decided to reward only the suggestion with the highest diagnostic value. However, if two motions — or two hardness categories — received an equal value, then they were both considered correct.

The achievements of the analysts on the various aspects are shown in table 6, 7, 8 and 9. Despite the fact that the interpretations were validated by means of formal criteria, in some instances it was still difficult to make the right decision. Since some of the decisions require explication, a summary of the descriptions of the analysts and the subsequent interpretations of WAVES is given in the remaining of this paragraph. This will give an indication of how the rules have been applied, of the discussions that accompanied some of the decisions, of the grounds for some of the decisions, but mainly it is meant to illustrate the difficulties which the analysts encountered and to trace the cause of the misinterpretations. The complete recordings of the analysts and the subsequent interpretations of WAVES are given in appendix IV and V, respectively.

Experiment 1

The first tool was rightaway one of the difficult ones. It is an unmodified blade which had been used for 20 minutes for cutting turnips, but which also shows some soil wear. These roots were classified as non-siliceous plants. The relative hardness was considered medium because the material had been more resistant than soft plants. The artefact showed well-developed traces: several edge removals, slight edge rounding and a considerable amount of polish.

Contact material: Already with the first test piece the descriptions of the analysts differed considerably, especially regarding the distribution, the topography and the width of the polish (see appendix IV). It is therefore not astonishing that exclusively the recordings of analyst III led to a correct conclusion. This is remarkable because it was this student's first piece to describe. He had never used a microscope before. Analysts I and II did not get any interpretation at all. Their descriptions did not match any of the wear patterns that WAVES has knowledge of. The former indicated that the polish was distributed in a band away from the edge. However, this does not correspond with a polish width of 5001 to 10.000 micron (class g), which is 0.5 to 1.0 cm. Personally they thought of harder materials (hard wood and antler). Analyst IV was very close with his personal inference as he assumed that the tool had been used on siliceous plants. This deviation may be explained by the difference between the silica contents of plants from our hemisphere and from Australia, with which the analyst was more familiar. His characterisation of the distribution of the polish as 'reticulated' caused the application to exclude the plants and to decide in favour of 'soft wood'. Although it was in the right direction, this answer was not rewarded (see table 6) because it did not include the plants.

Hardness: On 2 out of the 4 descriptions of the use retouch and edge rounding, the application confirmed that the tissue of the worked material was medium hard, although it assigned identical values to both 'medium hard' and 'soft' in the response that analyst II acquired. The wear recordings of analysts I and IV caused a preference for 'soft material' due to the fact that they indicated that the retouch was predominantly of the feathered type. This slightly deviated from the observation of the expert, since she had discovered some hinge terminations as well, which are indicative for more resistant materials. The reason that analyst II obtained a slightly higher diagnostic value for 'medium' than I and IV, is that he characterised the distribution of the retouch as 'close'. This caused WAVES to assign a bonus value to 'medium'.

Motion: In their personal inferences, all analysts correctly assumed a longitudinal motion, but with WAVES they could not obtain the same conclusion in all instances. For instance, the interpretation that analyst I got on the basis of the use retouch favoured a transverse motion. Furthermore, WAVES deduced on the description of the polish features by analyst II and IV also a transverse motion. It is remarkable, however, that with these three analysts the second component of the interpretation was correct. This means that all three of them obtained contradictory inferences on the applied motion. This illuminates one of the difficulties that the user of WAVES may be confronted with. Especially inexperienced analysts may have a problem when they are confronted with such conflicting results, because they probably cannot decide in favour of one of the two. Whenever this occurs, however, the user ought to take it as a warning and ought to conclude that he or she probably gave a conflicting description. In the case of analyst I for instance, the contradictory deduction was caused by the fact that she defined the orientation of the retouch as predominantly perpendicular, which WAVES considers to be a strong indication of a transverse motion, whereas she characterised the orientation (directionality) of the polish as predominantly parallel.

Table 6. The results the analysts obtained with WAVES in tracing the applied contact material. (1=correct answer, 0=incorrect answer)

exp.	activity	standard	I	II	III	IV
1	cutting roots (turnip)	1	0	0	1	0
2	butchering meat (roe deer)	1	0	1	1	1
3	scraping soaked antler (reindeer)	1	1	1	0	1
4	carving fresh bone	1	0	0	1	0
5	reaping cereals (emmer)	1	1	1	1	0
6	cutting dry grass	1	1	0	1	0
7	scraping fresh hide (hare)	1	0	1	0	0
8	scraping fresh hide (elk)	1	1	1	0	0
9	butchering fish (rudd)	1	0	0	1	1
10	splitting soft wood (willow)	1	0	1	0	1
11	scraping hide (swine) with flower	1	1	1	0	1
12	carving dry clay	1	0	0	0	0
13	scraping soft wood (birch bark)	1	0	1	1	0
14	carving fresh bone	1	0	1	1	1
15	sawing shell	0	0	0	0	0
	Total	14	5	9	8	6
	%	93.3	33.3	60.0	53.3	40.0

Table 7. The results the analysts obtained with WAVES on the interpretation of the hardness category. (1=correct answer, 0=incorrect answer, – = not applicable due to absence of wear indications)

exp.	relative hardness	standard (N=14)	I (N=14)	II (N=15)	III (N=15)	IV (N=15)
1	medium hard	1	0	1	1	0
2	soft/medium hard	1	1	1	1	1
3	medium hard	1	1	1	0	1
4	medium hard	1	0	1	1	0
5	medium hard	1	1	1	0	1
6	soft	1	0	0	0	1
7	medium hard	1	1	1	0	1
8	medium hard	1	1	1	1	1
9	medium hard	1	0	1	1	1
10	medium hard	1	–	1	1	1
11	medium hard	1	1	1	0	1
12	medium hard	1	1	1	1	1
13	medium hard	1	0	0	1	0
14	medium hard	1	0	1	1	1
15	hard	–	0	0	0	0
	Total	14	5	9	8	6
	%	93.3	33.3	60.0	53.3	40.0

Moreover, the edge was said to be convex rather than straight, which makes WAVES favour a transverse motion as well.

Experiment 2

This tool is an unmodified flake which had been used for butchering deer for 15 minutes. With this activity contact with the animals bones had occurred occasionally. It was expected that this would be a problematic piece, because the traces were not abundant and not particularly diagnostic. The correct answer would be 'meat and fish', which is synonymous for butchering in WAVES, and both 'medium' and 'soft' would be accepted as the relative hardness category because of the fact that the analysts might either decide to describe the traces that are characteristic for bone or meat working.

Contact material: The results of the analysis of the contact material are far better than was expected: except for that of analyst I, all descriptions led to a conclusion that included meat and fish. Since the traces were not very distinctive, it is

not astonishing that this category did not receive the highest value. Analyst II managed to get exactly the same conclusion as the standard description but with even better diagnostic values.[11] The description of analyst I was found to be indicative of wood: 'meat and fish' was excluded due to the characterisation of the polish topography as 'domed'. The interpretation did consist of both soaked and dry antler, which is said to be indistinguishable from bone (*cf*. Vaughan 1985: 31-34, 45-46). Nevertheless, this was not rewarded because the interpretation was far too heterogeneous and did not include bone at all. It is remarkable though that all four answers included a vegetal component (soft wood). Of the personal interpretations only that of analyst IV was exactly correct. Analyst II recognized traces of bone working but unjustly thought they were caused by a transverse motion.
Hardness: Since both 'soft' and 'medium' were accepted the success rate was optimal.
Motion: The movement that was involved in this experiment turned out to be difficult to discover: only 4 out of 8 interpretations turned out to be correct (see table 8 and 9). Although it should have been 'longitudinal', analyst II had personally been thinking of a transverse motion and this was also the conclusion of the application on the basis of his description. Analyst I had no personal idea and the motion WAVES inferred showed an absolute contradiction between the one that was based on the description of the micro traces and that of the macro traces. This time, the indication of the perpendicular retouch orientation led to the wrong answer. Analyst IV obtained no interpretation on the macro traces but a correct one on the micro traces. The reason for the former is that the retouch distribution was indicative of a dynamic and perpendicular motion, which conflicted with all other features.

