Appendix

The treatise of page III was originally intended as an introduction to the book

Logical Structures for Representation of Knowledge and Uncertainty,
by Ellen Hisdal,
Physica Verlag Heidelberg, A Springer Verlag Company.[1]

However, the 'Introduction' had to be abandoned because it grew to a size of more than 200 pages.

The present treatise is, however, valuable as an appendix to the Springer book. It explains several fundamental notions; such as,

- The operational definition of probabilities (chapter 7), and its comparison with axiomatic probabilities (section 7.5.3),
- Analytic versus synthetic truth (chapter 4),
- The semantic triangle (section 3.1),
- The assertion of a certainty as the assignment of a probability value 1 to an event (section 7.2),
- Bayes' postulate and its drawback, (section 8.3)
- The ambiguity of natural language concerning affirmation and negation (section 9.2),
- The updating of probabilities (chapters 8, 9),
- The role of quantification, classification and negation in logic (chapter 5),
- Probabilistic dependence versus cause and effect (section 10.2),

and a number of other concepts.

[1] An abstract of this book is given on the next page.

Abstract of Book

Logical Structures for Representation of Knowledge and Uncertainty,
by Ellen Hisdal,
Physica Verlag Heidelberg, A Springer Verlag Company.

A knowledge representation system must be able to draw inferences or answer questions concerning previously supplied knowledge and information. For this purpose the book uses a new truth table logic with built-in probabilities in addition to truth values. (Truth values are also called 'possibilities' in fuzzy set theory).

A conditional or *IF THEN* sentence is interpreted as a statement (or enquiry) concerning a conditional probability value.

A result of this probability logic, which is built on top of a yes-no logic, is that in contrast to first-order logic (propositional + predicate calculus) no 'predicate calculus' is needed to represent quantification and classification statements. The $\forall$ and $\exists$ symbols are superfluous. Conjunctions of quantification sentences are represented as conjunctions of *IF THEN* sentences with variables. This simplifies the processing of quantification and classification sentences.

The new 'truth tables' are called 'chain sets'. They consist of chains of 1's and 0's (*true* and *false* values) with a probability and a truth (possibility) value attached to each chain. Inferences in the chain set logic are often the same as in first order logic. However, due to a different inference procedure, inferences of first order which seem wrong according to the inference procedures of natural language (including those used in mathematics and the exact sciences), turn out to be different and more reasonable in the chain set logic. E.g., '$p \rightarrow \neg p$', '$(p \rightarrow q) \rightarrow (p \rightarrow \neg q)$', '$(p \rightarrow q) \rightarrow \neg(p \rightarrow q)$' are all contradictions in the chain set logic, but not in first order logic.

The chain set logic distinguishes between two types of updating of probabilities; and between *uncertainty* (concerning the occurrence or nonoccurrence of a given outcome) and *ignorance* (concerning the values of a probability distribution).

Sequentially supplied items of *IF THEN* information are stored in a special chain set called an *IF THEN* structure. The updating of an *IF THEN* structure by new *IF THEN* information is of 'type 1', while the updating of items of non-*IF THEN* information, as well as modus ponens updating, are of 'type 2'.

Logical Structures for Representation of Knowledge

A Unified Theory of Logic, Probability and Probabilistic Fuzzy Sets

Ellen Hisdal
Institute of Informatics, University of Oslo, Box 1080 Blindern, 0316 Oslo, Norway.

Started 25. 9. 91
Today's date September 5, 1997

Contents

List of Figures

Part I

Overview

Chapter 1

What this book is about

1.1 Introduction

This book deals with the problem of the representation of knowledge and information in a knowledge base. The representation should make it easy not only to retrieve the information in its original form, but also to deduce logically correct inferences. An inference is an answer to a question on the basis of previously stored information, the wording of the question being generally different from that of the originally supplied information. The inference or question-answering procedures are also an important part of the book. Surprisingly enough they take up less space than the knowledge representation part, both in the text of this book and in the question answering procedures of the Alex computer system described in chapterxxx. The reason for *xxx* this is, that once our knowledge base has a proper form, the inference procedures are quite simple.

The inferences can be certain or uncertain. Uncertain inferences are expressed in the form of the probability of occurrence of the event mentioned in the question, conditioned on the information stored in the knowledge base. The answer to an inference is thus a number in the interval [0,1]. A certain answer is expressed by the probability number '1' or '0', corresponding to a 'yes' or 'no' answer to the question respectively. Uncertain inferences have an intermediate probability value. They may be due to uncertainty in the supplied information or to complete ignorance due to absence of the desired information from the knowledge base. The answer to a question will then usually be in the form of a probability interval.

In sect. 1.3 we list the three main tools treated in this book. Of these the chain set system of logic is probably the newest one. This system can give probabilistic answers to questions in the case of uncertainty.

Chain sets have the form of 2-dimensional tables. They can be used for the representation of affirmation, negation and the AND, OR and IF THEN connectives. In contrast to the implication of mathematical logic, the IF THEN chain set structure represents the precise meaning of the extremely important IF THEN statement of natural language.

Classification and quantification structures are basically IF THEN constructions;

such as 'IF x is an instance of a dog THEN x is an instance of an animal'. Chain sets can therefore be used not only for the representation of the negation and the connectives, but also for the representation of the type of problems treated in the predicate calculus part of traditional logic. We believe that chain sets result in a more unified approach to logic than the traditional one, with its sharp bipartition into propositional calculus and predicate calculus. Probably they are also easier to process than the constructs of predicate calculus.

However, a system for representation of knowledge must have a wider frame which defines the structure of the knowledge base (also called lexicon here), the building-up procedures of the knowledge base through a dialog with the user, as well as the dialog facilities for answering questions. Both from the point of view of the user, and from the point of view of of the designer or knowledge engineer, the knowledge representation system must thus be able to run in two modes, the 'information supply mode', and the 'question mode'. Each of these two modes gives rise to a man-machine dialog. In the Alex knowledge representation system that we have built up in Oslo, the 'machine' or program writes its part of the dialog on the terminal under the name 'Max', and the user writes her or his answer or question on the terminal under the name 'Alex'. A printout of the dialog has thus the external form of a play with two participants. Each of the two modes has its own set of 'dialog procedures'. A dialog procedure contains, in part, a Max 'utterance' or 'speech' which is to be written on the terminal. Which particular speech is chosen by the program from the set of available ones depends usually on what Alex 'said' in the previous 'speech exchange'.

In the information supply mode part of the program, the information supplied by the 'man' or user 'Alex' is stored in the lexicon or knowledge base of the system, provided that the system accepts it.

In contrast, the knowledge base is not modified when the system runs in the question mode. In this mode the program assumes that the information stored in the knowledge base is true. And it answers the questions posed by Alex on the basis of this stored information, combined with the inference procedures of the system.

alex4/pinar /ocak/ dia250991 appendix ref-fxxx in footenote. Also xxx ref to spec. info term in footnote add special info terms to fig-dialex!

To give the reader an idea of what we are talking about, we show a partial printout of an Alex info-supply dialog in fig. 1.1. Fig. 1.2 shows the resulting lexicon or database entries and fig. 1.3 a short question mode run of the system.

A knowledge representation system should preferably satisfy, among others, the following basic requirements.

Requirement R 1 ***The information retrieval requirement.*** *The knowledge or data base must have a form which makes it possible, and preferably easy, to find the different items of information that were received and stored in the information supply mode. The form that one decides upon will, of course, affect decisively the procedures of both the information supply and the question mode.*

Requirement R 2 ***The logical and semantic structure requirement.*** *The form of the knowledge base must be such that it facilitates both the checking of logi-*

cal and semantic consistency between different items of received information, and the drawing of logical inferences.

In my experience, the programming of a computer system must start out by restricting oneself initially to simple cases. It is probably impossible to program a man-machine dialog and a knowledge base which, from the beginning, take into account all the different cases which may occur later on in praxis.

However, restricting oneself initially to the programming of simple special cases is not equivalent to the programming of procedures and database structures which work only in these special cases and may have to be redesigned completely when the system is extended to a more comprehensive one.

We have thus the following requirement.

Requirement R 3 ***The flexibility requirement for AI programs, or the requirement of programming-efficiency with respect to extensions of the system.*** *The procedures of the system should be flexible in the sense that the they need not be completely redesigned when the system is gradually extended. Neither should it be necessary to build up a completely new knowledge base in connection with an extension.*

There exist two additional efficiency requirements.

Requirement R 4 ***The requirement of efficient procedures.*** *The procedures of the system should, if possible, be efficient with respect to programming and processing time.*

Requirement R 5 ***The requirement of efficiency with respect to storage space.*** *The structure of the database should be such that it does not require too much storage space.*
How to satisfy this requirement depends on the type of information and dialog that we expect that our system will be applied to. In conventional databases it is usually expected that every entry of a given type will be supplied with the same items of information; such as, e.g., the address, and the date of birth, and . . . , of every employee in a company.

In contrast, in a knowledge base that is being built up on the basis of a dialog that is nearer to a natural language situation, we will have varied types of entries, and varied items of information that are supplied for even one type of entry. Usually we do not even know in advance all the different types of information items that the user may wish to specify for a given entry in the course of time. In this case, which is the one treated in the present book, we have the following requirement.

Requirement R 6 ***The requirement of flexibility with respect to the size of a given database entry.*** *The space allotted to a given database entry should not be of a constant, maximum size that has to accommodate all the different types of*

information that may be supplied in the course of many information supply dialogs. Instead it should, at a given point of time t, accommodate exactly the information that has been supplied up to this time.
E.g., it may be that John's height value has been supplied at time t, and Bill's has not. The database entry for John should then contain this specified height. In Bill's entry, the attribute 'height' should not take up any space. If no attributes at all have been specified for Bill, then there should be no attribute field in his entry.

The efficiency requirement with respect to programming flexibility is, in the case of AI programs, often more important than the requirements of efficiency with respect to processing time and storage space. When we already have a program that functions satisfactorily, the efficiency with respect to processing time and required storage space can often be improved upon without making major changes in the overall layout.

Requirement R 7 ***The natural language interface requirement.*** *The man-machine dialogue should be such that the user Alex need understand only natural language in order to take part in the dialogue. It should not be necessary for the user to understand special logical or computer languages.*
There are, of course, different degrees of natural language compatibility. In the Alex system we have chosen the usual form of a man-machine dialog in which the instructions or questions written on the terminal by the machine 'Max' always have a natural language form. This form is, however, such that the man or user Alex is instructed to write only a word or phrase on the terminal, usually not a whole sentence. In this way the program restricts Alex to those cases that it can handle. Furthermore the program need not make a syntactic and semantic analysis of arbitrary natural language sentences that Alex may decide to write on the terminal.

Requirement R 8 ***The natural language compatibility requirement of the logical structures and procedures.*** *By this we mean that there should exist logical structures and procedures in our system that are equivalent to good logical formulations and inferences in natural language. Here we think especially of the extremely important IF THEN connective when it is used both in statements and in questions in natural language. The meaning of the implication of mathematical logic is not a complete equivalent of the clear and straightforward meaning of the conditional sentence of natural language.*

refer to places in this book

The reason for the last requirement is that we cannot formulate first– and higher–order logical languages without using natural language as our basic metalanguage. If we do not understand the formulations, and the functioning of inference procedures in this metalanguage, we have no means of constructing meaningful and relevant higher–order logical languages.

Requirement R 9 ***The requirement for the processing of uncertainty.*** *The program must be able to represent uncertain information, and to draw the best possible inferences from this information, such that the degree of uncertainty of the inference is specified and has a clear meaning.*

```
Max : Please enter a term.
Alex: dog@

Max : What is the syntactic category of DOG?
Alex: count noun@

Max : DOG is a?
Alex: animal@

Max : A new subclass DOG has been created under ANIMAL.

Max : Please enter a term.
Alex: cat@

Max : What is the syntactic category of CAT?
Alex: count noun@

Max : CAT is a?
Alex: animal@

Max : Is CAT a subclass of DOG?
Alex: n@

Max : Is DOG a subclass of CAT?
Alex: n@

Max : Do you want to create a new partition for CAT under ANIMAL?
Alex: n@

Max : Is CAT disjoint from DOG?
Alex: y@

Max : A new subclass CAT has been created under ANIMAL.

Max : Please enter a term.
Alex: poodle@

Max : What is the syntactic category of POODLE?
Alex: count noun@

Max : POODLE is a?
Alex: dog@

Max : A new subclass POODLE has been created under DOG.
```

Figure 1.1: *Partial printout of a run of the Alex system in information supply mode. Whatever follows 'Alex:' on a line is written on the terminal by the user 'Alex'. Everything else is written on the terminal by the 'machine' or program 'Max'. Before the dialog started, the lexicon or database of the system already contained an entry for 'animal', but without any 'may be a' pointers.* ***figdiainfo***

```
ANIMAL                  [[alex]]
  IS                    [[system]]
    IS A                [[system]]
      CONCEPT           [[alex]]
    MAY BE A            [[system]]
      CAT               [[alex]]
      DOG               [[alex]]
  SYNTACTIC CATEGORY    [[system]]
    COUNT NOUN          [[alex]]

CAT                     [[alex]]
  IS                    [[system]]
    IS A                [[system]]
      ANIMAL            [[alex]]
    MAY BE A            [[system]]
  SYNTACTIC CATEGORY    [[system]]
    COUNT NOUN          [[alex]]

DOG                     [[alex]]
  IS                    [[system]]
    IS A                [[system]]
      ANIMAL            [[alex]]
    MAY BE A            [[system]]
      POODLE            [[alex]]
  SYNTACTIC CATEGORY    [[system]]
    COUNT NOUN          [[alex]]

POODLE                  [[alex]]
  IS                    [[system]]
    IS A                [[system]]
      DOG               [[alex]]
    MAY BE A            [[system]]
  SYNTACTIC CATEGORY    [[system]]
    COUNT NOUN          [[alex]]
```

Figure 1.2: *Four lexicon or database entries of the Alex system after the dialog of fig. 1.1. The entries for 'cat' 'dog' and 'poodle' are new. The 'may be a' pointers of the entry for 'animal' are also new. (The entries have been slightly processed. In the knowledge base itself the indentations are replaced by parentheses. Furthermore the 'special information terms' (in double square brackets here) are stored in a more structured way. See xxxx)* ***figdialex***

```
Max : What would you like to know about POODLE?
Alex: is an animal?@

Max : YES, a POODLE is an animal.

Max : What would you like to know about POODLE?
Alex: is a cat?@

Max : NO, a POODLE is not a CAT.
```

Figure 1.3: *Partial printout of a run of the Alex system in question mode, after the lexicon entries are those of fig. 1.2. Notice that none of the two questions of this run are merely an interrogative form of a supplied item of information of fig. 1.1. They can therefore be answered only by invoking inference procedures. In connection with the second question, the program the program assumes that the classes of one 'may be a partition' (that of 'animal') are disjoint, see the dialog of fig. 1.1.* ***figdiaqu***

1.2 Existing Tools for Representation of Knowledge

At present, the main existing tools for dealing with the knowledge representation and inference problem are those listed in fig. 1.4.

The first item in this list, conventional databases, pay most attention to the information retrieval requirement R1 and to the requirement of efficient procedures R4 of sect. 1.1. The types of information for which a conventional database makes allowance are extremely restricted. Each type of information, e.g. the age of an employee, is assigned a 'field' in the entry of the employee. Both the presence, and the location, and the length of the field are predetermined in advance.

The second item of fig. 1.4, propositional and predicate calculus, reverses the priorities of the first one. Traditional logic makes no provisions at all for the information retrieval problem but pays most attention to the *logical* structure part of requirement R2. However, the efficiency requirement R4 concerning the possibility of efficient programming and processing is probably not satisfied by traditional mathematical logic. Thus Ramsay [47, p. 2] says about propositional and predicate calculus that they possess a number of desirable properties, but are both inadequate and disappointing in some way or other. Winograd [64, p. 97] criticizes predicate calculus for not being amenable to efficient computation.

Neither does traditional mathematical logic satisfy the natural language compatibility requirement R8. We have already mentioned that the implication of logic is not equivalent to the IF THEN statement of natural language. IF THEN statements are handled much better by the expert systems techniques of forward and backward chaining.

Furthermore the requirement R9 concerning the possibility of processing uncertainty fails completely in traditional 2-valued logic. There is no place for probabilistic answers to questions in this system.

Item 3, fuzzy set theory and other variants of many valued logic, are interpolations of 2-valued logic. Instead of working solely with the two truth values FALSE and TRUE, or 0 and 1, they interpolate the operations of 2-valued logic so that the truth of a statement can assume values in the whole real interval [0,1].

1. Conventional data bases.
2. 2-valued logic in the form of propositional and predicate calculus.
3. The different types of many-valued logic, including fuzzy set theory.
4. Semantic networks (also called 'associative networks').
5. The theory of probability.
6. The expert systems techniques of forward and backward chaining.

Figure 1.4: Some of the existing tools for dealing with the knowledge representation and inference problem. ***figtools***

There are two disadvantages to this interpolation procedure. The first one is that although one may watch out that the new operations, such as those for the AND, OR and IF THEN connectives, give the correct answers in the limiting nonfuzzy case, in which solely the two truth values 0 and 1 are allowed, one cannot be certain that the interpolated theory is correct in the fuzzy case. The interpolation procedure has resulted in a multiplicity of operators in the different variants of many valued logic. A number of these are discussed in [3]and [4].

Fuzzy set theory relied originally mainly on the so-called 'noninteractive' max and min operators to represent the OR and AND connectives respectively [66]. Because of dissatisfaction with these operators it has, in cases where this was more expedient, allowed other operators, e.g. $+$ and $\cdot$. During the last years, one of the big subsubjects of fuzzy set theory has been to find other operators. These are required to be so-called t-conorms and t-norms [63]. One is thus left with sets of operators none of which is completely satisfactory.

As concerns the implication, the many different types of many-valued logic use again an interpolation from the implication of 2-valued logic which, as we have already noted under requirement R8, is not generally equivalent to the IF THEN statement of natural language. Zadeh's two(!) fuzzy set IF THEN relations have partially the meaning of the implication of 2-valued logic, and partially that of the IF THEN relation of natural language. In special cases this mixture of meanings gives rise to unacceptable numerical results[23, fig. 15].

The second disadvantage of the intermediate truth values of the different types of many-valued logic is that the meaning of the term 'truth value' is explained solely in terms of other words such as possibility, compatibility, grade of membership etc. [25, p. 94]. Many fuzzy set theoreticians reject completely a probabilistic interpretation of grades of membership or truth values [67]. Neither do other variants of many-valued logic use such an interpretation.

Fuzzy set theory interprets a grade of membership such as $\mu_{tall}(170\,\mathrm{cm})$ both as the *possibility* of u=175 cm for an object labeled 'tall', and as the *possibility* of the
in this book? label 'tall' for an object of height u=170 cm [25, p. 95]. However in [25], [24] and [27] it is shown that the direction of the conditioning is essential for the numerical value. And we argue that $\mu_{tall}(170\,\mathrm{cm})$ should be interpreted as the estimate by a subject of the *probability* that, e.g., a woman of height 170 cm will be assigned the label 'λ=tall' in the presence of different sources of uncertainty, the label λ being an element of a label *set* Λ.

The differentiation between distributions of $(\lambda|u)$ versus distributions of $(u|\lambda)$ has turned out to be essential not only for the interpretation of the grades of membership
xxx of fuzzy set theory, but also in the chain set logic of chapters xxx here. In the this logic we use the name 'possibility' for $P(\lambda|u)$, and 'probability' for $P(u|\lambda)$, where u is a 'chain' *ch* of the chain set (a column of 0's and 1's).

Item 4 of fig. 1.4, semantic or associative networks, have been tried out in many
xxx systems of knowledge representation [17]. The Alex system described in chaptersxxx here incorporates a semantic network which has a very strictly-defined, tree-based structure.

The last two items of fig. 1.4 are only partial tools as far as a system of knowledge representation is concerned.

1.3 The Three Tools of this Book.

The present book contains elements of all the theories and techniques of fig. 1.4, but combines them in new ways.

It presents three main systems namely

The Alex system. This is a system for knowledge representation and retrieval that has been implemented at the Institute of Informatics of the University of Oslo. The implementation incorporates the pure and multiply-partitioned classification system of the next item. Figures 1.1 and 1.3 illustrate two small runs of the Alex system.

A Tree Logic for Classification and Quantification Problems. This logic can also handle multiply-partitioned trees such that a given class can be partitioned into subclasses in more than one way.

The Chain set System of logic. In contrast to the interpolation techniques of the various kinds of many-valued logic, the chain set structure builds a many-valued logic on top of a two-valued one without using any interpolations. The formulas and procedures follow from the meaning of the chains and the attached probability values.

Chain sets are two-dimensional tables. Each column of the table consists of a chain of zeros and ones. In addition, a probability value and/or a possibility value (with a clear probabilistic meaning) are attached to each chain.

Chain sets are especially adapted to the representation of the negation and of AND, OR and IF THEN connectives; where the meaning of the IF THEN connective coincides with its meaning in natural language. The IF THEN chain sets work very well not only for pure, but also for multiply-partitioned classification and quantification structures. Furthermore chain sets can represent several types of partial information which can be represented in a tree logic only in an extremely inefficient and complicated way.

Inferences or answers to questions are given in the chain set system in the form of the probability of occurrence of the event mentioned in the question, based on the information that has been stored in the data base. A '1' answer can be replaced by 'yes', and a '0' answer by 'no'.

Although a pure tree structure is intuitively more transparent in the tree logic than in the chain set logic, the latter has much greater flexibility in the case of departures from a pure-tree structure, as well as in the case when we lack information concerning the exact position of one or more nodes in the tree.

1.4 Declaration of Policy

Our guideline in the course of this work has been to prefer depth to breadth, or quality to quantity if you wish. Boole's declaration of intent, of trying to discover laws, not to create them (see eq. (2.4) below), has been our criterion for quality. This means that the theoretical system has constantly and ruthlessly been subjected to the test of agreement with the reasoning used in the natural language parts of, e.g., good textbooks on scientific and mathematical subjects. It is reasonable to assume that those logical systems which are best adapted to the built-in mechanisms and constraints of our brains are the ones that will function best also for scientific purposes. This does not mean that we cannot transcend the mathematical structures which we use for the purpose of reasoning in natural language, and which are possibly stored as "library programs" in our brains. But if we wish to transcend them, we must again use the basic tools that are at the disposal of our human brains. It is therefore very important to find mathematical structures with whose aid we can simulate our built-in logical reasoning processes. In addition the structures should also be as simple as possible. Once we have a good foundation, it will be much easier to extend the system to more complicated cases.

A 'depth-first treatment' requires that we can distinguish the basic or primary problems from the more secondary ones. It is the deep or basic problems which we then try to solve first. The decision as to which problems are primary, and which are secondary is up to the intuition of the 'knowledge engineer'. The opinions concerning this distinction may vary. E.g., Kempson [35, p. 4] says,

> *In order to have any claims to adequacy, a semantic theory must fulfil at least three conditions:*
>
> 1. *It must capture the nature of word meaning and sentence meaning, and explain the nature of the relation between them.*
> 2. *It must be able to predict the ambiguities in the forms of a language, whether in words or sentences.*
> 3. *It must characterize and explain the systematic relations between words and between sentences of a language – i.e. it must give some explicit account of the relations of synonymy, logical inclusion, entailment, contradiction, etc.*
>
> *Any theory which fails to capture these relations, either at all, or in particular cases making the wrong predictions, must be inadequate, either in principle or in some detail of the theory.* (1.1)

We agree with Kempson's statements in points 1 and 3 (excepting the synonymy part of point 3). These points are discussed in more detail in chapters 2 and 3 here. However, Kempson's point 2 concerning ambiguous meaning, as well as the synonymy part of point 3 are, in our opinion, implicitly given too high a priority by including them in such a general list of requirements for a semantic theory.

That the same word or sentence can have two meanings is an unfortunate and weak point of natural, as well as of artificial languages.[1] Which meaning is the intended one must either be defined explicitly, or must be divined from the context.

A possible representation of two meanings of a word in the Alex lexicon is discussed in sect. **??**. The principle of this representation is that it facilitates the conversion of *xxx* the lexicon entry for the ambiguous word (with, e.g., three meanings) to a lexicon entry of one of the meanings at a time. Each of the latter three entries has the same structure as that of an ordinary entry for a nonambiguous word. Such a structure then confirms our classification of the ambiguity of meaning as a secondary problem.

Similarly it is possible to find a solution to the secondary problem of synonyms, i.e. to the reverse problem of two different words or sentences which have the same meaning. In chapters 2 and 3 we therefore pretend that the problems of ambiguous meanings and of synonyms do not exist, and talk about mappings as if they were always one-to-one. We do keep in mind, however, that our computer system must be such, that it can be adapted to the solution of the secondary problems.

Relativistic linguists such as W. von Humboldt (1767-1835) and B.L. Whorf (1897-1941) have put forward the view that a speaker's language determines his view of the world (or 'Weltanschauung') through the grammatical categories and semantic classifications that are possible in the linguistic system that he has inherited together with his native culture [21, under 'relativity'].

Although we do not wish to dispute this point, we also relegate it to a second rank as concerns our ranging of priorities. E.g., the translation of the words 'animal', 'plant', 'human', 'dog', 'cat', into other languages results in words with precisely the same meaning for most, and probably for all, languages.

Similarly, there exist whole sentences in different languages whose meaning is the same, i.e. they correspond to the same proposition, although their sequential, word-for-word translations may differ appreciably. Hurford and Heasley [32, pp. 21, 22] give the example of fig. 1.5,

	Sentence	Word-For-Word Translation to English
English	*I am cold*	*I am cold*
French	*J'ai froid*	*I have cold*
German	*Mir ist kalt*	*Me is cold*
Hebrew	*Kar li*	*Cold me*

Figure 1.5: *A single proposition expressed in four different languages. The example is taken from Hurford and Heasley [32, pp. 21, 22]. The last line was added by the present author.* ***figiamcold***

They indicate that they consider the four sentences to correspond to the same proposition, i.e., that they have identical meanings; they then go on to say: "*One may*

[1] In connection with artificial languages, think, e.g., of the many meanings of the symbol 'x' in different branches of mathematics.

```
MARY                        [[alex]]
  ATTRIBUTE                 [[system]]
    SPECIF OF OWN TEMP      [[sysalex]]
      ATTRIBUTE VALUE       [[system]]
        COLD                [[alex]]

              (a)
```

```
MARIE                                [[alex]]
  ATTRIBUTE                          [[system]]
    ANGABE VON EIGENER TEMP          [[sysalex]]
      ATTRIBUTE VALUE                [[system]]
        KALT                         [[alex]]

                    (b)
```

Figure 1.6: *Identity of structure of entries in an English and German Alex lexicon respectively for sentences having the same meaning but different word-for-word translation, see fig. 1.5. (It is assumed that the name of the information supplier is 'Mary'.)**figenglishfrench***

question whether perfect translation between languages is ever possible. We shall assume that in some, possibly very few, cases, perfect translation IS possible."

It is such primary, language-independent concepts and propositions whose meaning we wish to capture in our knowledge representation system. The problem of words and sentences whose finer nuances of meaning cannot be translated from one natural language to another is left to a later stage of the investigation. It is not treated in this book.

Suppose that we have two sentences in two different natural languages, but with identical meaning. What we do wish to achieve at this stage is to make the *structure* of the lexicon entries for these sentences as independent as possible of the particular natural language, even though the syntactic forms of the sentences may be quite different. For the first entry row of fig. 1.5, we could then have the lexicon entry of fig. 1.6 (a) when the lexicon is created by a man-machine dialog with an English-speaking user Alex; and the entry of fig. 1.6 (b) when the lexicon is created by a user who supplies the German sentence of the third row of fig. 1.5.

see xxx in footnote

We see that the English and German entries of fig. 1.6 have identical structures, the difference lying merely in the translation of the [[alex]] lines from English to German[2]. The price that we must pay for the structural identities of corresponding entries in an English and German knowledge base is that these entries are much less elegant than the original natural language sentences. However, the structure of the entries is such that it *can* be understood by English and German users respectively who have no special training in the Alex system. Figures 1.1 and 1.3 illustrate that even less understanding is required of Alex, the user of the Alex system, who needs no knowledge whatsoever concerning the structure of the lexicon entries.

1.5 The Contents of the Book

The remaining chapters of part I give a more general overview of the specific problems treated in the following parts. They discuss language, logic and representation of knowledge from a broader point of view.

[2]See section xxx concerning the meaning and translation of the [[alex]] versus [[system]] lines

Chapter 2

Logic, Language, Phonology and Syntax

2.1 Aristotelean and Boolean vs. Modern Logic

Probably more than for any other field of science, an overview of the field of logic should start with the Greeks, more specifically with Aristotle (384-322 B.C.), who gave the first formal treatment of logic and semantics in the *Organon*. This work consists of five parts, *Categories, De Interpretatione, Topics, Prior Analytics* and *Posterior Analytics.* An English translation of the first two of these can be found in [2], and of the last two in [1].

Reading these books, I am always astonished anew to rediscover how much of Aristotle's work in logic and semantics is still an important part of modern teachings. This is in stark contrast to his work in astronomy and physics which needed almost 2000 years to be finally rejected by Copernicus (1473-1543 A.C.), Galilei and Newton.

The reason for the superiority of Aristotle's work on logic and semantics over his work in physics is probably the lack and disdain of experiments in the Greek culture at the time of Plato and Aristotles [49, p. 186]; as well as Aristotle's philosophy of nature which assigned four different directions of movement to each of the four elements, earth, air, fire, water, of which all terrestrial things are made according to Aristotles. Furthermore his philosophy denied that any of these four elements are part of the celestial bodies. The Platonists and their Christian successors held the notion that the earth was tainted and somehow nasty, while the heavens were perfect and divine [49, p. 188].

According to Stigen [57, vol. 1, p. 191], the belief in Aristotle's physics and cosmology prevented the astronomer Ptolemy ($\approx$100 A.C.) from accepting the theory of Aristarchus of Samos who proposed already in the third century before Christ that the earth and planets rotate about the sun. (See also Sagan [49, p. 188] concerning Aristarchus.)

In semantics and logic, the lack of experiments and a wrong theory of physics are far less serious. Workers in these fields have the big advantage over physicists that they get their experimental material practically free as long as they know their language; and as long as they are honest and don't try to adjust the experimental facts, i.e. the use of language and logic, to a preconceived theory.

Aristotle's work on logic and inferences is contained in the 'Prior and Posterior Analytics' [1]. He deals solely with a type of inference that he calls syllogism. This is defined as follows [1, p. 4],

> *A syllogism is a form of speech in which, certain things being laid down, something else follows of necessity from them. By this last phrase I mean that they produce the conclusion, i.e. that no further term is needed to justify it.* (2.1)

Although Aristotle does not say so explicitly, the cases which he treats in [1] are all restricted to what we nowadays would call quantification problems, combined with possible negations. An example of an Aristotelian syllogism from [1, p. 128] is,

$$\begin{array}{c} \textit{All } B \textit{ is } A \qquad\qquad \textit{No } C \textit{ is } A \\ \text{and therefore} \\ \textit{No } C \textit{ is } B. \end{array} \tag{2.2}$$

Also the current meaning of the word *syllogism* refers solely to premisses and conclusions in connection with quantification and classification problems. These problems are, however, an important part of the human reasoning apparatus. In the Introduction to [1, p. viii], John Warrington says about this subject:

> *It is probable that Aristotle discovered the syllogism—and he did no less than that—through his critical appraisal of Plato's recognition of chains of classes, in which each class is a specification of the one above it in the chain [44]. If this is so, then, as Sir David Ross observes, 'Aristotle's translation of Plato's metaphysical doctrine into a doctrine from which the whole of formal logic was to develop is a most remarkable example of the fertilization of one brilliant mind by another.'* (2.3)

The first subject treated in modern textbooks on logic is, however, not the predicate calculus of universal and existential quantification but propositional calculus. This field deals with truth tables for the negation and the AND, OR and IF THEN connectives.

George Boole, the father of modern formal logic, also starts his discussion of logic with the AND and OR connectives [7, pp. 28-33]. However, he does not treat these with the aid of truth tables, but by connecting them up with two operations on sets or classes. Nowadays these operations are called intersection and union. We see that Boole works with a tight connection between logic and the theory of

sets. This important connection has been considerably loosened in later versions of mathematical logic.

eck The present book reinstates much of this connection. Part ?? starts out by treating the core of a classification problem as a hierarchy of classes and subclasses. More complicated cases are then treated as combinations of hierarchical structures.

The basic hierarchial structure of quantification problems is not quite as intuitively obvious in the chain set tables of part ??. But the inference procedures of the chain set system make it possible to trace the single or multiple tree structures contained in the hierarchy.

Both Aristotle and Boole take upon themselves the task of finding the steps used by humans to perform inferences from information in the form of statements in natural language. Boole says explicitly [7, p. 11],

> *It is the business of science not to create laws, but to discover them. We do not originate the constitution of our own minds, greatly as it may be in our power to modify their character. And as the laws of the human intellect do not depend upon our will, so the forms of science, of which they constitute the basis, are in all essential regards independent of individual choice.* (2.4)

This is in stark contrast to the attitude of a not inconsiderable part ot the modern mathematical-logic community. It is, for example, almost generally admitted that the material and other forms of the implication of mathematical logic are not equivalent to the IF THEN connective of natural language. However, in spite of all the ensuing difficulties, mathematical logicians continue to use these implications. Bandler and Kohout [4, p. 767] explain this phenomenon, as being connected with the principle of truth functionality, i.e. with the assumption that the truth value of a composite statement exists for all combinations of truth values of its components, and depends solely on this combination without further reference to the contents of the components. They say,

> *The principle of truth functionality is adhered to despite the anomalies to which it leads in connection with material implication. This principle is so convenient that it looms as an object of desire in the many-valued case also.* (2.5)

In connection with the implication, we shall depart in this book from that part of the principle of truth functionality which asserts that we must be able to specify the truth value of a composite statement for every combination of truth values of its components. The reason for this departure is that this principle forces a truth value on 'A implies B' also in those cases in which A is false. An example of a strange and false conclusion which is the result of this forced method is shown in sect. ??. *xxx*

2.2 Artificial Intelligence and Logic

The present book adheres strictly to Boole's declaration of intent (2.4). In connection with the implication, we then notice that in natural language, the statement 'IF A THEN B' tells us that B is true when A is true. It does not tell us anything about the truth of B or of the composite statement 'IF A THEN B' itself when A is false. Our representation of the implication, both in part **??** and in part **??** is in strict agreement with this interpretation. This means that we do not follow the path of convenience and least resistance expressed so nicely by (2.5).

To use Boole's term, the *discovery* of the inference procedures used in natural language is a typical artificial intelligence problem faced by the 'knowledge engineer' of an expert system. The expert system itself is a system for the representation of knowledge, and for drawing inferences. However, in contrast to expert systems like MYCIN [54] or PROSPECTOR [13], the expertise of our expert lies not in the field of medicine or geology, but in the field of logic. The criterion for an expert in our context is thus a person who can formulate her- or himself precisely, e.g. an author of a good textbook on mathematics or physics or mathematical logic. Let us remember in this connection that no such textbook can be based solely on artificially defined 'object languages', such as the language of propositional calculus. Every object language must, in its turn, be previously defined by successively lower level languages, each of which is a metalanguage relative to the next higher (object) language. The lowest level metalanguage is, necessarily, the natural language in which the concepts and symbols used by the first object language are defined.

Every natural language contains logical modifiers and connectives such as the English NOT, IF THEN, AND, OR. Without these, we would have no means of defining the higher level object languages. As an illustration, the greater part of Kleene's book on mathematical logic [37] consists of English sentences. Neither he, nor any other author of a mathematical or scientific text can manage without making use of the important IF THEN connective of natural language. E.g., in the beginning of his book [37, p. 5] Kleene says,

$$\begin{array}{l}\text{IF } A \text{ is a given } \textit{formula}\\ \text{THEN } \neg A \text{ is a } \textit{composite formula.}\end{array} \tag{2.6}$$

(The capitalizing of the words 'if' and 'then' is our own.)

sectionrefs In sections xxx we discuss the interesting fact that the meaning of the basic logical words, and consequently also the basic logical procedures are largely independent of the specific natural language in which the logical relationships are expressed. This is in stark contrast to the strong language dependence of the syntactical, and even more of the phonological components of natural language. In dealing with logic we therefore deal with processes which may originate from sources that are more fundamental than language itself.

2.3 Language, Phonology and Syntax-Semantics.

2.3.1 What is Language?

Hartmann Stork [21, p. 132] say that the three major branches of linguistics, the field of study of language, are phonology, grammar (morphology and syntax) and lexicology. This classification hides, however, two extremely important aspects namely 1) The differences between languages and 2) The uses of language for various purposes, where different languages are basically used for the same purposes. The differences between individual languages are better taken account of in Vogt's definition of language as a system of 'artificial signs' or 'symbols' [60, pp. 1-3].

A sign is something which stands for something else than itself. Thus a 'natural sign', such as a certain type of cloud, may be a sign of rain, and may induce us to wear a raincoat. A 'symbol' or 'artificial sign', such as red light, stands for danger in the context of traffic. Artificial signs are arbitrarily chosen and based on mutual agreement. They may thus be exchanged by another sign, e.g. blue light, provided that the agreement is changed.

The words of human languages are symbols which stand for something quite different from the combination of sounds or letters which make up the particular word. The choice of the particular combination or symbol is arbitrary, and based completely on agreement within a given language community. The famous story of Babel's tower in chapter 11 of Genesis of the old testament illustrates the freedom of choice of symbols in the different languages, and the confusion resulting from an attempted communication between persons talking different languages.

The confusion is due to two mappings of language, both of which are extremely language dependent. The first mapping is the phonological one, the second the syntactic-semantic one.

2.3.2 The Phonological Mapping of Language

The phonological, and the corresponding written mapping of language is from single objects in the external world to words or terms. The word 'object' is here used in a very general sense. Thus we have, e.g., the mapping of a creature with certain defined characteristics on the term 'human being' in English, 'menneske' in Norwegian and 'ben adam' in Hebrew. The action of an organism of propagating itself quickly upon the surface of the earth by moving its feet is also considered as an individual object in this connection, and is mapped or translated into the English word 'run'. In short, the first language mapping is that which an ordinary dictionary, such as the Webster or Oxford ones, tries to express. The trouble is, of course, that a dictionary has no other ways of pointing to objects in the external world than that of making use of language itself. Thus it must define a concept as a specialization of a more general one; or it must use circular definitions. E.g., 'human' is defined in Webster's dictionary as 'a person', and 'person' as an 'individual human being'. Originally the basic, first mapping of language is, however, from the set of objects in the external world to the set of words in the given language.

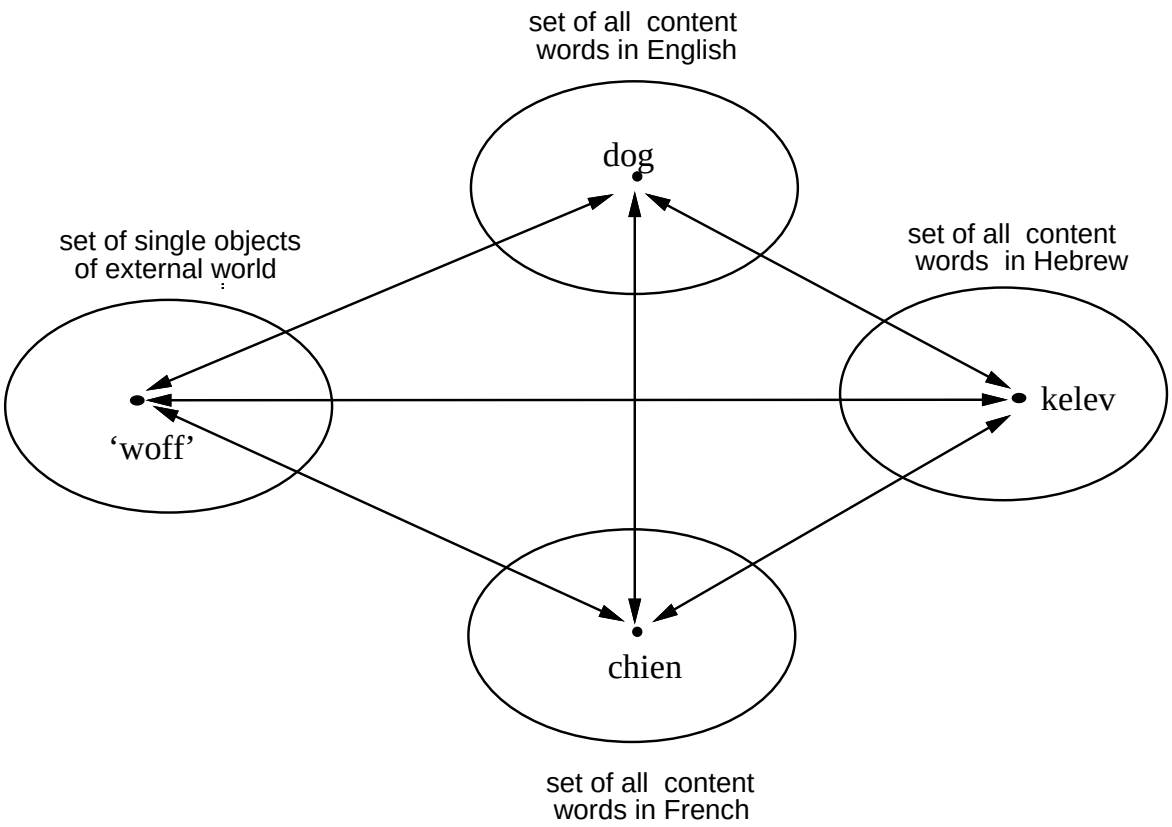


Figure 2.1: The phonological mapping from the set of *objects* of the external world to the set of 'content words' in different languages; and the resultant mapping of the content words of one language on those of another. **figmapphono**

A mapping of the objects of the external world into a second language then gives rise to a mapping between the two languages such as is expressed in a French-English dictionary. The phonological mapping is thus extremely language dependent. The situation is depicted in fig. 2.1.[1]

2.3.3 The Syntactic-Semantic Mapping of Language

We have just seen that the phonological mapping of language is from a collection of objects of the external world to a collection of words. In sect. 2.3.1 we have, however, already cited Vogt who defines language not merely as a *collection*, but as a *system* of symbols. Such a *system* is defined as a collection of symbols in which there exist relationships between the individual symbols. The syntactic mapping of language illustrates this relationship aspect. It is from situations in the external world involving more than one external object to phrases or sentences in the language, i.e. to sequences of words. The construction of these sequences depends upon the

[1] The phonological mapping from single objects of the external world is actually only to the set of 'content words' of the given language. A content word is one which has a full lexical meaning, e.g. chair, as opposed to a function word, e.g. 'the', 'to'. See Hartmann & Stork [21].

syntactic rules of the given language. However, in all languages, these rules keep to the following general pattern. The set of words in the language is partitioned into subsets, each of which represents a specific syntactic category or 'part of speech'; e.g., the subset of all count nouns or of all proper nouns or of all adjectives etc. . Each language then has its specific syntactic rules for the allowed order of words belonging to generally different syntactic categories.[2]

Assuming that we have a syntactically correct phrase or sentence in a given language, we can always replace a word or phrase by another one belonging to the same syntactic category without destroying the *syntactic* correctness of the sequence. In general we will, however, destroy its semantic correctness. An example of such a syntactically correct but semantically wrong sentence is Chomsky's [10, p. 15],

Colorless green ideas sleep furiously. (2.7)

Changing the order of the words to

Furiously sleep ideas green colorless. (2.8)

makes the sentence also syntactically wrong.