Experiment 3

Tool number 3 is a small retouched scraper that had been used for scraping soaked antler during 15 minutes. The relative hardness category was considered medium hard.
Contact material: Although the traces of antler working are said to be difficult to distinguish, WAVES could deduce the right conclusion on 3 out of the 4 descriptions. In one instance 'antler' was the sole suggestion, but in two others it did not get the highest diagnostic value. In these instances WAVES was rather persistent that the traces were more diagnostic of hard wood. This can be explained by the fact that both analysts (I and IV) described a bright polish with a smooth/matt texture and a domed topography, which is both observed on implements used for wood working and for antler working. Surprisingly, the blank student (analyst III) completely failed on this tool: despite WAVES' warnings, he described the intentional retouch on the dorsal face as if it was caused by use and none of his characterisations of the polish were related to antler working either. Analyst I and II both gave perfect personal interpretations. Especially in the case of analyst II this is highly remarkable as he was not a very experienced student. Analyst IV did not manage to give a correct interpretation of the material himself, even though his description led the application to include antler. Perhaps he did not take antler working into consideration because of a lack of experience with this material: roe deer and reindeer do not belong to the Australian wild life.
Hardness: With all four analysts the hardness category was acknowledged by WAVES, but that of analyst III was not rewarded because he described the wrong traces, *i.e.* the intentionally manufactured retouch.
Motion: Due to the absence of macroscopic indications, suggestions as to the applied motion were only obtained on the basis of the polish features. Analyst III did get a conclusion, but it was incorrect because, like with the relative hardness, he described the intentional retouch. The polish clearly was diagnostic for a transverse movement: all analysts obtained the correct answer.

Experiment 4

Tool number 4 is an unmodified blade of which a point had been used for carving soaked bone for 26 minutes. The relative hardness of the contact material was considered to be medium, because the bone had been soaked in water. Despite the fact that it showed considerable and characteristic traces, both regarding the edge damage and the polish, this tool caused some serious problems for the analysts. The distal end could not be studied optimally because, before the test had began, the top had broken during the transport to Australia.
Contact material: Only in two instances bone was part of the conclusion and it did not get the highest diagnostic value. Although the interpretation of analyst II included 'butchering' (meat and fish), this was not rewarded because this tool showed definitely more traces than can be expected from just occasional contact with bone during butchering. Since it is apparent that a large variety of descriptions was given and that they all differed considerable from the standard, it is assumed that the reason for these moderate results is situated in the descriptions. Especially because the standard description was interpreted correctly. For instance, the description of the participants of the distribution of the polish varied from 'a line along the edge' to 'isolated spots'; the brightness from 'very bright' to 'dull'. Moreover, some of them were convinced that the polish exceeded the retouch, whereas others described the opposite. The analysts' personal suggestions were slightly better: two believed that the tool had been used for butchering, one thought of wood working.

Table 8. The results the analysts obtained with WAVES in interpreting the applied motion on the basis of the micro traces. (1=correct answer, 0=incorrect answer)

exp.	motion	standard (N=15)	I (N=15)	II (N=15)	III (N=15)	IV (N=15)
1	longitudinal	1	1	0	1	0
2	longitudinal	1	1	0	1	1
3	transverse	1	1	1	1	1
4	carving	0	0	1	0	0
5	longitudinal	1	1	1	0	1
6	longitudinal	1	1	1	1	1
7	transverse	1	1	1	0	0
8	transverse	1	1	1	1	1
9	longitudinal	1	1	1	1	0
10	carving	1	1	0	0	1
11	transverse	1	1	1	1	1
12	carving	1	1	1	1	1
13	transverse	1	0	0	0	0
14	carving	1	0	0	0	1
15	longitudinal	1	1	1	0	0
	Total	14	12	10	8	9
	%	93.3	80.0	66.7	53.3	60.0

Table 9. Test results of WAVES on the interpretation of the applied motion on the basis of the macro traces. The number of interpretations varies because some analysts did not find any indications for the applied motion on some of the tools. These were not included for the calculation of the number of correct answers.
(1=correct answer, 0=incorrect answer, – = not applicable due to absence of wear indications)

exp.	motion	standard (N=7)	I (N=9)	II (N=10)	III (N=15)	IV (N=12)
1	longitudinal	1	0	1	1	1
2	longitudinal	1	0	0	1	0
3	transverse	–	–	–	0	–
4	carving	0	0	0	0	0
5	longitudinal	–	1	0	1	1
6	longitudinal	1	0	0	1	1
7	transverse	–	–	–	1	1
8	transverse	–	–	–	1	–
9	longitudinal	1	1	0	1	1
10	carving	–	–	0	0	0
11	transverse	0	0	–	1	1
12	carving	–	–	0	0	–
13	transverse	0	1	1	0	1
14	carving	–	0	–	0	0
15	longitudinal	–	–	1	1	0
	Total	4	3	3	9	7
	%	57.1	33.3	30.0	60.0	58.3

Hardness: The right hardness category was deduced from the recordings of analyst II and III. Despite the fact that the others obtained equally high values on ‘medium’ as well, their descriptions turned out to be more diagnostic for soft materials.
Motion: Although the tool had been employed in a longitudinal fashion, the exact interpretation had to be ‘carving’. Even though 6 out of the 8 conclusions included both motions, in none of them ‘carving’ received the highest diagnostic value. Only analyst II obtained an equal value on both motions on the description of the polish features. Consequently, there was just one positive result.

Experiment 5

The fifth implement is an unretouched blade which had been used for reaping cereals for half an hour. It showed abundant wear traces, in particular an extensive polish. Regarding the resistance of the tissue it was considered to be a relatively soft material, but the wear traces turned out to be more characteristic for a medium hard material. Since the interpretation of the standard description favoured the medium hard material, it was decided that this would be the only conclusion on which the other analysts would yield a positive assessment.

Contact material: Except on the presence of striations, the descriptions were not very divergent and the results turned out to be rather good. In no less than three of the four analyses the outcome included cereals and it even gained the highest value with two of them. With the other (analyst III), the wear seemed only in third instance diagnostic for cereals. This was still considered correct because the answer had a rather homogeneous composition and it predominantly consisted of vegetal materials. Solely the interpretation that was based on the description of analyst IV was not approved. Like he personally thought of siliceous plants or soft wood, WAVES also strongly suggested soft wood as the only possibility. Even though soft wood is also a vegetal material, this was not rewarded, because the other vegetal materials were excluded. Moreover, the positive results of the other analysts showed that the traces were clear enough to allow for a correct answer. The analysts themselves were also on the right track: analyst II made a perfect deduction, while analyst I and IV assumed 'siliceous plants'.

Hardness: Concerning the edge damage there was also a remarkable disagreement: the expert did not find any evidence, while all other analysts did. They were, however, hardly unanimous about the location of these traces: one of the analysts located them on one side only and another on both sides equally. Nonetheless, the majority of their descriptions led to correct suggestions of the relative hardness (3 out of 4) and of the motion (3 out of 4).

Motion: With reference to the applied motion, a longitudinal motion was favourite, although 'diagonal' was a popular second best. This corresponds exactly with the way the tool was used, for in reaping cereals the tool is not moved in an absolute longitudinal fashion, but slightly diagonal as well. Altogether, 6 out of 8 interpretations could be rewarded (see table 8 and 9).

Experiment 6

Experiment number 6 had been used for cutting dry grass (= siliceous plants) for 30 minutes, which had yielded a clear band of polish but minor edge damage. The relative hardness was 'soft'.

Contact material: All analysts personally suggested that the tool had been used on a vegetal material, although the exact material ranged from soft wood to siliceous plants. WAVES did not cause any surprise either as the results pointed rather homogeneously towards vegetal materials: in two instances siliceous plants were included, in one non-siliceous plants and in the fourth soft wood. Nonetheless, the latter two were not rewarded because siliceous plants had been excluded. The descriptions showed no extraordinary dissimilarity. Even on the topography of the polish there was a remarkable agreement. Still, this could not prevent that two interpretations slightly deviated. With analyst IV this was caused by the fact that he, like with experiment 5, characterized the distribution of the polish as 'reticulated'. The combination of a rough&matt texture with a medium brightness of the polish made WAVES exclude the siliceous plants in the suggestion to analyst II.

Hardness: Unfortunately, in nearly all cases WAVES believed that a medium hardness was most likely. Solely the result that analyst IV obtained on the basis of his recordings of the macro traces coincided with that of the standard description. Although the observations of the inexperienced student also matched that of the expert remarkably, his choice for 'heavy edge rounding' excluded the soft material.

Motion: All analysts correctly deduced a longitudinal motion themselves. WAVES was also convinced of a longitudinal motion on the basis of the polish features, but preferred in two instances a transverse motion because the orientation of the retouch was said to be perpendicular.

Experiment 7

In contrast to the analysis of the previous tool, that of number 7 hardly yielded good results. Despite the fact that this intentionally retouched scraper had been used for one hour on fresh hare hide, it only showed a thin line of polish and a slightly rounded edge.

Contact material: According to the expert this tool showed only minor signs of wear, but three of the other analysts did not agree with this. They claimed to have seen extensive bands of polish. One of them even observed a polish that extended eight times as far onto the edge as the polish which the expert had observed. However, the characterisations of the distribution of the polish are responsible for the poor results. Notwithstanding the fact that all personal suggestions were correct (although analyst IV had not been absolutely sure), solely analyst II obtained a correct and convincing interpretation from WAVES. The description of analyst I did not lead to an interpretation at all, and those of analysts III and IV turned out to be indicative for butchering rather than hide working. These are indeed vegetal materials but were not rewarded.

Hardness: The results on the basis of the macro traces were better than those of the micro traces: 3 out of 4 favoured a medium hard material, which was correct. Like the expert, two analysts did not detect any use retouch, whereas analysts III and IV did. All participants recognized a slight edge rounding. Again, analyst III mistakenly described edge removals that had been produced in manufacture, despite the fact that WAVES warns for this.

Motion: Except for the description of the polish features of analysts III and IV, all other descriptions made WAVES correctly suggest a transverse motion. Nevertheless, only four points were gained because two analysts could not give any more indications due to the absence of edge damage. Obviously such missing answers have been excluded from the calculation of the success rates, because they would have unjustly affected it negatively.

Experiment 8

Again this was an experiment with hide working. The tool had been intentionally retouched and used for scraping fresh elk hide for 60 minutes. It showed considerable edge rounding, and a distinctive but not very extensive polish. Instead of edge scarring it had incurred severe edge rounding.

Contact material: Even though all personal interpretations were correct, one analyst gave a description which led to a wrong interpretation and another obtained no suggestions at all. The two remaining participants received correct answers. Once more the inexperienced student (analyst III) described the intentional retouch as being caused by use. He did not recognize the heavy edge rounding and the directionality within the polish either. WAVES related the traces that he described both to animal and vegetal materials, but excluded hide because this is not characterized by a smooth&matt texture. Analyst IV described the traces almost in perfect harmony with the expert, but the selection of a bevelled distribution in combination with the other hide-characteristics was fatal for the interpretation. This example shows the limitation of WAVES. The combination of the observed features must correspond with the wear patterns it knows in order to allow for an interpretation. Human analysts, however, are far more flexible: they can doubt their observations, but may still (analyst IV) reach a correct conclusion on the basis of the other features.

Hardness: All interpretations were unanimous with respect to the medium hardness category.

Motion: The participants did not find abundant indications for the applied motion, but the ones that they recorded all caused the application to correctly infer a transverse movement.

Experiment 9

Tool number 9 was an unmodified blade which had been used for butchering fish for 35 minutes. It was expected that this would cause some problems as it did not show extensive wear and no well-developed polish. Similar to experiment 2, it was decided that both ‘meat and fish’ and ‘bone’ would be rewarded, because it depends on the analyst whether he or she describes the traces caused by the soft tissue or by contact with the bones of the fish.

Contact material: The descriptions as well as WAVES’ suggestions varied considerably and only two interpretations could be accepted. Analyst I gave a perfect personal interpretation, but did not manage to give an interpretable description. She had trouble to distinguish between the bone working and the meat working traces: some variables describe a bone polish, others a meat polish. This inconsequence confused WAVES and gradually excluded all materials, because the resulting wear pattern matched none of the patterns in its knowledge base. Although bone and antler are said to be hardly distinguishable, the answer that analyst II obtained (soaked antler) was rejected because the diagnostic value was not very high and it excluded all alternative suggestions. On the other hand, the interpretation received by analyst III was considered correct: it gave preference to bone but included both soaked and dry antler as well. This interpretation clearly shows, however, the difficulty with the analysis of non-diagnostic traces. In such situations it may be feasible to exclude some options, but certainly not to identify the specific contact material. In particular the heterogeneous composition of the conclusion that WAVES deduced from the recordings of analyst III would have made it almost impossible to infer the right answer.

Hardness: Due to the fact that it had been decided to accept responses that included either ‘soft’ or ‘medium hard’, no less than 3 out of the 4 suggestions on the hardness category could be considered correct.

Motion: It did not seem to be hard to deduce the applied motion: in 6 out of the 8 conclusions the longitudinal movement was favourite. For some inexplicable reason, analyst II erroneously described the orientation of the retouch scars as perpendicular and forfeited a correct answer on this element. The personal contributions were just perfect. The sole fact that analyst IV did not specify the directionality within the polish made that there were insufficient indications for a longitudinal motion.

Experiment 10

The piece that had been used for experiment 10 is not an intentionally retouched tool. Since it has a sharp point it turned out to be useful for splitting branches of willow. It had been used for 20 minutes, but showed only minor traces. Willow is one of the soft woods and its relative hardness is considered to be ‘medium’.

Contact material: In their personal conclusions, all analysts suggested a vegetal material, although two of them favoured hard wood. Despite this close interpretation, two analysts did not manage to withdraw the right conclusion from WAVES. Analyst I gave a description of the micro traces that highly deviated from that of the expert and analyst III described the polish as being bright with a pitted topography. Unfortunately, the latter combination is considered to be diagnostic for traces caused by bone and excluded wood working. They all agreed that the traces were minimal.

Hardness: The interpretations of the hardness category were better: 3 out of 4 were correct. Again, analyst I failed to get an answer. She did not observe any edge damage or rounding.
Motion: With respect to the applied motion only two interpretations were validated positively. Analyst I could not give any more indications due to the alleged absence of edge damage and all other indications either led to a wrong conclusion or to no interpretation at all. With the other three analysts the combination of the location of the micro traces together with the shape of the edge was conflicting.

Experiment 11
Tool number 11 is a scraper which had been used for working hide for 115 minutes. Since it concerned a hide of a swine that was extremely greasy, flower was used as an abrasive. The tool had been employed until it had become completely blunt. Consequently, the edge showed heavy rounding. On the non-retouched ventral side some use retouch had developed as well. Moreover, the tool displayed a distinctly polished surface.
Contact material: This tool turned out to be one of the easiest to interpret. All personal suggestions clearly indicated hide working and in three instances this answer was also received from WAVES. The inexperienced student, however, gave rather deviating indications. Since he described a polish that is characteristic for wood, hide working was excluded.
Hardness: With respect to the hardness category, a medium hard material was correctly concluded in three cases. Analyst III also forfeited this interpretation by not recognizing the heavy rounding. It must be stressed however, that this time he described the right retouch.
Motion: Although the correct motion was given in 6 out of the 8 cases, especially the use retouch made it difficult to deduce. This time it was analyst II who could not find any edge damage and, therefore, missed an answer from WAVES. Moreover, the characteristics of the location, distribution and orientation of the scars that was given by analyst I, suggested a longitudinal motion.

Experiment 12
Experiment number 12 was also one of the more difficult ones. It had been used for carving leather-dry clay for 20 minutes, but neither WAVES nor WARP had been able to interpret its traces correctly in the first test. In the meantime, the knowledge base of WAVES had been adapted, but it was still an absolute surprise how the traces would be recorded by other analysts. The tool displayed heavy edge rounding and a considerably extended, though not a very characteristic polish.
Contact material: The interpretation of the standard description illustrates that WAVES was now able to recognize the traces, but with the exception of analyst I, none of the participants recognized the traces personally. Unfortunately, none of them managed to get a correct interpretation from WAVES either. The descriptions of analysts I and II turned out to be not interpretable at all: their patterns did not match any of the application's. From the description of analyst III the application deduced bone working or butchering and from that of analyst IV hide working. It is peculiar that this analyst personally thought of hide and — surprisingly — WAVES conclude the same. This illustrates that an analyst may influence the system's interpretation by his own assumptions. If he or she is convinced of a hypothesis, than he may — unconsciously — describe his observations in a way that this hypothesis is confirmed.
Hardness: It is remarkable that this implement belongs to the small group on which a correct interpretation of the relative hardness was deduced by all four analysts. It is even more peculiar that in all cases the application was convinced of a medium hard material and that no alternatives were assumed to be possible.
Motion: On the basis of the descriptions of the polish, again all analysts obtained a correct interpretation: they almost exclusively deduced a carving motion. On the basis of the edge damage, however, the results were the opposite. Two analysts did not find any indications, one gave conflicting indications and one obtained a wrong answer.