Suppose that we have a syntactically and semantically correct sentence containing two countnouns. Exchanging these two will leave the sentence syntactically intact, but may radically change its meaning as is illustrated by the two sentences

The dog bites the man. *(German: Der Hund beisst den Mann.)* (2.9)
The man bites the dog. *(German: Den Hund beisst der Mann.)* (2.10)

The two sentences have quite different meanings in English. However, in German the word order *dog, bites, man* can be used both in the case when the dog does the biting and in the case when the man does it. The two situations are distinguished by the form of the definite articles in front of 'dog' and 'man' respectively. The two forms *der, den* indicate the subject and the object of the sentence respectively.

We thus see that there exists a connection between syntax and semantics. The two English sentences of (2.9), (2.10) differ merely by the exchange, inside the sentence, of two words belonging to the same syntactic category. In spite of that they are mappings of two completely different situations in the external world. Syntactic rules consist therefore not only of the specification of the syntactically *allowed* sequences of words according to their syntactic category. In addition we have syntactic rules which ensure a semantically correct mapping.

A comparison of the English and German sentences of (2.9), (2.10) illustrates the immense interlanguage variability of the syntactic rules. Even two such related

[2]According to Hartmann & Stork [21, under 'part of speech'], transformational grammar uses a reverse definitional technique. It assigns words to categories based on constituent structure in the deep grammar of a sentence; all words which can operate as the same constituent of a sentence are members of the same class.

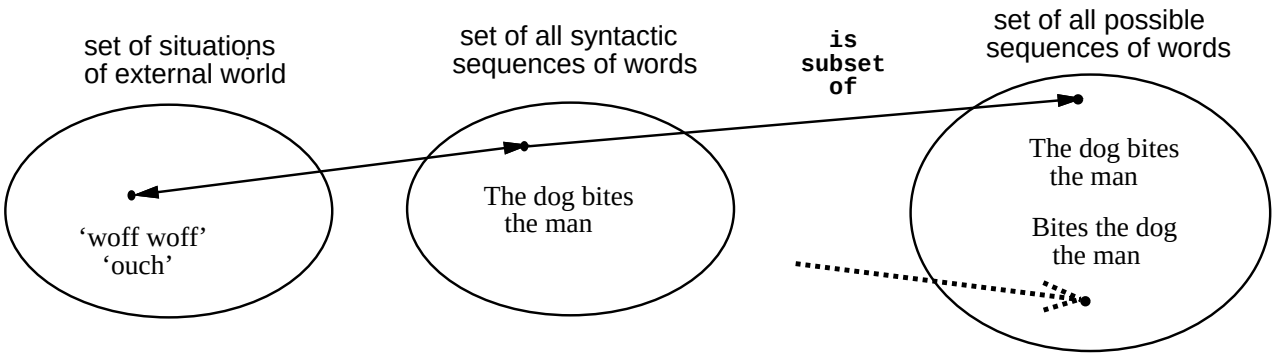


Figure 2.2: *The syntactic mapping from* situations *in the external world to syntactic sequences of words in a given language, here natural language English. The sequence 'Bites the dog the man', which is nearer to the syntax used in predicate calculus, is not a syntactic one in English.* ***figmapsyntax***

languages as English and German use completely different rules to distinguish two situations in the external world. English can only use word order for the purpose of distinguishing the subject from the object of the biting. In contrast, German uses a different inflexion of the definite article attached to the subject as compared with that attached to the object. The order in which the subject and the object appear in the sentence is, however, optional.

Another example of the interlanguage variability of syntax is the use and placement of the definite and indefinite articles. In English both of these are placed in front of the common noun which they modify, e.g., 'the house', 'a house'. In contrast, Norwegian uses the same word for both articles, but attaches the definite article as a suffix to the end of the word, 'huset', while the indefinite article is placed in front of the word, 'et hus'. Hebrew attaches the definite article 'ha' as a prefix to the word, 'habayit', while the indefinite article is indicated by the absence of a modifier, 'bayit'. Other languages have no article at all, e.g., Latin and Russian [21, under 'article'].

The situation concerning the syntax of languages can be summarized as follows. The syntactic rules of a language specify a subset of all possible sequences of words of that language. The subset consists of those sequences which are correct according to the rules of syntax of the given language. Furthermore these rules specify a mapping from situations in the external world to phrases and sentences in the given language. The mapping is illustrated in fig. 2.2.

2.4 The Purpose of the Two Mappings

What is the purpose of the two arbitrarily chosen codes, the phonological and the syntactic-semantic ones? We will mention three purposes here.

1. Communication. (Phonological and syntactic-semantic mapping.)

2. Aid to memory. (Phonological mapping.)

3. Eliminating Misunderstandings. (Syntactic mapping.)

Communication with other humans is probably the main purpose of the two codes. Suppose that we did not have any phonological mapping, and I wanted to ask my husband to buy some milk. I would then have to draw a tall, rectangular box for him; and in addition a cow, otherwise he might come back with a box of apple juice instead. In addition I might have to draw a picture of a supermarket so that he will understand that I am asking him to *buy* milk, not to *drink* milk. Such a direct visual code is, of course, quite possible, but both slow and prone to misunderstandings. A written list or message is also a visual code. However, this code is a remapping of the sounds of the visual code unto the written code.

Although communication probably was the primary purpose of the phonological and syntactic-semantic mapping, I believe that these two mappings have important spin-offs; namely the aid to memory due to the data compression in the phonological map; and the error correcting function of the syntactic map.

Suppose again, that we had no phonological mapping, and that I had drawn a picture of a milk box and a cow for my husband. In addition, I also ask him to buy potatoes and bread by attempting to draw these items for him. He would now have to keep the pictures of these items in his memory until he comes to the supermarket. It is well known that the storing of pictures uses up a lot of memory space. (The transmission of television pictures needs a much greater channel capacity than that of telephone conversations). In addition, a visual object can give rise to practically an infinity of pixel combinations on the retina, depending on the distance and angle from which it is seen. By the time he gets to the supermarket, he has probably forgotten most of the items. However, if he tries to repeat the words 'milk', 'potatoes' and 'bread' on the way, then he has a much better chance of remembering the items when he comes to the supermarket. To make things still easier for his memory, he can use the written code and make a list of the items he wants to buy. Having arrived at the supermarket, he then translates the written or phonological code back to the visual one in order to locate the items. The great superiority of the human reasoning apparatus as compared with that of animals is, in part, due to the phonological map which greatly reduces the number of bits necessary to describe a given situation.

The task of syntax is to avoid ambiguities and misunderstandings. Words are often ambiguous, and their phonological sound patterns are not at all as unique as we might wish them to be. In addition the hearing of the listener may be faulty. Finally we have already seen that a collection of words, such as' bites, dog, man', may be interpreted in more than one way without syntactic rules. The structuring which syntax imposes on a sentence, either by word-order or by inflection, helps the listener (or reader) to identify the message which the speaker (or writer) tries to convey in this 'noisy channel' situation.

In information theoretical terms, we can say that the phonological code effects data compression or reduction of redundancy, while the syntactic code has an error correcting function.

Chapter 3

The Semantic Mapping and Representation of Knowledge

3.1 Introduction

Semantics is defined as the system and study of meaning in language [21], [40, p. 136]. Many different theories of meaning have been put forward by different authors and schools. Our very short description of semantics here is basically probably nearest to Ogden and Richard's view [42, p. 11] that linguistic meaning can be explained in terms of a triadic relationship between

1. The right apex of the baseline of the triangle signifying the set of things or objects or instances to which reference is being made, and
2. The symbol or name used to refer to this set at the left apex of the baseline and
3. The idea, mental image or sense which the sumbol has for a speaker or hearer at the top apex.

This relationship is illustrated if fig. 3.1(a)-(d).

In this book we will give the name 'sensation' or 'point of pattern set' to Ogden and Richard's top apex, see fig. 3.1(d). In sect. 3.2, the point of the pattern set is considered to be an intermediate station between the object or referent on the one hand, and the symbol or word on the other. This station is finally eliminated by replacing the word by its complete 'lexicon entry' consisting of the word itself as well as a formalized description of the corresponding cluster of points in the high-dimensional pattern set.

Furthermore we shall see that the lexicon entry usually defines only a partial meaning of the word. The *procedures* that are invoked when the entry is created, or when additional information is inserted into it, should also be considered a part of the semantic description[1] of the word. This description therefore consists of the lexicon

[1]The semantic description or representation is often called 'the internal representation' in Artificial Intelligence. See, e.g., Charniak and McDermott [9, p. 169].

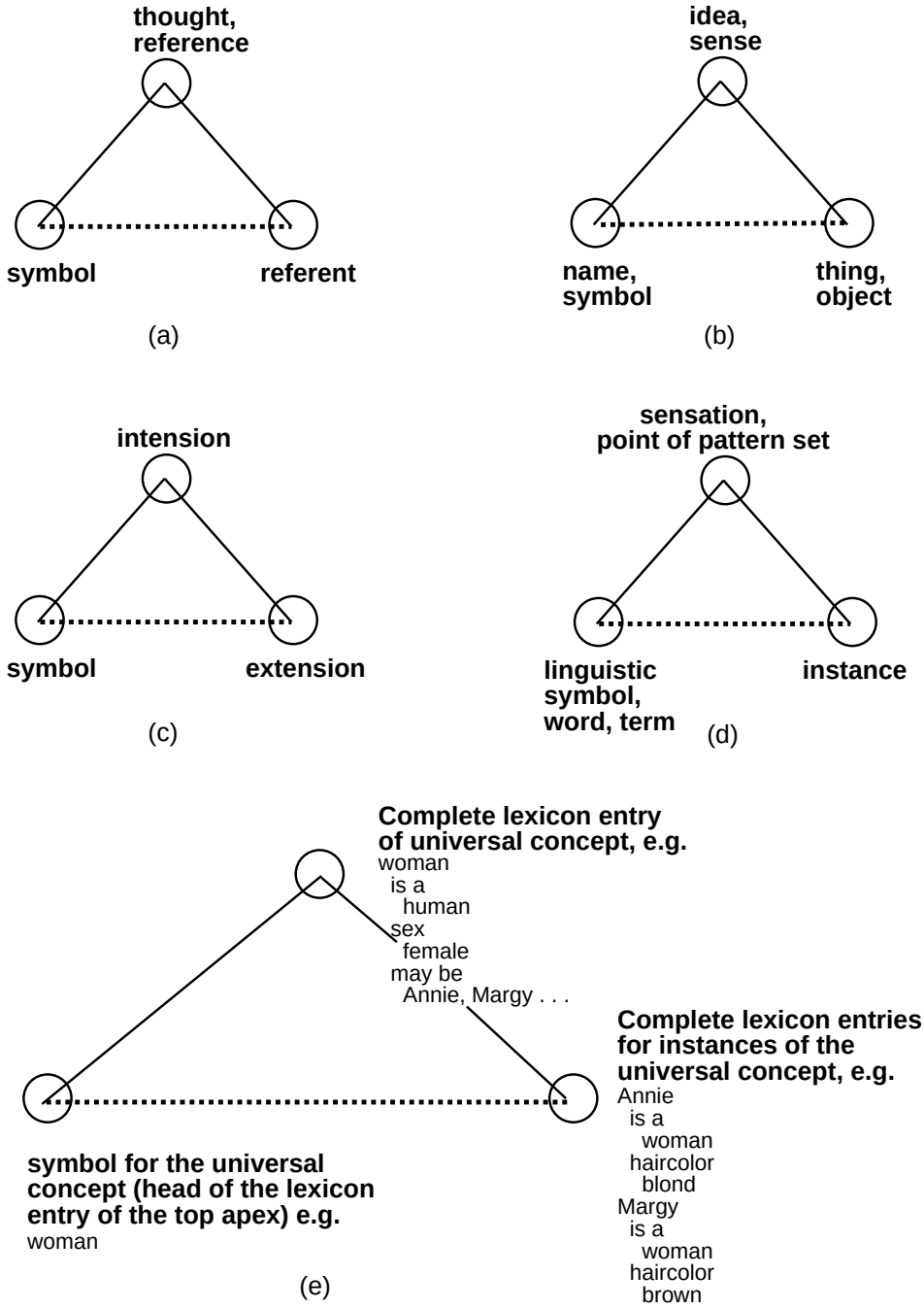


Figure 3.1: The semantic triangle. (a) With the original naming of the apexes by Ogden and Richards [42, p. 11]. (b) With the naming by Hartmann and Stork [21, under 'semantic triangle']. (c) With the naming of the extensional and intensional definitions of a set or class (see [21, under 'extension'] or [9, p. 375]). (d) With the naming used in this book. (a)-(d) refer to the human brain and the external objects. Each of the four figures describe the same relationship, the difference being solely in the synonymous labeling of the corresponding apexes by different authors. (e) illustrates the corresponding triadic relationship in the Alex knowledge representation system in which the brain (top apex), as well as the objects(right apex), are replaced by their representations in the lexicon of the database system. The 'is a' and 'may be' lines in the lexicon entries signify the relationship represented by the upward and downwards directions respectively of the right side of the triangle. E.g. 'Annie is a woman', 'A woman may be Annie. Sect. 3.2 explains in more detail how figure (e) comes into being. The lexicon entries are oversimplified in (e). See xxx for the form of the complete entries. **figsemtri**

entry together with the procedures which set up this entry. Some of these procedures may give rise to the modification of other lexicon entries. The important point about the formalized description in the lexicon concerns the goal that we already mentioned in chapter **??**. The *structure* of a lexicon entry should be the same, no matter what natural language is used by the lexicon. For basic concepts like 'animal', 'cat', 'walk', 'bite' it should be possible to attain the ideal goal of complete language independence of the structure of the entries.

Facilitating translation and communication between informational items expressed in different natural languages is not the only reason for the desirability of a structurally unique semantic representation. Every language can express the same information in more than one way. E.g. in English, each of the sentences,

$$\text{Every dog is a mammal ,} \tag{3.1}$$

$$\text{All dogs are mammals ,} \tag{3.2}$$

$$\text{A dog is a mammal ,} \tag{3.3}$$

conveys the same information. If a knowledge base had stored such sentences in their original form, then the question answering procedures would be very difficult. E.g., if the sentence (3.1) had been received in the information supply mode of the system, and if the system later on had to answer the question 'Is a dog a mammal?', then it would have to rewrite the interrogative sentence into all the three declarative forms eqs. (3.1)-(3.3), and try to find one of these in the database. Furthermore we would anyhow have to analyse the received sentence into its syntactic components in order to know under which word to store it; unless we stored the first sentence under the main lexicon entry 'every' the second under 'all', and the third under 'a'. A sentence such as 'A man whom I met yesterday said that the earth is flat' would then also be stored under the main entry 'a'. Such a storage of completely different types of sentences under the same main lexicon word would complicate enormously the information retrieval process as well as the correct answering of questions.

3.2 The Intermediate Pattern Recognition Step

Considering that, e.g., a dog, who does not have any detailed language, can recognize another dog or a human, or a cat or bird, it is reasonable to assume that the mapping of fig. 2.1 is a composite one with an intermediate pattern recognition station in the brain between the external world on the input side and the set of content words at the output side.

Suppose now that we had at our disposal, just as the brain does, various measuring apparatuses for the different sensory signals, then the composite mapping from the objects of the external world to the set of English words could be described roughly by the diagram of fig. 3.2. Each (measurement) point of the pattern set consists of the combination of different visual, auditory, olfactory (smell) and other sensory attribute values which characterize an object in the external world. The pattern set is thus a set of points in a high dimensional universe. One (composite) dimension,

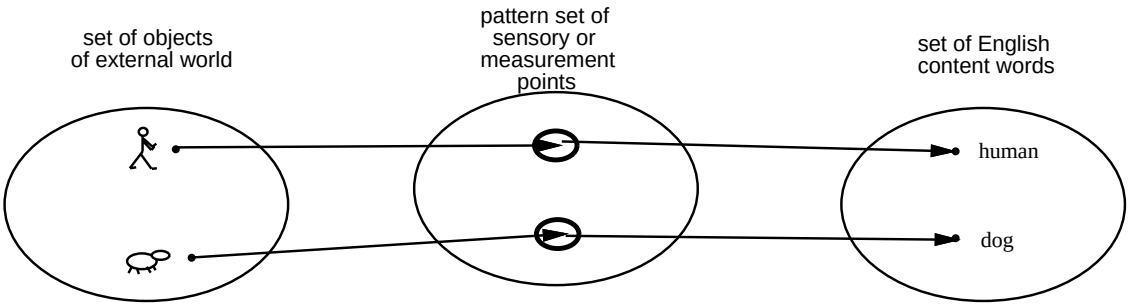


Figure 3.2: The intermediate pattern recognition step between the objects in the external world and the set of English content words in fig. 2.1. The 'pattern set' in the middle illustrates this step. In the brain of an animal, the points of this set consist of the combination of values of the various visual, auditory, smell and other sensory signals which characterize the corresponding object of the external world. In an artificial pattern recognition system a cluster of measurement points in the pattern set corresponds to each object. The pattern set in the middle applies to the sensory signals of all animals which can recognize a human or a dog. The word set on the right hand side applies only to human beings. **figmappat**

or axis in a coordinate system, could be for the values of the visual signals, one for those of the auditory signals etc. Each of these will usually again be a point in a high dimensional universe. Especially the visual signals are extremely complex. Furthermore many visual signals can correspond to the same object in the external world. The 'points' of the pattern set are therefore actually clusters of points, where each cluster corresponds to a given type of object, e.g. the one labeled 'woman', in the external world.

Neither a traditional dictionary nor our knowledge representation system have means of pointing at the objects of the external world or at the points of the pattern set. Instead, the dictionary or lexicon of the knowledge representation system describe in words the cluster of points of the pattern set belonging to a language symbol such as 'human'. The description takes the form of the attribute values which an object must have in order to be admitted to membership in the class named 'human'. These attribute values are, in turn, assigned linguistic or numerical symbols in everyday language. If the description is not to be circular, we must assume that the words for the attributes and their values are connected up directly with a given sensation, or a given result of a measurement, as illustrated by the middle set in fig. 3.2. E.g., there is a general agreement between human beings as to what colors are to be named 'red'. The agreement can be based either on the color sensation of the individual, combined with a learning process during childhood as to what colors are called 'red'. Equivalently, the agreement can be based on a measurement of all the mixtures of electromagnetic wavelengths which give rise to this sensation according to experiments performed by physicists.

Finally we leave out the intermediate pattern set in fig. 3.2, and replace the content words of the right hand set by their complete lexicon entries. This results in the 'semantic mapping' from objects of the external world to the lexicon entries. Fig. 3.3

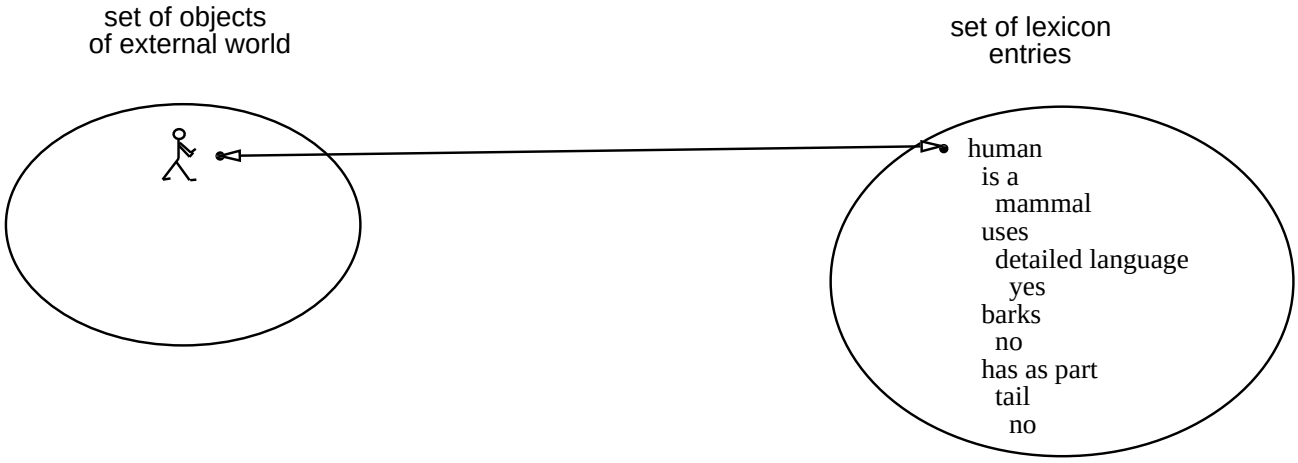


Figure 3.3: The semantic mapping. The intermediate pattern set step of fig. 3.2 has been removed. Instead, the points of the word set have been replaced by complete lexicon entries describing attribute values of the objects to which the words refer. **figmapsem**

illustrates this mapping.

Fig. 3.4 shows an abbreviated sketch of possible lexicon entries for the words 'animal', 'vertebrate', 'mammal', 'human'. Each entry consists of a 'head', namely the word or linguistic symbol for the given class, and a tail. The tail consists of the complete linguistic-semantic descritption of the class. We see that the structure of these entries is much more language independent than the syntactic structure of natural language sentences.

The reader may have noticed the possibility of saving appreciable storage space in the lexicon entries of fig. 3.4 in which subset relations exist between the classes. The description of such saving is one of the tasks of logic. It is discussed in sections **??**, **??**. In fig. 3.3 we have already used such a storage-efficient lexicon entry for 'human' *2x xxx* by using the 'is a' construct of both natural languages and tree logic, assuming that the lexicon contains the entry for 'mammal'.

3.3 Consistency Checks and Logical Procedures

The entries of fig. 3.4 allow us to perform consistency checks between newly supplied information, and information which is already stored in the lexicon. E.g., if the lexicon contains the entry 'human' of fig. 3.4, and if we now are supplied with the two additional pieces of information 'John is a human' and 'John is waving his tail', then the logical procedures of a really good knowledge representation system should send us a warning that John, being a human, does not have a tail, and therefore he cannot be waving it. Likewise Chomsky's sentence (2.7), concerning the colorless green ideas, should be recognized by the consistency checking procedures of the system as being unacceptable, although the sentence is syntactically correct. The check is performed on the basis of the lexicon entry for 'green' which says that it is an attribute value of

```
animal
  organism
    yes
  capable of sensation
    yes
  makes its food by photosynthesis
    no
```

```
vertebrate
  organism
    yes
  capable of sensation
    yes
  makes its food by photosynthesis
    no
  has backbone
    yes
```

```
mammal
  organism
    yes
  capable of sensation
    yes
  makes its food by photosynthesis
    no
  has backbone
    yes
  female has milk glands
    yes
```

```
human
  organism
    yes
  capable of sensation
    yes
  makes its food by photosynthesis
    no
  has backbone
    yes
  female has milk glands
    yes
  uses detailed language
    yes
  barks
    no
  has as part
    tail
      no
```

Figure 3.4: *Four purely semantic, simplified lexicon descriptions. The lines which are repeated in the different entries are later eliminated by the use of the 'is a' construct of the semantic tree logic, see fig. 5.1.* ***figentries***

the attribute 'color'. While the lexicon entry for 'color' contains information which says that it is an attribute which applies only to physical objects, material and light. Furthermore the implied semantic tree structure of the classes in the lexicon (see sectxxx) would show that an idea is neither a physical object, nor a material, *xxx* nor light. The consistency checking procedures of the knowlege base system must therefore protest when the phrase 'green ideas' is part of a sentence whose contents the system is supposed to store. If the phrase had been 'colorless green grass', then the system should not come with a warning concerning green grass, but it should warn us when it comes to the processing of the 'colorless' modification of 'green grass'; and so on for the rest of sentence (2.7).

All these system warnings are based on semantic information contained in the lexicon, and on the procedures for the processing of this information. We will say that the latter are logical procedures, although this use of the word 'logical' goes slightly, but not very much, beyond its traditional use. In our case we store attribute value sets for the different attributes such that the elements of each attribute value set are mutually exclusive. If we have stored in the lexicon the attribute value set {colorless, red, yellow, green, blue, violet} for the attribute 'color', then the warning due to the description of an object as being both colorless and green is equivalent to a warning given by a system based on traditional logic when an object is described as being both 'green' and 'NOT green'.

3.4 Meaning of Sentences versus Meaning of Contained Words

Sentences, and sequences of sentences in natural languages are wonderful tools to describe complete situations involving one or more instances of classes. Equivalent semantic desriptions can, however, become difficult unless we have very simple sentences. This problem has been attacked by Schank with his conceptual dependency theory [51], [52], and by Sowa [55], [56]. More general semantic networks than the strict structures described in parttparttree of this book have also been used for this *xxx* purpose [45]. A good overview of the early literature can be found in chapter 1 of [55]. The book by Charniak and McDermott [9] discusses memory organization based on predicate calculus, as well as on semantic networks.

Below we discuss the connection between the meaning of a phrase or sentence and the meaning of the individual words contained in the sentence only from a bird's-eye point of view. The detailed connections are extremely varied and numerous, and many of them are also language dependent. We must therefore expect decennia of theoretical and programming work to build up a knowledge base system for natural language sentences comprising the knowledge base of even a 5-year old child. Our ambition for this book is merely to describe a foundation which will, hopefully, support all the many and varied intricate, high-level representations.

According to Davidson [11], most philosophers of language concede that a satisfactory theory of meaning must give an account of how the meanings of phrases and sentences depend upon the meanings of the individual words in the phrase. An

exception to this point of view is Davidson himself who claims that we have no use for the meanings of words in a theory of meanings of sentences [11, pp. 306, 307]. I cannot really subsribe to this point of view. There is a clear difference of meaning between the sentences,

The dog bites the man. (3.4)

The dog bites the cat. (3.5)

The dog loves the man. (3.6)

And this difference is due to the difference between the meanings of the words 'man' and 'cat' and and the words 'bites' and 'loves' respectively.

Davidson's own example is the phrase 'the father of Annette' about which he says the following [11, p. 305],

> *Think of the infinite class of expressions formed by writing 'the father of' zero or more times in front of 'Annette'. It is easy to supply a theory that tells, for an arbitrary one of these singular terms, what it refers to: if the term is 'Annette' it refers to Annette, while if the term is complex, consisting of 'the father of' prefixed to a singular term t, then it refers to the father of the person to whom t refers. It is obvious that no entity corresponding to 'the father of' is, or needs to be, mentioned in stating this theory.* (3.7)

To analyse Davidson's last assertion, we start by noting, that a not inessential part of the meaning of the concept named 'father' is transferred to the meaning of the instance of a father called 'the father of Annette'. Namely that part of the meaning which says that an instance of a father must be an animal, and that it must have the attribute value 'male' for the attribute 'sex', and 'adult' for 'age'. We will say that this is the *lexical part* of the meaning of 'father'. In contrast, Davidson's analysis emphasizes the *procedural part* of the meaning.

Instead of saying, as Davidson does, that the meaning of 'the father of Annette' is independent of the meaning of 'father', we will therefore say that one part of the meaning of 'father' is given by the procedure invoked when a sentence such as

John is the father of Annette (3.8)

is presented to the knowledge representation system. In a system which uses LISP as the programming language we can, if we wish, even store this procedure as 'data' in the lexicon entry of 'father'. Assuming that the instances 'John' and 'Annette' are already stored in the lexicon in the forms of fig. 3.5(a), the effect of the father *procedure* which is invoked when the system is presented with the information (3.8), would be to give rise to the modification of these entries illustrated in fig. 3.5(b).

However, before performing the final modification of the lexicon entries, the father procedure would first perform the following checks:

1. Check whether John is an animal and Annette is an animal.

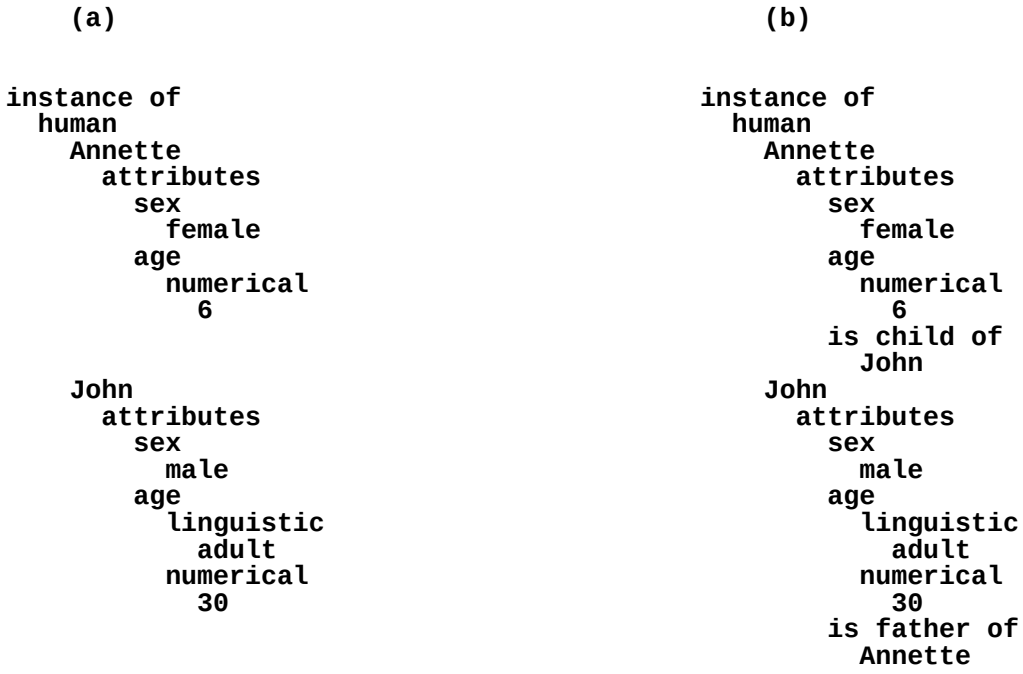


Figure 3.5: Two abbreviated lexicon entries connected by a father-child relationship. (a) Before the information 'John is the father of Annette' was received by the knowledge base. (b) After this information has been received. **figfather**

2. Check whether the species of John is the same as that of Annette.

3. Check whether the age of John is bigger than that of Annette.

All these checking procedures must also be considered to be a part of the meaning of 'father'. Words like 'mother', 'sister', 'brother', 'cousin', etc., denoting other family relationships, will invoke similar checking procedures in connection with phrases such as 'the sister of Annette'. Family relationships are quite specific. There is therefore no way of getting around the necessity of having a special procedure for each such relationship. However, the procedures for, e.g., 'mother', 'father', could make use of a more general procedure for 'parent' before adding an additional procedure for, e.g., 'mother' which first calls the parent-procedure, and then adds a check for the sex of the parent.

Our father-procedure will now continue to work very nicely when the system, after having received the information (3.8), receives the information 'Bill is the father of John' etc. .

In summary, the meaning of, e.g., 'father' is given by the complete lexicon entry of 'father' as well as by the procedure 'father'. In the case of the information supply (3.8), the father-procedure will add information to the lexicon entries of both 'John' and 'Annette'. To the entry for 'John' it would add the information '(is father of (Annette))', and to the 'Annette' entry it would add '(is child of (John))'.

The same general scheme of procedures which create special types of pointers between words, in analogy to the 'is father of', 'is child of' pointers between John and Annette, applies also to the semantic description of situations whose natural language English description is in the form of a 'subject – transitive verb – object' sentence. Fig.3.6 shows a possible semantic description of the 'The dog bites the man' situation depicted on the left hand side of fig.2.2. Here it is the verb of the sentence, 'bite', which is the pointer between two instances. Whether we decide to store the information in the lexicon entry for 'dog', or in that for 'man', or in the 'bite' or 'bitten' entry, or in several of these, depends on the type of questions which the system will have to answer, and on a trade-off between storage space and the complexity and time required for the question-answering procedures.

Be that as it may, our procedures for inserting the meaning of a word into the lexicon will often include instructions for the addition of information to the lexicon entry of another word due to a mutual relationship between the two. We thus end up with a more dynamic definition of the meaning of a word than the usual one consisting solely of the insertion of a single lexicon entry for that word.

The lexicon-modifying procedures used in connection with a specific sentence depend on the syntactic category of the words appearing in the sentence, and on the syntactic structure of the sentence. E.g., the sentences 'Every physical object has weight', 'The dog bites the man' and 'John is the father of Annette' will give rise to quite different procedures. The first sentence will add the attribute 'weight' to the lexicon entry of 'physical object', as well as an 'applies to (physical object)' subentry to the word 'weight'. The second sentence may add information to the

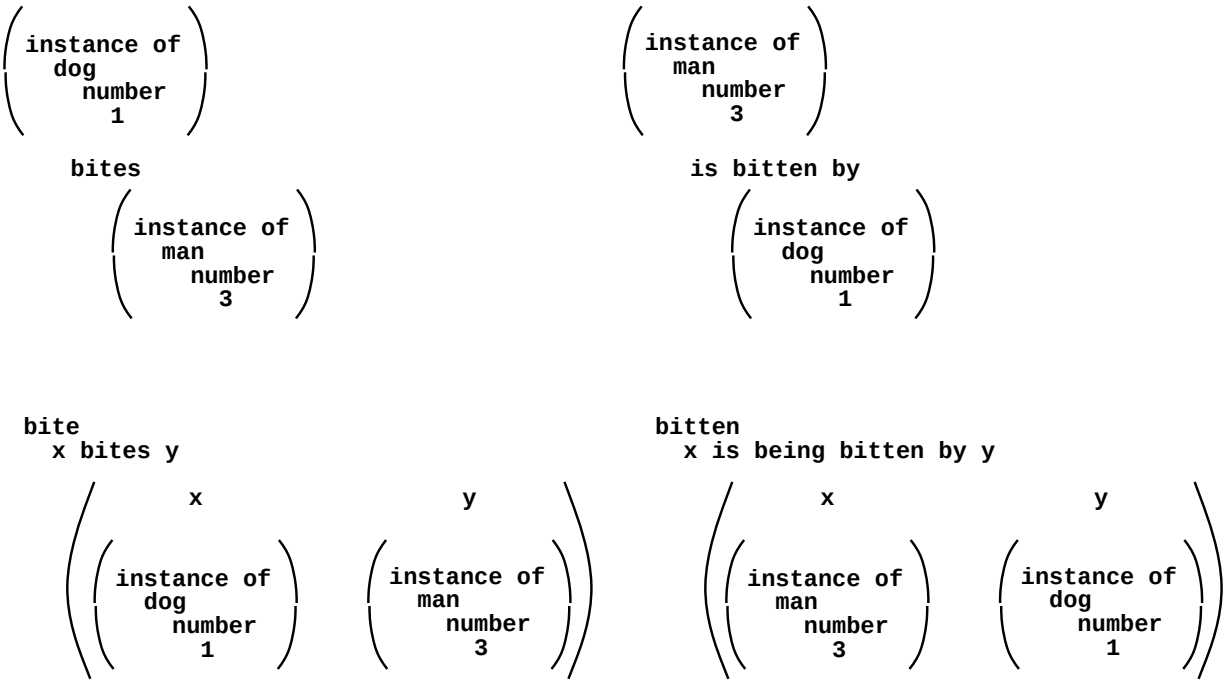


Figure 3.6: Four abbreviated lexicon entries, all of which are the semantic representation of "The dog bites the man". In addition the lexicon will contain an entry for the common noun 'dog', and one for 'man'. It is assumed that we have a procedure for assigning numbers to unnamed instances. **figdogbitesman**

lexicon entry of a particular instance of a dog, to that of a particular instance of a man, and, if we wish, to that of 'bite'. The procedure in connection with the third sentence also modifies the entries of two instances, but it must, in addition, have specific information that the reciprocal pointer of 'is father of' is called 'is child of' (see fig. 3.5).

We see that the lexicon entry for each syntactic category, such as common noun, proper noun, transitive verb, intransitive verb, has a unique form. In addition there are special common nouns, e.g., those denoting attributes or family relationships, which may require the addition of still more specialized information.

Again we emphasize the importance of making the semantic description as independent as possible of the particular natural language to which the lexicon word refers; where 'semantic description' includes both the structure of the lexicon entry, and the procedures for creating and modifying it. Such independence is essential for the logical procedures. It also simplifies enormously the task of translating a knowledge representation system to another natural language. We come back to the translation question in sect. **??**. *xxx*

Chapter 4

Truth

4.1 Introduction

A knowledge base is supposed to contain only representations of true statements. It is also supposed to be able to answer questions truthfully, based on the information which it contains.. The notion of truth is therefore an important one in connection with a knowledge representation system.

Philosophers consider the definition of truth to be an extremely difficult and many-faceted problem. (See, e.g., [20, pp. 86-134]). Here we will present shortly two classifications of truth. The first one, presented in sect. 4.2, is due to Kant. The second one is discussed in sect. 4.3. It concerns mathematical versus scientific truth.

Something about logic in a sectseclogic? *xxx*

4.2 Analytic, Synthetic and Logical Truth

According to Kant, there exist two types of truth. These have been burdened by him with the unfortunate names *analytic truth* and *synthetic truth*. We will also call them *meaning-related truth* versus *factual truth* respectively. They are supplied by meaning-related versus factual statements. An elementary but good discussion of analytic or meaning-related versus synthetic or factual statements can be found in

Hurford and Heasley [32, p. 91-94]. They say,[1]

> *An ANALYTIC sentence is one that is necessarily TRUE, as a result of the senses of the words in it. An analytic sentence, therefore, reflects a tacit agreement by speakers of the language about the senses of the words in it. . . .*
> *A SYNTHETIC sentence is one which is NOT analytic, but may be either true or false, depending on the way the world is.* (4.1)
> *Synthetic sentences are potentially informative in real-world situations, whereas analytic sentences and contradictions are not informative to anyone who already knows the meaning of the words in them.*

An example of an analytic statement is

All elephants are animals. (4.2)

This statement is true according to the meaning of the words 'elephant' and 'animal' in English.

The following is an example of a synthetic statement,

John is from Ireland. (4.3)

There is nothing in the senses of *John* or *Ireland* or *from* which makes (4.3) necessarily true or false.

A more sophisticated discussion of analytic versus synthetic truth is given by Quine [46]. Quine argues in this article that the difference between analytic and synthetic truth cannot be upheld.

Although it may be difficult to agree with all of Quine's arguments, his article
xxx contains interesting examples. In sect. **??** we show how the difficulties posed by some of these *can* be resolved in a knowledge representation system.

Quine (p. 21) defines a statement as analytic when it is true by virtue of meanings and independently of fact. On p. 23, he then subdivides analytic statements into *logically true* ones and a second class for which he has no name. We will call the latter class *semantically true*. A logically true statement is, according to Quine, typified by the sentence

No unmaried man is married. (4.4)

The sentence

No bachelor is married (4.5)

[1] Hurford and Heasley's definition (4.1) makes use of the concept 'sense'. This is defined by them on p. 91 as the indispensable hard core of meaning of an expression. Furthermore they define a contradiction on p. 93 as a sentence that is necessarily FALSE, as a result of the senses of the words in it.

is of the semantically true type.

The relevant feature of the logically true statement (4.4) is, according to Quine (p. 23),

> *that it is not merely true as it stands, but remains true under any and all reinterpretations of 'man' and 'married'. If we suppose a prior inventory of logical particles, comprising 'no', 'un-', 'not', 'if', 'then', 'and', etc., then in general a logical truth is a statement which is true and remains true under all reinterpretations of its components other than the logical particles.* (4.6)

The characteristic of a statement of type (4.5) is according to Quine,

> *that it can be turned into a logical truth by putting synonyms for synonyms; thus (4.5) can be turned into (4.4) by putting 'unmarried man' for its synonym 'bachelor'.* (4.7)

This definition of the characteristic of a semantically true statement is, not quite satisfactory. It cannot be applied directly to the simple example (4.2) of a semantically true analytic statement.

We will therefore, now, complement the definition of an analytic statement in (4.1) by a definition of a semantically true analytic statement which refers directly to the representation-of-knowledge language in the knowledge base system,

> *A statement is a semantically true, analytic one with respect to a given knowledge base when it is a partial or complete repetition of the lexicon entry of a word which denotes a class of objects. The partial or complete repetition of the entry can be retranslated into a natural language sentence, and is then a semantically true statement in natural language. So is any natural-language reformulation of this statement as long as its meaning is conserved.* (4.8)

As an illustration, suppose that the following entry exists in the lexicon,

```
woman
    is a
        human
    attributes
        sufficient set
            sex
                female
            age
                grownup    .
```
(4.9)

This entry is the translation into the knowledge base language of the natural language sentence

A woman is a female, grownup human. (4.10)

presented to the system in the information supply mode.

The following two statements in the knowledge representation language are then semantically true with respect to the lexicon of the knowledge base[2],

$$\begin{array}{l} woman \\ \quad is\ a \\ \qquad human \quad , \end{array} \tag{4.11}$$

and

$$\begin{array}{l} woman \\ \quad is\ a \\ \qquad human \\ \quad attributes \\ \qquad age \\ \qquad\quad grownup \quad . \end{array} \tag{4.12}$$

Retranslated into natural language English, these two statements become the semantically true analytic statements: 'A woman is a human' and 'A woman is a grownup human'.

A result of the definition of a factual or synthetic statement in (4.1) is that synthetic statements usually involve one or more *instances* of a class. This is illustrated by example (4.3) which contains both an instance of a person and an instance of a country.

In the Alex-, and probably most other-, knowledge base systems we do not have to worry explicitly whether a statement is analytic or synthetic. The meaning of words is specified by the user Alex in the man-machine dialog. When Alex presents a word that is not yet stored in the lexicon to the 'machine' then the first action of the Alex program is to inquire about the syntactic category of the word[3]. The program then enters a dialog mode which depends upon this syntactic category. In the common noun mode, the program then instructs the user Alex to specify the 'is a' pointer of the word, and, if she wishes, to specify attributes and their values. If Alex answers all questions posed by Max in a consistent way, then the common noun dialog proceeds to its end, and the supplied information is stored in the lexicon.

For words whose syntactic category is 'proper noun' the dialog is similar but not quite the same. And the storage procedures invoked in the proper noun mode are different from those in the common noun mode because they result in the insertion of *instances* of a class into the lexicon. For an adjective word, the program enters the adjective dialog which is quite different from those of the common noun and proper noun modes etc.

4.3 Truth in Science and Logic

In sect. 4.2 we discussed two types of truth. The analytic one concerning the meaning of words, and the syntactic or factual one concerning the correct description of a

[2]The '(syntactic category(common noun))' has here been left out from the two entries.

[3]See the run of the Alex program in chapter ??

The Requirements for a Mathematically versus Scientifically True Theory

1. Internal Logical Consistency.
2. Economy with respect to logically independent elements (basic concepts and axioms).
3. Agreement with Experience.
   The truth content of a system is measured by the certainty and completeness with which it can be made to correspond with the totality of experience.
   A correct theorem borrows its 'truth' from the truth content of the system to which it belongs.

Figure 4.1: *The Requirements for a Mathematically True and a Scientifically True Theory. Requirements 1 and 2 are sufficient for a mathematically true theory. For a scientifically true theory, all three requirements must be satisfied. The three requirements are taken from Einstein's discussion of his 'epistemological credo' (see [15, p. 12]). Einstein does not try to differentiate between mathematics and the natural sciences.***figtruth**

situation in the external world. All the entries in the knowledge base, as well as the cross references between them, are supposed to be truthful in these respects. Whether they really *are* truthful depends on the information supplier, called 'Alex' in the Alex system.

There remains, however, one additional truth requirement, namely that concerning the truth of the logical procedures. These procedures are initially built into the program of the system. They consist both of the inference or question-answering procedures, and of the procedures for checking the logical consistency of newly supplied information with information which is already stored in the lexicon. This subject is intimately connected with the requirement of truth in the natural sciences. It is discussed by Einstein in [15, pp. 10-13].

The three requirements which Einstein mentions for the truthfulness of a scientific theory are summed up in fig. 4.1. They say 1) that the theory must be logically consistent and 2) that it must be economical with respect to undefined concepts and axioms. The third item requires correspondence with experience. Although Einstein does not say so, it seems to be accepted nowadays that for a purely mathematical theory, the first two requirements are sufficient.

Mathematicians and scientists (excepting mathematical logicians) have always assumed implicitly that the laws governing tests for logical consistency and for logical inferences are well established. In this they have been amply justified. A logical error in the proof of a theorem, derived from axioms or other theorems, is practically always

discovered. Simple errors of computation belong to this class.

Scientists have usually not troubled with the explicit formulation of the laws of logic. They have always intuitively understood the correct use of IF THEN, NOT, AND, OR, ALL etc. in natural languages. And they have made use of natural language in formulating their axioms and the proof of theorems. The task of formulating the laws of logic explicitly was left to philosophers and mathematical logicians. Because the latter were mostly mathematically oriented, they did not require strict adherence to point 3 of fig. 4.1. Indeed many of them will wonder that this point has anything to do with a theory of logic. In the following we argue that the neglect of the third requirement in such a theory is not justified.

No theory of logic can be formulated without making use of natural language, including the use of the logical particles, as a metalanguage. Maybe the most important of the logical particles is the IF THEN connective. The explanation of the use of any table, including a truth table of logic, depends on the use of this connective in the natural language metalanguage. E.g., the truth table for $p \wedge q$ says, in part, that IF p has the truth value T, AND q has the truth value T, THEN $p \wedge q$ has the truth value T. A theory of logic must therefore be able to simulate the functioning in natural language of the IF THEN connective, as well as of all the other logical particles and of the ways of expressing uncertainty.

There is, however, as essential difference between the workings of fig. 4.1 for, e.g., a theory of physics, such as Newton's theory of mechanics, as compared with its workings in a theory of logic. In a physical theory, item 1 of fig. 4.1 is independent of item 3. No matter what the subject of the theory is, whether it deals with gravitation or thermodynamics, the laws of item 1 for establishing logical consistency are the same. They are one of the tests for the correctness of the physical theory. In contrast, in a theory of logic, item 3 is the basic subset and foundation of item 1. Item 1 is thus highly dependent on item 3. Such a theory must therefore start out with investigating item 3 of fig. 4.1, namely the laws which govern the functioning of logic in natural languages, and use these results in connection with item 1, the testing for logical consistency in specific, usually more complicated applications. There are good chances that item 2, economy with respect to basic conepts and axioms, will automatically be satisfied by such a procedure.

If we should find a specific case in which the tentative logical theory does not satisfy the agreementwith-experience-requirement 3, then we must modify, or possibly even abandon, not only the theory of item 3, but also of item 1.

Sophisticated theories, such as all the various mathematical and physical theories, e.g., the calculus or the theory of relativity, can then be built up with the aid of the basic logical apparatus. The logic of natural language is our assembly language without which the higher-level logical languages of mathematics and physics, as well as of computer science, would not exist.

xxx In xxx we show that the failure of mathematical logic to find a correct representation of the IF THEN connective of natural language can, in some cases, give rise to serious difficulties in the iterated testing scheme that we have just described.

Chapter 5

Logic

5.1 Introduction

We finally come to the chapter that has been one of the main goals of part ?? of this book. Namely to define the tasks and purpose of logic. Since logic is supposed to be logical, I assumed initially that the definition of logic would be an easy task. It is an understatement to say that this optimistic assumption turned out to be quite wrong. The reason why the present chapter is preceded by three others is, in part, due to the difficulty of defining the task of logic. Each time I embarked on this project, I found that the ground was not yet prepared.

According to the Webster dictionary, the word *logic* comes from the Latin *logica*, Greek *logike*, feminine of *logikas*, pertaining to speech, reason. The dictionary then goes on to define logic as the science of correct reasoning. Although this definition is correct, it does not bring us very far because we must now ask what *reasoning* is.

Boole's book [7], which was the start of modern logic, is named 'The Laws of Thought'. This title neglects important ingredients of human thinking, e.g., the association in a metaphor of two or more ideas whose literal meanings are unconnected. In spite of this, the phrase 'The Laws of Thought' contains a word which is extremely pertinent for logic, namely 'laws', or somewhat less pompously 'rules'. The representation of information in a logical language, the checking of logical consistency, and the drawing of inferences follow certain rules which are independent of the discussed subject, and even of the particular natural language used. The inferences are true conclusions which can be drawn from given statements, assuming that these statements are true; the formulation of the conclusion being generally different from that of the given statements or the given *information.*

The question now remains as to the purpose of the reformulation of the information in the form of the conclusion, instead of being satisfied with the original formulation.

Before we answer this question, let us look at a 'definition' of mathematical logic which is not uncommon in the literature. Hurford and Heasley [32, p. 133] say that *logical statements are centered around a small set of words, the 'logical vocabulary', namely and, or, not, if, every, some. It is the concepts behind these words that*

logicians have singled out for special attention. In (4.6) we have already cited a somewhat similar statement by Quine concerning logic.

On page 134, Hurford and Heasley then go on to say,

> *The kind of meaning that is involved in words such as 'and', 'or','not' is structural, i.e. it deals with the whole structure of propositions, rather than with individual items within propositions such as names and predicates. . . . The logical words are topic-free and hence more basic or general than individual items within propositions, such as names and predicates. A topic-free meaning is one that can be involved in discourse on any topic whatever, without restriction.* (5.1)

Also this statement is included in Quine's definition which, however, concerns only a logically true statement such as (4.4). Such a 'tautological' statement carries no information with it.

Based on the above discussion, Let us now try the following definition of logic.

Definition 5.1.1 *Logic consists of the following components, both in natural languages, and in more formal logical languages,*

1. *Rules for representing information which is assumed to be true. In natural languages the logical representation of information makes use of the 'logical particles'. The most important of these are* and, or, *if . . . then,* not, every, some. *The first three of these are called connectives. Each of the particles has a specific task and specific rules which apply to it. The rules for different particles may be mutually dependent. Each rule can be applied to a great class of cases, independent of the topic in which the particle occurs.*

2. *Logic also consists of rules for checking whether two items of supposedly true information are equivalent, compatible or contradictory. Equivalent items of information can often be expressed with the aid of different logical particles.*

   Furthermore logic consists of rules for finding truthful answers to questions based on previously supplied information. The question may also contain logical particles, and its formulation may be different from the formulation of the originally supplied information. In the latter case the answer to the question is called an inference.

In the following we try to discover, with the aid of examples, the task of each of the logical particles in the formulation or reformulation of information.

5.2 The Role of Quantification

We start with the role of the particles 'all' or 'every'; that is we start with statements that are called 'universal quantifications' in traditional predicate calculus. These give

rise to the simplest of all Aristotelian syllogisms,

$$\begin{array}{c} \textit{all } A\textit{'s are } B\textit{'s} \qquad\qquad \textit{all } B\textit{'s are } C\textit{'s} \\ \text{and therefore} \\ \textit{all } A\textit{'s are } C\textit{'s}\ , \end{array} \tag{5.2}$$

or equivalently

$$\begin{array}{c} \textit{every } A \textit{ is a } B \qquad\qquad \textit{every } B \textit{ is a } C \\ \text{and therefore} \\ \textit{every } A \textit{ is a } C. \end{array} \tag{5.3}$$

We shall come to the surprising conclusion that such universal quantification is a data compression code which eases the load on the human-, or machine- memory. We must, however, pay a price for this saving of storage space. Instead of being able to to retrieve the original information by using a simple look-up procedure in the lexicon, we must now use a decoding procedure. The decoding rules are called 'inference rules' in the case of a logical language. These rules are also an important part of the field of logic.

The logic of quantification is thus a code similar to the data compression codes of information theory; e.g., the Shannon Fano or Huffman codes. However, instead of making use of statistical properties of the sequence of symbols (letters) in the data representing the original information, logic makes use of structural properties of the information. In the case of syllogisms, these properties can be represented as a tree structure as we shall see in chapxxx.. Because humans have a very well *xxx* developed visual sense, it is much easier for us to remember such a tree structure and the inference rules which it implies than the complete list of attribute values for each entry in, e.g., fig. 3.4.

Universal quantification structures depend on the foresight of natural languages in giving special names to classes which are successive subclasses of each other; such as 'organism', 'animal', 'dog'. There is thus a mutual feedback between the structural properties of quantification and classification schemes and the choice of common properties of objects which are then denoted by common noun words in natural languages.

Probably all natural languages have recognized the advantage of making use of the structural properties of classification schemes in deciding what classes should be supplied with their own names. At the same time they also recognize the necessity of different classification structures for different purposes. The complications to which this gives rise in the tree structure (but not in the chain set structure!) are discussed in chpterxxx. *xxx*

To justify our claim that universal quantification is a data compression code, suppose that the 'all' and 'every' particles did not exist, and that we wanted to define the words 'animal', 'vertebrate', 'mammal', 'human'. For each of these conepts, we

would then have to make a list of the attributes which apply to them as shown in fig. 3.4. The list grows for each successive entry, and is quite long for the entry 'human'.

In natural language we would have the same phenomenon. We would have to say,

A human is an organism
which is capable of sensation
and does not make its food by photosynthesis
and has a backbone
and its female has milk glands
and uses detailed language
and does not bark
and has no tail, (5.4)

and similarly for 'mammal', 'vertebrate' and 'animal'.

Thus, whether we store the definition of 'human' in a structure of the type represented in fig. 3.4, or whether we store it in the natural language form (5.4), it is a heavy load for the memory.

Universal quantification reduces this load considerably. Fig. 5.1 shows the universal quantification form of the lexicon entries of fig. 3.4 in three logical languages, namely 1) Natural language English, 2) Predicate calculus, and 3) The 'is a' form used in many systems for representation of knowledge, including the Alex system of
xxx chapterxxx.

We see that because of the universal quantification device, represented in natural language by 'all' or 'every', in predicate calculus by $\forall$, and in the semantic tree language by the 'is a' line, we have been able to omit many of the entry sublines of fig. 3.4. For example, the entry for 'human' has been reduced by four attributes and their values. The reason that we can carry out such a reduction without loosing any information is that human, mammal, vertebrate, animal are successive subclasses of each other. This can be divined from fig. 3.4 by noting that, e.g., the class 'vertebrate' has all the attribute values of 'animal.' Furthermore the possible values of an additional attribute, namely that of having, or not having, a backbone, are restricted from the interval 'yes, no' to 'yes' only. Similarly mammal has all the attribute values of vertebrate, and vertebrate has all those of animal.[1]

We say that the attributes, and the restriction of their possible values, are inherited downwards in the semantic tree. Whenever such a subclass structure holds between two classes, with a corresponding subset relation between their extensions[2], then we are allowed to use the universal quantifier 'all' or some equivalent expression.

Existential quantification uses the logical particle 'some' in natural language, the symbol $\exists$ in predicate calculus, and the 'may be a' pointer in the tree logic. The

[1] We assume that the attribute values listed in fig. 3.4 are sufficient for the characterization of each class.

[2] The extension of a class denoted by a common noun is the set of instances which belong to this class.

NATURAL LANGUAGE

An animal IS AN organism WHICH
is capable of sensation AND
makes its food by photosynthesis.

A vertebrate IS AN animal WHICH
has a backbone.

(a)

A mammal IS A vertebrate WHOSE
female has milk glands.

A human IS A mammal WHICH
uses detailed language AND
does NOT bark AND
does NOT have a tail.

PREDICATE CALCULUS

$(\forall\ x)$
$(\text{ANIMAL}(x) \Longrightarrow$
$\quad(\text{ORGANISM}(x)\ \wedge$
$\quad\ \text{CAPABLE-OF-SENSATION}(x)\ \wedge$
$\quad\ \neg\ \text{MAKES-ITS-FOOD-BY-PHOTOSYNTHESIS}(x)))$

$(\forall\ x)$
$(\text{VERTEBRATE}(x) \Longrightarrow$
$\quad(\text{ANIMAL}(x)\ \wedge$
$\quad\ \text{HAS-BACKBONE}(x)))$

(b)

$(\forall\ x)$
$(\text{MAMMAL}(x) \Longrightarrow$
$\quad(\text{VERTEBRATE}(x)\ \wedge$
$\quad\ \text{FEMALE-HAS-MILK-GLANDS}(x)))$

$(\forall\ x)$
$(\text{HUMAN}(x) \Longrightarrow$
$\quad(\text{MAMMAL}(x)\ \wedge$
$\quad\ \text{USES-DETAILED-LANGUAGE}(x)\ \wedge$
$\quad\ \neg\ \text{BARKS}(x)\ \wedge$
$\quad\ \neg\ \text{HAS-AS-PART-TAIL}(x)))$

SEMANTIC-TREE LOGIC

```
animal
  is a
    organism
  attributes
    capable of sensation
      yes
    makes its food by photosynthesis
      no
```

```
vertebrate
  is a
    animal
  attributes
    has backbone
      yes
```

(c)

```
mammal
  is a
    vertebrate
  attributes
    female has milk glands
      yes
```

```
human
  is a
    mammal
  attributes
    uses detailed language
      yes
    barks
      no
    has as part
      tail
        no
```

Figure 5.1: *The data compression achieved in three logical languages, as compared with fig. 3.4, when the universal quantification mechanism is used in (a) natural language English, (b) predicate calculus, (c) the semantic tree logic with the 'is a' construct. The predicate calculus notation follows that of Nilsson [41, e.g. on p. 136]. Note that the* 'is a <parent node> which . . .' *notation of natural languages, as well as the* 'is a' *notation of semantic tree logic, but not the predicate calculus notation, indicate which 'attribute value' is considered to be the parent node in the semantic tree.* ***figisa***

latter has the opposite direction of the 'is a' pointer. Inferences in the form of the transitive law are valid in this direction also. E.g., from

$$\begin{array}{c} \textit{some animals are vertebrates} \qquad\qquad \textit{some vertebrates are mammals} \\ \text{we infer that} \\ \textit{some animal are mammals .} \end{array} \tag{5.5}$$

To retrieve the original attribute information stored in fig. 3.4 from the compressed code in fig. 5.1(c), e.g. the information that every mammal is capable of sensation, we use an inference in the form of the transitive law for the 'is a' pointers, in combination with the attribute values listed for the different nodes in fig. 5.1(c). In our example this procedure consists of applying the transitive law of eq. (5.3), setting A=mammal, B=vertebrate, and C=animal. We then find that 'every mammal is an animal'. Looking up the attributes of 'animal' in fig. 5.1(c), we find that every mammal is capable of sensation.

We thus have to pay for the saving of memory space in a universal quantification structure by being forced to apply an inference procedure for answering a question concerning one of the original attribute items under 'mammal' which has been deleted in the compressed logical language. This procedure is more complicated than the straightforward retrieval of information procedure which we could have applied to the more storage-demanding structure of the 'mammal' entry in fig. 3.4.

The most obvious saving of storage space in connection with the universal quantifier 'all' concerns the *instances* of a class. By saying 'John is human', we automatically ascribe to John all the attribute values listed in the 'human' node, as well as all those listed in all the nodes on the rootpath of 'human' in the semantic tree.

5.3 The Role of Classification and Negation

A *classification* or *taxonomy* tree is a special case of a quantification tree. Fig. 5.2(a) shows a quantification tree which is not a classification tree. It can be decomposed into the two separate classification trees (b) and (c) of fig. 5.2.

The requirement for a classification tree is the following.

Definition 5.3.1 The necessary disjointness requirement for a classification tree-structure. *In a classification tree-structure, all the subclasses of a given class (i.e. all the children of a given node in the tree) must be mutually disjoint.*

A consequence of this definition is that two nodes in a classification tree can never have the same name. This requirement is violated by fig. 5.2(a).

Definition 5.3.2 below is a reformulation of the requirement of definition 5.3.1. It makes use of the concept of negation instead of that of disjointness. In the sect. 5.4 below we discuss the relationship between these two formulations.

Definition 5.3.2 The necessary negation requirement for a classification tree-structure. *Let $n1$ and $n2$ be any two distinct nodes in a quantification structure which satisfy the condition that*

$$\begin{array}{c} n1 \text{ is NOT on the rootpath of } n2, \quad \text{AND} \\ n2 \text{ is NOT on the rootpath of } n1. \end{array} \tag{5.6}$$

If the quantification structure is to be a classification structure, then the classes denoted by these two nodes must be disjoint, and consequently we must have that

$$\begin{array}{c} \text{No } n1 \text{ is an } n2, \quad \text{AND} \\ \text{No } n2 \text{ is an } n1. \end{array} \tag{5.7}$$

This requirement does not hold for fig.5.2(a), e.g. for $n1$=vertebrate and $n2$=sea animal.

It is because of the negation requirement (5.7) for cousin nodes in the tree structure that classification trees are much more powerful instruments for inferences than general quantification structures. In contrast to the latter, classification structures can also be used for negative inferences of the type of (5.7).

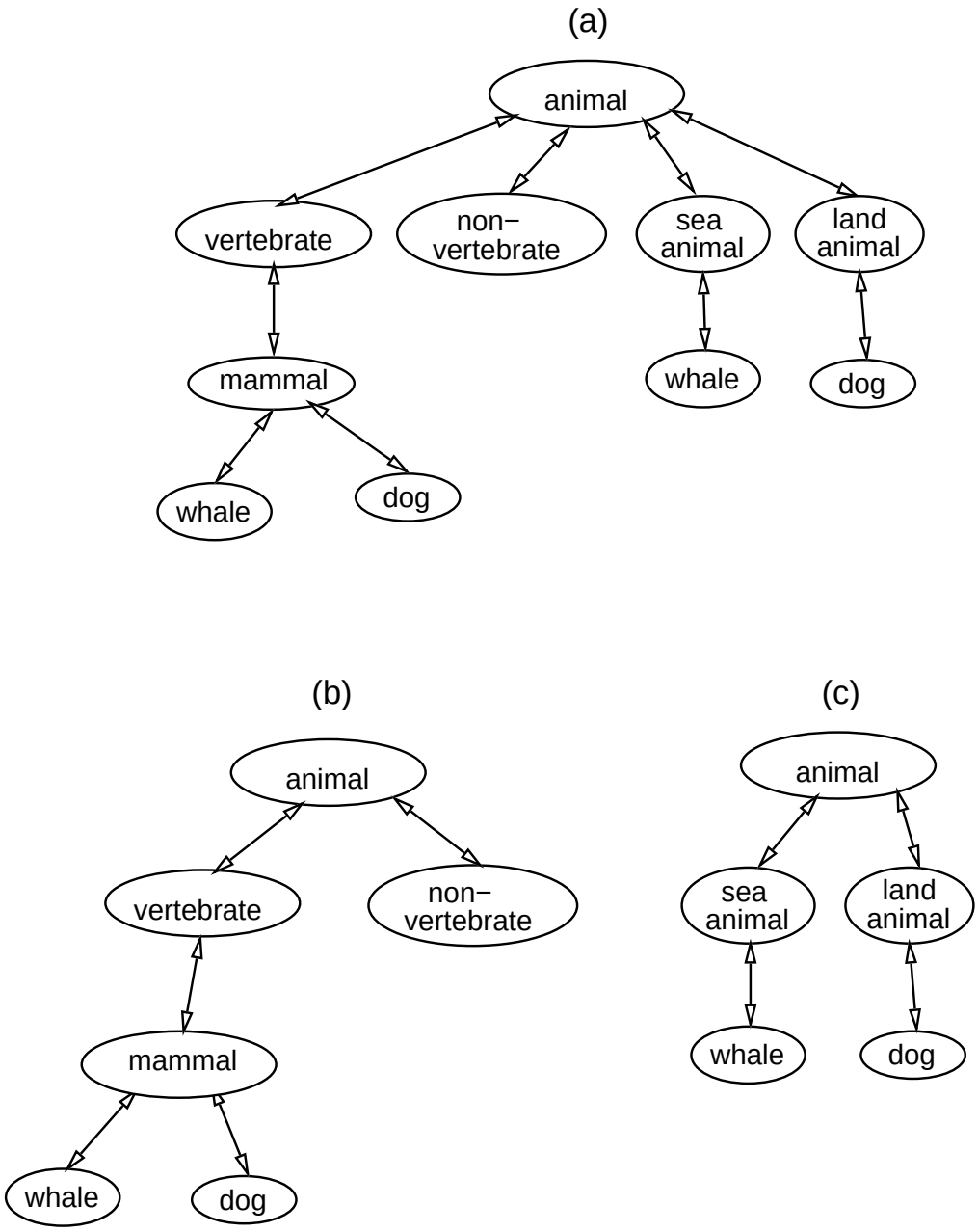


Figure 5.2: Classification versus quantification structures. All three figures are quantification trees. The meaning of the pointers is 'is a' in the upwards direction and 'may be a' in the downwards direction. Note that (a) has some nodes with identical names. It can be used for transitive inferences in the upward and downward directions, but not for negative inferences. (b) and (c) are the decomposition of (a) into two classification structures. Each of these can be used for transitive, as well as for negative inferences. An example of a negative inference drawn from figure (b) is 'A whale is NOT a non-vertebrate'. The requirement for a classification structure or 'taxonomy' is that all classes belonging to a sibling family must be disjoint. ***figdogwhale***

5.4 Negation, Complementation and Disjointness

The negation, just like logic itself, is quite difficult to define. Consider the following tentative definition,

Tentative definition of the negation. *To express that a given statement is not true, we use one of the negation particles NO or NOT, e.g.*

$$\begin{array}{c} \textit{'It is NOT true that the drawer contains forks.'}, \\ \textit{OR} \\ \textit{'The drawer does NOT contain forks.'}. \end{array} \tag{5.8}$$

However, to say that the negation expresses that a statement is NOT true, is a circular definition, and is therefore unacceptable.

I have not found it possible to define the negation without using the word 'NO' or 'NOT', unless one makes use of the concepts of complementarity or disjointness as the more basic ones. We have already followed this course in the formulation of definition 5.3.1 of a classification tree which makes use of the concept of disjointness. The equivalent definition 5.3.2, which uses the word 'NOT', was introduced only as a verbal reformulation of definition 5.3.1.

Definitions 5.4.1, 5.4.2 below of complementation and disjointness respectively show that these concepts *can* be defined without making use of the negation. The negation is then defined as a verbal symbolization of complementation or disjointness.

Definition 5.4.1 Complementation and Negation of Noun and Adjective Phrases. *Let S_1 be the extension of a class C_1. E.g., C_1 may be the class symbolized by the word 'human' in English, and S_1 the collection or set of all persons.*

Complementation and negation always refer to some universe of discourse whose class and extension we will denote by C_U and S_U respectively. S_1 must be a subset of S_U. We will then say that C_1 is a subclass of C_U.

In connection with the complementation or negation of noun and adjective phrases, the universe of discourse is often ambiguous. We therefore require that it must be explicitly specified. For example, we might have

C_1=*the class 'human',*	S_1=*the set of all humans,*
C_U=*the class 'animal',*	S_U=*the set of all animals,*
$C_1 \subset C_U$	$S_1 \subset S_U$.

The complement $\overline{S_1}$ of S_1 with respect to S_U is the set which is left over from S_U when all objects belonging to S_1 are removed from it. The class whose extension is $\overline{S_1}$ is denoted by $\overline{C_1}$ and symbolized in English by 'non-C_1 C_U', e.g. 'non-human animal'. The notation $\overline{C_1}$ must be supplemented by the additional qualification that the complementation refers to the universal class C_U=animal.

The 'non-' prefix can be used in natural language English only in connection with nouns and adjectives. In sect. 5.5 we shall see, however, that most words or phrases

which are part of a sentence, even prepositions, can be negated by the use of the word NO or NOT, combined with the stressing of the negated phrase. In our written text we will indicate the stressed word or phrase by a bold font. In this notation, 'non-human animal' is equivalent to 'NOT ***human*** *animal'.*

Similarly, 'non-female human' is expressed by 'NOT ***female*** *human', and is the complement of the set of all female humans with respect to the set of all humans. It is equivalent to 'male human'; while 'non-human female'='NOT* ***human*** *female' is the complement of the set of all human females with respect to the set of all females. It includes, e.g., female dogs but not women.*

The criterion for which elements of S_U belong also to S_1, and should therefore be removed in the construction of $\overline{S_1}$, defines the class C_1. In the lexicon this criterion takes the form of one or more attribute values which characterize C_1 and are thus either an *addition to,* or a *narrowing down of,* the set of attribute values characterizing an object of C_U. In the example 'C_1=female human' the attribute value set {female, male} of the attribute 'sex' of the class C_U=human is narrowed down to {female} for the class C_1=female human.

Since the set {female, male} of the possible attribute values of 'human' for 'sex' has only two elements, the modification of 'human' by the adjective phrase 'NOT **female**' is equivalent to the modification by 'male'. If we replace 'human' by 'animal' and assume that the set of attribute values of 'sex' for 'animal' consists of three elements, {female, hermaphrodite, male}, then the complement of 'female animal' with respect to all animals is symbolized by 'NOT **female** animal'. This is equivalent to '(hermaphrodite OR male) animal'.

xxx in footnote

A condition for the consistency of the above definitions is that the elements of the value set of the relevant attribute are disjoint.[3] We must therefore also define the concept of disjointness without making use of the negation.

Definition 5.4.2 Disjointness and Negation. *Let C_U be a class with extension S_U. And let C_1 and C_2 be subclasses of C_U with extension S_1 and S_2 respectively, the latter two being subsets of S_U. S_1 and S_2, as well as C_1 and C_2, are said to be mutually disjoint if and only if S_2 is a proper or nonproper subset of $\overline{S_1}$, the complement of S_1 defined in definition definition 5.4.1. The disjointness condition is reformulated in natural language by saying that S_1 and S_2 are disjoint if and only if they contain NO common elements.*

[3]To say that the elements of a set are disjoint is inconsistent according to the terminology of set theory unless these elements are themselves considered to be sets. That this is a natural way of looking at things is easiest to understand in connection with numerical attribute values. E.g., in connection with the attribute height we might use an attribute value set {[0 cm, 159 cm], [160 cm, 179 cm], [180 cm, 250 cm]}. (See xxx for what happens to these originally disjoint values in the fuzzy case when we designate the value set by, e.g., {short, medium, tall}.) Even in the case of 'sex' it is not unnatural to partially define 'female' and 'male' by numerical intervals, e.g. of hormon levels.

5.5 Negation of Words or Phrases

In sect. 5.4 we considered solely the negation of sets and classes denoted by common-noun phrases. In the present section we demonstrate that almost any component of a sentence, irrespective of its syntactic category, can be negated in spoken natural language by the device of stressing it.

The negation of a whole sentence, without stressing any of its components, leaves open a host of possible situations corresponding to the union of the possible situations left open by the separate negation of the different components of the sentence. Each of these separate negations describes a set of situations which is disjoint from the situation set described by the unnegated sentence.

Vice versa, the stressing of the word or phrase to which the negation applies narrows down considerably the set of possible situations, as compared with the set of situations when the negation applies to the whole sentence, without any stressed components. The stress therefore increases considerably the amount of information supplied by the sentence.

Consider the following sentence,

There is a red apple on the table, (5.9)

and its negation,

There is NO red apple on the table, (5.10)

or,

It is NOT true that there is a red apple on the table. (5.11)

The set of possibilities left open by sentence (5.10) or (5.11) is the union of the set of possibilities left open by each of the nine sentences in Fig. 5.3. In each of these sentences, the negation acts on a different word or phrase which is written in bold font. For each of the sentences 1-8 we have indicated in parenthesis one or more of the alternatives that are left open.

These examples indicate amply the insufficiency of working solely with the negation of whole sentences. All the enclosures in the parentheses of fig. 5.3 describe situations which are disjoint from that of sentence (5.9), and which are therefore possible situations in the case when sentences (5.10) and (5.11) are true. And all of them except sentence 9 are more specific, and therefore more information-bearing descriptions of situations, than sentences (5.10), (5.11).

It is a reasonable requirement of a system of logic that it should have representations which differentiate between the above nine cases of negation, thereby preserving exactly the qualitative and quantitative information which each of the eight sentences supplies.

xxx at end of caption of fignotalex

Fig. 5.4 shows how the differentiation between the sentences of fig. 5.3 can be per-

1. There **is** NO red apple on the table. (But there **was / will be**.)
2. There is NO **red** apple on the table. (The apple is **green / yellow**.)
3. There is NO red **apple** on the table. (It is a red **grapefruit / tomato / ball**.)
4. There is NO **red apple** on the table. (It is a **blue ball / red tomato / green apple**.)
5. There is NO red apple **on** the table. (It is **under / next to** the table.)
6. There is NO red apple on **the** table. (It is on **another** table, NOT on the one that we were talking about.)
7. There is NO red apple on the **table**. (It is on the **chair / TV**.)
8. There is NO red apple **on the table**. (It is **in the closet**.)
9. There **is** NO **red apple on the table.**

Figure 5.3: The semantic ambiguity of declaring a sentence as false, i.e. of the scope of the negation. Nine sentences consisting of the same sequence of words and containing a negation. Their meaning depends on the word or words that are stressed by the speaker in order to indicate the scope of the negation. Sentence 9, in which the negation can apply to every component of the sentence, is equivalent to the statement (5.11). ***figredapple***

formed in the Alex system. Fig. 5.4 (a) represents the affirmed sentence (5.9). To negate this sentence we can leave the entry (a) intact, except that one of its 'alex' lines is negated at a time. Fig. 5.4 (b) indicates six different entries, each of which is identical with that in the left hand figure (a) except for the modification of the particular line. All the six modified entries have different meanings, although all of them represent the same sequence of words (5.10) or (5.11). The numbers in parenthesis in the right hand column of fig. 5.4 refer to the particular sentence of fig. 5.3 which the modification of that line in fig. 5.4 (b) represents. E.g., the modification of the 5-th subline 'red' in fig. 5.4 (a) to '((red) 0)' represents sentence 2 in fig. 5.3, 'There is NO **red**' apple on the table'. This notation is actually an abbreviation of the chain set notation of part **??** of this book which can be used for chain sets whose ground universe consists of one element only. In our example this is the element 'red'. To make the system completely consistent, we should use the corresponding notation for affirmation, denoting the affirmation of 'red' by the subline '((red) 1)'. Here we simply assume that our default notation of fig. 5.4 (a) is equivalent to affirmation.

Line 8 of fig. 5.4 (a), representing the time to which the location of the apple refers (indicated by the present-tense verb 'is' in (5.9)), is written in a more conventional database style, assuming that its entries are defined in terms of year, month, day, hour, minute. We could, of course, have written out this line in the 'long-hand' Alex style, and have enclosed it in a box, see fig. 5.5 (a).

xxx The 'enclosing in a box' of several lines of an entry (see) is an important device in the Alex system. In connection with the negation it allows us, e.g., to enclose

There is a red apple on the table

```
1  instance of  [[control]]
2   apple  [[alex]]
3    attributes  [[system]]
4     color  [[sysalex]]
5      red  [[alex]]
6     location  [[sysalex]]
7      time  [[control]]
8       1992 03 19 0700  [[alex]]
9        on  [[alex]]
10         instance of  [[control]]
11          table  [[alex]]
12           number  [[system]]
13            1  [[alex]]
```

(a)

There is NO red apple on the table

2	((apple) 0) [[alex]]	**(3)**
5	((red) 0) [[alex]]	**(2)**
8	((1992 03 19 0700) 0) [[alex]]	**(1)**
9	((on) 0) [[alex]]	**(5)**
11	((table) 0) [[alex]]	**(7)**
13	((1) 0) [[alex]]	**(6)**

(b)

Figure 5.4: The negated sentence 'There is NO red apple on the table' at the top right can have different meanings, depending on which word(s) of the left hand sentence are intended to be negated. The figure demonstrates the representation in the Alex system of six such different meanings all of which correspond to a sentence consisting of the same sequence of words. In natural language these different meanings are differentiated by stressing the item which is intended to be negated. (a) The representation of the sentence without negations. (b) The representation of the negated sentence is the same as that of the unnegated one except that one line is modified at a time. The natural language sentence with the corresponding stress is indicated by the number in bold font in the right hand column. The number refers to the number of the sentence in fig. 5.3. Note that only the [[alex]] lines can be negated. The simultaneous modification of two or more lines is also allowed. It corresponds to the simultaneous negation of two or more items. The representation in fig. 5.4 is somewhat abbreviated. The full representation contains, e.g., a [[control]] line before the negated one to warn the system that it must expect the next subline to have a structure which differs from the default one. The 'time' line # 7 (representing the time at which Alex supplied the information concerning the location of the apple) is an example of a control line (see sect.xxx). ***fignotalex***

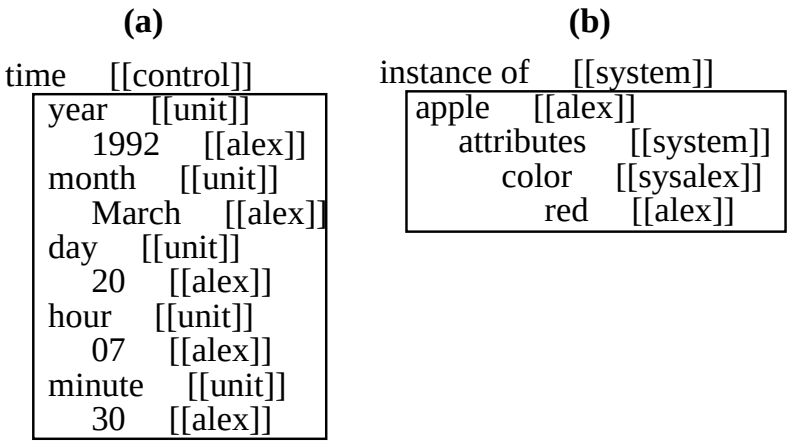


Figure 5.5: The enclosing in a box of two or more sublines of a lexicon entry. The whole box is then formally considered to be a single subline in the outer structure of the lexicon entry. ***fignotalex2***

lines 2,3,4,5 in fig. 5.4 (a) in a box, see fig. 5.5 (b). This box is then formally considered to be a single subline of the 'instance of' line 1 in fig. 5.3 (a) which can now be negated in the usual way. Thus, if we denote the whole box of fig. 5.5 (b) by the symbol x, then the single subline x replaces sublines 2-5 of fig. 5.4. In its negated form this line would now be '((x) 0)', and would represent sentence 4 of fig. 5.3.

The meaning of the negation of a whole box is that at least one of the 'alex' lines in the box is negated. This is in contrast to the separate negation of two or more 'alex' lines which are not enclosed in a box, indicating the simultaneous negation of both lines.

xxx in footnote

The representation of the union of *all* possible cases of negation, i.e. of all possible situations desribed by sentence (5.10) or (5.11), can be accomplished by enclosing all the sublines 2-13 of fig. 5.4 (a) in a box. Calling this box x, the negated sentence (5.10) or (5.11) will now be represented by a lexicon entry consisting of solely two lines on the outer level, namely 'instance of' and ((x) 0) respectively.

Finally we remark that negation can also operate on an expression that is already negated. Since negation of a class or set is equivalent to replacing it by its complement, and since complementation of a complement brings us back to the original class or set, it follows that double negation brings us back to the original. For example, 'NOT (NOT human)'='human'. It follows by induction that NOT$^n C$is equivalent to C when n is an even integer, and to *'NOT C'* when n is odd.

Summarizing the results of sect. 5.5, we note that it is possible to find unambiguous data structures to represent each of the sentences in fig. 5.3. All these data structures are derived by negating one or more 'alex' lines of the basic data structure of fig. 5.4 (a). The various resultant structures differentiate between all the different negated possibilities which the sentence 'There is no red apple on the table' leaves open.

5.6 The Role of- and Summary of- the Negation

Looking back upon our finding of the meaning of the negation as a symbolization of the complement of a situation set, we discover that not only quantification, but also negation fulfils the task of abbreviating or compressing data. In contrast to the quantification case, the compression will often lead to a loss of *detailed* information in the case of the negation. However, the details of the information can be quite irrelevant for the information seeker.

E.g., if you ask the question 'Is there an **apple** in the icebox?' and get the answer 'no', there are lots of possibilities open for things which *may* be in the icebox; such as oranges or butter or cake, all of which are disjoint from $\{apple\}$. However, if you are interested solely in eating an apple, it is irrelevant for you to know whatever else there may be in the icebox.

To take a less materialistic example, suppose that you live in Paris and are planning a business trip to London, where your friend Margy lives, on May 5-th. You send a fax to Margy's office, asking whether she will be in London on this date, and get the answer 'no'. This answer implies that on May 5-th Margy will be in any place in this world that is disjoint from London. She may be in Rome or New York or New Delhi or Sydney, or any one of many thousands of other places. But for you it is irrelevant at which of these places she will be. You know that she will not be in London, and therefore you will not set off time to visit with her on that day.

As we have illustrated in fig. 5.3 and fig. 5.4, not only nouns, but almost any word or phrase in a sentence can be negated. This includes adjectives, verbs, adverbs and even prepositions; the implicit assumption being that the sentence as a whole is unchanged except for the particular item that is being negated. The formal structure of this notation in fig. 5.4 is much nearer to that used in all natural languages than the notation of traditional logic which allows only the negation of complete sentences. Following the notation in Nilsson [41, chapter 4], the sentence 'There is a red apple on the table' would have to be decomposed into the following three sentences in traditional logic, each of which could be negated separately,

$$\begin{aligned}(\exists x)\, ON(x, TABLE-1)\ \wedge\\ RED(x)\ \wedge\\ APPLE(x)\,.\end{aligned} \tag{5.12}$$

Here we have not yet indicated the time to which this description refers. Nilsson does not say anything about a notation in traditional logic for the important temporal of information.