Experiment 13
Also the next experiment was experienced as a problematic one. The tool is a non-retouched blade that had been used for scraping birch bark for 20 minutes. It showed only slight edge damage, no rounding and not very extensive polish. Soft wood was classified as a medium hard material, but the edge damage was so minimal that it was tempting to decide that 'soft' would be accepted as well.
Contact material: All analysts had serious troubles deducing the applied contact material. Two personally thought of an animal material, the others did not give an interpretation. Very remarkable is the result that the blank student obtained, because WAVES deduced exactly the right answer. The diagnostic value, however, was not very high as he did not typify the topography of the polish, which is usually one of the most diagnostic features. Analyst II obtained a correct interpretation as well, although his recordings deviated considerably from that of the standard description. The observations of the other two analysts (I and IV) led to incorrect answers, but it is striking that these answers matched exactly their personal interpretations. Since this can hardly be coincidental, I am inclined to think that they unconsciously affected the reasoning process of WAVES by their own assumptions. This hypothesis is supported by the fact that the blank student (analyst III) was not hampered by

his own knowledge: he followed the guidelines that the application offered and carried out a successful analysis.
Hardness: The same applies to the deduction of the hardness category. Except for analyst III, none of the participants received the right interpretation of WAVES. The suggestions that were given were totally in line with those concerning the contact material.
Motion: The transverse motion was not recognised in all cases either. It turned out that three analysts gave correct descriptions of the macro traces only and that the fourth (analyst III) recorded features that the application considers to be diagnostic for a longitudinal movement.

Experiment 14
Tool number 14 is a point which had been employed for carving bone that had been softened by soaking it in water. This activity lasted 45 minutes. Even though the edge rounding was minor and the expert did not distinguish use retouch, a well-developed polish had evolved.
Contact material: The analysts did not agree on the absence or presence of edge removals: the expert and analyst II did not observe any use retouch, but the other three analysts did. They all agreed firmly, however, on the activity that had been carried out: bone carving. WAVES almost totally went along as well. It only did not reach a conclusion on the description of analyst I, because it does not relate a domed topography to bone working. The inference failed irrespective of the fact that the other features also clearly pointed at bone.
Hardness: WAVES only once suggested 'a hard material': on the basis of all other descriptions it preferred a medium hard material, which nicely matched with the indication of 'bone'.
Motion: In contrast with the former two elements, the interpretations of the motion were not very convincing. The correct answer was deduced only once (analyst IV): one analyst (II) merely found some indications in the polish due to the absence of use retouch, but obtained a false interpretation anyhow. The other two described macro traces that turned out to be uninterpretable and subsequently received also a misinterpretation on the basis of the polish.

Experiment 15
Once more, this was an experiment that confronted the analysts with a problem. Tool number 15 had been part of the first test, but both applications had failed to interpret its traces correctly. It had been used for sawing shell for only 10 minutes and although it displays no edge rounding, a clear band of polish is present. The difficulty, however, is that the polish is characteristic for antler working rather than shell working. The relative hardness of the contact material was considered to be medium hard because of the lack of severe edge damage.
Contact material: The non-diagnostic character of the polish was clearly represented by the analysis results: none of the analysts either indicated shell working personally or obtained a correct interpretation with WAVES. Moreover, this was the only implement that was not rightly interpreted by WAVES on the basis of the standard description. It is remarkable, however, that not only all the personal suggestions included bone or antler working, but all answers of the application as well. This could not prevent, however, that the answers were not rewarded. Since the blank analyst obtained this result as well, this seems to justify the extrapolation that these interpretations acknowledge the observation that the traces actually resemble an antler/bone polish. From the personal achievements it can be concluded that none of the analysts was familiar with the traces on this artefact.
Hardness: In two instances the right conclusion on the hardness category was given (analyst I and II). The interpretation that the third participant obtained, unjustly indicated soft material and analyst IV described conflicting characteristics and did not receive any suggestion at all.
Motion: Only four times the right motion received the highest diagnostic value: based on the description of the macro traces of analysts II and III and on the description of the micro traces of analyst I and, again, participant II. This means that only the latter scored on both elements.

7.4.4 CONCLUSION

From the success rates that were obtained on the standard description (tables 6-9), it can be concluded that the achievements of WAVES were considerably improved in comparison with the first test (table 4), despite the fact that it involved equally difficult wear traces. At first the application had not been able to interpret eleven tools as a result of insufficient knowledge, but it now demonstrated the enhancement of its knowledge base: its success rate on the deduction of the exact contact material increased from 58.3% to 93.3% (see standard description in table 6). It even gave correct interpretations of traces that had been caused by working antler or dry clay. The only aspect that remained problematic was the reconstruction of shell working. Although the concerning misinterpretation can partly be ascribed to the atypical characteristics of the polish on that particular implement, it remains a fact that the knowledge that is available on this type of contact material is too limited.
The achievements of the application on the deduction of the relative hardness (table 7) cannot be taken into account because the interpretations were used to establish the standard against which the responses of the other participants could be compared rather than that they were compared with a predefined standard.
With reference to the applied motion, it turned out that the micro traces provided diagnostic and well-interpretable

indications: WAVES obtained again a success rate of 93.3% (table 8). From the macro traces, on the other hand, it was much more difficult to deduce the applied motion: only seven tools displayed indications and even three of these were misinterpreted. Consequently, this yielded a success rate of only 57% (table 9). The reason why the analysis of the macro traces yields less reliable results is not entirely clear. It is unlikely that this is caused by the number of variables on which the interpretations are based. The interpretation of the motion that is based on the micro traces involves even less variables, but produces highly accurate results. Presumably, the attributes of the macro traces that we recorded do not have sufficient diagnostic value. Alternatively, the knowledge rules need to be improved or sharpened. Clearly, at this point one should not rely too much on this aspect of the interpretation alone. It should always be judged in conjunction with the part of the interpretation that is based on micro traces. If both answers confirm each other, it is likely that they are correct. If not, this is an indication that the descriptions of the traces are conflicting or that the traces are caused by complex or multiple activities or are affected by, for instance, post-depositional processes.

Despite the moderate achievement on the reconstruction of the applied motion by means of the macro traces, it seems to be justified to say that the theoretical validation of the knowledge base of WAVES can be judged positively. The results have demonstrated that when the application is provided with adequate descriptions, it can give highly accurate answers in return. The main goal of this test, however, was not a theoretical validation of the knowledge base but an evaluation of the application's practical functionality. Before any conclusion from the above results is drawn, I once more want to emphasize that this must be done with care. This test has been far from exhaustive and the results were affected by various aspects, such as the composition of the test set, the differences in the scientific backgrounds of the participants, etc. Therefore, the results do not demonstrate what WAVES or the participants are capable of, but they merely give an impression of the performance of WAVES in the hands of independent users and vice versa.

In relation to the functionality of the application, the achievements of the participants have revealed two important aspects. The first is that there are considerable dissimilarities between the results on the standard description and on the recordings of the analysts, and the second is that there are major discrepancies between the descriptions of the analysts.

The dissimilarities in the achievements is illustrated by the fact that the analysts performed *less well* than the standard description on recognizing the exact contact material, but *better* than the standard description on deducing the applied motion from the use retouch (see table 10). Their highest success rate was 60% on both, that of the standard description 93.3% and 57% respectively. From the fact that the recordings of the analysts yielded lower success rates on the reconstruction of the exact contact material than the standard description, I am inclined to conclude that the practical functionality of the application is not yet optimal. The reason for this conclusion is that the incorrect interpretations which the participants obtained from WAVES cannot be ascribed to the application's incapability or to biases in its knowledge base. This is proven by the satisfying results with the standard description, which demonstrated that the knowledge base is functioning properly. Alternatively, the lower success rates are probably caused by discrepancies between the recordings of the analysts and of the expert.[12]

Table 15 shows how many of the descriptions of the analysts deviate from the standard description. It is evident that numerous variables yielded many differences. For instance, it is conspicuous that the distribution of the polish is difficult to describe. In no less than 44 out of the 60 times (73.3%) it was characterised differently than in the standard description. Naturally, it was expected that the participants would record some wear features differently, especially since many of the variables have a highly subjective character and may be difficult to describe. This was one of the reasons why special adaptions were made in the design of WAVES. It was tried to prevent that a description of a user cannot be interpreted if it slightly deviates from that of the expert. The prime remedy was to validate the use wear on its individual features rather than the pattern as a whole (chapter 5). That this approach has a positive effect is illustrated by some of the recordings of analysts II and III: even though these differed repeatedly from the standard description, the analysts obtained still many correct interpretations.

It was not expected, however, that the descriptions would also differ on the more simple variables, like the location of the traces. There were even large variations in the descriptions of the metrical variables like the length of the edge removals and the width of the polish, the edge angle, the shape of the edge, etc. In a few instances the differences were extreme: in one case the recording of the width of a polish varied from 250 micron to 2000 micron. The exact reason for this is unclear. Perhaps the instructions in WAVES exhibit deficiencies or perhaps the analysts did not measure precisely enough. Nevertheless, the occurrence of such large discrepancies gives reason to believe that the description is one of the aspects of the analysis process that needs adjustments.

From table 15 it can be deduced which variables need some extra attention. For instance the recordings of the polish location, distribution, brightness and width deviated many times from the standard description. Possibly, the guidelines for measuring some of the variables have not been clear

Table 10. Success rates (in percentages) that the participants obtained with WAVES on the various aspects. Their personal success rates are given in parenthesis. *Due to the inexperience of analyst III, no personal interpretations were given.

	analyst I	analyst II	analyst III*	analyst IV
exact material	33.3 (53.3)	60.0 (60.0)	53.8	40.0 (53.3)
relative hardness	50.0	80.0	60.0	73.3
motion (from micro traces)	80.0 (86.6)	66.7 (86.6)	53.3	60.0 (93.3)
motion (from macro traces)	33.3	30.0	60.0	58.3
Total	49.2 (69.9)	59.2 (73.3)	56.8	57.9 (73.3)

Table 11. The interpretations of the exact contact material that the analysts obtained from WAVES on the basis of the study of polish features. In parenthesis the number is given of the correct personal interpretations of the analysts.
* Due to the inexperience of analyst III, no personal interpretations were given.

Interpretation of exact material	standard	analyst I	analyst II	analyst III*	analyst IV
correct:	14	5 (8)	9 (9)	8	6 (8)
highest value	*9*	*4*	*6*	*3*	*2*
second value	*5*	*0*	*3*	*1*	*1*
third value	*0*	*1*	*0*	*4*	*2*
other	*0*	*0*	*0*	*0*	*1*
wrong	1	4 (6)	4 (6)	7	8 (7)
no interpretation	0	6 (1)	2 (0)	0	1 (0)

Table 12. The interpretations of the relative hardness that the analysts obtained from WAVES on the basis of the analysis of use retouch and edge rounding. The number of tools on which the results are based vary because some of the analysts did not find use retouch or edge rounding on all tools.

Interpretation of relative hardness	standard (N=14)	analyst I (N=14)	analyst II (N=15)	analyst III (N=15)	analyst IV (N=15)
correct:	14	11	13	12	14
highest value	*14*	*7*	*12*	*9*	*11*
second value	*0*	*4*	*1*	*3*	*3*
third value	*0*	*0*	*0*	*0*	*0*
wrong	0	3	2	3	0
no interpretation	0	0	0	0	1

Table 13. The interpretations of the applied motion that the analysts obtained from WAVES on the basis of their description of polish features (micro traces). In parenthesis are the personal achievements of the analysts given. *Due to the inexperience of analyst III, no personal interpretations were given.

Interpretation of motion from micro traces	standard (N=15)	analyst I (N=15)	analyst II (N=15)	analyst III* (N=15)	analyst IV (N=15)
correct:	14	14 (13)	10 (13)	11	14 (14)
highest value	*14*	*12*	*10*	*8*	*9*
second value	*0*	*2*	*1*	*3*	*5*
third value	*0*	*0*	*0*	*0*	*0*
other	*0*	*0*	*0*	*0*	*0*
wrong	1	1 (0)	4 (2)	3	1
no interpretation	0	0 (2)	0 (0)	1	0

Table 14. The interpretations of the applied motion that the analysts obtained from WAVES on the basis of their description of use retouch and edge rounding (macro traces). The number of tools on which the results are based vary because some of the analysts did not find use retouch or edge rounding on all tools.

Interpretation of motion from macro traces	standard (N=7)	analyst I (N=9)	analyst II (N=10)	analyst III (N=15)	analyst IV (N=12)
correct:	4	4	4	9	7
highest value	*4*	*3*	*3*	*9*	*7*
second value	*0*	*1*	*1*	*0*	*0*
third value	*0*	*0*	*0*	*0*	*0*
other	*0*	*0*	*0*	*0*	*0*
wrong	2	4	4	4	2
no interpretation	1	1	2	2	3

variables	analyst I	analyst II	analyst III	analyst IV	total
retouch location	5	5	7	7	24
retouch distribution	4	5	5	6	20
retouch orientation	6	5	4	4	19
retouch termination	6	6	5	5	21
retouch length	1	2	3	4	10
edge rounding	6	7	8	2	23
invasiveness	3	3	3	3	12
polish location	8	9	10	11	37
polish directionality	6	10	8	4	28
polish distribution	11	11	9	13	44
polish texture	5	8	5	7	25
polish brightness	6	11	9	8	34
polish topography	7	4	8	5	23
polish width	10	8	9	8	35
striations	10	4	4	8	26
grain size	8	6	9	10	23
edge angle	4	4	7	4	19
edge shape	2	4	1	1	8
Total	108	112	114	110	441

Table 15. The number of times that a description of an analyst differed from that of the standard description.

enough. Perhaps more basic explications are needed as well, such as what is meant by the 'dorsal' and the 'ventral' side of a tool. Fortunately, it also turned out that some variables are easy to handle. The recording by the analysts of the length of the edge removals, the invasiveness of the polish and use retouch, and the shape of the edge, for example, did not deviate much from the standard description.

Apparently, not all misinterpretations must be ascribed to functional failures of the application: the personal effect of the analysts should not be underestimated either. It may be clear that when a person describes a cat as an animal with a fur, two legs, a long tail and a pouch, not a single application would be able to deduce that the observer is describing a cat. It would rather assume that it is the description of a marsupial. The latter would be the right answer, but would not correspond with the expected answer. In such a case, however, the knowledge base nor the reasoning process of the application can be held responsible for this failure. Similar cases were encountered in the test with WAVES. For instance in experiment 3, analyst III received on the basis of his recordings of the macro traces a correct interpretation from the application on the relative hardness of the applied material. Unfortunately this answer had to be rejected because the analyst had described the retouch that originated from manufacturing rather than from use (appendix IV). Consequently, the participant's lower success rate cannot be ascribed to a functional failure of WAVES.

This example has introduced the second aspect that the results of our test has revealed, the dissimilarities between the analysts. It was noticed that not only their achievements differed but also that it was not the same participant that performed best on the various elements of the test. For instance, analyst II obtained the best success rate on the determination of the exact contact material and of the relative hardness, but analyst I and analyst III were better in interpreting the applied motion on the basis of the study of the micro traces (80%) and the macro traces (60%) respectively. In total, analyst II obtained the best results. This student from Leiden University, who had only been practising use-wear analysis for half a year when he volunteered for this test, achieved an overall success rate of 59.2% (table 10). This is remarkable, especially because his results surpassed that of the two more experienced analysts (I and IV). It must also be kept in mind, however, that he may have been in a privileged situation due to the fact that he had been trained by the same expert that had supervised the development of the application. Consequently, he was already slightly accustomed to the approach used in WAVES.