Furthermore Nilsson's notation assigns a particular identifier number through a device of adding the suffix '–1' to the word 'table'. In contrast, the 'instance of [[control]]' line in fig. 5.4(a) eliminates the necessity of modifying the subline 'table' itself by adding a suffix to it. The two main differences between the notational language of the Alex system versus that of predicate calculus are however the following.

The first difference consists in the obligatory explicit use of variables in predicate calculus in connection with quantification; such as the variable x in eq. (5.12) The explicit use of variables is not obligatory in the Alex system. This subject is discussed in more detail in sect. 5.7.

The second main difference between the two notations is connected with the first one and consists in the readability of the lexicon; and in the degree of ease with which a given item of information can be retrieved, as well as the degree of ease with which one can process the drawing of inferences, the answering of questions, and the checking for consistency when new information is added to the knowledge base.

According to Hatcher [22, p. 2] the negation of traditional logic is an operator whose effect is given by its truth table. The table tells us that if the sentence is true, then its negation is false and if the sentence is false, then its negation is true.

Here we have defined a different, more semantically oriented meaning of the negation. Namely that of leaving open any of the situations described by the replacement of the negated word by an unnegated one denoting an element of the complementary class to the class of the originally negated word. It is then implied that one of the situations corresponding to an element of the complementary set is a truthful, partial description of the state of the external world. The verbal symbolization of the universal set with respect to which we take the complement must be of the same syntactic and semantic category as that of the originally negated word. E.g., the negation of 'on' in the sentence 'There is a red apple **on** the table' must be a preposition. Furthermore not all prepositions can be used as elements of the universe of discourse in connection with the complementation operation, only those which indicate a location. This is due to the 'location' line 6 in fig. 5.4(a), this being a superline of 'on'. Thus the preposition 'of' does not belong to the universe. Finally the elements of the universe of discourse must refer to mutually disjoint situations. E.g., we cannot have both 'on' and 'on top of' as elements of the universe of discourse used in connection with the complementation operation. But we can have 'under' and 'next to' in addition to 'on'. Proper nouns and pronouns are considered to belong to the same syntactic category in the present context. This can be illustrated by the sentence '**John** was NOT there, but **I** was'.

The emphasizing in natural language of the negated phrase, and its equivalent representation in our data structure, e.g. in fig. 5.4(b), makes the negated sentence much more informative than when the whole sentence is negated, without indicating to which part of the sentence the negation refers. The declaration as false of a whole composite sentence without emphasizing a particular part, such as in 'It is NOT true that there is a red apple on the table', is an extreme case of uninformativeness. Our notation takes care of this case also by the use of the 'wrapping up in a box' device.

5.7 The Use of Variables in the Object Language

Sløyf?

Mathematical logicians lay very great stress on the distinction between 'meta-language' versus 'object language'. The metalanguage is the language used for the definition of a logical system. The language of the logical system itself is called the object language. Such an object language #1 can become a metalanguage with respect with respect to a higher order object language #2 when the latter is defined in language #1. E.g. eq. (5.12) is an example of the use of the object languages of predicate and propositional calculus. The lowest and most fundamental meta language is, necessarily, always natural language.

Actually I do not believe that the relations between object and metalanguages are quite as simple as that. Instead there is a certain feedback effect between the two.

Part II

Probabilities for Use in Logic

Chapter 6

Uncertainty, Probability and Logic

6.1 Overview

The present part II of this book prepares the ground for the representation of knowledge with the aid of the chain set tables and the conditional probability tables of part ??. Those who wish only to get acquainted with the probabilistic tools for the treatment of certainty and uncertainty in logic can probably skip chapter chapter 6 and, if they are acquainted with the most basic concepts of probabilities as limits of relative frequencies, also chapter 7.

The reason why we have included chapters 6 and chapter 7 here is that the scientific literature abounds in discussions concerning the best treatment of uncertainty. Different investigators recommend different tools; such as the use of probabilities in their frequency interpretation, subjective probabilities, belief functions, possibilities of fuzzy set theory, many-valued logics, Bayesian versus non-Bayesian methods, deductive versus inductive reasoning and more. Chapters 6 and chapter 7 are included in order to give the reader a small taste of the ongoing discussions.

In part II we lay a foundation for the use of probabilities in logic. The main points of this foundation are the following.

1. A clear differentiation between the updating of relative frequencies versus the updating and narrowing down of the set of possible underlying probability distributions, each element of which could have given rise to the observed relative frequencies. This subject is discussed in chapter 8, sections 8.4-8.9.

2. The updating under item 1 refers to a single underlying probability distribution which one tries to learn from a sequence of prolongations of an experimental series. The present item refers to the updating by additional information supply concerning the object set to which the probabilities refer. The object set is narrowed down, with a consequent change in the underlying probability distribution. This subject is discussed in sections sect. 9.1 and XXXX. *XXXX*

3. The necessary tools for distinguishing the conditional probability $P(u|v)$ from the conditional probability $P(v|u)$. Both the relations of traditional logic, and

the grades of membership or possibilities of fuzzy set theory have neglected the distinction, as well as the connection, between conditional probabilities (or possibilities) with different directions of conditioning. It turns out that the distinction between these two helps us to establish the connection between probabilities on the one hand, and the truth values of both 2-valued and many-valued logic (or the grades of membership of fuzzy set theory) on the other.

4. The use of Bayes *postulate* (not Bayes *theorem*!) for the representation of the state of ignorance has always been an extremely controversial subject (see, e.g., [18, p. 67 et seq]). Bayes postulate can help us part of the way. As shown in sect. 8.3, it does, however, include an element of inconsistency which makes it useless for the solution of more complicated problems. For a consistent representation of the state of complete or partial ignorance we must use the m-notation of sect. 8.4 and the updating rules of chapter 8.

6.2 The Treatment of Uncertainty

Most descriptions of real-world situations involve, necessarily, an element of uncertainty because the describing person lacks precise information. For example, a pedestrian who wants to cross a road, and sees an approaching car, must decide whether to cross before or after the car has passed. Her decision will be based on a rough estimate of the velocity and distance of the car.

The main tool for treating situations involving uncertainty is the theory of probability. Outside the theory of probability we have different systems of many valued logics, including fuzzy set theory, as well as modal logic.

The majority direction inside the theory of probability interprets probabilities as limits of long-run relative frequencies. In addition we have the theory of subjective probabilities (see sect. 6.4.3) which does not use such an interpretation, but identifies probabilities with degrees of confidence or belief (see, e.g., Jeffreys' book [34, pp. 15, 369]. According to the preface in that book there is, however, very good agreement between the methods following from the theory of subjective probabilities and those recommended in statistical praxis.

Dempster Shafer's theory of evidence occupies an intermediate position between probabilistic and non-probabilistic methods (see sect. 6.4.3).

Unfortunately the different directions, both within and outside the theory of probability, seem to defend their positions with almost religious zeal.

6.3 Bayesians versus non-Bayesians

Both within the theory of probability, and in connection with fuzzy set theory as contrasted with probability theory, the differences of opinion concerning the best method for the analysis of data are often represented in the form of being a 'Bayesian'

versus a 'non-Bayesian'. This categorical formulation severely beclouds the issues because it is far from clear where being a Bayesian starts and stops.

The main criterion for being a Bayesian is probably that of making use of Bayes' formula for finding the best estimate of an underlying, unknown probability distribution from the observation of a series of data to which this distribution gave rise (see *check sect.* 10.6). Bayes estimate of a probability distribution from a series of observed data is also called the *posterior distribution*. This is in contrast to the *prior*, unconditioned probability distribution before the data were observed.

To find the estimated *posterior* distribution, Bayes makes use of the law of compound probabilities. This basic formula connects different probabilities in a two- or higher-dimensional universe or space $U = V \times W$. It is stated and proved in sect. ?? here. Bayes applies the law of compound probabilities to the special case in which a point of V is a series of N observations to which the sought after probability distribution gave rise, and a point of W is a possible value of a parameter of the underlying distribution.

In addition to Bayes' *formula* we have Bayes' *postulate* which says that when the prior probability distribution is unknown, then we should postulate that it is uniform. This postulate is more debatable than Bayes' formula. The difficulty to which it gives rise is discussed in sect. 8.3 here. The m-notation of sections 8.4-9.1 resolves the difficulty.

Jaynes [33] and Laviolette & Seaman [39] compare 'orthodox' and 'suboptimal' statistical methods with Bayesian ones. Jaynes demonstrates that the Bayesian method is easier to apply, and yields the same or better results. And Laviolette and Seaman say,

> *Being statisticians, we are acquainted with procedures, such as classical statistical inference, which are suboptimal with respect to criteria founded on Bayesian statistical decision theory, but which remain very useful. Indeed in the absence of an existing Bayesian method for approaching a particular problem, and lacking the time to develop one, we have gladly employed "suboptimal" procedures. However, the need for expediency does not preclude the development of other, superior, solutions.* (6.1)

The majority of fuzzy set theoreticians work completely outside the theory of probability and scorn the Bayesian point of view.

In this book we do not make use of the special 'Bayes-application' of the law of compound probabilities. We do, however, apply the law of compound probabilities in connection with the differentiating between $P(v|w)$ versus $P(w|v)$. In addition we discuss the insufficiency of Bayes *postulate* in sect. 8.3 and in sectionXXX. The *m-notation* of sect. 8.4 avoids the use of Bayes postulate and is able to solve problems for which this postulate is inadequate.

quantification problems in semantic tree section

Although we do not make use of Bayes *formula*, i.e. of the special Bayes application of the law of compound probabilities, we explain it summarily in sectionXXXX. *XXXX*

We have several purposes with this explanation. 1) If the reader is a confirmed non-Bayesian, we wish to make it clear that the law of compound probabilities has no connection with Bayesianism. This law can be proved by elementary methods in the theory of probability.

2) The second purpose with the presentation of Bayes formula is to clarify the question of deductive versus inductive reasoning in connection with a data base. Although the data base procedures suggested in this book make use solely of deductive logic, the person who supplies factual information to the data base must, necessarily, rely on inductive reasoning. We feel therefore that a discussion of the problem of induction which Bayes famous paper [5] is said to attack is necessary for a more complete picture of the issues involved in representation of knowledge.

3) The third purpose with presenting Bayes formula is to give a formal, deductive proof that in a data base which makes use solely of deductive reasoning, a specified probability value 0 for the occurrence of an event can never be updated later on to a different value. The same is true for a specified probability value 1. Completely different laws hold, however, for the updating of maximum likelihood probability estimates; i.e. for the updating of relative frequencies based on an experimental sequence of finite length. If we wish to store such estimates, we must therefore assign to them separate fields in the data base in addition to the specified probability values. These fields can be used to falsify the specification of impossibilities (probability values 0) and certainties (probability values 1).

We feel strongly that names such as 'prior probabilities', 'posterior probabilities', 'being a Bayesian' etc. have become slogans which make mysteries and even villains out of simple and straightforward concepts. At the same time the aversion against the use of the indisputable, basic law of compound probabilities makes a mystery out of the meaning of the grade of membership concept of fuzzy set theory. It has been shown that when one does use this law, then the grade of membership concept can be clearly defined [27]. This is in contrast to Zadeh's fuzzy set theory which lacks a clear interpretation of the grade of membership concept [25, sect. 3], [31].

Closely connected with the interpretation of the grade of membership concept is the differentiation between the probability versus the truth value or possibility row of a chain set. The connection between these two is is also established by the use of
XXXX the law of compound probabilities (see sect. 6.4.2 and sectXXXX).

6.4 Probability and Logic

6.4.1 Introduction

Already George Boole, the father of mathematical logic, indicates that logical reasoning must include reasoning with probabilities. The second part of his book [7, chapters 16-21] is devoted exclusively to the theory of probability, making use of the frequency definition of probabilities. It does not, however, establish a connection with the first part.

The very great majority of subsequent work on logic has, unfortunately, completely ignored the use of probabilities. Instead it has worked solely with the truth values t and f (or equivalently 1 and 0) of propositions. In 1920 Łukasiewicz extended this '2-valued logic' to a '3-valued' one by introducing an intermediate truth value which he denoted by $\frac{1}{2}$. He then performed a further extension to an n-valued logic L_n, and finally to an infinite valued one L_∞ whose truth values may be any number in the real interval [0,1]. (See [8, pp. 87, 140,173], also [20, p. 206].) Nowadays Zadeh's fuzzy set theory (see [66] and XXXX in this book) is probably the most well-known system of infinite-valued logic. *XXXX*

Both Boole and Łukasiewicz believed in a connection between truth values and probabilities. The latter says [8, p. 173],

> *The relation of the infinite-valued system to the calculus of probabilities awaits further inquiry.* (6.2)

While the following quotation from Boole's book [7, p. 248] shows that he already in 1854, made a straightforward connection between probabilities and truth values,

> *Instead of considering the numerical fraction p as expressing the probability of the occurrence of an event E, let it be viewed as representing the probability of the truth of the proposition X, which asserts that the event E will occur.* (6.3)

The chain set system of part **??** of this book uses probabilities, not merely as an addition to its logical system, but as an integral part of the logical structure.

6.4.2 Truth Values or Grades of Membership versus Probabilities

The chain set system is in stark contrast to Zadeh's version of fuzzy set theory. The latter has completely abandoned any vision of a connection between logic and probabilities [67]. A not inconsiderable part of the fuzzy set community is strongly opposed to a probabilitsic interpretation of the 'truth values', or 'grades of membership', or 'possibilities' of fuzzy set theory.

However, The TEE model for grades of membership, as well as the chain set system, demonstrate that a probabilistic interpretation of truth values is not only possible, but leads to more consistent and reasonable results than Zadeh's theory. [1] *see XXXX in footnote*

To make a connection between the truth values of logic on the one hand, and probabilities on the other, the main stumbling block that had to be overcome was the differentiation, with the aid of the law of compound probabilities, between two types of probabilities. Both of these are conditional probabilities but with opposite directions of conditioning. This is illustrated by the example of sentence (6.4) below.

One of the tools of natural language for describing a situation of uncertainty is the OR connective. E.g., the statement

$$\lambda = \text{Margy will be at home on Sunday OR Monday ,} \tag{6.4}$$

[1] See [27], [25], [26], [28], [29], and XXXX here for more details concerning the probabilistic interpretation of grades of membership.

supplies us with uncertain information. It tells us that Margy will be at home on at least one of the two days. Supposing that the OR connective is intended in its inclusive sense, the statement leaves open three possibilities, namely that she will be at home on

'Su AND NOT Mo', on 'Mo AND NOT Su', or on 'Su AND Mo' . (6.5)

Which of these three cases is the correct one will be unknown to the listener unless she receives some information in addition to (6.4). It should also be unknown to the informant unless she intentionally misleads the listener to believe that she is in possession of less information than she actually does have.

According to the information available to the listener after having received the statement (6.4), each of the three events of (6.5) has a probability that is neither 0 nor 1. While the event that she is at home on 'NEITHER Su NOR on Mo ' has a probability 0.

see XXX in new footnote

To describe this state of information we can invoke Bayes' *postulate* and assign to each of the three events the probability $\frac{1}{3}$. [2] Alternatively the m-notation of sect. 8.4 assigns the probability value m to each of the three possible outcomes of eq. (6.5). The exact numerical values of the probabilities of the three events in (6.5) are usually unimportant. The important knowledge is that none of the three probabilities is initially 0 or 1; and that they sum up to exactly 1, so that the remaining event 'NOT Su AND NOT Mo' has the exact probability of occurence 0.

At first sight it seems strange that although each of the three events of (6.5) has a probability value that is different from 1, the truth table for the OR connective assigns the truth value t to each of them in the traditional propositional calculus of mathematical logic.

The solution to this apparent paradox is the following. The probability value 1/3 for the first event in (6.5) refers to the probability that 'Su AND NOT Mo' will occur when we are given the statement λ of (6.4),

$$P[(\text{Su AND NOT Mo}) \mid (\text{Su OR Mo})] = 1/3 \ . \tag{6.6}$$

While the truth value t of the truth table for λ=Su OR Mo (for the row Su=t, Mo=f) is identified with the probability of a 'yes' answer to the question 'λ?'='Su OR Mo?', given that she *is* at home on 'Su BUT NOT Mo'='Su AND NOT Mo',

$$P\,[\text{yes-(Su OR Mo)} \mid (\text{Su AND NOT Mo})\,] = 1 \ . \tag{6.7}$$

The distinction between P(event | yes-λ) versus P(yes-λ | event) is essential for a consistent theory of logic that makes use of probabilities.

We shall see in part **??** that the chain set system makes use of both probabilities and likelihoods, where the likelihoods have the probabilistic interpretation of P(yes-λ | event). The relation between P(yes-λ | event) and P(event | yes-λ) is derived in XXXX. An analgous distinction is also essential in fuzzy set theory according to the TEE model [27], [29], [30].

XXXX

[2] The sense in which probabilities are to be understood for a sentence such as eq. (6.4) is discussed in section XXX.

6.4.3 Belief Functions

As has been emphasized by Giles [19], a statement reflects the belief of the informant. Already Poisson identified a probability with a measure of belief. He says (cited in [7, p.244]),

> *The probability of an event is the reason we have to believe that it has taken place, or that it will take place. . . .*
> *The measure of the probability of an event is the ratio of the number of cases favourable to that event, to the total number of cases favourable or contrary, and all equally possible (equally likely to happen).* (6.8)

Nowadays uncertainty is sometimes treated with the aid of belief functions which, in contrast to Poisson's formulation, are not directly identified with probabilities, and which may even give rise to formulas that are not identical with the probabilistic ones. E.g., the 'summing-up-to-1' formula of probabilities does not hold generally for Dempster and Shafer's belief functions. For these functions the sum is only required to be smaller than or equal to 1. However, the formulas for their belief functions contain the Bayesian formulas as a special case [53, pp.6, 4].

Working with probabilities instead of more general belief functions has the big advantage that their numerical values have, in contrast to more general belief functions, a clear definition. They can therefore be combined with the values of other probabilities according to derived instead of postulated rules.

The last statement holds only if we assign to probabilities their basic meaning of long-run limits of relative frequencies. It does not hold for their interpretation as 'subjective probabilities' in the technical sense in which this term is used nowadays. (See, e.g. Jeffreys [34], DeFinetti [12], Savage [50] for the theory of subjective probabilities.) The theory of subjective probabilities is in need of a not inconsiderable number of postulates or axioms. (See Jeffreys [34, p.16 et seq.].)

It is not always easy to find a frequency-probabilistic interpretation of initially loosely defined quantities such as 'belief functions', or 'subjective probabilities', as well as of the 'possibilities' or 'grades of membership' of fuzzy set theory. However, when one takes the trouble to find it, then one is rewarded by ending up with a theory which is necessarily consistent and needs no axioms. Furhtermore, in many limiting special cases the results of the formulas derived from a probabilistic interpretation of the loosely defined quantities are in better agreement with the expected results than the postulated formulas of the non-probabilistic theories of uncertainty. Frequency-probabilistic interpretations of the grades of membership of fuzzy set theory are given in [27] and [26], and in XXXX here. *XXXX*

Examples 7.7.1, 7.7.2 and 8.1.2 illustrate possible frequency interpretations of belief functions or subjective probabilities. They refer to the description of situations which, at first sight, seem to be non-repetitive and therefore to defy a relative-frequency interpretation.

The informant herself is often unable to identify the sources of uncertainty which make a frequency interpretation possible. This is, however, no argument against such

an interpretation. We know from work with expert systems that one of the most difficult tasks which faces the system engineer is to identify the criteria on which the experts base their conclusions. The experts are, in many cases, not consciously aware of the criteria that they use.

6.4.4 Modal Logic

6.5 Deductive Reasoning

Philosophers distinguish between deductive and inductive reasoning. These two modes of reasoning are described in the present and next section respectively. In connection with the theory of probability, *deductive* reasoning starts out by *assuming* a given probability distribution, and deriving consequences from this assumption in the form of predicting the probability of different events. In contrast, *inductive* probabilistic reasoning starts out with observed statistical data, and tries to 'derive' the probability distribution that gave rise to these data.

In non-probabilistic problems, deductive reasoning starts out with one or more definitions and, if necessary, with one or more assumptions. The initial, supposedly true assumptions are called axioms or postulates.

From the definitions and axioms one then derives theorems which must necessarily also be true according to the inference rules of the particular deductive logic or 'sentential calculus' as it is sometimes called. All systems of mathematical logic are purely deductive ones except, maybe, nonmonotonic logics [9, p. 369].

In the above deductive inference scheme or *argument*, the assumptions are called the *premisses* of the argument, and the derived theorem its *conclusion* [22, p. 7]. The initial premisses are always axioms which cannot be proved by the deductive logic. The Aristotelian syllogisms mentioned in sect. 2.1 are typical deductive inference schemes or arguments. E.g., in eq. (2.2), the two sentences in the first line are the premisses or assumptions, and the last line is an inferred conclusion.

Not only mathematical *logic*, but also the whole field of mathematics is a purely deductive logical system. All mathematical derivations and proofs are inferred deductively, starting out with definitions and axioms formulated as sentences in a natural language. Other true sentences are then derived deductively by the logical rules of natural languages. In this process one makes use mostly of the IF THEN connective, the negation, the AND and occasionally the OR connective, and quantification expressions like ALL, SOME, NO. In a subsequent argument, the new premisses can be former conclusions or inferences from the axioms. In this way new theorems can be derived which can again be used as premisses in a new argument, etc. .

In any natural language, the inference process of an argument can be expressed by saying that the argument shows that "IF the premisses of the argument are true, THEN its conclusion is true". An equivalent statement in terms of probabilities is: $P[(\text{conclusion} = t) \mid (\text{premisses} = t)] = 1$. An analogous statement cannot be made in the language of, e.g., propositional calculus because the material implication of this calculus is not a complete equivalent of the IF THEN connective of natural language.

(See sect paris romeXXXX) *XXXX*

6.6 Inductive Reasoning

Inductive reasoning or learning proceeds from particular observed data to more generally valid laws. It is the typical reasoning used by children, animals and, last but not least, scientists to orient themselves in this world and set up laws of nature. In contrast to deductive reasoning, inductive reasoning is not guaranteed to be correct. But no animal, including humans, could survive without it.

The adage "The burnt child fears the fire" illustrates the inductive learning and reasoning processes of children. A child that has been hurt once or a few times by putting her hand in the fire assumes that this action will always lead to pain and refrains from repeating it. Similarly a dog whose master uses an electric leash to give him a small shock each time he pulls at the leash infers that the pain is due to the leash-pulling, and does not repeat this action. A scientist who works with quantities such as the time interval between two events, or the mass of a body, whose numerical values turn out to be the same for every experiment that he has ever performed assumes that this constancy holds generally.

The above three examples of inductive learning inferences illustrate that such learning may be correct. But it may also be incorrect, i.e. an over-generalization. The child makes a generally correct inference concerning the fire. In contrast, the inference of the leash-pulling dog may be wrong. The dog-owner may, later on, buy an ordinary leash and the dog will still be afraid to pull because he has learned inductively, but now falsely, that leash-pulling always results in an unpleasant shock.

The history of science abounds in theories which turn out to be over-generalizations, and thus generally wrong, although they may hold in certain cases. Thus the theory of special relativity had to give up the inductive, Newtonian assumption of the constancy of time intervals or of the mass of a body. These physical magnitudes turned out to depend on the velocity of the coordinate system in which they are measured, especially for velocities which approach that of light.

6.7 Statistical Inference

There exist two main ways of ascertaining the values of a probability distribution function. The first of these depends on prior knowledge concerning the objects to which the distribution refers. The probability values deduced from the prior knowledge can then be specified directly to the data base. Such specifications are described in sections 8.6 and 7.7. An example is the probability of drawing a king from a complete pack of 52 cards. Our prior knowledge concerning such a pack tells us that this probability is equal to $4/52=1/13$. When such prior knowledge is not available we must turn to the second way, namely that of statistical inference.

The field of statistics deals with the inference or 'learning' of an underlying probability distribution from observed data. It is therefore usually said to be inductive.

E.g., Barnard says in his preface to the 1958 reproduction of Bayes famous paper that it is the first expression in precise, quantitative form of one of the modes of inductive inference. And Price, who communicated this paper to the Royal Society of London in the year 1763, after Bayes death, says that it gives a clear account of the strength of inductive reasoning. [5, pp.293, 297].

In sect. 7.5.2 we define the probability of an event as the $N \to \infty$ limit of the relative frequency of occurrence of the event in a series of N single experiments. In order to infer precisely an underlying probability distribution from an observation, one would therefore have to observe an infinitely long series of data. Since this is not possible, a statistically inferred probability distribution can never be guaranteed to be the correct one. It can, however, be made to approach the correct one with increasing degree of precision by increasing the length of the series of observations. In addition one must, of course, take other measures of precaution, such as ensuring that the particular sample of data is a random one.

Because the inferred probability distribution cannot be guaranteed to be completely correct, one often gets the impression from the literature that no steps in statistical reasoning can be deductive. E.g., Jeffreys says that the fact that deductive logic provides no explanation of the choice of the simplest law is an absolute proof that deductive logic is grossly inadequate to cover scientific and practical requirements; and that he rejects the attempt to reduce induction to deduction [34, pp.5, IX]. In the present section we wish to point out that there can be many deductive steps in an inductive inference.

In the first place, we can always make *some* deductive inferences from the data. E.g., suppose we have a die about which it is unknown whether it is loaded or honest. We throw the die once, and the number 5 occurs. Already from this single observation we *can* infer deductively, and with certainty, that $P(5) > 0$. The reason is that if the probability of 5 in the underlying distribution had been 0, then 5 could not have occurred.

Suppose that the face 4 turns up at the second throw. This additional evidence lets us infer deductively that $P(4) > 0$, and in addition that $P(5) < 1$ and $P(4) < 1$; because if, e.g., $P(5)$ had ben equal to 1, then P(4) would have to be 0, and 4 could never have occurred.

We thus see that even one or two throws eliminate a number of probability distribution which were possible a priori (i.e. before any observation). In our example we have, among others, eliminated the case in which the die is so loaded that the face 5 always turns up; corresponding to the probability distribution $P(5) = 1$, $P(1) = P(2) = P(3) = P(4) = P(6) = 0$. This elimination has been inferred from the two observed data by pure deductive logic, making use of the frequency definition of probabilities.

In addition to the obvious elimination of such limiting cases, other deductive steps can be used in statistical reasoning, resulting in a set of possible distributions; in the sense that if we assume that any particular one of these distributions is *the* correct underlying one, then it *can* give rise to the observed data. It is only when we insist that one of these possible distribution is *the* correct one, that we perform a step

that cannot be guaranteed to be correct according to deductive logic. It is a matter of definition whether we say that only this last step is an inductive one; or whether we say that the whole reasoning procedure is inductive although all of its steps are deductive except for the last one. The last definition is probably preferable because it conforms to the common use of the term induction as being a mode of reasoning from observed data to underlying laws. There is no reason why inductive reasoning cannot contain deductive steps. These deductive steps may follow from the frequency definition of probabilities or, for those who prefer it, from the subjectivists definition.

There exist two main, deductively correct, reasoning procedures in statistics which say something not only about the possibility that a given probability distribution is the correct one, but in addition make some probabilistic statement concerning this distribution. Both of these procedures operate with probabilities of probabilities or of parameters of probability distributions. They are called the likelihood procedure and the Bayesian a posteriori procedure respectively. These procedures are discussed in sections sect. 10.4 and sect. 10.6 respectively. Before we can discuss these two subjects we must give a short summary of the subject of probabilities.

rewrite or delete this paragraph? max likelihood defined in 7.5.4

Chapter 7

Basic Probability Theory

7.1 Introduction

The present chapter is not intended to be a text in the theory of probability. Its purpose is to help the reader to grasp the most basic concepts of the theory, and to derive the most basic formulas from these concepts.

This cannot be accomplished by starting out with the postulates of the axiomatic theory of probability. Instead, one must try to connect the definition of probabilities with their measurement or estimation in practical life. The only way to establish such a connecting link is to use the original definition of probabilities as long-run relative frequencies. The most important formulas of the theory of probability, e.g. the 'summing-up to-1' law of eq. (7.37) and the law of compound probabilities of eq. (10.1), are derived from this definition.

One *can* then set up an axiomatic theory in which one attaches numbers, called probabilities, to points in the 'universe' or 'sample space'. The numbers are required to satisfy certain rules or axioms which are chosen in such a way that the probabilities have just those properties that we want uniquely-defined, long-run relative frequencies to have.

While writing the present chapter, I was astonished to discover how lightly some of the textbooks in my bookcase on the theory of probability pass over the connection between the theory and statistical experiments. Thus Kingman and Taylor's book on measure and probability [36] mentions the frequency interpretation of probabilities only after 261 pages of difficult theory.

Feller [16] discusses in his introduction the necessity to distinguish between three aspects of the theory: (a) The formal logical content, (b) the intuitive background and (c) The applications. He then gives examples of problems treated by the theory of probability, without ever mentioning the word 'frequency'. However, when he defines conditional probabilities on page 114, he suddenly introduces relative frequencies into his formulas in order to justify his definition. This is done without ever having mentioned the terms 'frequency' or 'relative frequency' previously.

Jeffreys [34] completely rejects the frequency definition of probabilities on p. 369. On p. 370 he then gives an example which, he claims, shows the fallacy of the fre-

XXXX quency interpretation. We will, however, show in sectionXXXX that this example completely misrepresents the frequency interpretation.

A more balanced view is given in the books by Renyi [48] and Sverdrup [58]. Both of these authors discuss relative frequencies in the beginning of their books. Renyi then uses the name 'intuitive probability' for the interpretation of probabilities in terms of long-run relative frequencies. While Sverdrup says that probabilities are idealized frequencies. They thus recognize that the axiomatic theory must not let us forget that the postulates of measure theory are not religious doctrines. They are the result of the need for a tool for analysing situations of uncertainty with the aid of experimentally found frequencies. Once this tool has been established, we can reverse the order of reasoning and try to treat situations of uncertainty with the aid of an axiomatic theory whose axioms have, however, been carefully set up so that the value of the relative frequency of an event tends stochastically to that of its probability. (see theorem 7.5.1, eq. (7.38)).

Such a sequence of reasoning makes statistics fit into the pattern of other fields of science. Measurements are interpreted in terms of theories. The theories are then used to predict new phenomena which had not previously been subjected to an analysis in terms of that particular theory. However, the operational basis of the theory must never be neglected. E.g., Einstein's explanation of the relativity of length intervals, and his prediction of the relativity of time intervals, were based on a precise analysis of the procedures for measuring such intervals.

A textbook on probability which neglects the link between the *theory* of probability on the one hand, and the *measurement* of probability on the other, can be compared to a textbook on length measurement which consists of many complicated chapters on optics and interferometry without ever mentioning that a rough length measurement can be performed with a measuring tape; or to a textbook on time measurement which devotes hundreds of pages to atomic theory, the vibration of quartz crystals, and coincidence counters before it ever mentions an ordinary clock. The measurement of relative frequencies is by no means such an everyday undertaking as looking at a clock or using a measuring tape. In the case of probabilities it is therefore even more necessary to show the connection between theory and praxis.

Ideally precise measurements are nonexistent in any field of science; and the uncertainty principle of quantum theory tells us that a precise measurement cannot exist unless we pay the price of infinite uncertainty in some other quantity. In principle there is thus no difference between the imprecision in the measurement of probabilities and that of any physical quantity.

The book by von Mises [61] is a notable exception to the other books on probability. Mises discusses the connection between probabilities and relative frequencies in detail, as well as other difficult questions that we shall merely touch upon, such as the definition of a random sequence. However, in his eagerness to concentrate upon an operational definition of probabilities, Mises goes to the other extreme. On pages 11-15 he gives a number of examples to which the probability concept has been wrongly applied in his opinion. One of his examples is the historical accuracy of the narratives in the bible. He does not agree with Markoff whom he cites as saying

that such an application is possible. In sect.XXXX we come back to this and other examples that have to do with non-precise estimates of probabilities. *XXXX*

For every example that I have ever worked with, I have always succeeded, although not on the spur of the moment, to find not only a probabilistic, but also a frequency-probabilistic interpretation of uncertainty. E.g., it turns out that the grades of membership of fuzzy set theory, which are claimed to have no probabilistic interpretation, can be given a straightforward such interpretation as the estimate by a subject of the probability of assignment of a given label λ (out of a given set of labels Λ) to objects of a given attribute value u; such as the probability of assignment of the label 'tall man' to a man of height 175 cm. (See, e.g., [26], [28], [29], [27], [30] and XXXXhere). *XXXX*

A frequency probabilistic interpretation of uncertainty has the enormous advantage that it allows us to derive formulas, e.g. of grades of membership of composite labels such as 'λ=tall OR medium', instead of having to postulate them. Also the relation between the probability and the possibility (or truth value) rows of a chain set (see part **??**) can be based on a frequency probabilistic interpretation of both.

However, there is no reason why those who prefer the subjectivist point of view should not make use of subjective probabilities. As far as I understand from Jeffreys book [34], they will end up with the same formulas.

The few definitions and derivations which are needed for this book are very basic and, at the same time, quite simple. Even those who have not taken a course in the theory of probability should be able to follow them.

We need only a single postulate for a frequency-probabilistic theory of probability. This postulate says that in a sequence of N drawings of an object from a bag containing M physically identical objects, each object will occur approximately equally often when $N \gg M$ (see definition 7.3.4 for a more precise formulation). The other definitions of sect. 7.3 are for the purpose of defining randomness, and of being able to define nonuniform probability distributions, without using expressions like 'probability', 'chance', 'being equally likely' as forerunners to a definition of probability. I have gone to some trouble to lay the ground for a logically sound, noncircular definition of probabilities with the aid of frequencies of occurrence of an outcome[1]. All the important laws of the theory of probability can easily be derived from this definition. It is then unnecessary to take a course in axiomatic probability theory or measure theory in order to be able to work with probabilities. All one needs is a knowledge of the operations of addition, multiplication and division for real numbers. We will also assume that the reader is familiar with the notion of a traditional set, and with

[1]The definition of a random sample is a prerequisite to a definition of probabilities. Many, otherwise excellent books on probability and statistics use a circular definition for this purpose. For example, Wonnacott & Wonnacott [65, pp. 4, 86] say: "For a very simple random sample (VRS) all individuals in the population are equally likely to be observed." In this context the expression *'are equally likely'* is, however, a synonym for *'have the same probability.'* Thus a random sample is defined with the aid of the probability concept, and probabilities are defined with the aid of the concept of a random sample. A similar circular definition of a random choice is made by Feller who uses straightout the expression 'equally probable', see our eq. (7.27) below.

the operations of union and intersection of sets.

I believe that those researchers who keep the straightforward frequency definition of probabilities in mind will have a better foundation for separating the significant problems from the insignificant ones. However, there is no reason why those who prefer the axiomatic theory of probability should not formulate there problems in this language, provided they do not forget to tell the reader how their probabilities are connected with data obtained in praxis.

7.2 The Description of Certainty

Human Language uses the means of affirmation and negation to describe situations of certainty; such as the statements,

$$1/u = \text{I will be at home tomorrow at 10 a.m. ,} \tag{7.1}$$

$$0/u = \text{I will NOT be at home tomorrow at 10 a.m. ,} \tag{7.2}$$

respectively.

(7.2) is the negation of (7.1) and vice versa. Furthermore, no possibilities other than (7.1) or (7.2) are left, provided that we define in advance a situation such as standing in the door opening at 10 o'clock, as belonging to one or the other of the two situations.

In the usual notation of mathematical logic, one describes the negation of a sentence u by $\neg u$. Our notation 1/u, 0/u for a sentence and its negation respectively is more adapted to the chain set notation of part **??**. It emphasizes the symmetry between afirmation and negation.

In natural language a declarative sentence, such as the right hand side of (7.1) or (7.2), asserts the factual (synthetic) truth of the sentence in the real world. Each of the two sentences describes a belief of the informant in a certainty, with no room for any doubt. E.g., the statement on the right hand side of (7.1) predicts that the event described by the sentence does, or will, indeed occur in the real world, and that the occurrence of the event described by (7.2) is impossible in that world.

In contrast, mathematical logic deals with the relations between truth values of different propositions. A proposition is something that may be either true or false. The assertion that a proposition is synthetically true in our world is of no interest to mathematical logic. It is, however, of enormous interest in a data base.

Let us suppose that $1/u$ on the left hand side of (7.1) merely denotes the *proposition* on the right hand side in the sense of mathematical logic, not the assertion that this proposition is true. Then the notation of the theory of probability provides us with an excellent tool for asserting that the event described by the proposition is certain to occur or not to occur in our world. The assertion of the sentence on the right hand side of (7.1) can be written as

$$Prob(1/u) = 1 \qquad \text{or equivalently} \qquad Prob(0/u) = 0 \ . \tag{7.3}$$

And the assertion of the truth of the right hand side of (7.2) can be written as

$$Prob(0/u) = 1 \qquad \text{or equivalently} \qquad Prob(1/u) = 0 \ . \tag{7.4}$$

Thus certainties are expressed by probability values of either 1 or 0. The general definition of probabilities in sect. 7.5 allows us to operate also with intermediate probability values in order to express uncertainty, e.g.,

$$Prob(1/u) = 0.6 \qquad \text{or equivalently} \qquad Prob(0/u) = 0.4 \ . \tag{7.5}$$

For the time being equations (7.4), (7.5) are simply a symbolization, in terms of a special type of code, for asserting the degree of factual truth in the real world of the sentences on the right hand sides of (7.1), (7.2) respectively. We come back to the meaning of (7.4), (7.5) in sect. 7.7.

7.3 Experiments, Universes, Object Sets and Randomness

7.3.1 Marble Sets and Random Choices, Samples and Sequences

The present section lays the theoretical foundation for a precise, noncircular, frequency definition of probabilities. Those readers who are used to the definition of probabilities as relative frequencies, and who are mainly concerned with practical applications, can probably read the definitions of sect. 7.3 rather superficially.

In sect. 7.2 we used affirmation and negation to denote the occurrence of two mutually exclusive events which exhaust all possibilities. If $1/u$ occurs then $0/u$ cannot occur and vice versa. This is a consequence of the definition of the negation in sect. 5.4.

The theory of probability deals with the everyday situation in which there exists some ignorance concerning the happening or nonhappening of an event. To define such a situation more formally, the theory defines a 'universe' or 'space' U [2],

$$U = \{u_1, \ldots, u_i, \ldots u_I\} = \{u_i\}, \qquad i = 1, \ldots, I \tag{7.6}$$

of I possible outcomes of an 'experiment' or a 'random experiment'. I is a positive integer which may now be bigger than 2. The I outcomes are assumed to be mutually exclusive. They are also assumed to be a complete set of outcomes in the sense that one of them must necessarily occur as the result of the experiment.

In addition to an experiment which results in one of the I 'outcomes' or 'sample points', i.e. in one of the I elements of the universe or space U, the theory deals also with formally defined events. An 'event' E refers to a subset of the universe, $E \subseteq U$. We say that the event E has occurred in a given experiment if and only if the outcome of the experiment is one of the elements of E. An outcome is a special case of an event for which the subset E consists of one element only.

In order to be more concrete, we will illustrate our definitions with two examples.

[2] The expression 'space' or 'sample space' is used in the theory of probability. Here we follow the praxis of traditional and fuzzy set theory and use the expression 'universe' for the same concept.

Example 7.3.1 *The experiment consists in the throwing of a die with an unknown degree of loading. The universe for the possible outcomes of the throw has then I=6 elements, namely the six faces of the die,*

$$U = \{u_1,\ u_2,\ u_3,\ u_4,\ u_5,\ u_6\} = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}\ , \qquad I = 6 \tag{7.7}$$

$E \in \{1,\ 2,\ 3\}$ *is an example of en event in this universe. The event E occurs whenever the outcome is a number smaller than, or equal to 3.*

In the next example, expressions containing the word 'random' appear several times in parentheses. The example is valid when the parentheses, including their contents, are erased everywhere. After the definitions 7.3.1-7.3.3 of randomness the example can be reread as it is now, only the parentheses themselves being removed.

Example 7.3.2 *The experiment consists in the (random) choice of an (adult) man from the population of men in the world, and the measuring and noting of his height. U is now a height universe with a number of elements that depends on the 'quantization' of the height values. We could, e.g., work with quantization intervals of 10 cm,*

$$u_1 \in [0,10)\ cm,\ u_2 \in [10,20)\ cm, \ldots,\ u_I \in [290,300)\ cm, \quad U = \{u_i\} \quad I = 30. \tag{7.8}$$

To simplify the notation, we will abbreviate this to the form

$$u_1 = 5\ cm,\ u_2 = 15\,cm, \ldots,\ u_I = 295\,cm, \quad U = \{u_i\} \quad I = 30\ . \tag{7.9}$$

Note that the experiment of example 7.3.2 is conceptually somewhat different from that of example 7.3.1. The height example 7.3.2 goes through an extra step to determine the sample point or outcome of the experiment. It first chooses (at random) a man from the 'population' of all men. The population will also be called the 'object set' OB, of all objects ob,

$$OB = \{ob_1, \ldots, ob_m, \ldots, ob_M\}\ , \tag{7.10}$$

each ob being a man. Since each man has a definite height, we have a mapping from OB to U. The 'sample point' or 'outcome' of the (random) experiment is the measured height $u_i \in \{U\}$ of the (randomly) chosen man.

Experiments of the type of those of examples 7.3.1 and 7.3.2 in which the outcome cannot be predicted are said to be 'stochastic'. There exist many additional illustrations of stochastic situations, both in science, and in technology, and in everyday life. For example, the number of radioactive disintegrations in a time interval of a given length, from a given amount of a specified radioactive material, varies stochastically. So does the life time of a light bulb as well as the number of days with sunshine in Oslo in July. The purpose of the theory of probability is to give a quantitative description of such phenomena. Stochastic phenomena are made up of random choices or selections as defined in definition 7.3.2 below.

I have not been able to find a satisfactory definition of a random sample in the literature. Feller [16, p. 30] says: *The word "random" is not well defined, but when applied to samples or selections it has a unique meaning. The term* random choice *is meant to imply that all outcomes are equally probable.*

We face two problems in connection with Feller's definition. The first one is that we wish to use the concept of a random sample for the purpose of defining probabilities. We are then not allowed to use the word 'probability' to define a random choice or sample.

The second problem is that we need an object set in order to define a random sample. However, in the typical stochastic die-throwing example 7.3.1 there exists no immediate parallel to the object set *OB* of all men in example 7.3.2. We will therefore define an imagined equivalent object set for every experiment. For reasons which will appear below, we call the equivalent object set the 'equivalent marble set' or simply the 'marble set' of the experiment. It is the marble set from which the random choices are made and the random samples are taken.

In the height example 7.3.2, the marble set can be identified with the real world object set of men. In the die example 7.3.1 we will use the outcomes obtained in a sequence of die throws to obtain an approximation to the equivalent marble set of the experiment. As the length N of the sequence is increased, we obtain progressively better approximations to the marble set.

Definition 7.3.1 *of the 'marble set' OB^o corresponding to an object set OB. Consider an object set OB, such as the set of all men in the world in example 7.3.2 and eq. (7.10). The number of elements in OB is M,*

$$OB = \{ob_1^{i1},\ ob_2^{i2},\ \ldots,\ ob_m^{im},\ \ldots\ ob_M^{iM}\}\ . \tag{7.11}$$

In the expression on the right hand side of this equation the subscripts denote the number of the object (in some arbitrary order) and the superscripts 'im' correspond to the attribute value $u_{im} \in U$ of the m-th object.

Let OB_i be the subset of OB consisting of all objects (men) with the attribute (height) value u_i. We then have,

$$OB = \cup_{i=1}^{I}\ OB_i\ , \qquad OB_i \cap OB_{i'} = \emptyset \ \text{ for } i \neq i'\ . \tag{7.12}$$

The marble set OB^o corresponding to OB has M elements just like OB. It is defined as follows.

Perform a one to one mapping of OB on a set OB^o. The m-th element of OB is mapped on the m-the element of OB^o and vice versa. All the elements of OB^o are physical bodies with identical properties. They can, for example, be round marbles all of which have the same size, are made of the same material, and have the same surface polish. Each element is, however marked with the number 'm' of the element $ob_m \in OB$ to which it corresponds and with the number 'i' of the attribute value u_i of ob_m,

$$OB^o = \{ob^{o\ i1}_{\ 1},\ ob^{o\ i2}_{\ 2},\ \ldots,\ ob^{o\ im}_{\ m},\ \ldots\ ob^{o\ iM}_{\ M}\}\ . \tag{7.13}$$

In order to have a name for OB^o we will call it the 'marble set' although, of course, it need not consist of marbles. It might just as well consist of M chips.

For the marble set we have an equation corresponding to eq. (7.12)

$$OB^o = \cup_{i=1}^{I} \; OB^o{}_i \; , \qquad OB^o{}_i \cap OB^o{}_{i'} = \emptyset \;\; \text{for } i \neq i' \; . \tag{7.14}$$

We continue with the definitions of a random choice and a random sample.[3]

Definition 7.3.2 *of a random choice or a random selection. Let OB be a set of M objects, and let OB^o be the marble set of definition 7.3.1 corresponding to OB.*

To perform a random choice of an element of OB^o, and of the corresponding element of OB, all the M marbles of OB^o are put into a box. The box is closed and is thoroughly rotated and shaken. A door, slightly bigger than a marble, is then opened in the bottom of the box until a marble falls out. Denote this marble by $ob^o{}_m$. We then say that $ob^o{}_m$ is a randomly selected *element of OB^o. The corresponding randomly selected element of OB is the object ob_m to which it corresponds. Before the random choice is completed, the attribute value u_i of ob_m must be noted.*

Definition 7.3.3 *of a random sample and of a random sequence (or stochastic process) of attribute values u_i from an object set or population OB, each element of which has an attribute value $u_i \in U$. To select a random sample of objects from an object set OB we work with the marble set OB^o corresponding to OB, definition 7.3.1. A marble is selected at random from this set (see definition 7.3.2), its attribute identifier 'i' is noted, the marble is returned to the box, and a new random selection can now performed. For a random sample of size N we perform N random selections from OB^o. Each* random sample *of size N gives rise to a* random sequence *of N attribute values $u_i \in U$,*

$$\mathbf{x} = \langle u_{i1}, \ldots, u_{in}, \ldots, u_{iN} \rangle = \langle x_1, \ldots, x_n, \ldots, x_N \rangle \; . \tag{7.15}$$

A given element of U may occur more than once in the sequence. The x_n instead of u_{in} notation is used in the next section because it has simpler subscripts. The sequence (7.15) is also called synonymously a 'stochastic process' (see, e.g., Feller [16, p. 419]).

The size N of the sample may be smaller than, equal to, or bigger than the size M of the object set OB. In the height example 7.3.2, N will usually be much smaller than the number M of men in the world, and it will almost never happen that the same man is selected more than once. The sample from OB^o can then be mapped upon a subset of the set of men OB. The possibility of N being bigger than M is due to our use of a random choice 'with replacement'. This means that the randomly selected

[3] There exist, of course, other equivalent methods to those of definitions 7.3.2, 7.3.3 for obtaining a random or nonbiased sample. The two definitions are only intended to clarify the principle. It would indeed be difficult to manufacture and shake a marble set consisting of billions of marbles, corresponding to the number of men in the world. However, if we assign successive integers to the men we could, e.g., use a roulette type wheel to determine the first, second, ... digit in the ordinal number of a selected man. At one time such a method of selection was, indeed, implemented in the drawing machines of the Norwegian state lottery.

marble is returned to the box before a new choice is performed. A small set OB can thus result in a sample of a much bigger size.

For each attribute value u_i we denote the number of elements in the sequence having this attribute value by n_i,

$$\sum_{i=1}^{I} n_i = N \ . \tag{7.16}$$

7.3.2 The Basic Postulate versus Bayes Postulate

We need only one single postulate in connection with a theory which defines probabilities as limits of relative frequencies. This postulate refers to the random choice definition 7.3.2. In common language it says that each marble has the same chance of being selected in a random choice. Since the word 'chance' has actually the meaning of 'probabilty' in this connection, and since we wish to avoid the use of the word 'probability' in the definitions of the present section, we refer the postulate to definition 7.3.3 of a random sequence of objects.

Definition 7.3.4 *of the basic postulate of the frequency oriented theory of probability. Let OB^o be a marble set of M marbles with identical physical properties. A random sequence of length $N \gg M$ is selected from this set according to the procedure of definition 7.3.3,*

$$\textit{random sequence of } N \textit{ marbles} = \langle ob^o{}_{m1},\ ob^o{}_{m2},\ \ldots,\ ob^o{}_{mn},\ \ldots\ ob^o{}_{mN} \rangle \ . \tag{7.17}$$

The symbol for a given marble may occur more than once in this sequence due to our 'sampling with replacement', see definition 7.3.3.

Let n_m be the number of times that the symbol for the object (i.e. the marble) $ob^o{}_m$ occurs in the random sequence of eq. (7.17). We then have,

$$\sum_{m=1}^{M} n_m = N \ . \tag{7.18}$$

The basic postulate now says that for $N \gg M$, the number of times n_m that the symbol for a given marble will occur in this sequence is approximately the same for all the M marbles, $n_m \approx N/M$, and the relative *frequency of ocurrence of a marble in the sequence, n_m/N, is therefore approximated by $1/M$. More precisely,*

$$\lim_{N \gg M} \frac{n_m}{N} = \frac{1}{M} = \textit{const independent of the particular marble } ob^o{}_m . \tag{7.19}$$

Note that the basic postulate refers only to marbles and their ordinal numbers. The attribute value markings have no significance for the basic postulate.

The postulate of definition 7.3.4 has a close connection with Bayes postulate discussed in sect. 8.3, but is more fundamental than the latter. There are two differences between the two postulates. In the first place, the physical definitions of a marble

set and of a random choice make the basic postulate a very reasonable one. Since all marbles are physically identical, there is no reason why one marble should appear more often than another in a long random sequence. In contrast, in the case of Bayes postulate, there is no reason to assume a priori that an unknown probability distribution is a uniform one. The second difference is that Bayes postulate assumes that probabilities have already been defined. The basic postulate does not make use of probabilities. Instead it is used for the purpose of defining probabilities.

7.3.3 Experiments with No Natural Object Set

We now know how to obtain a random sequence of height values in example 7.3.2 by selecting a random sample and sequence from the marble set OB^o corresponding to the set OB of all men. However, in the (generally loaded) die example 7.3.1 there exists no natural object set. Without such a set we cannot define the equivalent marble set, and without the latter we cannot define a random sample and sequence. What we wish is to be able to define a random sequence of throws of a loaded die in analogy to our definition 7.3.2 of a random sequence of height values from the set of all men. The definition should apply not only to a situation of die throwing, but to any stochastic experiment without or with a naturally given object set OB.

Let us first assume that a marble set corresponding to the experimental situation is given, each marble being marked with the attribute value u_i to which it refers. And let m_i be the number of times that a marble marked with the attribute value u_i occurs in the marble set on the right hand side of eq. (7.13). We then have

$$\sum_{i=1}^{I} m_i = M \ , \tag{7.20}$$

and consequently

$$\sum_{i=1}^{I} \frac{m_i}{M} = 1 \ . \tag{7.21}$$

E.g., in the case of the loaded die experiment each of the marbles would be marked with one of the attribute values u_i taken from the set $U=\{1,2,3,4,5,6\}$. If the marble set represents a *loaded* die experiment, then at least two m_i's will not be equal; e.g., $m_3 \neq m_6$, the number of marbles marked with $u_i = 3$ is not equal to the number of marbles marked with $u_i = 6$.

As before, let n_m be the number of times that marble number m appears in a random sequence of N selections from the marble set. And let m_i be the number of marbles in the marble set with attribute value u_i. Then the number n_i of times that the attribute value u_i occurs in the random sequence is given by

$$n_i = m_i \cdot n_m \ . \tag{7.22}$$

Reversing the left and right hand sides of this equation and dividing by N we get,

$$m_i \cdot \frac{n_m}{N} = \frac{n_i}{N} \ . \tag{7.23}$$

Finally we take the $N \gg M$ limit on both sides, equate n_m/M in this limit to $1/M$ according to the basic postulate of definition 7.3.4, and obtain

$$\frac{m_i}{M} = \lim_{N \gg M} \frac{n_i}{N} \ , \tag{7.24}$$

or

$$\frac{m_i}{M} = \lim_{n_i \gg 1} \frac{n_i}{N} \ . \tag{7.25}$$

In the last equation, the limiting operation has been changed from $N \gg M$ to $n_i \gg 1$. The equation emphasizes that our interest has shifted from the frequency of occurrence of a given marble ob_m in the sequence to the frequency of occurrence of a given attribute value u_i in the same sequence.

The correct limit is determined by the type of error of estimation that we consider. E.g., an error in m_i relative to M, $\Delta m_i/M$, versus $\Delta m_i/m_i$. For a given upper limit in the expected value of these two, we need a longer sequence N for the latter quantity than for the former, especially when $m_i \ll M$. From the Poisson distribution we know that the standard deviation of n_i is the square root of its expectation value.

It thus follows from the basic postulate of definition 7.3.4 that *when the marble set of an experiment is given*, then the relative frequency of occurrence of a given attribute value u_i in a long random sequence of N experiments approaches m_i/M, the relative frequency of occurrence of marbles with the attribute values u_i in the marble set. This relation holds whether the experiment is one of die throwing, or of the measurment of height values, or of the observation of the number of days with sunshine in Oslo in July over a period of many years. In the height and sunshine experiments we have natural object sets (of men and of recorded years respectively), and can therefore immediately construct the equivalent marble sets.

In the die example we have no natural object set. However, eq. (7.25) tells us that in any stochastic experiment we can approximate the fractions m_i/M in the equivalent, but unknown, marble set by the observed fractions n_i/N in an experimental sequence. The longer the sequence, the better the approximation is expected to be.

Thus, suppose that we base our approximation to the marble set on the observation of a sequence of N=210 die throws. If there are 40 occurrences of the face 5 in this sequence, then our approximation to the marble set will contain 40 marbles marked 'u_i=5', and our approximation to m_5/M in the true marble set of the experiment is 40/210. If there are 46 occurrences of 5 in the sequence on which we base our approximation to the marble set, then our approximation to m_5/M is 46/210.

Any prediction concerning a future random sequence is now based on a random selection from the approximated marble set performed according to definition 7.3.2 and definition 7.3.3. The reliability of the prediction is expected to increase with the length of the original experimental sequence as well as with the length of the predicted experimental sequence.

The number of elements in an approximation to the marble set constructed according to the above procedure is equal to the length of the experimental sequence on which the approximation is based. This may seem disturbing unless we realize that

the significant quantities for the description of stochastic uncertainty are not the absolute numbers of occurrences of a given outcome, but the relative numbers. Indeed we shall see later that probabilities depend only on the ratio between the different n_i's in the sequence, not on their absolute values. For example, when we are given a marble set with M marbles, we can construct an equivalent marble set of $2M$ marbles without changing the values of the probabilities. This is done simply by duplicating each marble in OB^o together with its attribute value. These considerations lead us to the following definition.

Definition 7.3.5 *Definition of equivalent marble sets. Two marble sets OB^o, OB^o' are said to be equivalent if and only if*

$$m_i/m'_i = const, \quad \text{where the constant is the same for all attribute values } u_i. \tag{7.26}$$

m_i and m'_i are the number of marbles with attribute value u_i in OB^o and OB^o' respectively.

7.3.4 Nonuniform Distributions, Object Sets vs. Attribute Universes

Both in probability and in fuzzy set theory there exists a confusion, and a resultant equivocation between the 'universes' OB and U. Thus Feller [16, p. 30] says,

The term random choice *is meant to imply that all outcomes are equally probable.* (7.27)

We have already discussed the circularity of this definition in the first footnote of the present chapter. What we are concerned with here is that Feller's definition of a random choice rules out the possibility of a stochastic process with a nonuniform distribution of attribute values. Since an outcome is an element u_i of U, (7.27) implies that all die faces have equal probability of occurrence in connection with our die example, or that all height values have equal probability of occurrence in connection with the height example 7.3.2. We would thus be unable to define a probability scheme with different probabilities for different outcomes.

What Feller should have said is that in a random choice, all elements of OB or OB^o, not of U, have the same probability of being drawn. In connection with a coin tossing experiment Feller [16, pp. 29, 30] calls the set $\{head, tail\}$ the 'population'. However, a long sequence of random choices from this set would always result in approximately 50 % heads and 50 % tails. If we want to have the possibility of describing experiments with a biased coin also, then we must introduce the equivalent marble set or population OB^o, and make random choices from this set.

Also in fuzzy set theory there exists a certain confusion between the 'universes' OB and U. Thus the fuzzy set of 'tall men' can be defined either as a fuzzy subset of all height values, or as a fuzzy subset of a set of men. For example, in [66], Zadeh works with fuzzy sets, such as 'tall man', which are fuzzy subsets of the attribute (height) universe U. There is, however, one exception namely eqs. (2.13), (2.14). Eq. (2.13) works, suddenly, with a universe U={Tom, Dick, Harry }, and the fuzzy set 'agile'

is defined as a fuzzy subset of this U which should rather have been denoted by OB. The result of the confusion is not quite so disturbing as in the theory of probability because the experimental determination of grades of membership does not depend on the choice of a random sequence from OB or OB^o, but rather on the estimate, by the subject who specifies the membership value, of the uncertainty in the assignment of the label, e.g., 'tall' versus 'medium height'. The sources of randomness are, in this case, only indirectly connected with the object set (see Hisdal [26] and XXXX in this *XXXX* book).

Thus an experiment is characterized not only by the universe of possible outcomes, but also by the reference object set OB or its equivalent marble set OB^o. We will say that two experiments are identical if their outcomes belong to the same universe and if the marble sets are identical or equivalent for both experiments. Otherwise the experiments are nonidentical or different.

Consider our die example with the universe $U = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$. The experiment consists of the throw of the die, and the observation of the uppermost face. A random experiment consisting of a second throw of the same die is then identical with the first experiment, although its outcome may be different. If the second experiment had been the throw of a different die, then it would be a different experiment although the two experiments have the same attribute universes. However, the reference marble sets of the two experiments differ because they refer to different dice which may be differently loaded.

Books on statistics describe tests as to whether a given experimental sequence is a random sequence of mutually independent outcomes from a given marble set. E.g., in connection with the number of days of sunshine in Oslo in July, it might be that the marble set, and therefore the quantities m_i/M, change with time due to permanent changes in the weather of the earth. Or it may be that the outcome of a single experiment depends on the previous $K-1$ outcomes. Such a stochastic process is called a Markoff process provided that K is finite. To describe such a process, each marble of the marble set would have to correspond to K sequential outcomes, and would therefore have to be marked with a sequence of K attribute values. We can, however, still approximate the 'marginal marble set' for a single outcome as before, using eq. (7.25), which yields,

$$\text{approximation to } m_i = n_i \frac{M}{N} \ , \tag{7.28}$$

provided that the experimental sequence from which it was approximated is long enough. (It must be at least by a factor of K or more longer than the length of a random sequence without interdependence between successive outcomes.)

Markoff processes are not treated in this book. However, experiments in which each outcome is a sequence of N mutually independent single outcomes are needed, for example, for finding the maximum likelihood estimate of an underlying probability (see sect. 7.5.4). Especially the axiomatic theory of probability makes extensive use of such 'higher dimensional universes', e.g. in connection with the probability-frequency law of theorem 7.5.1. This brings us to the subject of composite experiments.

7.4 Composite Experiments with Identical Components

An experiment in which each outcome is characterized by a sequence of outcomes from two or more universes is called a composite experiment. When the sequence is of lenght N, then we will say that we have an N-dimensional composite experiment. We will see that an N-dimensional composite experiment can be described as a noncomposite experiment with an N-dimensional universe or space.

In the present section we assume that the N component experiments are identical and have therefore identical universes and identical reference object sets OB. We thus consider, as before, a random sequence of N single or noncomposite experiments,

$$\mathbf{x} = \langle x_1, \ldots, x_n, \ldots, x_N \rangle, \qquad x_1 \in U, \ \ldots, \ x_n \in U, \ \ldots, \ x_N \in U \ . \tag{7.29}$$

Each element of the sequence consists of an outcome $u_i \in U$. Eq. (7.29) represents the same random sequence of outcomes that we considered in sect. 7.3, eq. (7.15). However, we now look upon the whole sequence as a single unit which can be represented by a single point in the N-dimensional universe

$$\mathbf{X} = U^N = \{\mathbf{x}_1, \ldots, \mathbf{x}_{I^N}\} \ . \tag{7.30}$$

The elements of the universe $\mathbf{X}$ are thus sequences of N u-values. The notation U^N for $\mathbf{X}$ in (7.30) is a symbolic one. It indicates that $\mathbf{X}$ can be considered as an N dimensional coordinate space with the I points $u_1, \ \ldots \ , u_I$ on each axis. Each outcome $\mathbf{x}$ of the composite experiment corresponds to one, and only one, point of this space. Since each $x_n \in U$ can be chosen in I ways, there are in all I^N sample points in the space $\mathbf{X}$.

For example, if the composite experiment consists of two single experiments, then each sample point is a sequence $\mathbf{x} = \langle x_1, x_2 \rangle$ of two u values. It corresponds to a point in a 2 dimensional coordinate system with I^2 elements. x_1, the first element of the sequence $\langle x_1, x_2 \rangle$, denotes the outcome of the first single experiment, x_2 that of the second one.

In connection with the die example 7.3.1, the composite experiment would describe two throws of the die. The outcome of the first throw is denoted by x_1, that of the second throw by x_2. The sample space $\mathbf{X}$ consists of 6^2=36 points.

In connection with the height example 7.3.2, the '2 dimensional composite experiment' consists of (1) The random choice of a man from the set of all men, and the noting of his height $x_1 \in U$. The man is returned to the set of all men, and a second man is chosen at random. The measured height of this man is $x_2 \in U$. The outcome of the experiment is the ordered pair of height values $\langle x_1, x_2 \rangle$.

7.5 Interpretative versus Axiomatic Probabilities

7.5.1 Introduction

The long-run relative frequency definition of probabilities is also called by mathematicians the 'intuitive definition' (see, e.g., Renyi [48, sect. 2.1]). This definition connects the theoretical concept of probabilities with their experimental measurement in the form of relative frequencies. Actually it is therefore not only far from being intuitive, it is the step which converts intuition concerning probabilities to the procedures on which their estimation is based; just as the system engineer of an expert system often must convert the intuitively performed procedures of the experts to their precise definition. We will therefore use the name 'interpretative' instead of 'intuitive' for that direction in the theory of probability which *interprets* probabilities as long-run relative frequencies; and which bases its derivation of different probabilistic laws on this interpretation. In more modern times this operational point of view has been developed in detail by von Mises [62].

The experimentally found numerical values of relative frequencies in two different experimental sequences of length N are, in general, not precisely equal, although the relative spread in values decreases with increasing N. From a mathematical point of view, the interpretative definition has therefore the disadvantage of having to assume, without proof, the existence of a unique numerical limit for relative frequencies in different experimental sequences whose length N goes towards infinity.

The axiomatic theory of probability, based on measure theory, was first presented in an orderly, complete and very elegant fashion by Kolmogoroff [38]. It attaches a unique numerical value, $Prob(u_i) \in [0,1]$, to each element u_i of the universe without attaching any meaning to it. Certain formulas are then postulated to hold for these meaningless 'probabilities'. The postulates or axioms of the theory are, however, chosen in such a way that certain basic mathematical laws which hold for relative frequencies are required to be satisfied for these 'probabilities' also.

The axiomatic probabilities have the disadvantage of having no a priori connection with an experimental situation. This lacking connection is clearly illustrated in many textbooks on probability which do not mention the term 'frequency' or 'relative frequency' before presenting hundreds of pages of difficult theory.

None of the books on the axiomatic theory can *prove* a connection between *their* probabilities and relative frequencies in experimental sequences of randomly chosen elements of the universe. All of them are thus forced to use specific examples, such as coin or die tossings, to illustrate that their probabilities have the same properties as those which we desire for the interpretatively defined ones in the $N \to \infty$ limit. Thus the axiomaticians must use their intuition to establish a connection between their probabilities and reality, while the experimentalists use their intuition to assume the existence of limiting values for relative frequencies. (See theorem 7.5.1 for a more precise formulation concerning this 'limiting value'.)

Unless we are concerned with epsilons, and with rare situations for which there exists no $N \to \infty$ limit of relative frequencies, it is justified to assume that certain

mathematical formulas which hold for relative frequencies in every specific experimental sequence also hold generally for their $N \to \infty$ limit, i.e. for probabilities.

The relative frequency proofs are usually very simple. In this book we need two basic formulas which we will prove for relative frequencies. We will then assume that they hold also in the $N \to \infty$ limit, i.e. for probabilities. These are 1) The summing-up-to-one law for probabilities, eq. (7.37), and 2) The law of compound probabilities

XXXX equationXXXX.

In addition to these two important laws which are easily derived in the frequency interpretation of probabilities, it turns out that *conditional* probabilities can easily be defined in a meaningful way in the frequency interpretation. In contrast, conditional probabilities are defined in the axiomatic theory without assigning any meaning to them. Thus Kolmogoroff [38, eq. (5)] defines the conditional probability of the event B under the condition A as

$$Prob(B|A) = \frac{Prob(A \cap B)}{Prob(A)} \ , \tag{7.31}$$

without giving any reason for this definition; except that with the above definition he can prove from his axioms that $Prob(B|A)$ has, for a given A, all the properties which are required of a probability function.

In sect. 7.5.2 we define relative frequencies and interpretative probabilities. The methods of the axiomatic theory of probability are discussed superficially in sect. 7.5.3. In the same section we discuss an important theorem which is a special application of the law of large numbers of the axiomatic theory. The theorem says that the frequency of occurrence of the outcome u_i in an experimental sequence of length N tends stochastically towards $Prob(u_i)$. Assuming that the probabilities of the axiomatic theory are connected with the results of statistical experiments, we have thus a proof that the desired $N \to \infty$ limit of relative frequencies exists in a stochastic sense.

7.5.2 Probabilities and Frequencies

We start this subsection with the definition of the frequency and the relative frequency of an event in a specific experimental sequence of length N.

Definition 7.5.1 *of frequency and relative frequency. Consider a composite experiment each outcome of which is a sequence of N identical, random, noncomposite experiments referring to the same universe U. The outcome of the composite experiment is denoted by $\mathbf{x} = \langle x_1, \ldots, x_n, \ldots, x_N \rangle$, see eq. (7.29). Each x_n is thus an element of U. Let n_i be the number of times that the outcome u_i occurs in the sequence $\mathbf{x}$. n_i is called the frequency of occurrence of u_i in the sequence, and the fraction*

$$Freq(u_i) = \frac{n_i}{N} \tag{7.32}$$

is called the relative frequency of u_i.

From the definition of n_i it follows that

$$\sum_{i=1}^{I} n_i = N \ , \tag{7.33}$$

and therefore

$$\sum_{i=1}^{I} Freq(u_i) = 1 \ . \tag{7.34}$$

Let us now imagine that a composite experiment, resulting in a sequence of N u-values, $\mathbf{x} = \langle x_1, \ldots, x_n, \ldots, x_N \rangle$, is repeated a number of times. For each repetition, the relative frequency $Freq(u_i)$ of the outcome u_i is computed. We then find experimentally that the values of the $Freq(u_i)$ are not precisely the same for each sequence. Instead, they fluctuate about some mean value.

However, we also find experimentally that for the random or stochastic sequences of definition 7.3.3 there exist many phenomena for which the deviation of the $Freq(u_i)$'s from their mean value decreases, on the average, when the length N of each experimental sequence is increased.

Because of the average decrease of the deviation with increasing N, one assumes that there exists, for stochastic phenomena, an underlying, uniquely-valued relative frequency $Freq(u_i)$ which we would find if the length N of each experimental sequence were infinite. It is this limiting $Freq(u_i)$ for $N \to \infty$ which is called $Prob(u_i)$, the probability of u_i (see theorem 7.5.1 for a more precise formulation).

Definition 7.5.2 *The interpretative definition of probability. $Prob\,(u_i)$, the probability of the outcome u_i in a single experiment, is defined as the limit, for $N \to \infty$, of the relative frequency in an experimental sequence of length N,*

$$Prob\,(u_i) = \lim_{N\to\infty} Freq(u_i) = \lim_{N\to\infty} \frac{n_i}{N} \qquad \forall i \in \{1, \ \ldots, \ I\} \ . \tag{7.35}$$

Because both n_i and N are nonnegative, it follows from this definition that

$$Prob(u_i) \geq 0 \qquad \forall i \in \{1, \ \ldots, \ I\} \ . \tag{7.36}$$

Assuming that eq. (7.34) holds also in the limit when $N \to \infty$, we have in addition that

$$\sum_{i=1}^{I} Prob\,(u_i) = 1 \ . \tag{7.37}$$

We see that this fundamental 'summing up to 1' law for probabilities is a direct consequence of the definition of probabilities as long-run relative frequencies.

Definition 7.5.2 agrees with Poisson's original definition of probabilities, cited in eq. (6.8) here.

A condition for the correctness of definition 7.5.2 is that the sequence $\mathbf{x}$ represents an unbiased sample. Thus, in connection with example 7.3.2, we must not choose all the N men from the same country because the height distribution for men of that country may differ somewhat from that for men over the whole globe.

7.5.3 Axiomatic Probabilities

The axiomatic theory of probability avoids the necessity and difficulty of letting N go to infinity in definition 7.5.2. It works with a σ-algebra of events, and assigns to each event E a numerical 'probability' value $Prob(E) \in [0,1]$. The axioms of the theory are chosen so that the most important formulas which hold for relative frequencies also hold for probabilities. One of the main axioms is that the probability of an event $E = E_1 \cap E_2$ is equal to the sum of the probabilities of E_1 and E_2 respectively when E_1 and E_2 are disjoint. Another axiom says that the probability of the event U (i.e. of the whole universe) is equal to 1. From these two axioms it follows that the summing-up-to-1 law for interpretative probabilities, eq. (7.37), holds also for the axiomatically defined ones.

By working in the universe U^N whose elements are sequences $\mathbf{x}$ of outcomes of N identical, noncomposite experiments, see eq. (7.30), one can derive the 'law of large numbers' in the axiomatic theory. This law refers to random variables. It sets up a relation between the expectation value of a random variable and its mean value in an experimental sequence of length $N \to \infty$. An important connection between probabilities and relative frequencies is derived by the following special application of this law.

We consider only whether the outcome of the experiment is the event $\{u_i\}$ or its complement $U - \{u_i\}$, i.e. '$NOT\ u_i$'. The outcome is assigned the number 1 in the former case, and the number 0 in the latter. Our experimental sequence consists now of a sequence of N 1's and 0's.

The random sequence of N outcomes has now been converted to a random sequence of N numbers (namely 1's and 0's). In probability theory one then says that we have a sequence of N values assumed by a 'random variable'. In the present case the random variable can assume values solely in the set $\{0, 1\}$.

The mean value of the random variable of the new sequence is equal to the sum of all elements of the sequence divided by N. However, the sum of all elements is equal to the number of 1's in the new sequence and therefore to the number n_i of u_i outcomes in the original sequence. Consequently the mean value of the random variable of the new sequence is equal to the relative frequency of u_i in the original sequence $\mathbf{x}$. Applying the law of large numbers to this case one then finds the following theorem,

Theorem 7.5.1

$$Prob\ \{\, |Freq(u_i) - Prob(u_i)\,| \ > \ \epsilon \,\} \ \to \ 0\ . \tag{7.38}$$

We will use the name p-f law *(probability-frequency law) for the relation expressed by eq. (7.38). The following is a reformulation of the p-f law in words.*

Consider the absolute value of the difference between the relative frequency of an event in a specific experimental sequence of length N and its probability. The p-f law says that the probability that this difference exceeds a small positve number ϵ goes towards 0 with increasing N. Notice that the p-f law does not say that the difference goes towards 0 for every experimental series, only that it goes towards 0 for the great majority of them. The difference tends stochastically *towards 0.*

Thus the law of large numbers finally establishes a connection between an axiomatically defined probability of u_i and its relative frequency in a long sequence.

The proof of the law of large numbers and of its application in the p-f form can be found, e.g., in Feller [16, p. 152].

Every experimental determination of a probability from a finite sequence of experiments can only be an estimate of the probability. However, theorem 7.5.1 tells us that the longer the sequence, the better is the chance that the estimate is within an ϵ of the true, underlying probability.

7.5.4 The Maximum Likelihood Estimate of Probabilities

Suppose that we are given an experimental sequence of length N with n_i u_i outcomes, and consequently N–n_i $NOT u_i$ outcomes (outcomes which are different from u_i). We wish to estimate $Prob(u_i)$ from this sequence.

To facilitate the computation, we will use the abbreviated notation

$$p = Prob(u_i) \qquad n = n_i \ . \tag{7.39}$$

Definition 7.5.3 *The maximum likelihood estimate of the underlying probability $p = Prob(u_i)$ is that value of p for which the probability of the observed experimental sequence is a maximum.*

Assuming the value p for the probability of u_i, a particular sequence of length N and n u_i outcomes has the probability

$$p^n(1-p)^{N-n} \ , \tag{7.40}$$

provided that successive outcomes are stochastically independent (see sectionXXXX). *XXXX*

To find the value of p which maximizes this expression, we differentiate it with respect to p and equate the result to 0,

$$np^{n-1}(1-p)^{N-n} + p^n(N-n)(1-p)^{N-n-1}\frac{d(1-p)}{dp} = 0 \ . \tag{7.41}$$

Dividing this equation by $p^{n-1}\,(1-p)^{N-n-1}$, and setting $d(1-p)/dp = -1$, we find

$$p = \frac{n}{N} \ . \tag{7.42}$$

This important result wil be given the status of a theorem,

Theorem 7.5.2 *Consider an experimental sequence of N mutually independent outcomes from a universe U. Then, for every $u_i \in U$, the maximum likelihood estimate $Prob^{ml-est}(u_i)$ of the underlying $Prob(u_i)$ is equal to the relative frequency of u_i in the sequence,*

$$Prob^{ml-est}(u_i) = \frac{n_i}{N} = Freq(u_i) \ , \qquad \forall i \in \{1, \ \ldots, \ I\} \ , \tag{7.43}$$

where n_i is the number of u_i outcomes in the sequence.

7.5.5 Summary

We sum up sect. 7.5 by noting that we have presented three connecting links between probabilities and relative frequencies.

The first of these is the interpretative definition 7.5.2 of probabilities as the limit of relative frequencies.

The second connecting link is the special application of the law of large numbers in the form of the p-f law of theorem 7.5.1. This law reasons from probabilities to relative frequencies. The latter are shown to tend stochastically towards probabilities.

Finally the maximum likelihood theorem 7.5.2 reasons from relative frequencies to estimates of the underlying probability distribution.

I must admit that in spite of the consistency between these three connecting links, I am not satisfied that I have succeeded in presenting a completely consistent connection between interpretative and axiomatic probabilities. My consolation is that apparently none of the experts on probability have succeeded in this project either; and that most textbooks on the theory of probability skip elegantly over this difficult gap instead of trying to bridge it.

We *can* say, however, that the axioms which are required to hold for probabilities in the axiomatic theory, as well as the theorems which can be derived from these axioms, are just those laws which can be shown to hold for relative frequencies in the interpretative theory. It is this agreement between the probabilities of the axiomatic theory and the experimentally found long-run relative frequencies which justifies the use of the axiomatic theory of probability. Without this agreement the axiomatic theory would be of no use to statisticians.

7.6 Precisely Known Probabilities

7.6.1 Certainties as Limits of Probabilities

Since we wish to present a unified theory of probability and logic in this book, it is important to consider the transition from probabilities between 0 and 1 to the degenerate case of probabilities whose value is exactly 0 or 1. In propositional calculus, an event with probability 1 would be assigned the truth value t; and an event with probability 0 would by assigned the truth value f. The discussion in the present section thus closes the circle started in sect. 7.2 where we introduced probabilities of 1 and 0 as a means of describing certainty of occurrence and of non-occurrence of an event respectively. And we said that the general definition of probabiliities allows us to operate also with intermediate probability values. Here we start out with the much more complicated case of general probabilities and analyse what happens to them in the limit when uncertainties go over to certainties.

To illustrate how probability distributions degenerate to certainty distributions we will assume that we have prior information concerning the certainty of an outcome or event. As an illustration, let us refer back to the die example 7.3.1, and the height example 7.3.2. Suppose that we know that a die is so heavily loaded that the

same face always turns up, e.g., the face 6. Every experimental sequence of length N will then consist of $n_6{=}N$ 6's only. In the height example 7.3.2, suppose that we know that the height of all objects in the set OB of (adult) men lies in the interval $u_{18} \in [170, 180)$ cm. Every random sample taken from this set will then result in the sequence $\mathbf{x} = \langle u_{18}, \ldots, u_{18} \rangle$ of N u_{18} outcomes. Both of these cases describe certain knowledge concerning the occurrence of the same outcome in every repetition of the noncomposite experiment; as well as certain knowledge concerning the non-occurence, or impossibility, of any of the other outcomes.

Let us now consider the *general* case of an experiment for which only one outcome can occur. Denoting this outcome by $u_{i'}$, we then have,

$$\begin{array}{ll} n_i = N & \text{for } \; i = i' \\ n_i = 0 & \text{for } \; i \neq i' \; . \end{array} \tag{7.44}$$

This holds for any length $N \geq 1$ of the experimental sequence.

In sect. 7.5.2 we considered the case that a composite experiment, resulting in a sequence $\mathbf{x} = \langle x_1, \ldots, x_n, \ldots, x_N \rangle$ of N u-values, is repeated a number of times. And we said that n_i/N, the relative frequency of u_i computed from each composite experiment, fluctuates about some mean value. However, for a random or stochastic sequence, the deviation of $Freq(u_i)$ from $Prob(u_i)$ tends stochastically towards 0 according to theorem 7.5.1.

For the certain outcome $u_{i'}$ it follows from eq. (7.44) that,

$$\begin{array}{ll} Freq(u_i) = n_i/N = N/N = 1 & \text{for } \; i = i' \\ Freq(u_i) = n_i/N = 0/N = 0 & \text{for } \; i \neq i' \; . \end{array} \tag{7.45}$$

There are thus no fluctuations of $Freq(u_i)$ about a mean value. The relative frequency of a certain event is always 1, and that of an impossible event (an event that is certain not to occur) is always 0, independent of the length N of the sequence.

In the interpretative definition of eq. (7.35) of probabilities we therefore need no longer take any $N \to \infty$ limit. For a certain event, as well as for an impossible one, we find from (7.35) and (7.45) that

$$\begin{array}{ll} Prob(u_i) = Freq(u_i) = n_i/N = N/N = 1 & \text{for } \; i = i' \\ Prob(u_i) = Freq(u_i) = n_i/N = 0/N = 0 & \text{for } \; i \neq i' \; . \end{array} \tag{7.46}$$

From eq. (7.46) it follows that for a distribution whose outcome is certain,

$$|Freq(u_i) - Prob(u_i)| = 0 \qquad \forall u_i \in U \quad \text{and} \quad \forall N \geq 1 \; . \tag{7.47}$$

We thus find that in the p-f law of eq. (7.38), the difference between $Freq(u_i)$ and $Prob(u_i)$ not only *goes towards 0* with increasing N, it *is 0* for every sequence, independent of the value of N.

Our end result is thus that for any outcome u_i of a distribution for which only one specific outcome $u_{i'}$ can occur we have that the relative frequency of each of the

outcomes, as computed from an experimental sequence of length N, is precisely equal to the probability of that outcome for any $N \geq 1$.

We end this subsection with formal definitions of a 'certainty distribution' and of a 'certain event and an 'impossible event.

Definition 7.6.1 *Consider a probability distribution* $Prob(u_i)$, $u_i \in U = \{u_1, \ldots, u_i, \ldots, u_I\}$. *We say that this distribution is a certainty distribution iff all outcomes of any experimental sequence are identical and equal to* $u_{i'}$; *or equivalently iff*

$$\begin{array}{ll} Prob(u_i) = Freq(u_i) = 1 & for \;\; i = i' \\ Prob(u_i) = Freq(u_i) = 0 & for \;\; i \neq i' \; . \end{array} \qquad \forall N \geq 1 \; . \tag{7.48}$$

The outcome $u_{i'}$ *is called the certain outcome, and each of the other outcomes are called impossible outcomes, or outcomes that are certain not to occur.*

The word 'outcome' can be replaced by 'event' in the above formulation, the event being a subset of the original U. *The formulation refers now to a universe* U' *of two outcomes, namely* E *and NOT* E, E *being the symbol for the event.*

Note that a certain *outcome* always implies a certainty *distribution* because of the summing-up-to-1 law for probabilities. In contrast, an impossible outcome does not necessarily imply a certainty distribution. It may well be that one or more outcomes of a noncomposite experiment are impossible (have probability 0), but that none of the remaining outcomes is certain (has probability 1).

In conclusion we note that the transition from probabilities between 0 and 1 to a certainty distribution completely changes the character of the stochastic process. The former fluctuations of relative frequencies (in different sequences of length N) are reduced to 0; and consequently it is no longer necessary to make use of the difficult $N \to \infty$ limit.

Seen in this light, a 2-valued logic which uses solely the truth values t and f is a very primitive one as compared with a logic which operates also with intermediate probability values and the resultant possibility of fluctuating experimental sequences.

7.6.2 Other Precisely Known Probabilities

According to the p-f law of theorem 7.5.1, the estimation of a probability value from a random sample can never be relied upon to be completely precise. We have, however, seen in the previous subsection that for prior information concerning a certainty distribution the probabilities need not be estimated from a sample but are known with complete precision.

Another case in which the probabilities are known with complete precision is that of a finite object set, the attribute value of each object being known. An example is the random drawing of a sequence of cards from a complete pack of 52 cards. After each drawing the drawn card is returned to the pack, and the pack is reshuffled. The

probability of drawing a given card, e.g. the queen of hearts, in a single draw is then exactly 1/52. And the probability of drawing a queen is exactly 4/52.

This case does, however, differ from that of a certainty distribution in that we do have fluctuations in the relative frequencies of occurrence of a given event or outcome for different random sequences. But we do not need to make use of such sequences for the purpose of estimating the probabilities which *are* precisely known a priori.

7.7 Specification of Single-Instance Probabilities

7.7.1 Meaning of Single-Instance Probabilities

Sect. 7.7 has three purposes. The first is to define and clarify the meaning of the probability of occurrence of a given outcome in a single instance of an experiment. The second purpose is to show how such probabilities can be deductively updated to 1 or 0; this is in contrast to the deductive updating of a given underlying probability by prolongation of an experimental sequence. Finally we mention some natural language modifiers which serve the purpose of specifying certainties and probability values.

We have seen that both the interpretative and the axiomatic theory of probability establish a connection between probabilities on the one hand, and long-run relative frequencies in a sequence of random choices on the other. In the interpretative theory the connection is established by definition, in the axiomatic theory by assumption.