In particular the success rates of participant I stayed behind expectations. While she personally identified 53.3% of the applied contact materials correctly (table 10, 11), her descriptions enabled WAVES to deduce only five correct answers (33.3%). On the determination of the relative hardness she obtained slightly more correct interpretations, but still less in comparison with the others. The most striking result however, was that the totally blank student (analyst III) obtained a mean success rate of 56.8% while he had never been analysing wear traces before. From 8 out of the 15 descriptions (53.3%) the exact contact material could be deduced and of the assumed motions even 60% was correct.

This proves the ability of WAVES to give reliable results when its indications are followed and it is provided with adequate descriptions.

There may be various reasons for the differences in the achievements of the participants. One reason may be found in the educational backgrounds and the various traditions of describing traces: it can hardly be coincidental that exactly the two analysts that originate from different methodical schools (I and IV) obtained a lower success rate. An additional argument for this assumption is that they obtained better personal interpretations: analyst IV even interpreted 93.3% of the applied motion correctly (table 10, 13), while with WAVES he only reached success rates of 60% and 58% (table 10). It is also remarkable that, for instance in the descriptions of the striations, there is a large discrepancy between the analysts from Leiden (II and III) and the foreign participants (I and IV). The recordings of the former two were more in accordance with the standard description. Another example is provided by the recordings of the topography of the polish by analyst I: while the standard description contains only three instances of a 'domed topography', analyst I had a clear preference for this and used it nine times.

Apart from the educational background, the level of experience of the analysts seems to be of crucial importance as well. Presumably, the experienced analysts could not give objective descriptions of their observations because they were hampered by their own assumptions: unconsciously they gave descriptions that directed the reasoning process towards their personal conclusion. This is demonstrated by the fact that if their personal interpretation was wrong, the application often gave exactly the same wrong response. An additional argument is that the least experienced analyst (III) turned out to be the most susceptible to the guidelines of WAVES. For instance, his descriptions showed more variance than some of the other analysts. Since he was hampered less by a particular repertoire for descriptions, this presumably enabled him to adjust more easily to that of WAVES. This could not prevent, however, that the blank student made some 'beginners mistakes' that the others did not. For instance, he repeatedly confused intentionally applied retouch with use retouch and did not distinguish bright from very bright polishes. In comparison with the other participants he also declared more frequently that the topography of the polish was indistinct: analyst I zero times, analyst II two times, analyst III nine times and analyst IV six times.

Probably there are various other reasons for the differences between the participants. Bamforth once argued that personal differences can also relate to factors like native abilities, experimental backgrounds, mental and physical conditions, etc. (Bamforth 1988: 20). Like the differences in the experimental backgrounds explains why analysts that are familiar with artefacts made of coarse flint materials validated the raw materials of the experimentally used tools differently than our expert, all of these factors may explain a particular part of the observed dissimilarities. Since it will never be possible to rule out the influence of these human factors, it will be difficult to optimize the functionality of the application. Even if WAVES would give perfect guidelines on describing wear traces, analysts might still make erroneous recordings due to the simple fact that they will remain responsible for the decision of the location of a measurement and for the actual measurement. Nevertheless, it is crucial for the functionality of WAVES that the users are made aware of the importance of accurate descriptions and that they are willing to act accordingly. They should realise that they use the description of their observations merely as a reminder of the image, while WAVES totally depends on it since it cannot contemplate the image as a whole.

7.4.5 COMPARISON WITH OTHER BLIND TESTS

In order to put the results that were obtained by WAVES in the right perspective, a comparison was made with other blind tests (table 16). We must, however, be careful with the conclusions that we draw from this comparison, because the tests that have hitherto been carried out are highly different. There are differences in the composition of the test set (involved tool types, contact materials, activities, raw materials of the artefacts, duration of the experiments), in the goals of the tests, in the criteria for rewarding interpretations, in test conditions, in expertise of the participating analysts, in the number of participants (individual scores versus average scores of several analysts), in involved methods of analysis (low-power versus high-power approach), in involved wear categories, in the prehistoric period from which the activities are simulated, in microscopic equipment, in used magnifications, in cleaning procedures, etc. Consequently, the complexity of the various tests differs considerably. Furthermore, the validation of these tests is complicated by the fact that the results were not always reported similarly: sometimes only the success rates are given, sometimes detailed descriptions of the experiments and the interpretations are provided as well. Especially the relative hardness of the worked material was only occasionally an integral part of the analysis.

Due to these difficulties, the figures in table 16 require explanation before any conclusion can be drawn from them. In table 16 several figures do not totally correspond with the results in the original publications. Differences in the criteria for rewarding interpretations made it necessary to adjust some of the achievements in order to make them comparable. As far as the data in the publications made it possible, they were subjected to the same criteria that had been employed in the second test with WAVES. Additionally, a distinction has been made between average scores and individual

Table 16. Comparison of other blind test results with those obtained with WAVES.
'mean' indicates an average success rate of a group of analysts, 'indiv.' represents the highest individual achievements.
* This test involved the low-power approach exclusively.

	exact contact material		relative hardness		motion	
	mean	indiv.	mean	indiv.	mean	indiv.
Keeley & Newcomer (1977)	–	60.0%	–	–	–	80.0%
Odell & Odell-Vereecken (1980)*	–	38.7%	–	67.7%	–	69.4%
Gendel & Pirney (1982)	–	73.9%	–	–	–	82.6%
Unrath *et al.* (1986)	49.0%	83.0%	–	–	59.0%	78.0%
Newcomer *et al.* (1986)	33.7%	60.0%	–	–	50.0%	–
Bamforth *et al.* (1990)	–	66.7%	–	–	–	78.6%
WAVES	46.7%	60.0%	65.7%	80.0%	65.0%	80.0%

scores. Some tests were executed by a group of analysts, whereas others involved only one expert. Since the achievements of these individuals were generally better than the average of a group of analysts, it has also been tried to withdraw the best personal achievement from the participants in the group tests.

The first test by which our results were compared was that of Keeley and Newcomer from 1977. They composed a test with tools "*...used for specific tasks thought to be relevant to prehistoric hunters*." (Keeley & Newcomer 1977: 29). These were artefacts that could be expected in Lower or Middle Palaeolithic assemblages. The materials on which they had been used consisted of various kinds of wood, bone, meat (both fresh and frozen), siliceous plants (bracken) and hide. The implements were employed in boring, cutting, sawing, scraping, whittling and chopping motions. Their analyses were based on the high-power approach.

Although they claimed a success rate of 62.5%, *i.e.* 10 out of 16 for the worked material and 75% (12 out of 16) for the tool movement (*ibid.*: 60), these results had to be adjusted slightly when my criteria were imposed. The score on the contact material dropped slightly to 60% (9 out of 15)[13], but on the motion it even increased to 80% (12 out of 15). The authors had already been rather strict, but I considered an interpretation saying 'unknown, possibly vegetable matter or meat' (tool no.4) correct when the tool had indeed been used for meat. If WAVES would have given an interpretation consisting of two materials with only low diagnostic values, it would have been rewarded too. In one instance (tool number 10) Keeley and Newcomer rejected the interpretation of the motion because it was a guess and the underlying reasoning was incorrect. However, as this is not an argument in the validation of the results of WAVES, it was considered correct. With these adjustments, this test seems pretty comparably with ours, especially since it was a relatively complex test, just like ours. There is, however, one major difference: their test was performed by one analyst only who, in addition, was rather experienced.

A second test was carried out by Odell and Odell-Vereecken (1980). In contrast with the previous one, they focused on the low-power approach (magnifications up to 100×). Their test consisted of 31 tools, which had been used in a large variety of tasks. Since they did not confine the experiments to a particular prehistoric cultural stage, several 'unusual' tasks were involved like sawing yams, digging in the ground, chopping hemlock, crushing hazelnuts and pounding hazelnut shells. The tools were made of basalt instead of flint, and the average duration of the experiments had only been 13 minutes. In my opinion, this test belongs to the category of high complexity. Fortunately the authors presented the answers of the analysts, so they could be validated against the criteria used for the test with WAVES. But since all results corresponded completely, no adjustments had to be made.