In sect. 7.5 we saw that the values of the probabilities of the different outcomes of an experiment can be estimated from the frequency of the outcome in a long experimental sequence. Alternatively the probabilities can be specified in advance, the specification being based upon knowledge concerning the object set to which the experiment refers. Examples of such specifications were given in subsections 7.6.1, 7.6.2. Other examples of such specifications are predictions based upon laws of nature which we believe to be true. E.g., a physicist may predict the occurrence of an eclipse of the sun at a certain location and at a certain time t_2 with what she considers to be a complete certainty, i.e. with probability 1. This prediction is based upon her knowledge of-, and belief in-, the laws of gravitation, as well as on her measurement of the positions and velocities of the earth, sun and moon at a previous point of time $t_1 = t_2 - \Delta t$. If she considers that her measurements at time t_1 are subject to inaccuracies, then she may change her estimate of the probability of an eclipse at the given location, and at time t_2, from 1 to, e.g., 0.98.

Note that we are now talking about $Prob(u_i)$ in a single instance of an experiment, the instance referring to a given time of occurrence. Another example of a single-instance probability was given in sect. 7.2, namely the probability of my being at home tomorrow at 10 a.m. .

In the following we define the probability of an outcome in a single instance of an experiment. In examples 7.7.1, 7.7.2 we then illustrate the meaning of this definition.

Definition 7.7.1 *$Prob(u_i)$, the probability of the outcome u_i in a single instance of an experiment is defined to be numerically equal to the value of the probability of*

that outcome in an experimental sequence, provided that both probabilities refer to the same object set, i.e. to the same underlying probability distribution.

The reason why we emphasize that we are talking about the probability of an outcome *in a single instance of an experiment* is that when one talks about the *updating* of a single-instance probability, it is easy to forget that one usually refers to a change in the reference object set, the new object set being a subset of the original one. It then seems very puzzling why the updating rules for such probabilities
XXXX (see sections 9.1 and XXXX) are completely different from the updating rules of the estimates of probabilities based on stepwise prolongations of an experimental sequence (see sections 8.1-8.9) in which the object set is kept invariant.

If the values of probabilities are to be meaningful, we must be able to define the object set to which they refer. In the following we use a modification of the 'at home' example of sect. 7.2 to clarify the meaning of single-instance probabilities.

Example 7.7.1 *of a single-instance probability. Consider that on Saturday, June 4, 1994, an informant S1 utters the statement,*

The probability that Ruth is at home at 10 a.m. on Sunday, June 5, 1994, is x , (7.49)

where x is some number in the real interval [0,1].

Assume that S1 uses the value x=0.8 in the statement (7.49). A prediction of this sort is always based on estimates of the probability of occurrence of certain factors which influence Ruth's being, or not-being, at home. Let us assume that S1 knows that, as a rule, Ruth takes long walks on Sunday and is consequently not at home. However, Ruth has just told S1 that she will probably not be able to take a walk because she feels that she is probably developing a cold. S1 interprets this in the sense that that in a fraction 0.8 of all previous cases in which Ruth felt just as she is feeling today, she was unable to take a walk on the next day and stayed at home. And she expresses this estimate in the form of sentence (7.49).

From S1's point of view the object set now consists of all days on which Ruth has felt just as she is feeling today. S1 marks a fraction 0.2 of these objects 'able to take a walk on next day'='1/(able to take a walk on next day)', and a fraction 0.8 'not able to take a walk on next day'='0/(able to take a walk on next day)'. The two alternative notations for affirmation and for negation on the left and right hand sides of each of the two equality signs correspond to the natural language notation, and to the slightly more formal one used often in this book, respectively.

The listener who hears S1 utter the statement (7.49), and who wants to visit Ruth on Sunday June 5-th, does not know S1's or Ruth's reason for the specification of the probability value x=0.8. For her the object set is a collection of days on which an informant makes a statement of the type of (7.49). And she interprets (7.49) in the sense that in a fraction x of all cases in which such a statement is uttered by a truthful informant Ruth is indeed at home on the specified day, and in a fraction $1-x$ she is not.

Example 7.7.2 *of the updating of a single-instance probability. We refer again to the statement (7.49) which S1 made on June 4-th. On June 6-th, S1 knows with certainty that Ruth was not at home on the previous day because the cold did not materialize. On June 6-th, S1 therefore states,*

The probability that Ruth was at home at 10 a.m. on Sunday, June 5, 1994, is 0 . (7.50)

S1's reference object set has thus changed to the object set consisting of the single object 'Sunday, June 5-th'. Or, equivalently, to the set of all Sundays on which Ruth was not at home at 10 a.m. Resulting in the updating of x from 0.8 to 0; i.e. to the certainty that the outcome does not occur or, equivalently, to the impossibiltiy of the outcome.

The updating of the probability of a single instance of occurrence of an outcome, and, more generally, the updating due to additional information which narrows down the object set, are discussed again in in sections (9.1) and XXXX. *XXXX*

7.7.2 Natural Language Probability-Modifiers

Natural languages have a number of means for specifying approximate probability values, as well as probability intervals. These means consist of the attachment of a modifier to an originally affirmed or negated declarative sentence. The modifier modifies the probability value implied by the unmodified declarative sentence. Some of the modifiers used in English, as well as the probability values which they induce, are listed in fig. 7.1. All values in the last column of fig. 7.1 refer to the probability of the outcome of line 1, i.e. to the probability of 'I am at home at time t' as specified by the sentence in the preceding column.

The probability values 1 or 0 of lines 1, 2 and 9-11 denote certainty values. These values are the only correct interpretations of the corresponding sentences. To assign the outcome 'I am home at time t' different probability values from those in the last column would violate the correct use of English.

Lines 4a,b, 5 and 6 refer to the modifiers 'probably', 'may be' and 'may possibly be'. The values in the last column of these lines are subjective estimates by the author of the numerical probability values induced by the three modifiers. Note that when the same modifiers are applied to the affirmed and to the negated sentence respectively, then the corresponding probability values must add up to 1. This holds for each of the following pairs of lines in fig. 7.1: (1, 10), (4a,b, 13) and (6, 12). The reason is that a given modifier always induces the same probability value of the outcome to which it is attached. Thus, e.g., 'probably not at home' induces the probability 0.85 for the outcome 'not at home', just as 'probably at home' induced the probability 0.85 for 'at home'. However, when 'not at home' has probability 0.85, then 'at home' must have probability 0.15 (when the given sentence is 'probably not at home') because '$1/u$=at home' and '$0/u$=not at home' make up a complete universe or space U=$\{1/u,\ 0/u\}$ of possible outcomes.

	Modifier	Sentence	Probability(at home)
1.		$1/u$=I am at home at time t	1
2.	*It is certain that*	*It is certain that* I am at home at time t	1
3.	*It is not certain that*	*It is not certain that* I am at home at time t	[0,1)=0m
4a.	*probably*	I am *probably* at home at time t	0.85=m
4b.	*It is probable that*	*It is probable that* I am at home at time t	0.85=m
5.	*may be*	I *may be* at home at time t	0.5=m
6.	*may possibly be*	I *may possibly be* at home at time t	0.2=m
7.	*There is a possibility that*	*There is a possibility that* I am at home at time t	(0,1]=m1
8.	*It is not impossible that*	*It is not impossible that* I am at home at time t	(0,1]=m1
9.	*It is impossible that*	*It is impossible that* I am at home at time t	0
10.	*not*	$0/u$=I am *not* at home at time t	0
11.	*It is certain that not*	*It is certain that I am not* at home at time t	0
12.	*may possibly not be*	I *may possibly not be* at home at time t	0.8=m
13.	*probably not*	I am *probably not* at home at time t	0.15=m
14.	*It is not probable that*	*It is not probable that* I am at home at time t	$[0,1]-\{0.85\}$

Figure 7.1: *Suggested probability values or intervals corresponding to different natural language modifiers. The probability values refer to the outcome specified in line 1, i.e. to 'I am at home at time t'. For the meaning of the m-expressions in the last column, see sect. 8.4.* ***figprobmodif***

For a sentence with a modifier there exists, however, a different possibility of negation, namely the negation of the modifier itself. In this case we compare two identical sentences, except that the modifier of the second sentcence is a negation of the modifier of the first sentence. E.g., consider the modifier 'It is certain that' in line 2 which induces the probability value 1 just like sentence 1. 'It is not certain that I am at home' tells us that the probability of my being at home is not equal to 1. The universe of probability values is the real interval [0,1], The negation *of the probability value 1*, (not of the outcome 'at home'!) corresponds therefore to the complementation of {1} with respect to the universe of possible probability values [0,1]. This leaves open the whole real interval [0,1] except the value 1. The resulting interval of possible probability values is denoted in line 3 by [0,1) or by $0m$ in the m-notation of sect. 8.4.

Similarly, 'It is impossible' induces the probability value 0 in line 9, and 'It is not impossible' of line 8 induces the complement of {0} with respect to [0,1], namely (0,1]=m1. The probability *intervals* belonging to the modifiers 'not certain' and 'not impossible' are thus objective intervals with no intersubject variation, just like the probability *values* induced by 'certain' and 'impossible'.

The same principle applies to lines 4a and 14 pertaining to the modifiers 'probable' and 'not probable' respectively. If we assume that the first one induces the probability 0.85, then its negation leaves open all values in the real interval [0,1] except 0.85.

All of the three linguistic modifiers of lines 4-6 induce probability values between 0 and 1. The exact numerical values corresponding to these modifiers will vary somewhat from subject to subject, but their ordering will probably be the same for all subjects.

Because of the intersubject variability, the assignment of a precise and unique numerical value to the probability of an outcome induced by a sentence using, e.g., the modifier 'probably', is thus unsatisfactory, and may even lead to formal inconsistencies. E.g., the information supply 'Prob(at home)=0.8' may be considered inconsistent with the information supply 'Prob(at home)=0.9'. Furthermore a probability 'interval' such as that of line 14, consisting of all points in [0,1] except for the point 0.85, is an ugly mathematical creation.

For these reasons it is better to replace the three precise numerical probability values (*between* 0 and 1) of lines 4-6 by fuzzy sets over the probability interval [0,1]. The subjective fuzzy set for, e.g., 'probably' would then have the value 1 over the abcissa interval 0.85-1, and fall off to 0 in an S-shaped manner on the left hand side of the 0.85 point. Even if the subjective fuzzy sets of two subjects are slightly displaced with respect to each other, the fuzzy set representation does not lead to a complete intersubject inconsistency. The negation of a modifier then induces a fuzzy set which is the complement of the original one. (For fuzzy sets and their negation, see XXXX *XXXX* and XXXX.) *XXXX*

The fuzzy set solution applies neither to straightforward affirmation and negation, nor to the modifiers 'it is certain that', 'it is impossible that'. In all these cases there exists no intersubject variation concerning the induced probability values.

We sum up the conclusions of this subsection concerning affirmed and negated sentences with modifiers by the following two theorems.

Theorem 7.7.1 *concerning the modified probability of an outcome and of its negation. When the same modifier is applied to an outcome and to its negation respectively, then the corresponding two probabilities of the affirmed outcome add up to 1.*

Theorem 7.7.2 *concerning the modified probability of a given outcome, using a modifier and its negation respectively. Consider two sentences pertaining to the same outcome. Both sentences make use of a probability modifier, the two modifiers being negations of each other. The probability intervals for the given outcome which are induced by the two sentences respectively are then complements of each other with respect to the probability interval [0,1]. When we say that the sentence induces a probability* interval*, we mean that any value in the interval is an acceptable one as far as that sentence is concerned. If we use fuzzy sets over the probability interval [0,1] (instead of point-valued or interval-valued probabilities) then the fuzzy sets induced by the two modifiers are complements of each other. Certainties and impossibilities always correspond to the precise probability values 1 and 0 respectively and are therefore not represented by fuzzy sets.*

Chapter 8

Updating of Probability Values, Type 1 of Updating

8.1 Introduction

This chapter deals with the updating of information concerning an underlying probability distribution. The main emphasis of the chapter is upon the following distinctions.

1. The important distinction of certain knowledge concerning the occurrence or nonoccurrence of an event from uncertain knowledge. A certainty has the probability value 1 or 0, uncertainty has the value $m \in (0,1)$. We will call these three values 'unique values' although the value m is actually interval-valued. The 'uniqueness' refers to the distinction between certainty versus uncertainty.

2. The distinction of ignorance concerning the values of a probability distribution from the knowledge that the probability distribution is a uniform one. Bayes' postulate (see sect. 8.3), which has been debated since 1763, does not make this distinction. In the m-notation of sect. 8.4, ignorance concerning the value of the probability of an event is expressed by the 'set-valued' probability values $0m1=\{0, m, 1\}$, $0m=\{0, m\}$, $m1=\{m, 1\}$, and $01=\{0, 1\}$. $0m1$ corresponds to complete ignorance concerning the probability value of the event. $0m$ corresponds to ignorance as to whether the event will never occur or sometimes occur, i.e. to ignorance as to whether the event is impossible or uncertain. $m1$ corresponds to ignorance as to whether the event will sometimes occur or always occur, i.e. to ignorance as to whether the event is uncertain or certain to occur. And 01 corresponds to ignorance as to whether the event will never occur or always occur according to some specified information which excludes the possibility that it will sometimes occur and sometimes not.

3. The distinction between the updating or 'learning' of a single underlying probability distribution by successive prolongations of an observed experimental sequence (sections 8.5-8.9); versus the updating of the probability values by

additional information supply and a consequent narrowing down of the object set to which the probability distribution refers. The most extreme case of such updating concerns the observation of the occurrence or nonoccurrence of the event in a single instance of an experiment (section 7.7 and chapter 9).

In the first updating case of item 3 one finds that a probability value m can never be updated to 1 or 0; in contrast to the second case in which the specification of a certainty (probability 1 or 0) overrides the specification of an intermediate probability value m for the same outcome or event. Unless one keeps the distinction of item 3 in mind, it is very confusing why we can have two such seemingly contradictory updating rules.

Updating of type 1 is summed up by theorem 8.5.2 and in figs. 8.5 and 8.7. It is treated in detail in sections 8.5-8.9, and is illustrated by example 8.1.1 below, and in sect. 8.6.3 in connection with quantification problems. Updating of type 2 is treated
XXXX in chapter 9 and more generally in chapter 10, sectXXXX. Example 8.1.2 below and the examples of chapter 9 illustrate updating of type 2. For the treatment of both types of updating we need the *m-notation* of sect. 8.4. m is an acronym for 'medium' or 'maybe'. It denotes any value in the real interval (0,1) (excluding the two end points). If the numerical value of m is known then one can, of course, replace m by this value.

Example 8.1.1 *We throw a (generally biased) die twenty times. The outcome 2 occurs once and only once in this sequence. We then know that Prob(2) for this sequence is bigger than 0 and smaller than 1, i.e., Prob(2)=m. This value will never be changed by prolongations of the experimental sequence because in every such sequence the first 2 outcome will still be present; so will the 19 outcomes that are different from 2. If the probability value had been 1, then all outcomes of the sequence would have been the face 2. If it had been 0, then no outcome would have been this face. The probability value m can therefore never be updated to 0 or 1.*

Example 8.1.2 (See also example 9.1.2)
At a point of time t_{n-1}, the data base is in possession of the information

$$info\,(t_{n-1}) = Drawer\ \#1\ contains\ knives\ OR\ forks\ , \tag{8.1}$$

where OR stands for the inclusive or the exclusive disjunction.

At time t_n the knowledge base is supplied with the new item of information

$$newinfo\,(t_n) = Drawer\ \#1\ does\ NOT\ contain\ knives\ . \tag{8.2}$$

(8.1) leaves open the following three possible outcomes (using an abbreviated notation for the complete sentences),

$$\begin{array}{l} knives\ BUT\ NOT\ forks, \\ forks\ BUT\ NOT\ knives, \\ knives\ AND\ forks. \end{array} \tag{8.3}$$

Each of these three outcomes has the probability of occurrence m according to the state of knowledge (8.1) at time t_{n-1}. The probability of occurrence of each of the three labels of of (8.3) is here interpreted in the sense of the last paragraph of sect. 7.7.1. Namely in the sense that when a truthful subject makes the statement (8.1) on N different occasions ($N \gg 1$), then in a fraction m of these N cases, $0 < m < 1$, the outcome will consist of the affirmation of the first component and the negation of the second component of the disjunctive statement (first line of (8.3)). In a fraction m the outcome will consist of the affirmation of the second component and the negation of the first one (second line of (8.3)). And in a fraction m the outcome will consist of the affirmation of both components (third line of (8.3)). The three fractions must add up to 1.

However, the additional *newinfo* of (8.2) excludes all outcomes for which the first component is affirmed (first and last line of (8.3)). The reference object set of the N outcomes is thus narrowed down to solely those objects for which the outcome consists of the negation of the first component of (8.1) and the affirmation of the second (second line of (8.3)). This outcome has therefore the probability of occurrence 1, the other two lines of (8.3) having the probability 0. The particular instance of an outcome to which (8.1) and (8.2) refer has therefore also the probability 1 for the second line of (8.3). Examples 8.1.1, 8.1.2 are discussed again at the end of sect. 9.1.

Example 8.1.2 illustrates the general rule of sect. 9.1 that a specified probability value of 0 for a given *instance* of an outcome takes precedence over any other probability value specified for that outcome (excepting 1). The same holds for a specified probability value 1. If one item of information assigns the probability value 1 to the outcome of this instance, and another the probability 0, then the two items are contradictory. The 'instance of an outcome' refers here to the reference in the real world of the statements 8.1 and 8.2. This reference being drawer #1 whose contents are the same at times t_{n-1} and t_n. Only our state of knowledge concerning these contents has become more specific at time t_n.

8.2 Updating in Mathematical Logic

Updating of probability distributions is an important subject in the theory of probability. In a theory of logic, updating of information should also be one of the main subjects. That this is not quite so in traditional logic is due to the particular slant of propositional calculus. The truth tables of this calculus specify the t or f truth value of a composite proposition as a function of all possible combinations of the t or f truth values of its components, including those for which the composite proposition is false.

This is not the type of problem that we usually meet in a natural language discourse in which one person supplies information to another. Neither is it the problem that we try to solve in a data base system. In these cases the informant supplies information which the listener or the data base assumes to be true. The listener or database system tries, or should try, to store the information in a way such that 1) no

information is lost; 2) it is easy to retrieve the supplied information; 3) it is easy to draw inferences from the information. Of course it may be that the listener suspects the informant of being capable of lying. The best thing she can do in this case is probably to ignore the information supplied by this particular informant. In short, the data stored in the knowledge base should consist only of such information as the system believes to be true.

An additional desirable requirement (4) for a knowledge base system is that different formulations of the same information should, as far as possible, have the same representation in the database, this being the most efficient one from the viewpoint of the previous requirements.

The situation of eqs. (8.1), (8.2) in example 8.1.2, and illustrates this point. We would like the procedures of the knowledge base system to update automatically the information in (8.1), (8.2) to the form that it would have if (8.1), (8.2) had been replaced by the single item of information

$$info\,(t_n) = \text{Drawer } \#1 \text{ contains forks BUT NOT knives }, \tag{8.4}$$

Such an updating does not occur automatically in propositional calculus in which we have no automatic mechanism for deriving (8.4) from (8.1)and (8.2). The best we can do in propositional calculus is to show that the conjunction of eq. (8.1) and eq. (8.2) is equivalent to eq. (8.4). But this preassumes some unknown source of information which has supplied (8.4).

The chain set system of part **??** of this book combines propositional calculus with probabilities and their updating. In this system the chain set for the conjunction of (8.1) and (8.2) is, indeed, automatically updated to the same form as the chain set for (8.4).

8.3 Bayes Postulate and its Drawback

The classical tool for dealing with ignorance is the use of Bayes postulate. This postulate says that when nothing is known about the probability distribution in a given universe, then we should assign the same probability to each outcome. E.g., when we do not know whether a die is loaded or not, then we should assign the probability $\frac{1}{6}$ to each face.

Bayes postulate gives rise to a serious ambiguity because two different states of information result in the same representation in the knowledge base; namely the state of complete ignorance concerning a probability distribution versus the case in which the probability distribution is known to be a uniform one. In connection with the die example, the state of ignorance is the case when we do not know whether the die is unloaded or loaded, and if it is loaded we do not know to which degree. According to Bayes postulate this state is described by the initial assignment of the probability $\frac{1}{6}$ to each face, just as in the case when the die is known to be unloaded. The following example illustrates that we may incur a great loss of money when we apply the Bayes-postulate description of the state of complete ignorance in a betting situation.

Example 8.3.1 *Suppose that we have an opaque urn about which we are informed that it contains 1000 balls, each ball being either black or white. The balls are of equal size and weight.*

An experiment consisting of 100 consecutive random drawings of a ball from the urn is performed. After each drawing the color of the ball is noted, the ball is returned to the urn, and the urn is thoroughly shaken so that the balls are rerandomized.

In the following we distinguish between two cases of prior information.

Case 1. *We are not told anything in advance about the number of black balls and the number of white balls in the urn.*

Case 2. *We have the prior information, i.e. we are told in advance, that the urn contains 500 black and 500 white balls.*

We are now offered to make a bet that an experimental sequence of 100 drawings will result in a number of black balls lying in the interval 40 to 60, and a number of white balls lying in the interval 60 to 40.

In case 1, the application of Bayes postulate results in the estimate 0.5 for the probability that a drawn ball will be black and 0.5 for the probability of its being white. This probability distribution is the same as that of case 2.

However, in case 2 we should be willing to bet much more money in favor of the specified outcame of 100 drawings than in case 1. For all we know, the urn might contain 10 black balls and 990 white ones in case 1 so that the acual (but unknown) probability distribution is 0.01 for black and 0.99 for white. It is then very improbable that the specified outcome will occur.

The m-notation explained in sect. 8.4 has different descriptions of the above two prior states of information. The description of case 2 is unchanged, namely

$$P(\text{black}) = 0.5 \ , \qquad P(\text{white}) = 0.5 \ . \tag{8.5}$$

In contrast, the prior information in case 1 is described by

$$P(\text{black}) = m \ , \qquad P(\text{white}) = m \ , \tag{8.6}$$

or by

$$P(\text{black}) = 0m1 \ , \qquad P(\text{white}) = 0m1 \ . \tag{8.7}$$

The description (8.6) holds when we know a priori that the urn contains at least one black, and at least one white ball. (8.7) holds when the prior information leaves open the two cases in which the urn contains only black or only white balls. m is always an element of the open interval (0,1). However, two m values, e.g. in (8.6), need not be numerically equal.

In contrast to the formal description (8.5), which holds for both case 1 and case 2 according to Bayes postulate, the formal description of (8.6) or (8.7) will make us very wary of betting in favor of 40-60 black and white outcomes in case 1.

8.4 The m-Notation

The m-notation, of which we will make use in part **??**, uses the symbol m for an intermediate probability value that is neither 0 nor 1. It is a means of making the important distinction between certainties and uncertainties, and of being able to work with uncertainties in cases of ignorance when the precise numerical value of the probability is unknown. Both of the two updating types mentioned in sect. 8.1 can be used in combination with the m-notation. We shall see in connection with classification
XXXX and quantification structures ofXXXX, as well as in the case of a conjunction of IF THEN statements, that Bayes postulate can help us solve the most basic problems of pure tree structures. In the more complicated 'multiple partition case' (in which a partial overlap between two classes is allowed) Bayes postulate is inadequate. These problems can be solved only with the aid of the m-notation.

In sect. 8.3 we showed that Bayes postulate is ambiguous in the sense that it can result in the same description in the data base of two different states of information. Reasoning with the aid of the m-notation remedies this state of affairs. This reasoning does not make use of Bayes postulate. Neither does it preclude the use of a numerical probability value, such as 0.4, instead of m in a case in which this value is known.

We saw in sect. 7.7.2 that natural language converts an expression of certainty to an expression of uncertainty by using qualifiers such as 'probably', 'maybe', 'there is a small chance that'. These expressions correspond to decreasingly smaller numerical probability values, all of which lie inside our m-interval (0,1). In contrast, certainty concerning an outcome is expressed by the lack of a quantifier in natural language. And certainty concerning the absence of an outcome is expressed by the negation.

The m-notation follows the hint of natural language of making a basic distinction between certainty (probability 1 or 0) on the one hand and uncertainty (probability in interval (0.1)) on the other. It partitions the [0,1] probability interval for an outcome or event into three disjoint regions, namely

$$Prob(event) = 0 \qquad Prob(event) = m = (0,1) \qquad Prob(event) = 1\ . \tag{8.8}$$

(0,1) denotes the probability interval which includes all values in the real interval [0,1] except the two end point. When we say that the probability of the event *is interval-valued* and *equal to* m=(0,1), we mean that it lies in the interval (0,1).

The state of complete ignorance concerning the probability of the occurrence of a given event leaves the whole real interval [0,1] open. We will denote its probability by $0m1$ because it is the union of the interval m=(0,1) with the 'interval' consisting of the single point 0, and that consisting of the single point 1.

Information supply which excludes $Prob(event) = 1$ corresponds to the interval [0,1), this interval being closed on the left and open on the right. We will denote it by $0m$. Similarly information supply which excludes $Prob(event)$=0 corresponds to the interval $Prob(event)$=(0,1], this interval being open on the left and closed on the right. It will be denoted by $m1$. The m-notation thus operates with the interval-valued probability values listed in fig. 8.1.

$Prob(event) = 0m1 = \{0, m, 1\} = [0, 1]$	state of complete ignorance
$Prob(event) = 1 = \{1\}$	certainty
$Prob(event) = 0 = \{0\}$	impossibility (certainty that event does NOT occur)
$Prob(event) = m = \{m\} = (0, 1)$	certainty and impossibility excluded
$Prob(event) = 0m = \{0, m\} = [0, 1)$	certainty excluded
$Prob(event) = m1 = \{m, 1\} = (0, 1]$	impossibility excluded
.....	
$Prob(event) = 01 = \{0, 1\}$	either certainty or impossibility
$Prob(event) = \emptyset$	inconsistent information

Figure 8.1: *Possible point- or interval-values for probabilities in the m-notation. As an example, the specification of an interval-valued probability such as Prob(event)=m1=(0,1] means that we have information that the probability of the event can assume any value in the real interval [0,1] except 0. In each row, the value after the first equality sign is an abbreviated notation for the set after the second sign and for the interval-value after the third one. Concerning the last two rows, see sect. 8.6.4.* ***figupdatem1***

Definition 8.4.1 *of point-valued, interval-valued, unique, set-valued, and m-valued probability values; and of uncertainty versus ignorance. Numerical probability values such as 0, 0.4, 1 are called point-valued. Probability values which are defined to lie in a proper or improper subset of the real interval [0,1] are said to be interval-valued. The values* $0m1$, $0m$, m, $m1$ *and* $0m1$ *are all interval-valued. The values 0, m and 1 are called unique although m is interval-valued. However, these three values distinguish uniquely between certainty, uncertainty and impossibility of an event. All the probability values of fig. 8.1 are called m-values.*

We say that $0m$, $m1$, $0m1$ *and* 01 *are set-valued probability values. A set-valued probability will be said to describe a state of ignorance. This is in contrast to the unique probability value m which describes a state of uncertainty; while, e.g.,* $0m$ *describes ignorance as to whether the event is impossible or uncertain to occur. The state of complete ignorance concerning the probability of an event is characterized by the set-valued probability value* $0m1$.

8.5 Learning from Experience and Updating by Prolongation of Observed Sequence

8.5.1 Introduction

Both children and scientists learn from experience. Formally this learning process can be described by successive increases of the length N of the observed sequence. The learning can concern information of both the factual and the meaning-related

type (see chapter 4).

In sect. 8.5 we derive two updating procedures for probabilities, based on the observation of an experimental sequence and its prolongation. One procedure is for purely deductive learning, the other for inductive learning.

Because of statistical flucutations, numerical estimates of probabilities from an observed sequence, however long, can never be guaranteed to be correct. Instead of using such estimates we will therefore start out by assuming that the whole real interval [0,1] is initially open for the value of a probability. This interval, which describes the state of complete ignorance, will then be narrowed down or shrunk, by the observation of an experimental sequence. The updating due to the shrinkage of ignorance concerning a probability value can also be called a 'learning process'. The tables of figures 8.5, 8.7 describe this updating for deductive and inductive learning respectively. Sect. 8.5.5 desribes the updating process as an intersection of possible probability intervals.

The learning-induced shrinkage of the possible probability interval resulting from a prolongation of an experimental sequence has no connection with the updating of sect. 9.1 which shrinks the object set to which a probability value refers.

Consider a single random experiment with I mutually exclusive possible outcomes from the universe $U = \{u_1, \ldots, u_I\}$. And let this experiment be repeated N times, resulting in a random sequence of N outcomes. Following the notation of sect. 7.4, we denote this sequence by $\mathbf{x} = \langle x_1, \ldots, x_n, \ldots, x_N \rangle$, each x_n being an element of U.

Example 8.5.1 *As an example of learning a factual (synthetic) probability distribution from a sequence of length N, suppose that we want to learn the probability distribution of height intervals for all men labeled 'tall'. The latter make up our object set OB. A single experiment consists of drawing at random a man from this set of tall men, noting his height interval, and returning him to the set. The experiment is repeated N times, resulting in a sequence of N values from the universe U of height intervals.*

The following reverse example concerns the learning of a meaning-related (analytical) set of concepts. It illustrates the meaning of the grade of membership concept of fuzzy set theory according to the TEE model. (See [27], [25], [26], [28], [29], [30] *XXXX* for this model.)XXXX

Example 8.5.2 *Imagine that a child or program tries to learn the appropriateness of the labels 'short', 'medium', 'tall' to men whose height lies in a given height interval u_i. The* universe *is now a complete set of mutually exclusive labels, U={short, medium, tall}. We assume that each element of the set of all men has been assigned one of these three labels by some subject.*

The object set *OB is that subset of the set of all men whose height lies in the height interval u_i. The probability distribution concerns the probability that a randomly chosen man from the object set has been assigned a specified label from U, e.g. the*

label 'tall'. This $Prob(tall|u_i)$ is the definition of $\mu_{tall}(u_i)$, the grade of membership of the height u_i in the fuzzy set 'tall'.[1]

For any experiment of the type described in sections 7.3, 7.4 we have that the observed frequency of the outcome u_i is defined by the formula

$$Freq(u_i) = \frac{n_i}{N} \ , \tag{8.9}$$

where n_i denotes the number of times that u_i occurred in the experimental sequence of length N. We have therefore that

$$n_i \in \{0, 1, \ldots, N\} \ , \tag{8.10}$$

and consequently

$$\begin{array}{ll} Freq(u_i) = 0/0 & \text{for} \quad N = 0 \\ Freq(u_i) \in [0,1] & \text{for} \quad N \geq 1 \end{array} \ . \tag{8.11}$$

The first, N=0 row of this equation denotes the state of ignorance before any observation of an outcome. In the second row we only note that $Freq(u_i)$ has, for $N \geq 1$, some definite numerical value which always lies in the interval [0,1].

We will use the same partition for the possible $Freq(u_i)$ values as the partition for the underlying probability values in eq. (8.8),

$$Freq(event) = 0 \qquad Freq(event) = m = (0,1) \qquad Freq(event) = 1 \ . \tag{8.12}$$

8.5.2 Deductive Learning from Experience

Consider an experiment referring to a universe $U = \{u_1, \ldots, u_i, \ldots, u_I\}$ of I different outcomes with underlying probabilities $Prob(u_i)$, $i = 1, \ldots, I$. The experiment is repeated N times, resulting in a sequence of N outcomes from U. The relative frequency of a particular outcome u_i in this sequence is given by eq. (8.9).

From the observed relative frequency $Freq(u_i)$ we wish to deduce the possible $Prob(u_i)$ values which can have given rise to this frequency. In the terminology of the m-notation which operates with interval-values and set-values (see fig. 8.1 and definition 8.4.1) we can also say that we wish to deduce the set-value or unique value of the underlying $Prob(u_i)$ from the observed relative frequency.

Fig. 8.2 shows the possible values of the underlying $Prob(u_i)$ (column II) for a given value of the relative frequency of the outcome u_i in an experimental sequence (column I). All 'values' are in the m-notation of sect. 8.4.

[1] There can be different reasons why men of a given, precisely measured, height are not always assigned the same label. For three such 'sources of fuzziness', see [27] or [26]. Source of fuzziness # 1, refers to the case that a subject who knows the precisely measured height of the object takes into account that this height can give rise to different estimated height values in everyday life. Source of fuzziness # 3, is due to the subject's taking into account that the labels may be assigned by different persons whose height thresholds for the three labels may differ. The labels are mutually exclusive only for a given person.

N	I $\Longrightarrow$ $Freq(u_i)$=relative frequency of outcome u_i in observed sequence of length N	II possible values of $Prob(u_i)$=underlying probability of outcome u_i
≥ 1	0	$0m$
≥ 1	1	$m1$
≥ 2	m	m

Figure 8.2: *Deductive inference of underlying $Prob(u_i)$ (column II) from observed relative frequency, $Freq(u_i)$, (column I). E.g., the first row of entries tells us that a relative frequency 0 can be due to an underlying probability 0. It can also be due to an underlying probability m, but not 1. m is an element of the open interval (0,1) (not including the two end points). See sect. 8.4 for the m-notation.* ***figupdatefp***

N_1	I $\Longrightarrow$ $Freq(u_i)$ in observed sequence of length N_1	II possible values of $Freq(u_i)$ in sequence of length $N_2 > N_1$
≥ 1	0	0
≥ 1	0	m
≥ 1	1	m
≥ 1	1	1
≥ 2	m	m

Figure 8.3: *Updating of $Freq(u_i)$, the relative frequency of the outcome u_i by prolongation of the observed sequence. Note that, depending on the additional outcomes, a value of 1 or 0* can *be updated to $m \in (0,1)$.* ***figupdateff***

N_1	I $\Longrightarrow$ maximum likelihood estimate of $Prob(u_i)$ from observed sequence of length N_1	II possible values of maximum likelihood estimate of $Prob(u_i)$ from sequence of length $N_2 > N_1$
≥ 1	0	0
≥ 1	0	m
≥ 1	1	m
≥ 1	1	1
≥ 2	m	m

Figure 8.4: *Possible updating of $Prob^{est-ml}(u_i)$, the maximum likelihood estimate of the probability of u_i, based on the observed sequence of outcomes. Note that, depending on the addtional outcomes, a value of 1 or 0* can *be updated to $m \in (0,1)$.* ***figupdateml***

N_1	I $\Longrightarrow$ possible values of underlying $Prob(u_i)$ from sequence of length N_1	II possible values of underlying $Prob(u_i)$ from sequence of length $N_2 > N_1$	N_2
$= 0$	0m1	$0m$	≥ 1
$= 0$	0m1	$m1$	≥ 1
$= 0$	0m1	m	≥ 2
≥ 1	0m	$0m$	≥ 2
≥ 1	0m	m	≥ 2
≥ 1	m1	$m1$	≥ 2
≥ 1	m1	m	≥ 2
≥ 2	m	m	≥ 3

Figure 8.5: The Basic Deductive Updating Table for a Probability, Assuming a Single Underlying Numerical Probability Value. *Updating of possible set of values of the underlying $Prob(u_i)$, based on the observation of a sequence of N_1 outcomes (column 1), and of a prolongation of this sequence (column II). The table assumes deductive reasoning. Note that we can never infer an underlying probability value of 1 or 0 from such reasoning.(See fig. 8.1 for the meaning of the different intervals in the m-notation.)* ***figupdatepunder***

To prove the correctness of the first row of the table, we start with column II. An underlying $Prob(u_i)$=0 implies that the outcome u_i can never occur in an experimental sequence. Consequently $Freq(u_i)$ will always be 0, no matter how long the observed sequence is. However, with an underlying $Prob(u_i) = m \in (0,1)$, it may happen, by chance, that u_i does not occur in the particular experimental sequence either. A 0 value for $Freq(u_i)$ *can* therefore occur also for $Prob(u_i) = m$. The only value which $Prob(u_i)$ cannot have for $Freq(u_i)$=0 is 1, because $Prob(u_i)$=1 implies that every outcome in the sequence is u_i. This contradicts the $Freq(u_i)$=0 value in column I.

The second row of the table refers to the case that all outcomes of the observed sequence are u_i. This will always happen when $Prob(u_i)$=1. For $Prob(u_i)$<1 it may also happen, by chance, that all outcomes of the observed sequence are u_i, resulting again in $Freq(u_i)$=1; unless $Prob(u_i)$=0. In this case there would be no u_i outcome in the sequence.

Finally we have the case in which $Freq(u_i) = m$ in the observed sequence. This implies, according to eq. (8.9), that the sequence must have at least one u_i outcome, and one outcome that is different from u_i. But this cannot happen when $Prob(u_i)$ is either 0 or 1. Consequently $Prob(u_i)$ can have solely the value m in this case.

The same table of entries as that of fig. 8.2 can also be used with different headings. In the first place, a given value of an underlying $Prob(u_i)$ can give rise to the same value (in the m-notation) of $Freq(u_i)$ provided that the sequence is of sufficient length. The heading of column II can therefore be replaced by 'Possible values of $Freq(u_i)$ in prolonged experimental sequence'. This results in fig. 8.3. Column II of this figure shows the possible values which $Freq(u_i)$ may have in a prolonged sequence of length $N_2 > N_1$, given the frequency of column I, as computed on the basis of an observed sequence of length N_1.

According to theorem 7.5.2, the relative frequency of u_i is also equal to the maximum likelihood estimate of the probability of u_i, $Prob^{est-ml}(u_i)$, based on the given experimental sequence. We can therefore replace the headings of columns I and II in fig. 8.3 by this estimate, resulting in fig. 8.4

Note that a maximum likelihood estimate, as well as a relative frequency of 0 or 1 can be updated to m when the sequence is prolonged. This seems to be in direct contradiction to example 8.1.2 in which probabilities of 0 or 1 could never be updated; while a probability value m could be updated to 0 or 1. The solution to this seeming paradox is given in section sect. 9.1.

Fig. 8.4 illustrates the updating of those values of an assumed $Prob(u_i)$ for which the observed sequence has the maximum probability. Our main interest in the present section is, however, focused on the set of all possible values of the underlying $Prob(u_i)$ which could have given rise to the observed sequence, not merely on the single value of an assumed underlying $Prob(u_i)$ for which the probability of the sequence is a maximum. The updating of the set of possible values of $Prob(u_i)$ is summed up in fig. 8.5.

The first three entry rows of column I of this table refer to sequences of length

N_1=0, i.e. to the state of complete ignorance before the observation of any experimental sequence. In this state the underlying probability is completely unknown, any element of the real interval [0,1] is a possible candidate for this probability. The possible updated values of the underlying probability for sequences of length $N_2 \geq 1$ are shown in the first three rows of column II. These are the values of column II in fig. 8.2.

Using the first three values of column II in fig. 8.5 as initial values of the underlying $Prob(u_i)$ results in the last five rows of the same figure. E.g., the fourth row tells us that a value of $0m$ for $Prob(u_i)$, due to a sequence with no u_i's, is unchanged when the prolonged sequence does not contain any u_i either. If an outcome u_i does appeear in the prolonged sequence, then we know that the underlying probability cannot be 0; and consequently the value $0m$ is updated to m. This possibility is shown in row 5. Considering the new values of column II as initial values to be put into column I gives us nothing new because the set of the three possible values, $0m$, $m1$, m, in the last five rows of column II is identical with the set of values in column I for the same rows.

The most significant result of fig. 8.5 is that there are no 0 or 1 entries in column II. We sum up this result in the form of a theorem.

Theorem 8.5.1 *No observation of an experimental sequence, however long, can ever result in the deduction of a certainty concerning the occurrence, or nonoccurrence of an outcome u_i. Consequently it cannot result in the deduction of a certainty distribution either. (See definition 7.6.1 concerning the definition of certainties.)*

Any statistical experiment can be used to illustrate the tables of figures 8.2-8.5. E.g., it may happen in a political poll for a coming election, with the three candidates u_1, u_2, u_3 for presidency, that none of the first N_1=10 persons interviewed has given her or his vote for candidate u_3; resulting in a value of 0 in column I of figures 8.2-8.4. This does not necessarily mean that no vote will be given for u_3 when the number of interviewed persons is increased to N_2=1000. A $Prob^{est-ml}(u_i)$ value 0 *can* be updated to a value $m \in (0,1)$.

This kind of result is so obvious that the reader may consider the updating rules of our tables to be trivial. The rules do, however, have the following three important tasks.

1. The updating tables of probabilities and relative frequencies allow us to check whether new values which are to be stored in a data base are consistent with the values which have already been stored.

2. The rules clarify the seeming paradox between the updating of an estimated probability of 1 or 0 to m, versus the updating of a probability m to 1, such as in example 8.1.2. See sect. 9.1 for this clarification.

3. The use of the updating rules in the m-notation in connection with the representation of states of complete or partial ignorance, thus avoiding the ambiguities

N	I $\Longrightarrow$ $Freq(u_i)$=relative frequency of outcome u_i in observed sequence of length N	II possible values of $Prob(u_i)$=underlying probability of outcome u_i
≥ 1	0	$0m$
$\geq ?$	0	0
≥ 1	1	$m1$
$\geq ?$	1	1
≥ 2	m	m

Figure 8.6: *Inductive inference of underlying $Prob(u_i)$ (column II) from observed relative frequency, $Freq(u_i)$, (column I). Note that in contrast to the analogous fig. 8.2 for deductive reasoning, inductive reasoning allows the inference of underlying probabilities of 1 or 0 when the relative frequencies are 1 or 0 respectively.* ***figupdatefpind***

to which Bayes postulate can give rise according to sect. 8.3. Although we will use mainly the updating rules of fig. 8.5 in this connection, it is important to understand to what quantities the updating rules of the other tables of the present subsection apply.