In comparison with the other tests, the achievements described by Odell and Odell-Vereecken deviate on all categories. Especially considering the fact that these are individual scores rather than mean scores, the achievements are inferior to those obtained by the high-power method. The authors explained this low success rates by the fact that, at that time, they had not yet been very experienced with the low-power method. They could not obtain accurate results to a greater specificity than to a relative resistance of the worked material (Odell & Vereecken 1980: 116). It is difficult, however, to put these results in real perspective by comparing them with similar tests with the low-power method because these are simply not available. The only aspect of this test which is comparable with the WAVES test is the highest individual score on the relative hardness.

In the third test again the high-power approach was employed. Gendel and Pirnay claim a correctness score of almost 74% on the exact contact material and of 82.6% on the motion (Gendel & Pirnay 1982: 257). In an absolute sense these are the highest scores of all tests, but it must be stressed that their test was of a much lower complexity than the others. For example, they did not include less common contact materials like fish, meat, cereals, grass, roots, dry clay and shell. The only exception was an experiment on a fox tooth and, remarkably, this was one of the few tools of which the traces were misinterpreted. Their starting point was that "*…the raw materials utilized by the experimenter were to be comparable to those with which the wear analyst was familiar*" and that "*…the tools were to be used in ways thought to be reasonable in Palaeolithic and Mesolithic contexts.*" (*ibid.*: 251). This clearly narrowed down the range of activities the analysts had to consider in their analyses. Another difference with the test with WAVES and some of the others, is that they did not try to make a distinction between traces caused by bone or antler working. In the interpretations they constantly mentioned them together. This made it impossible for me to validate these answers in comparison with these of WAVES. Furthermore, the motions they exclusively involved were scraping and whittling (both transverse), and boring.

The blind test that was carried out by Unrath, Owen, Van Gijn, Moss, Plisson and Vaughan at the end of 1984 (Unrath *et al.* 1986) can also be placed within the category of tests with a high degree of complexity. Although they focused on activities typical for the Upper Palaeolithic of temperate Europe (*ibid.*: 122), they incorporated a broader range of activities and materials than most of the other tests. Besides the usual activities they involved traces from bag carrying and post-depositional processes like trampling, but also from ivory, shell and antler working and from all steps of hide preparation. Moreover, realistic activities were carried out rather than monotone motions. They performed their analyses by means of the high-power approach. Unfortunately, Unrath *et al.* did not provide the answers of the individual analysts, which made it more difficult to validate the answers according to the criteria employed for WAVES. Furthermore, the achievements of the individual analysts are lacking: they only presented the number of correct responses to the analysis of each implement. Nevertheless, by summing the specific answers and the indication of the motion or material group that they presented (Unrath *et al.* 1986: table 1, 150), it was possible to deduce the figures for table 16. This was believed to correspond best with the kind of answer that would have been rewarded in the test with WAVES. The authors did give the best individual success rate, but this figure is not totally comparable either since it includes scores for partly correct answers or correct interpretations of the relative hardness.

In the next test, that of Newcomer *et al.* (1986), four analysts participated and they focused on the high-magnification approach. Since they obtained rather low success rates, especially on the interpretation of the exact contact material (an average of 33.7%), they have received severe criticism on the composition of the test. For instance, the incorporated activities simulated not only realistic Upper Palaeolithic tasks (tools 1-10), but also the unrealistic motion of rubbing the ventral surface against several types of contact materials (tools 11-20) and monotone motions (tools 21-30). Although they only used frequently occurring materials (hide, wood, antler, bone and ferns), the unusual rubbing experiments and the average tool use of approximately 11 minutes makes this one of the most complex tests.[14] For the sake of comparison: the tools involved in the test with WAVES had been used for an average time of 34 minutes.

Unfortunately, the results presented by Newcomer *et al.* are difficult to compare. Firstly because they did not give all the answers of the analysts, especially not the most interesting ones, *i.e.* the reconstructions of the simulated Palaeolithic tasks. Therefore the achievements could not be validated against the criteria used for the test with WAVES. Secondly a comparison was complicated because their individual success rates included the scores on the identification of the used area, the motion and the material. Thirdly this test deviated because the analysts had been told in which movements tools 11-30 had been employed. Nevertheless, some comparisons are possible. For instance, the best personal performance seems to be in accordance with that of other analysts in other tests. Analyst IV even obtained a 60% success rate on the tools that had been rubbed against the contact materials for only 10 minutes (tools 11-20). The average result on the exact contact material is incongruous as it is clearly less than the others. Several analysts (Moss 1987; Bamforth 1988; Hurcombe 1988) have already commented on this result and gave various explanations, but since it has been impossible to assess these results on the criteria used for WAVES, an objective validation cannot be given. Nonetheless, it may be clear that the figures themselves indicative that this was a peculiar test.

Finally, Bamforth *et al.* (1990) performed a test with 20 replicated tools of Californian chert. They included all kinds of activities that were known to be carried out by hunter-gatherers. While the applied contact materials varied from bone, antler, hide and wood to meat, shell, plants, fish and plastic (!), the motions consisted of the more regular ones, *i.e.* scraping, cutting and drilling. Together with the fact that the average duration of the experiments was around 14 minutes, this test belongs to the category of medium complexity. They analyzed the resulting wear traces by means of the high-power method. Fortunately, they presented both the details of the experiments and the answers of the analyst, which made it possible to assess them with the criteria employed for the test with WAVES. It turned out that no adjustments were needed. Methodically this test was very well comparable to ours and also the achievements seemed to be in accordance.

With this explanation in mind, some conclusions can be drawn from table 16. It turned out that the results obtained with WAVES correspond with those of most of the other blind tests, regarding the mean scores as well as the highest individual scores. Moreover, the other results correspond with each other as well, except for the highest individual score on the exact contact material that was obtained by the low-power method 16 years ago. Since the analysts have gained experience, the knowledge on use-wear patterns has increased and the methods have been adjusted and refined, it is to be expected that the results would probably be quite different when this test would be repeated with the present state of knowledge.

From the figures in table 16 it can also be deduced that there are considerable differences in the achievements of individual analysts. In particular this is apparent from the multi-analyst tests. The WAVES test does not turn out to be an exception: while the mean score on the interpretation of the exact contact material is only 46.7%, the highest individual success rate is 60%. The same counts for the inference of the applied motion: 65% versus 80%. It is also remarkable that the achievements obtained by means of the low-power method are less than with the higher magnifications. With respect to the motion, the highest individual scores obtained with WAVES is 60% for macro traces and 80% for micro traces (average 45% (macro) en 82% (micro)).

Does this allow the conclusion that WAVES performs at the same level as other analysts? In answering this question it must be kept in mind that there is one major difference between our test and all the others: the results of WAVES were predominantly obtained by inexperienced analysts. An additional handicap is that some of them were not even familiar with the contact materials from our geographic region or with our flint material. In this respect it is, however, interesting to compare the achievements of our blank student. Although his individual success rates of 53.3% (exact contact material), 60.0% (relative hardness) and 60.0% (motion) are not the best, they are absolutely not out of place in table 16. The score on the contact material is better than all average scores, and the others are similar to the average achievements. Moreover, if the process of wear feature recording by the analysts can be improved by WAVES, I even expect these achievements to ameliorate.

7.5 Discussion

The tests that have been described in this chapter, in particular the second, were mainly meant to validate the functionality of WAVES. Although it has been impossible to validate its entire knowledge base and all other aspects, the tests give an impression of its achievements, of the range of wear traces it can cover and of its practical functionality.

One of the things that these tests have shown is that an expert system application cannot prevent analysts from describing the same traces differently. It turned out that many of the recordings of the analysts deviated as much from the standard description that they were uninterpretable for WAVES. At the time of the first test missing interpretations were not considered problematic because they could be explained. Moreover, it was expected to be obviated when the application would be expanded with knowledge and with photographs and drawings that would illustrate the variables. Moreover, it was assumed that the drawings and photos would give a clear impression of the meaning of the attributes and that this would facilitate the selection of variables. The second test, however, illustrated that this did not prevent discrepancies to occur in the descriptions of the different participants and demonstrated that the data input process needs more attention.

Still, this does not imply that all shortcomings exclusively relate to the analysis procedure employed in WAVES. The achievements of WARP for instance demonstrated that another approach not necessarily yields significantly better results. This shows that part of the functional restrictions of WAVES must be ascribed to the method of analysis itself. For example, some of the misinterpretations of the implements with relatively few traces indicate that the identification of non-diagnostic wear-patterns remains a problem.

It is an illusion to think that an automated approach can overcome these difficulties. Even the very best experts or the most intelligently designed artificial devices cannot work miracles when an artefact does not bear diagnostic remains of its function.

Secondly, these tests have shown that an expert system application cannot prevent that incorrect descriptions yield correct interpretations. For instance in the second test analyst III indicated in the case of experiment 1 that the dominant termination of the use retouch was of the step type. This was an incorrect observation, but it yielded a correct interpretation. Such examples illustrate the need for an expert or experienced analyst to supervise the automated analyses.

Thirdly it was experienced that the application could not prevent the more experienced analysts from unconsciously affecting the reasoning process by selecting those features that confirmed their own ideas. Since the blank student turned out to be the most susceptible for guidance he achieved better results on some aspects of the analyses than some of the other participants. It was therefore concluded that the automated approach works best for inexperienced users.

One of the questions that the evaluation of the functionality of the system was supposed to answer was the degree in which WAVES can substitute the support of the human expert in educating students. I believe that it can be concluded from the success rate of the blank participant that students can practise use-wear analysis without being supervised by a human teacher. But, to this I must immediately add that the personal progression of this student has not been measured. It has not been validated whether he is now able to make an analysis without the help of the application.

In my opinion, the strength of an application like WAVES should not be searched in this direction, however. It shows to full advantage when the expert employs it as an educational aid. For instance, the second test illustrates that the systematic analysis procedure reveals the weak spots of the apprentices at an early stage. Analyst III appeared to misunderstand the difference between intentionally manufactured retouch and retouch caused by tool use. This kind of systematic information not only makes it more easy for the teacher to make immediate corrections and to keep track of the students' progressions, but it also enables the student to validate his achievements. Since WAVES saves the descriptions of the wear features that the analyst gives and shows which feature was responsible for the exclusion of a contact material or motion from the interpretation and how it composed the diagnostic values (see chapter 5.7.3), the student experiences the consequences of his recordings.

One of the other questions that the second test had to answer concerned the students' appreciation of working with the application and whether they would be willing to accept it as a tutoring system. In general, all analysts were positive about the application as a whole. They characterised it as user-friendly and helpful. At some points they had some comments though: several photos were not optimal and some explanations on the meaning of the attributes could be refined. Their main difficulty, however, was to choose between a description of the traces on the ventral or the dorsal side. They were advised to describe the side on which the traces are developed best, but some wear features are most diagnostic on one side and others on the other. It was already argued in chapter 5 (paragraph 5.8.4) that it will be difficult to meet this shortcoming, since it not only requires a serious technical adjustment of the reasoning process but also because there is not enough knowledge available on this aspect. Up till now there has been no systematic recording of all variables on both sides in the experimental programme that WAVES is based upon.

Furthermore, the analysts indicated that they have difficulties with the fact that they can only select one attribute to describe a wear feature. WAVES is only interested in the most dominant characteristics, but in practice this is not always applicable. This can be illustrated by a wear pattern that is caused by butchering. Since it both shows traces of meat and bone working, it forces the analyst to make a choice in the description. Technically it is not very complex to indulge upon the wish from the users to allow for a description of less dominant features as well, but again it simultaneously requires relevant knowledge.

With respect to the question whether the analysts would be willing to employ this system as a standard aid for obtaining a second opinion on their interpretations or as an aid for training, only one of the participants was reserved. Fortunately, this was not the result of a bad experience with WAVES, but a personal preference since he in general did not appreciate computers. Although the others were positive, they also had some suggestions that would improve the application's functionality. One analyst, for instance, expressed the wish for a final conclusion or advice on the basis of the outcome of the analyses. Especially if the answer of WAVES consisted of a compound answer, this participant could not easily decide on the final conclusion. What was actually asked for, however, is an advice of the application of how to interpret its suggestions on the applied contact materials and motions. This would not only be complicated to program, because it requires a human insight, but it may not be sensible at all. The user's final conclusion on the suggestions of the system ought not to be influenced by the system, but rather by an independent adviser. It is for the same reason that a judge cannot ask an attorney how to compose a verdict.

Additionally, the blank student indicated his preference to get an introduction by the expert or another human analyst. He had felt a little bit lost without it. Since this probably represents the lack of self-confidence of any student that is confronted with something new, this is an important signal that ought to be kept in mind.

While the former question referred to the *possibility* to have students or other archaeologists working with the system without the help of the expert and without a basic introduction into use-wear analysis, one should also wonder whether such a development would indeed be *desirable*. I do not think so. It is recommendable not to deploy a system like WAVES unsupervised, because of the large degree of responsibility that users have in interpreting its outcome. The very existence of such a system may also provoke the impression that anybody can use it for the analysis of tools that he or she recently excavated. Such people, unfortunately, I must disappoint. First of all, the test has demonstrated that the weakest point remains the descriptions of the user. These must be absolutely adequate in order to get reliable interpretations. Moreover, once somebody is heading for a wrong direction in describing observed features, he or she can only be corrected by an experienced analyst. However, the system was also meant to offer experienced analysts a means to obtain a second opinion or to validate hypotheses. In view of the fact that the system performs comparable to human analysts (see table 16), I am therefore inclined to advocate the application of WAVES for practical purposes beyond the class room. Analysts who have finished their supervised training may appreciate some support from a knowledge-based application, if only for the mere fact that they can use it as a reference collection. The test has also

shown that analysts who are not familiar with the approach employed by WAVES, may have difficulties to adapt their own approach. For this reason it is recommendable to start using the system in an early stage of the learning process.
To conclude with, it must be kept in mind that the above presented tests only give an impression of the abilities and inabilities of WAVES under blind test circumstances. There may be considerable differences, however, between the achievements in these blind tests and in situations that are encountered in training students. It is also undeniable that the circumstances in real use-wear laboratories are even more demanding. For instance, Unrath *et al*. (1986: 165) have emphasized that blind test assemblages are always simplifications of the archaeological world, because the tools usually lack complex traces that are caused by multiple activities and they have not been influenced by post-depositional surface modifications.
Moreover, these tests do not answer questions like how WAVES performs when different analysts use it for the analysis of their own material, when it is employed to analyze traces on stone materials other than on flint from Northwest Europe, or when it is applied to complex wear traces that were caused by multiple activities. Since the application's knowledge base has not really been tuned to these circumstances, it can be predicted that the present version of WAVES will have difficulties in handling them. Despite its restrictions, however, the comparison of the results of our second blind test with former tests has shown that students can obtain similar achievements as experienced human analysts when they utilize WAVES.

notes

[1] The results of the first test have previously been published in Van den Dries 1993 and 1994.

[2] M.J. Schreurs and Ch. Nieuwenhuis.

[3] M.J. Schreurs.

[4] An interpretation is not very conclusive if it consists of several materials which all have low diagnostic values.

[5] Because of the diversity of the experiments the replicated tools showed a large variety of traces, including generic weak polishes. The archaeological tools, on the other hand, originated from a group of implements which, in its turn, had been selected by the analysts because they showed interpretable traces. Moreover, since it was impossible to revert to experimental evidence to assess the interpretations I had to include traces of which the origine would not be highly disputable.

[6] The composition of this procedure was discussed in more detail in Van Gijn 1989.

[7] This may seem rather fast, but they were already told which part (distal or proximal end) of the implement had been used, so most of them did not screen the entire tool extensively.

[8] Analyst III could not give personal interpretations since he had never seen wear traces before.

[9] The expert was A.L. van Gijn, the two other analysts M.J. Schreurs and Ch. Nieuwenhuis.

[10] The descriptions of the expert were never doubted. In some cases, however, they might have been checked because all other analysts agreed on a different feature. This has not been done since these concerned only minor differences which were of no crucial influence on the composition of the interpretation. At most they only affected the height of the diagnostic values.

[11] The differences in the height of the diagnostic values are caused by differences in the descriptions (see chapter 5.2.4).

[12] Not in all cases the analysts made mistakes in the true meaning of the word, sometimes they described wear patterns which mismatched those of the knowledge base of the application only slightly.

[13] The results of only 15 tools could be validated because Keeley and Newcomer described only 15 in stead of 16 in their paper.

[14] The calculation of the average duration of the experiments is based on the information that was provided for 27 tools. The other three tools, of which one had not been used at all, were not incorporated.