8.5.3 Inductive Learning from Experience

We have already mentioned inductive reasoning in sect. 6.6. Such reasoning refers to the learning from experience, just like the deductive reasoning of sect. 8.5.2. It is, however, less demanding and more naive, or trusting if you wish, than deductive reasoning in the sense that it allows to draw conclusions with complete certainty. In contrast to deductive reasoning, it thus allows the updating of probabilities to the values of 1 or 0. When a sequence of N outcomes of an experiment results in N identical values u_i, then inductive reasoning allows to draw the conclusion $Prob(u_i)$=1, assuming that any future experiment will also result in the outcome u_i; and that u_i has always occurred in any past experiment under the same set of conditions. Similarly, when an observed sequence of N outcomes does not contain u_i, then inductive reasoning allows the inference that $Prob(u_i)$=0, i.e. that u_i has never ocurred in the past, and will never occur in the future under the conditions to which the experiment refers. Any inference of a value of $m1$ or $m0$ in column II of figs. 8.2, 8.5 will thus allow an additional inference of 1 and 0 respectively for inductive reasoning. The result is shown in figs. 8.6 and 8.7. Note that although values of $0m$ and $m1$ for the possible underlying $Prob(u_i)$ *can* be inductively updated to 0 or 1, there exists no inductive updating to 0 or 1 from an initial value m.

The minimum value of N which a person requires for the inductive inference of

N_1	I $\Longrightarrow$ II possible values of underlying $Prob(u_i)$ from sequence of length N_1	possible values of underlying $Prob(u_i)$ from sequence of length $N_2 > N_1$	N_2
$= 0$	$0m1$	$0m$	≥ 1
$= 0$	$0m1$	0	$\geq ?$
$= 0$	$0m1$	$m1$	≥ 1
$= 0$	$0m1$	1	$\geq ?$
$= 0$	$0m1$	m	≥ 2
≥ 1	$0m$	$0m$	≥ 2
≥ 1	$0m$	0	$\geq ?$
≥ 1	$0m$	m	≥ 2
≥ 1	$m1$	$m1$	≥ 2
≥ 1	$m1$	1	$\geq ?$
≥ 1	$m1$	m	≥ 2
≥ 2	m	m	≥ 3

Figure 8.7: The Basic Inductive Updating Table for a Probability, Assuming a Single Underlying Numerical Probability Value. *Updating of possible set of values of the underlying $Prob(u_i)$, based on the observation of a sequence of N_1 outcomes (column 1), and of a prolongation of this sequence (column II). The table assumes inductive reasoning. Note that in contrast to the analogous fig. 8.5 for deductive reasoning, inductive reasoning allows the updating to certainty values, i.e. to underlying probabilities of 1 or 0. A value m can, however, never be updated to 1 or 0. An inductive updating to a probability of 1 or 0 on the basis of a prolonged observed sequence can never be guaranteed to be correct.* ***figupdatepunderind***

a certainty or impossibility is subjective. Persons who wish to be on the absolutely sure side will require $N{=}\infty$, i.e. they will never use any inductive reasoning, and will thus never infer a certainty from observed data. The less demanding a person is concerning the correctness of her inductive inference, the shorter is the length N of an experimental sequence which she will require for the inference of a certainty.

Consider the case of a sequence $\mathbf{x}$ with N identical outcomes u_i,

$$\mathbf{x} = \langle u_i, u_i, \ldots, u_i \rangle \ . \tag{8.13}$$

And let us denote the underlying $Prob(u_i)$ by $1{-}\epsilon$,

$$Prob(u_i) = 1 - \epsilon = m1 \in (0, 1] \ . \tag{8.14}$$

Assuming that successive outcomes are mutually independent, the probability of the sequence $\mathbf{x}$ is then

$$Prob\{\mathbf{x}|[Prob(u_i) = 1{-}\epsilon)]\} = (1{-}\epsilon)^N = 1{-}N\epsilon{+}\text{terms of higher order in } \epsilon \ . \tag{8.15}$$

It follows that unless ϵ is 0, i.e. unless $Prob(u_i){=}1$, the probability of the observed sequence $\mathbf{x}$ is extremely small if we only require a big enough N.

Similarly, suppose that $\mathbf{x}$ is a sequence of length N with no u_i outcomes. Denoting the underlying $Prob(u_i)$ by ϵ in this case,

$$Prob(u_i) = \epsilon = 0m \in [0, 1) \ , \tag{8.16}$$

we find that the probability of $\mathbf{x}$ is again

$$Prob\{\mathbf{x}|[Prob(u_i) = \epsilon)]\} = (1 - \epsilon)^N = 1 - N\epsilon + \text{terms of higher order in } \epsilon \ . \tag{8.17}$$

We conclude that although inductive reasoning can never be guaranteed to be correct, the likelihood that it is incorrect can be made as small as we wish by increasing the length of the observed sequence.

All these quantitative statements preassume that the experiment which results in the sequence $\mathbf{x}$ is based on an unbiased sample. Cases in which inductive reasoning concerning the correctness of a scientific theory fails are often due to samples which, at a later date, are considered to have been biased.

For example, Newton's theory which assumed that the mass of a body is independent of its velocity was always found to be correct as long as one did not have the technical tools for accelerating masses to velocities near the velocity of light; or for measuring masses moving with such a velocity. However, seen from a present-day point of view, the old experiments worked with biased samples of masses whose velocities were extremely small compared with the velocity of light. It is thus not the inductive reasoning which failed, but the requirement of an unbiased sample.

8.5.4 Inductive Learning of Intermediate Probability Values

In sect. 8.5.3 we considered inductive learning of impossibilities and certainties, i.e. of probabilities 0 or 1. This is the only type of inductive learning of which we make use in the present chapter 8. In chapterXXXX we will consider situations in which unique, intermediate, non-interval-valued, numerical probabilities are specified a priori to a knowledge base or to an information recipient. From our analysis in sect. 7.5.3, theorem 7.5.1, we know that unique values of probabilities can never be ascertained by deductive reasoning from an experimental sequence.

XXX

unless we use the whole object set as our sample

However, if the relative frequency values $Freq(u_i)$ computed from successive prolongations of the experimental sequence are found to be constant (i.e. up to specifications of two decimals), then we can infer *inductively* that we are certain that $Prob(u_i)$=$Freq(u_i)$. Unique specifications of intermediate probability values can therefore also be looked upon as being the result of inductive learning from an experimental sequence. Alternatively, such values can be based on a scientific theory combined with the inductive inference that this theory is correct. E.g., Boltzmann's kinetic theory of gases allows us to infer deductively the velocity distribution of molecules in a gas of a given temperature, assuming *inductively* that the theory is correct.

8.5.5 Conjunction of Set-Valued Probability Values

Consider the case that an underlying probability or certainty for an outcome or event is specified, in the m-notation, to a knowledge base at time t_1. At another point of time t_2, the underlying probability for the same event is specified anew in the m-notation.

We wish to set up rules which tell us whether the two specifications are consistent or inconsistent. If they are consistent, we wish to keep that distribution in the m-notation which leaves us with a minimum of ignorance.

Our criterion for consistency is that both distributions can have been learned inductively or deductively from two experimental sequences $\mathbf{x_1}$, $\mathbf{x_2}$ which are due to the same underlying probability distribution. If the two specified distributions are identical in the m-notation, then we store one of them. Otherwise we store that probability distribution, in the m-notation, which could have given rise to either of the sequences.

A set-value, in the m-notation, of the probability of u_i which underlies a given observed sequence $\mathbf{x}$ indicates that the sequence does not contain sufficient information to decide which of the elements of the set is the correct one. While any element of {0, m, 1} which is not contained in the set-value could not have given rise to the sequence. But $\mathbf{x_1}$ and $\mathbf{x_2}$ are due to the same underlying $Prob(u_i)$. We have therefore the following theorem.

Theorem 8.5.2 *Assume that two probability values of an event are specified, in the m-notation of fig. 8.1, for the same underlying probability distribution. And that each of the two specified probabilities is the result of inductive or deductive reasoning based on two experimental sequences of observations respectively. In the case of inductive*

0	m	1	0 ORE m	m ORE 1	(0 ORE m) AND (m ORE 1)
t	f	f	t	f	f
f	t	f	t	t	t
f	f	t	f	t	f

Figure 8.8: Conjunction of two set-valued probability values (defined by the headings of columns 4 and 5), using a truth table of traditional propositional calculus. The first three columns indicate which of the three possible underlying $Prob(u_i)$ values is the true one. Only the second row has the truth value t in the last column. Going backwards in this row, we infer that the underlying probability has the value m (because the m or second column is that column of the first three ones which has the truth value t in the second row). This complicated inference is equivalent to the simple intersection rule of theorem 8.5.2. ***figupdateconj***

reasoning, it is assumed that the same threshold is used in both experiments for the length N of the experimental sequence which allows the inference of a probability value 0. An analogous statement holds for the inference of a probability value 1. Each probability value is thus a subset of the universe $\{0, m, 1\}$.

The probability value to be stored in the knowledge base is then the intersection of the two specified (generally set-valued) probability values.

For example, if one specified value is $0m1$ and the other $m1$, then we store $m1$. If one is m and the other $m1$, then we store m. If one is 1, and the other $m1$, then we store 1. And if one is $0m$ and the other $m1$, then we store m. If the intersection is empty, then the two specifications are contradictory and cannot be accepted simultaneously by the knowledge base. This occurs, for example, for the two specifications m and 1, or m and 0, or 0 and $m1$. We see that the specification to the knowledge base of two different unique probability values (two different elements of $\{0, m, 1\}$) is inconsistent as long as they refer to a single underlying probability distribution.

The example below shows that the intersection rule of theorem 8.5.2 can be replaced by the following procedure which makes use of the truth table of traditional propositional calculus for the conjunction. However, the truth table procedure is considerably more complicated than the intersection procedure of theorem 8.5.2.

Example 8.5.3 *Suppose that two experimental sequences have resulted in the two interval-values $0m$ and $m1$ respectively for the underlying probability. These experimental results can be expressed by the sentence,*

$$[Prob(u_i) \text{ is } (0 \text{ ORE } m)] \quad AND \quad [Prob(u_i) \text{ is } (m \text{ ORE } 1)] \tag{8.18}$$

where ORE stands for the exclusive OR. Eq. (8.18) expresses a conjunction between two interval-valued probabilities for the same event. Fig. 8.8 shows the truth table for this conjunction. This table consists of three rows only because the underlying

$Prob(u_i)$ has always one, and only one of the three values 0, m 1 (see first three columns). If one of the three values is true, then the other two must be false. The last column of the table lists the truth value of (8.18). We see that this column has only one t value, namely in the second row. This is the row in which m (second column) has the truth value t. The requirement of the truth of the conjunction of eq. (8.18) is thus equivalent to the intersection rule stated in theorem 8.5.2. An analogous procedure holds also for any other combination of possible interval-valued probabilities.

8.6 M-Values for Specified Natural Language Quantifiers

8.6.1 Introduction

That part of mathematical logic which deals with the meaning of sentences containing one or more of the three *quantifiers* 'every' (or 'all'), 'some', 'no' is considered to be a part of *predicate calculus*. It treats the problem of the two and a half millenia old Aristotelian syllogisms which we defined in chapter 2, eq. (2.1).

In chapterXXXX we come back to a general treatment of classification and quan- *XXXX*
tification problems with the aid of conditional probability tables. We believe that this treatment, which erases the sharp division between propositional and predicate calculus, is simpler and, in some cases, also more reliable, than the treatment of traditional predicate calculus. It makes use of the natural-language meaning of the IF THEN connective, as well as of probabilities and their values and updating rules in the m-notation. Here we will merely illustrate the use of the m-notation and the updating rules of sect. 8.5 in a few simple cases.

In the examples of the next subsections we use the terminology that the knowledge base is supplied with certain items of information such as 'Every B is an A'. By this we mean either that the informant supplies this information directly to the knowledge base when it is running in the 'information supply mode'. Or that the corresponding question 'Is every B an A?' is directed by the knowledge base at the informant, and that the informant answers 'yes'. (See sect. 8.9 for more details.)

We have already emphasized that communication between an informant and a human information recipient or a data or knowledge base is usually based on the assumption that the information supplied by the informant is correct; except in so far as the system checks whether the newly supplied information is consistent with the information already stored in the knowledge base. One of the purposes of sect. 8.5 was to set up rules for testing such consistency.

The knowledge base is not concerned with the way in which the information concerning probability values was obtained by the informant; whether it was obtained by deductive or inductive reasoning from the observation of an experimental sequence, or by an inductive reliance on the correctness of a theory; or whether it was based, e.g., on a promise given by or to the informant such as 'I will be at home tomorrow'; or on meaning-related (analytic) information such as 'All dogs are animals', or 'Some animals are dogs'. Natural language has an almost infinite variety of linguistic tools for expressing different types of certain and uncertain information. The probability

modifiers of sect. 7.7.2 are one class of such tools. They are used to specify numerical or fuzzy probability values which lie within the m=(0,1) interval.

In the present section we illustrate that there is a very close correspondence between the quantifiers of natural language and the m-valued probability values of fig. 8.1. We shall also see how the intersection theorem 8.5.2 works for the purpose of consistency-checking and updating of quantified information. In addition we show in sect. 8.6.4 that it is advantageous to include the value 01=$\{0,1\}$='either 0 or 1' along with the other six set-valued probabilities in fig. 8.1. The updating results concerning a single underlying probability distribution are summarized in sect. 8.8.

To make the updating situation, as well as the use of terms such as 'supplied information', 'knowledge base' etc. more tangible we present in sect. 8.9 several box diagrams for updating or learning situations. These refer both to updating of knowledge base systems in computers and to the situation encountered by a human child when it learns the meaning of concepts.

8.6.2 Every and No

We give here two examples which illustrate the use of the natural language quantifiers 'every' and 'no'. These quantifiers correspond to the specification of the conditional probabilities 1 and 0 respectively. The 'some' and 'not every' quantifiers of sect. 8.6.3 correspond to the set-valued probabilities $m1$ and $0m$. And the 'some but not every' conjunction of these two quantifiers corresponds to the specification of a probability value m.

Example 8.6.1 *A knowledge base is supplied with the information*

$$\textit{Every } B \textit{ is an } A, \qquad \textit{e.g.,} \qquad \textit{Every dog is an animal.} \tag{8.19}$$

The 'universal quantification' sentence (8.19) is an abbreviation for

$$\textit{IF } x \textit{ is an instance of a } B \textit{ THEN } x \text{ is an } \textit{instance of an } A. \tag{8.20}$$

Sentence (8.20) states that the conditional probability that x is an instance of A, given that x is an instance of B, is equal to 1,[2, 3]

$$P(A|B) = 1 \qquad \text{or} \qquad B \subseteq A\ . \tag{8.21}$$

[2] A conditional probability is not different from any other probability except that one of the conditions on which the probability is premised is explicitly stated in its symbolic notation. Conditional probabilities are defined in chapter 10. From here on we use the symbol P for 'probability'.

[3] To avoid the proliferation of symbols, we use the symbols A, B etc. in more than one sense, depending on the context. All the meanings are, however, closely related. The *class* A is is a collective name for all objects with certain prescribed attribute values. The *intension* of the class A is the collection of these atribute values. The *set* A, or the *extension* of A, is the collection of all instances belonging to the class A. In (8.19), A is an abbreviation for 'instance of the class A'. In the conditional probability equation of (8.21), A is an abbreviation for 'x is an instance of the class A'. And in the subset equation of (8.21), A stands for the extension of the class A, i.e. for the set of all objects with the attribute values belonging to the intension of A. Similarly for B.

The Venn diagram corresponding to eqs. (8.19)-(8.21) is shown in fig. 8.9(a). It illustrates that the set of all instances of B is a subset of the set of all instances of A.

The opposite case of two disjoint classes is expressed by the quantifier 'no' in English. It is illustrated by the following example and fig. 8.9(c).

Example 8.6.2 *The knowledge base is supplied with the information*

$$\begin{array}{l} \textit{No } B \textit{ is an } A, \\ \textit{No dog is a cat.} \end{array} \qquad (8.22)$$

e.g.,

(8.22) is an abbreviation for

IF x is an instance of a B THEN x is NOT an *instance of an A.* (8.23)

In probabilistic and set notation, this sentence is represented by

$$P(A|B) = 0 \qquad \text{or} \qquad A \cap B = \emptyset \ . \qquad (8.24)$$

Examples 8.6.1, 8.6.2 make use of the natural language quantifiers *every (all)* and *no*. They are typical descriptions by a teacher or lexicon or expert to convey the relationship between the meanings of two classes A and B to a pupil or knowledge base. In the terminology of sect. 4.2 they convey a meaning-related, analytic truth. Both of these quantifiers convey *certain* information which can, consequently, be represented in the form of a probability value of either 1 or 0. In sect. 8.5.3 we saw that deductive reasoning from the observation of an experimental sequence can never result in the inference of one of these 'certainty values'. This will be illustrated later on in example 8.6.4.

8.6.3 'Some' and 'Not Every'

The natural language quantifier *some* can also be used to convey meaning-related information. As we shall see below, *some* and *not every* are, from a classificational point of view, more ambiguous than *every* and *no*.

The reason for this ambiguity is that the original purpose of *some* and *not every* is to convey the results of a learning situation of the type desribed in sect. 8.5, based on a sequence of experimental observations. However, the frequencies in the observed experimental sample may not be representative for the underlying probability distribution. This may give rise to the set-valued probabilities $m1$ and $0m$ of fig. 8.5. Each of these leaves open two possibilities between which we cannot distinguish on the basis of the observed sample sequence. *Some* describes correctly those cases in which the given experimental sequence lets us deduce only the probability value $m1$. This leaves open all probabilities which are not 0. While *not every* desribes those sequences from which we can deduce solely the probability value *0m.*, leaving open all probabilities which are not 1. In such cases it may happen that the correct classificational situation cannot be identified on the basis of the observed sequence.

The ambiguity of *some* is thus correct and necessary for a truthful description of the available information. We then need additional information for the identification of the correct classificational relationship between A and B.

In section 8.5.2 we did not refer to any observational sequence in connection with the quantifiers *every* and *no*. Instead we assumed that an informant supplies the knowledge base with information of the type of sentences (8.19) and (8.22). In the case of the identification of the elements of an experimental sequence, the task of the informant is much more limited. She does not give an all-embracing definition which is generally valid for the relationship between two classes, making use of quantifiers, negations and possibly connectives. All she does is to identify each object with which she is presented as to its belonging or not-belonging to the class A, and similarly for B. She leaves it to the experimenter or learner to gradually identify the attribute values characterizing the two classes by prolonging the sequence of observations.

The *some* quantifier deals with probabilities that are neither 0 nor necessarily 1. We will always understand its meaning to be 'one or more'.

Example 8.6.3 below and fig. 8.9 illustrate the classificational ambiguity of a single statement with a *some* quantifier. The application of sentences with quantifiers to describe a sequence of experimental observations is illustrated in example 8.6.4.

Example 8.6.3 *The knowledge base is supplied with the information*

$$\text{info1:} \qquad \textit{Some } B\textit{'s are } A\textit{'s,} \tag{8.25}$$

e.g.,

$$\textit{Some animals are dogs,} \tag{8.26}$$

or

$$\textit{Some sea animals are mammals.} \tag{8.27}$$

(The singular form of these sentences should also be included in the example, such as 'Some B is an A'.)

(8.25) is an abbreviation for

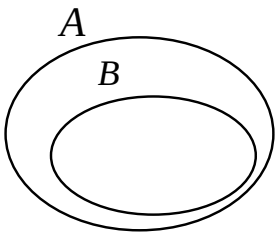
$$\textit{IF } x \textit{ is an instance of a } B \textit{ THEN } x \text{ may be an } \textit{instance of an } A. \tag{8.28}$$

These 'existential quantification sentences' give rise to a conditional probability that is bigger than 0,

$$P(A|B) > 0\ , \qquad \text{or, in the m-notation,} \qquad P(A|B) = m1\ . \tag{8.29}$$

The ambiguity of *some* follows from the fact that, surprisingly, Venn diagram (c) of fig. 8.9 is the only one for which no experimental sequence can ever fit eq. (8.25) or its equivalents (8.28), (8.29). This statement will be clarified in example 8.6.4.

We have thus proved our contention that the *some* quantifier is more a natural language tool for reporting the result of a learning experiment from the observation of an experimental sequence than a tool for the specification of one of the five classification situations in fig. 8.9. The specification $P(A|B) = m1$ or, equivalently, the

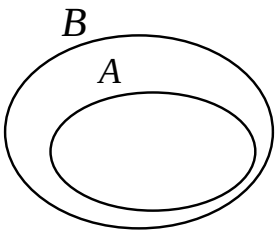


(a)

Every B is an A

IF B THEN A

$B \subseteq A \quad P(A|B)=1$

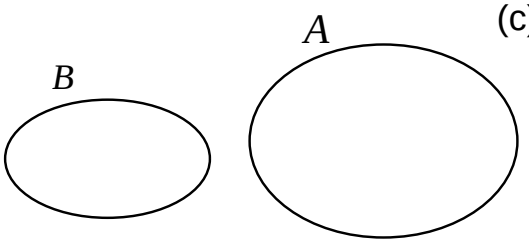


(b)

Every A is a B

IF A THEN B

$A \subseteq B \quad P(B|A)=1$

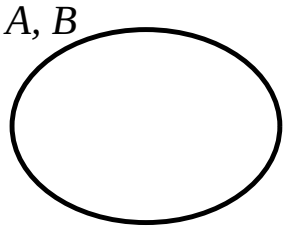


(c)

No B is an A

IF B THEN NOT A

$B \cap A = \emptyset \quad P(A|B)=0$

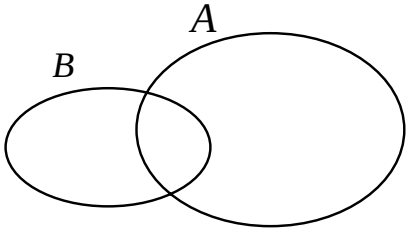


(d)

A is equivalent to B

IF B THEN A AND IF A THEN B

$A \equiv B \quad [P(A|B)=1 \text{ AND } P(B|A)=1]$



(e)

Some B's are A's AND Some A's are B's

IF B THEN maybe A AND IF A THEN maybe B

$[B \not\subseteq A \text{ AND } A \not\subseteq B \text{ AND } B \cap A \neq \emptyset] \qquad [P(A|B)=m1 \text{ AND } P(B|A)=m1]$

Figure 8.9: *Different descriptions of five basic quantification structures. For each of the figures (a)-(e) there are given the following five equivalent desriptions: 1) Venn diagram. 2) A quantification sentence. 3) An IF THEN sentence. 4) A set relationship. 5) One or more conditional probability values in the m-notation. Note that there exist many possibilities of observed, usually short, sequences which could be due to two or more of the diagrams (a)-(e). E.g., the first three rows of fig. 8.10 show a sequence of three outcomes in the universe* $\{1A, 0A\} \times \{1B, 0B\}$ *due to diagram (e). The same sequence could also be due to diagrams (a) or (b) or (d). To determine to which of the four situations the sequence is due, we must prolong the observed sequence. Even then we may have to use inductive reasoning to determine which diagram is the correct one. The equivalence diagram (d) is a limiting case of both (a) and (b) and (e).* $A \equiv B$ *means that the* concepts *A and B are identical. The* labels *may, however, be different. An example is: Label of* A=woman, *Label of* B=feminine, adult human. ***figupdateevery***

sentence 'Some B's are A's' is ambiguous as far as a distinction between diagrams (a), (b), (d), (e) of fig. 8.9 is concerned unless we are supplied with additional information.

Let us start with the additional information

$$\text{info2:} \qquad \textit{Not every } B \textit{ is an } A \ , \tag{8.30}$$

and consequently $P(A|B)$ can never have the value 1,

$$P(A|B) \neq 1 \qquad \text{or} \qquad P(A|B) = 0m \ . \tag{8.31}$$

It follows that

$$A \not\equiv B \ . \tag{8.32}$$

Info2 thus excludes diagram (d) and, as a consequence, also (a) as possible situations described by the conjunction of *info1* and *info2*.

According to the intersection theorem 8.5.2 we have from (8.31) and (8.29) that

$$P(A|B) = m1 \cap 0m = m \ . \tag{8.33}$$

We see that the probability value m corresponds to the conjunction of the quantifier 'some' with the quantifier 'not all', i.e. to the expression 'some but not all'.

After receiving *info1* and *info2*, we are still left with two possible classification situations, namely those of diagrams (b) and (e) in fig. 8.9. To distinguish between these two, the knowledge base directs the questions *'Are some A's B's?'* and *'Is every A a B'* at the informant. Suppose that the informant gives the answer 'yes' to the first question, and the answer 'no' to the second one. We then have the following two new information supply sentences which are analogous to the previous *info1* and *info2* with the roles of A and B reversed,

$$\text{info3:} \qquad \textit{Some } A\textit{'s are } B\textit{'s} \ , \tag{8.34}$$

and

$$\text{info4:} \qquad \textit{Not every } A \textit{ is a } B \ . \tag{8.35}$$

and deduce, in analogy to eq. (8.33),

$$P(B|A) = m \ . \tag{8.36}$$

However, in diagram (b) of fig. 8.9 we have that

$$P(B|A) = 1 \ . \tag{8.37}$$

From (8.36) and (8.37) we find, according to the intersection theorem, that

$$P(B|A) = m \cap 1 = \emptyset \ . \tag{8.38}$$

The intersection of the set-valued probability m with 1 is empty, implying that the conjunction of *info3* and *info4* contradicts fig. 8.9(b). However, for fig. 8.9(e) we have

that $P(B|A)$ is equal to $m1$. Intersecting this value with the value m of eq. (8.36) we find $P(B|A) = m$. Consequently [*info3* AND *info4*] do not contradict the previous information when we choose the Venn diagram (e). Thus the information in the conjunction of *info1* with [*info2* AND *info3* AND *info4*] removes the ambiguity in the *some* sentence of *info1*, eq. (8.25), leaving us unambigously with fig. 8.9(e).

We notice that the four successive items of information in example 8.6.3 have been learned from four experimental sequences referring to a single underlying probability distribution in the 2-dimensional universe $\{1A, 0A\} \times \{1B, 0B\}$.[4] This is illustrated by the following example and fig. 8.10.

Example 8.6.4 *Our object set OB is a set of animals. The experimental sequence consists of consecutive random selections of an object from this set, and the assignment by an expert in biology (who is now our informant) of one label from the pair* $\{1A, 0A\}$*, and one from the pair* $\{1B, 0B\}$ *to the selected object. The sequence of objects and their labels is our random sample. An illustration which fits the particular sequence of outcomes in the present example is* $1A$*=mammal,* $0A$*=non mammal,* $1B$*=sea-animal,* $0B$*=non sea-animal.*

The experimenter now collects all those objects in the sample to which the label $1B$ *has been assigned, and observes that one or more of these have been assigned the label* $1A$*. This information is conveyed to the knowledge base in the form of* info1*, eq. (8.25). E.g., the sample may contain in all three objects which are labeled* $1B$*, and all these three turn out to have been labeled* $1A$ *also (see objects 1, 2, 3 in fig. 8.10). Deductive reasoning then results in the value* $m1$ *for* $P(1A|1B)$*, and the storage of* info1 *in the form of eq. (8.29) in the knowledge base.*

In another random sample from the same object set, or in a prolongation of the first one, the experimenter finds in all three $1B$ *objects. All three have been assigned the label* $0A$ *by the expert. This information is conveyed to the knowledge base in the form of 'Some B's are non-A's, or of* info2*, eq. (8.30), resulting in the value* $0m$ *for* $P(1A|1B)$*, see objects 4, 5, 6 in fig. 8.10. Together the six objects yield the result* $P(1A|1B) = m$ *(bottom of last column of fig. 8.10), which leaves us with the two possibilities (b) and (e) of fig. 8.9.*

In order to distinguish between these two possibilities, the experimenter now performs a third experiment to elicit the value of $P(1B|1A)$*. In this experiment she selects from the random sample all the objects to which the label* $1A$ *has been assigned and finds that some of these have been assigned the label* $0B$*, and some* $1B$*, see objects 7-9 in fig. 8.10. The third experiment yields therefore* info3 *and* info4 *of example 8.6.3, or* $P(1B|1A) = m$*. This excludes the diagram (b) for which* $P(1B|1A) = 1$*. Our end result is thus that diagram (e) holds for the conjunction of* info1, info2, info3 *and* info4*.*

[4] $1A$ and $0A$ stand for the occurrence and non-occurrence respectively of an instance of A, and similarly for B. In this notation a probability such as $P(A|B)$ should be replaced by $P(1A|1B)$. See sections 9.2, 9.3 concerning this notation.

	Conditioning attribute value	Value of the other attribute	Deductive inference from exp. # i	Deductive inference from exps. # (0-i)
		Experiment # (i=0)		
no object			$P(1A\|1B) = 0m1$	$P(1A\|1B) = 0m1$
		Experiment # (i=1)		
object 1	$1B$	$1A$	$P(1A\|1B) > 0$	$P(1A\|1B)$
object 2	$1B$	$1A$	or in m-notation	$= 0m1 \cap m1$
object 3	$1B$	$1A$	$P(1A\|1B) = m1$	$= m1$
		Experiment # (i=2)		
object 4	$1B$	$0A$	$P(1A\|1B) < 1$	P(1A|1B)
object 5	$1B$	$0A$	or in m-notation	$= m1 \cap 0m$
object 6	$1B$	$0A$	$P(1A\|1B) = 0m$	$= m$
		Experiment # (i=3)		
object 7	$1A$	$0B$	$0 < P(1B\|1A) < 1$	
object 8	$1A$	$0B$	or in m-notation	
object 9	$1A$	$1B$	$P(1B\|1A) = m$	

Figure 8.10: *Narrowing down of the five possible classification situations of fig. 8.9 to the partial overlap situation of fig. 8.9(e) by deductive updating of conditional probability values in the m-notation (i.e. by narrowing down the set-valued conditional probability values to a single value). The successively updated values are here deduced from observed relative frequencies, see example 8.6.4.* ***figupdateexp***

8.6.4 The Probability Values 01 and Ø

We shall show here that there exist certain simple natural-language descriptions of situations which give rise to the set-valued probability value 01={0,1}='either 0 or 1'. Verbal information of this sort is so common, and so useful, that it seems reasonable to include its formal equivalent in the probabilistic description. Such an inclusion is also satisfying from an aesthetic point of view. The set of all possible probability values in the m-notation is then the set of all subsets of the complete ignorance value 0m1={0,m,1} which leaves open all the three possible unique values. 01 has therefore been included in the bottom line of fig. 8.1.

The probability value 01 is somewhat special in the sense that it must always be specified to the knowledge base, it is a value that will never be inferred solely on the basis of an experimental sequence. Without additional specified information of the type of (8.39) or (8.41) below, one can infer inductively either the probability value 1 (when the sequence consists only of affirmations of the outcome) or the value 0 (when the sequence consists only of negations of the outcome). When it consists of both affirmations and negations, then the value m, not the value 01, is inferred both deductively and inductively. The value 01 is due to a state of knowledge in which we are informed that the outcome of an (N-dimensional) experiment will always be either a sequence of N affirmations, or a sequence of N negations. Which of these two sequences is going to occur is, however, unknown.

The following two examples demonstrate the specification of 01 probability values in natural language.

Example 8.6.5 *Consider the sentence,*

> *In this semester Margy has a consultation hour either on every Tuesday or on every Thursday (I am not sure which).* (8.39)

The object set pertaining to this sentence consists of all the weeks of the semester, and the attribute value or outcome of each week object is the day of that week on which Margy has her consultation hour, namely either Tuesday or Thursday. The sentence (8.39) tells us that either P(Tuesday)=1 and P(Thursday)=0, or vice versa. Consequently we can assign the following set-valued probability value to Tuesday and Thursday,

$$P(\text{Tuesday})=01\ , \qquad P(\text{Thursday})=01\ . \tag{8.40}$$

To determine which of the two values 0, 1 is the correct one for Tuesday and for Thursday we need observe the outcome of only a single week. Once we have observed that Margy's consultation hour is on Tuesday of that week, then we know that the probability of Tuesday cannot be 0, and is therefore m1. From the intersection theorem 8.5.2 it then follows that the probability of Tuesday is the intersection of the two set valued probabilities 01 and m1, and therefore P(Tuesday)=1. From the exclusive OR sentence of (8.39) we then infer that P(Thursday)=0.

Another example of a statement describing a 01 probability is,

Example 8.6.6

I put all the apples
either into the upper drawer or into the lower drawer (8.41)
(I cannot remember which).

Considering the apples as our object set, and assuming that they are numbered, we can now talk about the probability that a randomly numbered apple is in the upper drawer, and the probability that it is in the lower one. The sum of these two probabilities must be 1. Again we have

$$P(\textit{upper drawer}) = 01, \qquad P(\textit{lower drawer}) = 01\ . \tag{8.42}$$

Suppose that a radio transmitter emitting 'beep' signals is attached to apple #10, and that the beeps let us locate apple #10 in the lower drawer. From this information concerning the location of a single apple we find, using the method of sect. 8.5.2 and fig. 8.2, that the probability that a randomly numbered apple is in the lower drawer is equal to $m1$*, and the probability that it is in the upper drawer is equal to* $0m$*. Intersecting these values with the 01 values of eq. (8.42), we obtain* $P(\textit{lower drawer}) = 1$*, and* $P(\textit{upper drawer}) = 0$ *for the location of a randomly numbered apple. Information concerning the location of a single apple allows us to deduce the location of each of the other apples on the basis of sentence (8.41).*

In addition to the probability value 01 we can, if we wish, include the empty set in fig. 8.1. *Prob(event)*$=\emptyset$ tells us that we have inconsistent information concerning the probability of the event. In example 8.6.3, eq. (8.38), we had an illustration of a case in which the previous information left open two possible classification situations. However, one of these, namely that of fig. 8.9(b), resulted in an empty m-value for its set-valued probability $P(B|A)$. The other possible classification situation, that of fig. 8.9(e), resulted in the nonempty value m for $P(B|A)$, and was thus deduced as the correct one according to the four information supply items.

8.7 Updating by Exactly Specified Probabilities

In sect. 8.6.2 we already considered two cases of exactly specified probability values, namely the two certainty values 1 and 0 induced by the quantifiers 'every' and 'no' respectively. In this case the probability values 1 and 0 are due to meaning-related or analytical truth (see sect. 4.2). In sect. 7.6 we gave examples of exactly known probability values due to factual (synthetic) truth. These can be certainties as well as exact numerical probability values between 0 and 1; e.g., the probability 1/52 for drawing a given card from a complete pack of 52 cards.

Such precisely known probabilities can also be specified to a data base. Since their values are exact, they always override interval-valued probabilities learned from an experimental sequence, and referring to the same object set.

The intersection theorem 8.5.2 for the specification of two probability values of the same event (referring to the same reference object set) holds also in the case when one value has been learned from an experimental sequence, and the other is a directly specified precise value. For the purpose of checking the consistency of two such probability values, one can replace a precisely specified value lying in the interval m=(0,1) by m. E.g. suppose that the precisely specified value is 0.8. This value is replaced by m for the purpose of consistency checking. If the experimentally found value also contains an m, then the two items are consistent. The finally stored value is then 0.8 because this has a bigger information content than m. If the experimentally found value is an inductively inferred 0 or 1, then it is inconsistent with the precisely specified value in the interval $m = (0, 1)$.

8.8 Summary of Updating Situations Concerning a Single Underlying Probability Distribution

In sections 8.5-8.7 we treated the updating of probabilistic information concerning a single underlying probability distribution. One or more experimental sequences of observations was considered to be the fundamental evidence upon which such updating is based in sect. 8.5.

In sect. 8.6 we showed that the result of the analysis of such sequences can be expressed by sentences in natural language containing quantifiers or, equivalently, by probability values in the m-notation, which are presented to the knowledge base.

Uncertainty means that the outcome of a single experiment, or of a sequence of single experiments, cannot be predicted with certainty. Vice versa, when the outcomes of an observed sequence are known, then elementary probability theory tells us which probability distributions could have given rise to this sequence, and which could not. E.g., if a coin with an unknown degree of bias is thrown five times, and the sequence consists of five tails, then one possible probability distribution which may have given rise to this sequence is a *certainty distribution* in which only tails can occur due to extreme bias of the coin. However, when we wish to find the set of probability distributions which could have given rise to this sequence, we must leave open the possibility that a head may occur in future throws of the coin. Once we have observed at least one head and one tail in a sequence of throws of this coin, we say that the probability of occurrence of both head and tail is $m \in (0, 1)$. This value can never be updated to 1 or 0. The estimation of *numerical* probability values on the basis of experimentally observed sequences is treated in the field of statistics.

in the file follows a comparison between probability and 2-valued logic

Although the m-notation combines all probability values in the interval (0,1) to a single symbol m, the consistency checking and updating rules of this chapter are based on considerations which follow from the traditional, elementary theory of probability. They are presented in the tables of figures 8.5 and 8.7 and in theorem 8.5.2.

There exist cases in which probability values can be specified directly to the knowledge base instead of having to estimate them from an experimental sequence. Sections 7.6 and 8.6, 8.7 treat such cases. The updating rules of figures 8.5, 8.7 and theorem 8.5.2 are required to hold also for such directly specified probabilities, as well as to mixed updating of experimentally estimated probabilities by specified ones and vice versa; provided that all probability values refer to the same underlying probability distribution or object set.

Directly specified probabilities can be presented to the knowledge base in the form of natural language sentences, or in the equivalent form of probability values in the m-notation. In section 8.6 we have demonstrated that natural languages have excellent tools for describing situations of certainty versus situations of uncertainty. In sect. 8.7 we have shown that the use of the m-notation does not preclude the use of numerical probability values from the interval (0,1) in cases when these are known.

In addition to the probability values 0, m, 1 there exist also probability values which are denoted by two or three elements of the set $\{0, m, 1\}$. These are said to be set-valued. They imply complete or partial ignorance as to whether the given outcome is certain to occur, will occur in a fraction of all cases, or will never occur. The state of complete ignorance concerning a probability value is denoted by $0m1$. Probabilities having the set-values $0m$, $m1$ and 01 imply partial ignorance. They can be directly specified by the natural language quantification sentences of sections 8.6.3, 8.6.4. The values $0m$ and $m1$ can also be deduced from experimental sequences according to the method of fig. 8.2. Additional information is needed to narrow down these set-valued probabilities to one of the 'unique' probability values 0, 1 or m. The value m can always be inferred on the basis of deductive reasoning from an experimental sequence. The values 0 and 1 can be inferred from an experimental sequence only on the basis of inductive reasoning. Alternatively these two probability values, as well as the value m, can be part of the definition of the concepts A and B in a quantification sentence such as 'Every A is a B'.

Once a probability value is supplied to the knowledge base, it is assumed to be correct. This holds both in the case when the knowledge base is presented by the experimental sequence itself and computes the probability values on the basis of this sequence; and in the case when the probability values are specified directly to the knowledge base, either in the form of natural language sentences with quantifiers, or in the form of probability values in the m-notation. The three unique probability values 0, m, 1 can never be updated. The assignment of two different elements from the set $\{0, m, 1\}$ to a probability value is therefore logically contradictory and cannot be accepted by the knowledge base. These statements hold only as long as we refer to a unique underlying probability distribution or object set. In sect. 9.1 we show that a probability value m of a single instance of an experiment *can* be updated to 1 or 0 when the *object set* is narrowed down. More generally, the narrowing down of
XXXX an object set is described by the conditioning of probabilities (see XXXX).

Classification structures are described in natural language by the quantification sentences of sections 8.6.2, 8.6.3. To define a classification structure completely, we must know which of the five diagrams of fig. 8.9 holds for every possible pair of labels.

Diagrams (a), (b) and (c) are the neatest ones and need only a single quantifier, or a single set equation, or a single probabilistic equation to define them. No single one of the diagrams (a)-(d) can be inferred deductively from an experimental sequence. They can, however, follow from the definitions of the concepts A and B. In this case they represent meaning related or analytic truth as defined in sect. 4.2.

The *some* and *not every* quantifiers are originally tools for describing the probability values deduced from an experimental sequence. These quantifiers summarize succintly the deductive information which can be gleaned from such a sequence. They correspond to the probability values $m1$ and $0m$, and can thus be updated by additional information to m or 1, and to 0 or m respectively.

There exists no single natural language quantifier to describe the partial overlap situation of diagram (e) in fig. 8.9. This situation can be described only by a conjunction of quantifier sentences or, alternatively, by the conjunction of the two probability statements $[P(B|A) = m$ AND $P(A|B) = m]$. The equality situation of diagram (d) is represented by $[P(B|A) = 1$ AND $P(A|B) = 1]$, or by the corresponding two *every* sentences.

Because quantifier sentences can be expressed by conditional probability values, they can also be expressed by IF THEN sentences in natural language. E.g., 'Every A is a B' can be expressed equivalently by 'If x is an instance of A then x is an instance of B'. The probability logic presented in this book does not make use of the implication of mathematical logic. Instead it makes use of the natural language meaning of IF THEN as a specification of a conditional probability value. As a result, propositional and predicate calculus are no longer two separate fields in the probability logic.

8.9 A Model for a Learning Situation

In sect. 8.6 we used terms such as 'informant', 'supplied information', 'knowledge base', 'expert', 'experimenter'. To make these terms somewhat more tangible we will outline here a possible top-level architecture for a man-machine computer system which localizes the function of the informant, the knowledge base etc. to different parts of the system. Figures 8.11(a), 8.12(a) show box diagrams for such a system. Analogous diagrams for the learning situation of a child are shown in figures 8.11(b), 8.12(b).

The box diagram of fig. 8.11(a) portraits a possible overall architecture of of an interactive computer system for the representation of classification knowledge. The diagram is based largely on experience with the Alex computer system developed at the Institute of Informatics, University of Oslo. The system is described in more detail in part **??**. It consists of, 1) Procedures and 2) A lexicon in which the information supplied by the informant is stored.

The Alex system can run in two main modes. When it runs in the question mode, the user Alex (i.e. the 'man' of the man machine system) writes questions on the keyboard. These questions are automatically processed by the knowledge base

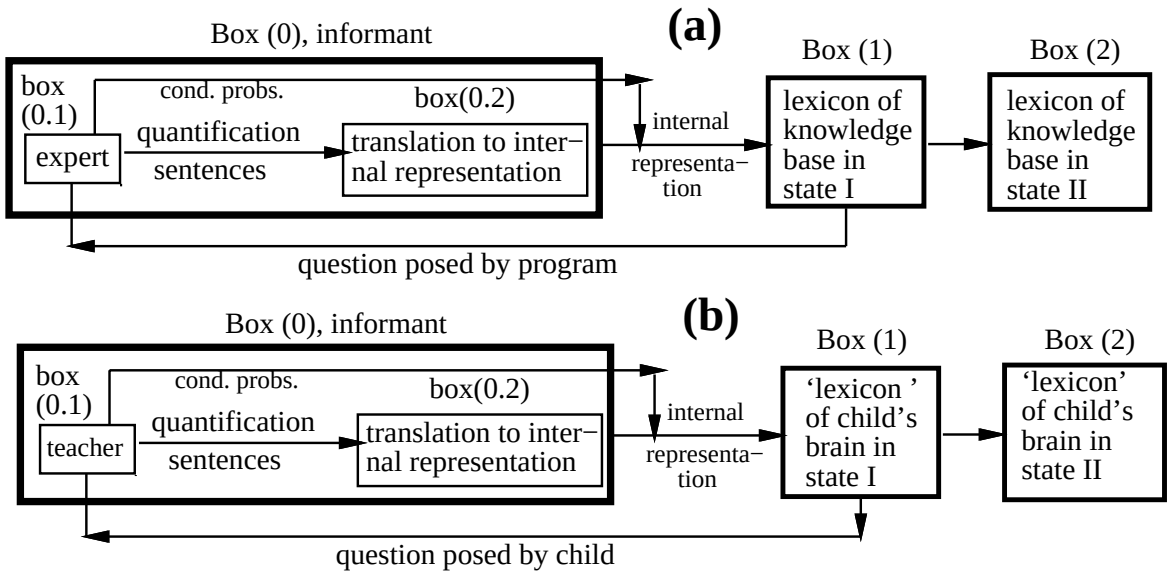


Figure 8.11: *Box diagram for the learning of a classification structure from information specified in the form of quantification sentences, or directly in the form of conditional probabilities in the m-notation. (a) Learning with the aid of an interactive computer system. (b) Analogous learning situation for a child.* ***figupdatebox1***

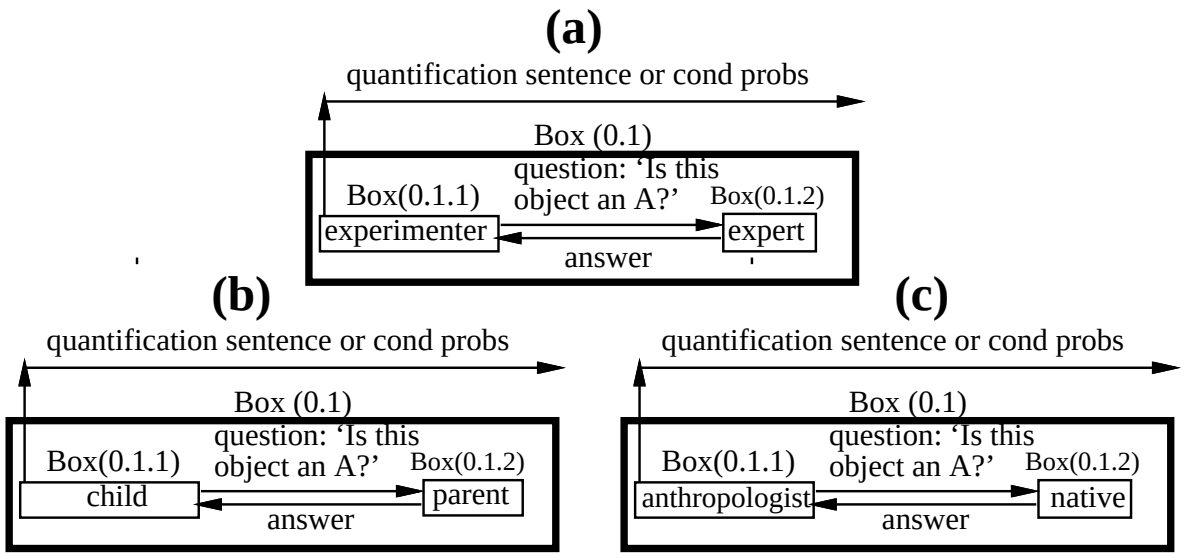


Figure 8.12: *Box diagrams for learning from an experimental sequence. Each of the three diagrams shows a magnification of box (0.1), fig. 8.11 in the case when the expert merely assigns labels to objects with which she is presented in an experimental sequence. From this sequence and the attached labels the experimenter then infers the correct quantification sentences, or directly the correct conditional probabilities in the m-notation, see example 8.6.4. (a) refers to an artificial, experimental learning situation, (b) to the learning situation of a child, and (c) to that of an anthropologist trying to learn the meaning of classification words in a native language. The output of all three diagrams goes to box (0.2) in fig. 8.11.* ***figupdatebox2***

and answered on the basis of the information stored in its lexicon. The answers are written on the screen.

The second mode in which the system can run is the information supply mode in which the user Alex supplies information to the knowledge base. As a result, the procedures of the knowledge base update the information stored in its lexicon unless the supplied information is already contained there; or unless the new information is inconsistent with the information which is already stored. Figures 8.11, 8.12 here assume that the system runs in the information supply mode.

Seen from the point of view of the computer system, the informant who supplies information to the system is represented by box (0) of fig. 8.11(a). This box consists of the two subboxes (0.1), (0.2). Box (0.1) represents the man Alex of the man-machine system.

When the system runs in the information supply mode, Alex is the expert who supplies information to the system. In the Oslo Alex system the information supplied by Alex has the form of quantification sentences in natural language English. These sentences are translated by box (0.2) into an internal representation which is passed on to box (1). The translation box (0.2) is a part of the program belonging to the Alex computer system.

In the present Alex system, which was started several years ago, the internal representation which is passed on to box (1) consists of terms (such as 'dog' and 'animal' or 'textbook in mathematics' and 'textbook') and of 'is a' and 'may be a' pointers between these lexicon terms (see chapter XXXX). The conditional probabilities and *XXXX* the m-notation of the present chapter suggest an internal representation in the form of entries in a conditional probability table (see chapterXXXX). In fig. 8.11 we also *XXXX* leave open the possibility that the expert of box (0.1) supplies information directly in the form of its internal representation. In this case box (0.2) is skipped over and the information goes from box (0.1) to box (1) without the intermediate station of box (0.2).

Let us assume that the sentence

Every dog is an animal (8.43)

is supplied by the expert, box (0.1), and that the term 'dog' is not yet contained in the lexicon of box (1). This term will now be added to the lexicon together with its relation to the already existing term 'animal'. As a result the lexicon, which was originally in state I, now makes a transition to state II, depicted by box (2). The new state is considered to be state I for the next information supply dialogue. The new dialogue will then convert it to a new state II which is again converted to state I in the context of the next dialogue etc.

In the case of many quantification sentences, the classification structure implied by the new sentence will be ambiguous. E.g., if the informant says in the next dialogue

Every cat is an animal , (8.44)

then several possibilities are left open namely 1) 'Every cat is a dog', 2) 'Every dog is a cat', 3) 'Some dogs are cats AND some cats are dogs', and finally the correct case 4) 'No dogs are cats'. The lack of knowledge as to which of these four possibilities is the correct one gives rise to non-unique, that is to set-valued conditional probabilities in the m-notation of the type $0m$ or $m1$. In such cases the knowledge base of box (1) sends one or more questions back to the informant in order to reduce the set-valued probabilities to unique ones, or equivalently in order to ascertain which of the possible classification situations applies to the new term. The lower, backward arrow from box (1) to box (0.1) depicts the questions which the knowledge base directs back at the informant. When the knowledge base system has no more questions to ask of the expert, then its procedures finally update the lexicon from state I to the new state II.

In the existing Alex system the informant *must* answer the questions directed at it by box (1) (i.e. by the program) before the knowledge base is updated. If she refuses to answer, then the current dialogue is neglected. In the new, set-valued probability logic it is not necessary to throw the new partial information that 'Every cat is an animal' away when the informant does not want to -, or is unable to -, answer such questions as 'Is every dog a cat?' posed by the knowledge base. The m-notation embodies exactly the specified information as well as the correct left-over ignorance. A later dialogue can then narrow down the set-valued probabilities to unique ones.

The term 'expert' for the human information supplier of box (0.1) is taken from the terminology of expert systems. We assume here that the expertise of the expert consists in supplying information concerning the mutual relations between terms that are to be learned by -, and stored in -, the lexicon of the knowledge base system. The meanings of the quantifiers themselves (e.g. of the words 'every', 'no') are preprogrammed into the storage and inference procedures of the knowledge base system.

The artificial experimental situation of fig. 8.12(a) can be used as a template for the typical school or book type of learning by a child. Fig. 8.12(b) refers to such a situation. In this case the informant, box (0.1), is the teacher or possibly a dictionary. Box (0.2) represents that part of the child's brain which translates the quantification sentences into their internal representation. Box (1) represents the lexicon part of the child's brain in state I, and box (2) the same part after the child has digested the information supplied by the teacher.

Fig. 8.12 refers to the case in which the original information is gleaned from a sequence of outcomes of an experiment which is then translated into quantification sentences, or directly into conditional probabilities in the m-notation. Box (0.1) of fig. 8.11(a) is then replaced by the two boxes (0.1.1), (0.1.2) of each of the three diagrams of fig. 8.12.

The case of an artificial system is depicted in diagram (a). Here we have an experimenter, represented by box (0.1.1). The experimenter should be thought of as a human who belongs to the staff of the computer system. Her task is to interview the expert of box (0.1.2). The experimenter presents the expert of box (0.1.2) with one object at a time and asks the question 'What is this object?' The expert answers

in the form of one or more labels which apply to the object. The answer is sent back to the experimenter who now presents the expert with a new object.

After having learned the names of the labels themselves, the experimenter asks her questions in the form 'Is this object an A?', 'Is this object a B?' etc., just as in example 8.6.4. These answers are processed by the experimenter and translated either into quantification sentences (in which case they are sent to box(0.2)), or directly to conditional probabilities (in which case they are sent directly to box(1)).

The artificial experimental situation of fig. 8.12(a) can serve as a template for a natural learning situation through observation of an experimental sequence, combined with the labeling of the observed object by an 'expert'. This is depicted in figures 8.12(b) and (c). Fig. 8.12(b) models the learning situation of a child (experimenter) who points at an object and asks her parent (expert) 'What is this?'. While fig. 8.12(c) models the learning situation of an anthropologist trying to learn the meaning of classification words in a native language. In this case the anthropologist is the experimenter, and the native is the teacher.

8.10 Summary of Updating of Type 1

In chapters 8, 9 we investigate the updating of probability values expressed in the m-notation. In this notation, the symbol m refers to a probability value in the open interval (0,1), while the values 1 and 0 express certainty concerning the occurrence and nonoccurrence respectively of an outcome or event. We call the updating of the present chapter 8 'updating of type 1', and that of chapter 9 'updating of type 2'. In contrast to type 1 updating, type 2 updating does not refer to a unique probability distribution.

The initial and updated probability values of the present chapter refer to the probabilities in a unique underlying distribution. There are two ways in which a person or database can obtain knowledge about these values. The basic way is to deduce, or possibly induce, the values from the observation of a random experimental sequence generated by the underlying distribution. Since probability values of 0 and 1 can never be *deduced* with certainty, we operate also with the 'set-valued' probabilities $0m \in [0,1)$ and $m1 \in (0,1]$. These express partial ignorance. Additional observations can update these values *deductively* to m, but only *inductively* to 0 or 1. Complete ignorance concerning a probability value is expressed by the set-valued probability $0m1 \in [0,1]$. The m-notation thus obviates the use of the much disputed Bayes' postulate (see sect. 8.3) to represent ingnorance concerning a probability value.

Updating of the set-valued probabilities by the observation of an additional experimental sequence from the same distribution is performed by intersecting the set values. E.g., if one experimental sequence results in the set value $0m$ for an outcome, and another sequence in the value $m1$ for the same outcome, then the probability of this outcome is updated to m.

The set values obtained from the deductive updating rules of type 1 are guaranteed to be correct. They are not maximum likelihood or other estimates of probabilities.

The price that we pay for this correctness is, of course, that a probability value m can be any value in the interval (0,1). We thus distinguish only between certainties, i.e. probability values of 1 or 0, and uncertainties (probabilityvalues m). However, this does not prevent us from replacing m by a smaller interval value, or even by a point value in the case when this is known. When we operate with the three probability values 1, 0, and m, we keep within the tradition of propositional calculus which distinguishes between inferences $A \rightarrow B$ which are tautologies, inferences which are contradictions (i.e. whose negation is a tautology), and inferences for which none of the above two cases holds.

Instead of *learning* the probability values from an experimental sequence, they may also be specified directly to the knowledge base by a reliable informant. In natural languages the *quantifiers*, 'every', 'no', 'some' etc. are used as means of specifying m-valued probabilities. This subject is discussed in sections 8.6.2, 8.6.3. The updating rules are the same whether the values are obtained from an experimental sequence or from direct specification, or from both.

Chapter 9

Type 2 Updating and the Connectives

9.1 Updating of a Single-Instance Probability or Updating of Type 2

Probabilistic updating is one of the strong connecting links between logic and probabilities. For type1 updating this concerns the quantifiers and, as we shall see in chapter **??**, the updating of and by definitional IF THEN information in general. For type2 updating it concerns the AND and OR connectives, as well as the 'general modus ponens' updating of IF THEN information. The latter includes the updating of quantification information by existence information.

In traditional logic, the conjunction *A AND B* of two atomic or composite statements A, B is treated with the aid of a truth table. In our probability logic we say that the assertion of '$info\,1 = A$' is type2 updated by the assertion of '$info\,2 = B$', or vice versa.

The inferences which can be drawn on the basis of traditional propositional calculus versus type2 updating of probabilities are, on the whole, equivalent except in connection with *IF THEN* statements and questions. Inferences in the probability logic keep strictly to the meaning of *IF THEN* in natural language as a specification of a conditional probability, or as a query concerning such a probability.

The final procedure for updating of type2 is stated by rule 9.1.1 below.

A more formal derivation of the updating rules of type2 is given in sectionXXXX *XXXX* where updating of probabilities due to a narrowing down the object set is treated as a transition from joint to conditional probabilities.

In sect. 7.7.1, definition 7.7.1 we defined $Prob(u_i)$, the probability of the outcome u_i in a single instance of an experiment, as being numerically equal to the underlying probability of u_i in a sequence of experiments. Most persons working with probabilities will probably consider this definition to be trivial, or even unnecessary. There are, however, exceptions to this rule. Von Mises, a well-known expert in the frequency interpretation of probabilities, objects to the use of statements like "Mr. X, now aged

forty, has the probability 0.011 of dying in the course of next year". Von Mises' *XXXX* opinions concerning this point are discussed in more detail in sectionXXXX.

Tracing the reasons that von Mises gives for his objection to the use of single-instance probabilities, it becomes clear that he has not clarified to himself that there exists a type of updating of probabilities which is completely different from the updating due to a prolongation of the experimental sequence. This second type of updating is due to additional information supply about, and a consequent narrowing down of, the object set to which $Prob(u_i)$ refers. As a result, we deal no longer with a single underlying probability distribution, but with new numerical values of the probabilities of the different outcomes. This is in contrast to the updating of sections 8.5-8.9 in which we only narrowed down the possible set-values of each $Prob(u_i)$, $i = 1, \ldots, I$, which could have given rise to one or more observed experimental sequences due to a single, unique underlying probability distribution. We will call the updating by prolongation of the experimental sequence 'updating of type 1', and the updating of the present section, due to a narrowing down of the reference object set, 'updating of type 2'. The most marked formal difference between the two types of updating is that a probability value m can never be deductively updated to 1 or 0 for updating of type 1. Such updating is, however, possible for updating of type 2.

We begin with the limiting case of type2 updating which occurs when the additional information consists in the observation or specification of the outcome $u_i = u_{io}$ of the single-instance experiment. For each $i \neq io$, $Prob(u_i)$ is then updated to 0. For $i = io$, $Prob(u_i)$ is updated to 1. We have already discussed such 'certainty distributions' in sect. 7.6.1. The probabilities of 1 and 0 refer now to a narrowed-down object set consisting of a single object, this being the given instance of the experiment. Or, equivalently, to those objects only for which the outcome u_{io} occurs. The object set to which we refer in this connection is discussed in more detail in sect. 9.4.

A probability value $Prob(u_i)$=$m \in (0, 1)$ is a tool for expressing uncertainty as to whether an instance of an experiment will, or will not, result in the outcome u_i. Once the outcome of the instance has been observed, the uncertainty is removed for the observer. The probability of the outcome in the particular instance of the experiment is then changed to 1 when the outcome occurred in this instance, and to 0 when it has not occurred. Again we refer to the probability value according to the information available to the observer, or to somebody to whom the observer has communicated her observation. It is thus inherent in the definition of probabilities that when the outcome of a given instance of an experiment is observed, then an initial intermediate probability value m, e.g. $Prob(u_i)$=0.7, is changed to either 1 or 0 for that instance. This holds for each outcome u_i, $i = 1, \ldots, I$.

The above considerations result in the updating rules of fig. 9.1. The m-valued probability values of column I of this figure refer to the unique, initial underlying $Prob(u_i)$ as inferred from an experimental sequence according to the methods of sect. 8.5. Alternatively these probability values can have been specified directly to the data base. According to definition 7.7.1, these values are also the initial probability values for the outcome of a single instance of an experiment.

I $Prob(u_i)$ (as stored in the knowledge base) for a sequence, and therefore also for an instance, according to definition 7.7.1	II $Prob(u_i)$ for a single instance of an experiment after the outcome of that instance has been observed	III Final $Prob(u_i)$ for the instance	IV Type 1 updated $Prob(u_i)$ for the original underlying distribution
1	1	1	1
1	0	*c=contradictory*	*c=contradictory*
0	1	*c=contradictory*	*c=contradictory*
0	0	0	0
$\underline{m}$	1	$\underline{1}$	m
$\underline{m}$	0	$\underline{0}$	m
$0m1$	1	1	$m1$
$0m1$	0	0	$0m$
$\underline{0m}$	1	$\underline{1}$	m
$0m$	0	0	$0m$
$m1$	1	1	$m1$
$\underline{m1}$	0	$\underline{0}$	m
01	1	1	1
01	0	0	0

Figure 9.1: *Limiting case of updating of type 2, from column I to column III, by observation or specification of the outcome of an instance. In contrast to the deductive updating of a unique underlying probability distribution by prolongation of an experimental sequence, a probability value m for a single instance* can *be updated to 1 or 0, see rows with underlined values. Column I lists $Prob(u_i)$ as stored in the knowledge base. According to definition 7.7.1, this is also $Prob(u_i)$ for a single instance of an experiment whose outcome has not yet been observed. Column II lists the single instance $Prob(u_i)$ after the outcome has been observed. $Prob(u_i)$=1 for that instance when the outcome was u_i, otherwise $Prob(u_i)$=0. The final, updated, single-instance $Prob(u_i)$ in column III is the same as that in column II, except in the cases when it is inconsistent with the value of the stored underlying distribution in column I. When the stored values of column I for the underlying distribution are set-valued, it may happen that they are type 1 updated by the observation of the new instance. The type 1 updated values of the underlying distribution are shown in column IV.* ***figupdatetype2***

Column II of fig. 9.1 refers to $Prob(u_i)$ for a given single instance of an experiment after the outcome of this experiment has been observed. These probabilities are therefore 1 when u_i has occurred in the single instance, and 0 when it has not occurred. Examples 9.1.1, 9.1.2 below, and especially sect. 9.4, illustrate and analyse the meaning of such probabilities and their updating.

Column III of fig. 9.1, represents the final probability of u_i in the single instance of the experiment. Since a certainty can never be updated, the entries of columns II and III are identical; except in the case when one of the first two columns has the entry 1, and the other the entry 0. In this case the two probability values are *inconsistent* or ***contradictory***. The reason is that when the underlying $Prob(u_i)$ has the value 1, then no instance of an experiment can have an outcome which is different from u_i. And when the underlying $Prob(u_i)$ has the value 0, then no instance of an experiment can result in the outcome u_i. The single-instance probabilities of columns I and III are underlined for those rows in which a probability m is updated to 1 or 0. Such updating is, as we know, forbidden for both deductive and inductive updating of type 1.

The outcome of the single-instance experiment can, in some cases, reduce the ignorance of the probability values in the first column. For example, when the value $Prob(u_i)$=0m was specified in the first column, indicating ignorance as to whether $Prob(u_i)$ has the value 0 or an intermediate value m, then $Prob(u_i)$=1 for the single instance is inconsistent with $Prob(u_i)$=0 for the initial underlying distribution. But it is not inconsistent with $Prob(u_i)$=m for that distribution. Consequently the set-valued 0m value for $Prob(u_i)$ in the underlying distribution is updated to the unique value m after the observation of the outcome of the single instance. The updated underlying value of $Prob(u_i)$ for the unique, underlying distribution, and therefore also for future outcomes of single instances of experiments, is listed in column IV.

We now consider the more general case in which the reference object set of a probability distribution is narrowed down, but not necessarily to a single object, or to objects with the single outcome or attribute value u_i.

Instead of observing, or specifying, which outcome does occur, the additional information narrows down the object set by specifying the probability value 0 for one or more outcomes u_i. Objects with these u_i values are thus eliminated from the object set. It may then happen, in contrast to the limiting case that we have just discussed, that we are left with the possibility of more than one outcome which could have occurred. We can thus have one or more outcomes u_i with $Prob(u_i)$=m in the original distribution for which $Prob(u_i)$ is type 2 updated to 0 without necessarily updating $Prob(u_i)$ to 1 for a specific u_i=u_{io}. These more general updating rules of type 2 are summarized in fig. 9.2 in which c stands for *contradictory* or *inconsistent*. The table assumes that the knowledge base contains two items of probabilistic information, *info* 1 and *info* 2, concerning the same instance of a happening or state in the world. The table entry shows the updated value of $Prob(u_i)$ after info 1 has been supplemented by *info* 2 or vice versa.

The formal difference between the updating from column I to column II in fig. 9.1

Probability of the outcome u_i for a given instance of an experiment according to *info* 1	Probability of the outcome u_i for the same instance of an experiment according to *info* 2			
	0	m	1	c
0	0	0	c	c
m	0	m	1	c
1	c	1	1	c
c	c	c	c	c

Figure 9.2: *General Updating of type 2. Mutual updating and modification of the underlying probability distribution by narrowing down the object set. The single-instance probability of the outcome u_i is type 2 updated on the basis of the $Prob(u_i)$ value according to info 1 in the left margin and according to info 2 in the upper margin. The table entry shows the updated $Prob(u_i)$ value. m stands for a probability value in the open interval (0,1), and c for a probability value whose specification according to info 1 and info 2 is contradictory. The table applies separately to each of the I possible outcomes u_i. Thus objects with specific outcomes may be eliminated (assigned the probability 0). In contrast to the limiting case of fig. 9.1 we may, however, be left with more than one possible outcome according to the information 'info 1 AND info 2', and thus with no outcome with probability 1.* ***figupdatetype2p2***

compared with the updating in fig. 9.2 is that in fig. 9.2 an initial $Prob(u_i)$=m value can be unmodified after the type2 updating.

Both fig. 9.1 and fig. 9.2 can be applied separately to $Prob(u_i)$ for each of the I possible outcomes $u_1, \ldots, u_i, \ldots, u_I$. The only constraint on the I applications is the 'summing up to 1' law for probabilities,

$$\sum_{i=1}^{I} Prob(u_i) = 1 \ . \tag{9.1}$$

The general updating of type2 of fig. 9.2 can be summarized as follows: *For an instance of an experiment, the specification of a $Prob(u_i)$=0 value by one item of supplied information overrides a $Prob(u_i)$=m value specified by another item of supplied information. If only one outcome u_i is left with a nonzero probability m, then its probability is updated to 1 according to the 'summing up to 1' constraint of the last equation.*

Fig. 9.2 shows that the 'updating of type2' operation is commutative, associative and idempotent'. 'Idempotent' means that when the probability values specified by *info* 1 and *info* 2 are identical, then the final value is equal to these two. For those who are acquainted with the terminology of group theory, we mention that the updating of type2 operation has the property of closure. Furthermore there exists a null element c, and an identity element m. All these terms are excellently explained on a few pages in sections 2.0-2.3 of Peterson [43].

When *info* 1 or *info* 2 is set-valued, then each element of the set is updated according to the rules of fig. 9.2. E.g., suppose that *info* 1 specifies the probability value $0m$. Assuming that *info* 2 specifies a probability 0 for the same instance, the final updated probability is then 0 because both 0 and m are updated by 0 to 0 according to fig. 9.2. For a probability m according to *info* 2, the final updated probability is $0m$. For a probability 1 according to *info* 2, the final updated probability is 1 because the value 0 of *info* 1 contradicts the value 1 of *info* 2 and is therefore eliminated; while m is updated to 1 by *info* 2. Finally a probability c of *info* 2, cannot be 'repaired' by either of the noncontradictory values 0 or m of *info* 1 according to fig. 9.2.

These considerations result in the following updating of type2 rule.

Rule 9.1.1 *Let each of info 1 and info 2 specify a unique or set-value (e.g. $0m$) for the probability of the outcome u_i in a given instance of an experiment. When the value is unique (e.g. m) we will also formally consider it to be a set-value, the set being a singleton (set with one element). Then each element of the first set value is type 2 updated according to the table of fig. 9.2 by each element of the second set value. The set of values so obtained is the initial set-value for the updated $Prob(u_i)$. The final type 2 updated value of $Prob(u_i)$ is obtained by eliminating possible c-values from the initial set-value. If the resulting set-value is empty, then the two items of information are contradictory.*

Examples of updating of type2 of single-instance probabilities are given in the following, and in more detail in sect. 9.3. We start with a variation of example 8.1.1

of the (generally biased) die. The example describes simultaneous updating of type 2 and type 1, making use of the table of fig. 9.1.

Example 9.1.1 *Consider successive throws of a die with an unknown degree of loading. Each of the six faces has initially a probability* $0m1$ *of turning up, i.e. we start out in a state of complete ignorance. The possibility of extreme bias of the die for which the same face will always turn up is included in this state. The probability that a given face will turn up in the first instance of throw of the die is also* $0m1$*,* $Prob_{first\ throw}(u_i)$=$0m1$ *for any* $i \in \{1,2,3,4,5,6\}$.

Suppose now that we note that the face 3 came out at the first throw. This new information updates $Prob_{first\ throw}(3)$ *to 1, and* $Prob_{first\ throw}(u_i)$ *to 0 for all the other faces,* $i \in \{1,2,4,5,6\}$. *At the same time, the underlying* $Prob(3)$ *is type 1 updated from* $0m1$ *to* $m1$ *because with* $Prob(3)$=0 *the first outcome could not have been 3. Intersecting this with the initial* $0m1$ *value we get* $Prob(3) = m1$. *More formally, the outcome 3 updates Prob(3) in the underlying distribution from* $0m1$ *to* $m1$ *according to columns I and IV of the seventh row of fig. 9.1. Likewise each of the five other probabilities of the underlying distribution is type 1 updated from* $0m1$ *to* $0m$*, according to the eigth row of fig. 9.1. These type 1 updated probability values are also the initial instance probabilities for the second throw.*

Suppose that the second instance of a throw results in the face 5, giving $Prob_{second\ throw}(5) = 1$ *and* $Prob_{second\ throw}(u_i) = 0$ *for each of the other five faces. These are then also the type 2 updated values for the a posteriori probabilitites of the six outcomes in the second throw. ('A posteriori' means* after *having obtained additional information concerning the outcome.) In the underlying distribution* $Prob(5)$ *is now type 1 updated from* $0m$ *to* m*, and* $Prob(3)$ *from* $m1$ *to* m. *For the next instance of a throw we have* $Prob_{third\ throw}(5)$=m *and* $Prob_{third\ throw}(3)$=m. *If the face 5 is observed to turn up at the third throw, then these two instance probabilities are type 2 are updated from* m *to 1 and from* m *to 0 respectively.*

The next example concerns updating by additional *specified* information. Such updating is discussed in greater detail in sect. 9.3. The *meaning* of the underlying m-values of the probabilities specified by an *OR* connective is discussed in sect. 9.4.

Example 9.1.2 *A typical updating of a probability value* m *to 1 or 0 occurs in the truth table of, e.g.,* $(A \vee B) \wedge \neg A$ = *(A OR B) AND (NOT A) in traditional propositional calculus, although it is never formulated in these terms. In sect. 9.4 we analyse the meaning of such type 2 updating in more detail.*

In our probability logic, the assertion of (A OR B) by itself assigns the probability m *to each of the three possible outcomes (1A, 0B), (0A, 1B) and (1A, 1B). These correspond to the three rows of the traditional truth table for the disjunction which have the truth value t for (A OR B); while the outcome (0A, 0B) is assigned the probability 0 by (A OR B), corresponding to the row with the truth value f in the last column of the truth table of (A OR B).*

We now update the assertion of (A OR B) by the assertion of (NOT A). Of the three nonzero-probability outcomes of (A OR B), only the outcome (0A, 1B) contains

the negation of A. The probability of this outcome is therefore updated from m to 1 by (NOT A), and the probability of the other two outcomes is updated from m to 0.
XXXX *In section XXXX we show that the chain set construction of part?? results in the same updating in a simpler and more automatic way. Note that the final Prob(u_i)=1 outcome, namely u_i=(0A, 1B), corresponds to the only row in the traditional truth table of of $(A \vee B) \wedge \neg A$ which has the truth value t in the last column.*

*The drawer example 8.1.2 of sect. 8.1 is a special case of the present example with '1A=*Drawer #1 contains knives*' and '1B=*Drawer #1 contains forks*'. We conclude that the probability value m of the outcome* 'Drawer #1 contains forks BUT NOT knives' *has been updated to 1 by* newinfo(t_n) *of eq. (8.2). For the probability of each of the other two possible outcomes according to* info(t_{n-1}) *of eq. (8.1) we have an updating from m to 0 by* newinfo(t_n).

9.2 The YN (yes-no) Notation

9.2.1 The Ambiguity of Natural Language Concerning Affirmation and Negation

The probabilistic point of view requires a somewhat different notation from that of 2-valued mathematical logic. Both of these notations differ again from the notation of natural language in which a declarative sentence A is mostly understood in the sense of an assertion.

However, as Frege has pointed out (see eq. (9.6) below), it can also be understood in the sense of an idea which can be either affirmed or negated.This ambiguity of natural language is especially transparent when a declarative sentence, for example,

$$A = \text{The drawer contains knives} , \tag{9.2}$$

is converted to its interrogative or question form

$$qu(A) = \text{Does the drawer contain knives?} , \tag{9.3}$$

which can be answered by either 'yes' or 'no'.

The negation of the sentence A,

$$B = NOT\ A = \neg A = \text{The drawer does NOT contain knives} , \tag{9.4}$$

has the question form

$$qu\,(B) = qu\,(NOT(A)\,) = \text{Does the drawer NOT contain knives?} . \tag{9.5}$$

It should be answered by 'no' in those cases in which the affirmed question (9.3) is answered by 'yes' and vice versa.

In natural language, the sentence (9.2) and its negation (9.4) can thus be understood in the sense of three different meanings,

1. As an idea which can be either affirmed or negated.

2. As the description of an outcome corresponding to the affirmed form of the sentence. This outcome may, or may not have occurred in the world.

3. As the assertion of the affirmed form; namely the assertion that the outcome of (9.2) occurs in the world to which we refer.

Usually the context in which the sentence appears in natural language determines which of the three meanings is intended.

In sect. 9.2.2 we explain a notation which distinguishes between these three meanings. Meaning 1 will be denoted by *yn-set(A)*, the 'yes-no set of A', meaning 2 by $1A$, and meaning 3 by '$P(1A) = 1$'.

The assertion or the denial of A by an informant are thus represented in the probability logic as specifications of a probability distribution over a universe consisting of two outcomes; these being $1A$=affirmation of A, and $0A$=negation of A.

Because probability distributions refer originally to a sequence of outcomes in successive repetitions of an experiment, and because we wish to make our definitions valid for any declarative sentence X, we will then in sect. 9.4 consider the utterance of a particular declarative sentence A as a particular instance or outcome of a sequence of utterances of generally different declarative sentences, each of these representing an instance 'A' or 'B' or 'C' or . . . of the variable sentence X.

The device of representing an assertion as a specification of a probability for an instance of an 'experiment' may seem artificial in connection with the atomic sentences of the present section for which the probability distributions degenerate to the certainty distributions of sect. 7.6.1. For a certainty distribution the probability of a given outcome is always equal to the probability of an instance of this outcome, also after the outcome of this instance has been obseverved.

However, for the composite sentences (sentences with connectives) of sect. 9.3, the specified probability distributions need no longer be certainty distributions. The probabilistic description is therefore a natural one for composite sentences. It is thus useful to consider atomic sentences as special cases of composite ones.

When there is no place for misunderstandings, we will sometimes follow natural language and use A (or B etc.) for any of the three meanings that we have listed. For example, the elements of the ground universe of a chain set (see left margin in the chain sets of fig. 9.6) should be more correctly denoted by *yn-set(A)* and *yn-set(B)* respectively because each such marginal element is affirmed or negated by the 1 or 0 entries of that row of the chain set table. To simplify the notation we denote it simply by A or B etc.

In other cases the affirmed form of A, which should actually be denoted by $1A$, is sometimes replaced by A; for example, when we write an *IF THEN* statement in the form *IF* A *THEN* B instead of the more correct form *IF* $1A$ *THEN* $1B$.

9.2.2 Explaining the Notation in Detail

In propositional calculus, the sentence A of e.g. eq. (9.2) is looked upon as being an idea which is supposed to convey a certain meaning. This idea may or may not fit the specified drawer; i.e. its truth value may be t or f. It is expressed by the truth table of A as shown in the second row of fig. 9.3.

In contrast, consider now that (9.2) is uttered by P_1 in a natural language dialog between two persons P_1 and P_2 in a context in which P_1 plays the role of an informant. The sentence (9.2) is then automatically understood by P_2 in the sense that the informant P_1 wishes to state that the sentence is true.

The distinction between A considered as an idea, and A considered as an assertion (judgement) is made by Frege [59, p. 11] in his 'Begriffsschrift'. He says,

> *A judgement will always be expressed by means of the sign* $\vdash\!\!\!-$ . . . . *If we omit the small vertical stroke, the judgement will be transformed into a mere combination of ideas of which the writer does not state whether he acknowledges it to be true or not. For example, let* $\vdash\!\!\!- A$ *stand for the judgement "Opposite magnetic poles attract each other"; then* $\overline{\quad} A$ *will not express this judgement; it is to produce in the reader merely the idea of mutual attraction of opposite magnetic poles, say in order to derive consequences from it and to test by means of these whether the thought is correct.* (9.6)

In our probability logic we will use the following terminology and notation.The concept or idea A, i.e., Frege's $\overline{\quad} A$, will be said to give rise to *yn-seq* (A). This is an (ordered) sequence consisting of two elements,

$$\textit{yn-seq}\,(A) = \langle 1A,\, 0A \rangle\ . \tag{9.7}$$

The first element of *yn-seq* (A) is the idea of an outcome described by the affirmation of A. It is denoted by $1A$ in the probability logic. The second element of *yn-seq* (A) is the idea of the outcome described by the negation of A. It is denoted by $0A$. The natural language symbol for the outcome $1A$ is identical with the declarative sentence A itself. E.g. in (9.7), $1A$ denotes the idea of the outcome "The drawer contains knives", and $0A$ the idea of the outcome "The drawer does not contain knives",

$$\begin{aligned} &\text{Affirmation of label } A = 1A = \text{The drawer contains knives}\ , \\ &\text{Negation of label } A = 0A = \text{The drawer does NOT contain knives}\ . \end{aligned} \tag{9.8}$$

The terminlology 'label A' instead of 'sentence A' in eq. (9.8) is intentional. In the chain set logic we allow a label such as 'λ=knives' or 'λ=knives OR forks', or the chain representation of such a label, to be a subentry in a knowledge base of an entry such as '(drawer #1 (contents of))'. Such a subentry can be negated to '$NOT\ \lambda$', e.g. 'NOT knives', or its chain set representation.

When A itself is a number, we can use the notation $1/A$ and $0/A$ instead of $1A$ and $0A$ in order to avoid misunderstandings.

Traditional Logic	Probability Logic	Probability Logic in Chain Set Representation
Assertion of A ($\vdash\!\!\!\!-\!\!\!- A$)		
A with truth value t	$Prob(1A) = 1$ $Prob(0A) = 0$	A | 1 likelihood | 1 probability | 1
Idea of A ($\!\!\!\!-\!\!\!- A$)		
Truth table for A, including the case when A is false: A | A t | t f | f	yn-sequence of A= $yn\text{-}seq\,(A) =$ $\langle 1A,\, 0A\rangle$	
Assertion of *NOT A* ($\vdash\!\!\!\!-\!\!\!- NOT\ A$)		
A with truth value f	$Prob(0A) = 1$ $Prob(1A) = 0$	A | 0 likelihood | 1 probability | 1
Idea of *NOT A* ($\!\!\!\!-\!\!\!- NOT\ A$)		
Truth table for $\neg A$, including the case when A is true: A | $\neg A$ t | f f | t	$yn\text{-}seq\,(NOT\ A) =$ $(-1)\ yn\text{-}seq\,(A) =$ $\langle 0A,\, 1A\rangle$	

Figure 9.3: *Notation in traditional versus probability logic.* $\vdash\!\!\!\!-\!\!\!- A$ *and* $\!\!\!\!-\!\!\!- A$ *are Frege's notation.* ***figupdate2notation***

The yn-sequence og of A, eq. (9.7) is identical with the 2-element set or universe

$$yn\text{-}set\,(A) = \{1A, 0A\}\ , \tag{9.9}$$

with the addition that the order of the two elements of the set is specified, the affirmed element being the first one. The notation $yn\,(\mathrm{A})$, the yn-table of A, is used when it is immaterial whether we refer to the yn-set or to the yn-sequence.

Frege's $\vdash\!\!\!\!- A$, i.e. the *assertion* of the affirmed form of A has, in the probability logic, the meaning of the specification of the probability distribution

$$Prob(1A) = 1, \qquad Prob(0A) = 0\ , \tag{9.10}$$

over the yn-sequence of A. We will say that the assertion of A *induces* this distribution.

The yn-sequence of B=$NOT\ A$, eq. (9.4), is derived from the yn-sequence of A, eq. (9.7), by reversing the order of the two elements,

$$yn\text{-}seq\,(B) = \langle 1B,\, 0B\rangle = \langle 0A,\, 1A\rangle\ , \tag{9.11}$$

where the following relation holds between the elements of the yn-sequences of the sentences A and B,

$$1B = 0A, \qquad 0B = 1A\ . \tag{9.12}$$

The assertion of $B = NOT\ A$ induces the probability distribution

$$Prob(1B) = Prob(0A) = 1, \qquad Prob(0B) = Prob(1A) = 0\ . \tag{9.13}$$

We say that the *assertion* of $B = NOT\ A$ is equivalent to the *denial* of A.

We note also that eq. (9.13) agrees with the general definition of the negation as a complementation operation in sect. 5.4. $\{0A\}$, the negated A event, is the complement of $\{1A\}$, the affirmed A event with respect to the yn-set of A, eq. (9.9).

Symbolically we can represent the negation '$\neg$', or 'NOT', as a multiplication by (-1), the multiplication being performed on the idea of A, i.e. on the yn-sequence of A, not on its affirmed or negated elements $1A$, $0A$. The (-1) 'multiplication' operator reverses the order of the elements of the yn-sequence. Or, equivalently, it leaves the order unchanged, but replaces 1 by 0 and 0 by 1. We call this operation *inversion*. Thus an even number $2n$ of negations of a sentence, leaves the idea of A unchanged. The idea of

$$C = NOT\ (NOT\ A) \tag{9.14}$$

is represented by

$$yn\text{-}seq\,(C) = (-1)\ \cdot\ (\ (-1)\cdot yn\text{-}seq\,(A)\ ) = (-1)^2\ yn\text{-}seq\,(A) = yn\text{-}seq\,(A)\ . \tag{9.15}$$

We must therefore distinguish between the *negation of a yn-sequence* which is an operator that reverses the order of the elements of the sequence; versus the *choice of the negated element* of $yn\text{-}seq\,(A)$, this being the second element, namely $0A$.

Similarly *affirmation of a yn-sequence* is an identity operator on this sequence which leaves the order of the two elements intact; while the affirmed element of *yn-seq* (A) is the first one, namely $1A$.

The reader may wonder why we do not designate the negated element of *ynseq* (A) by $(-1)\,A$, or $-A$, instead of using the symbol $0A$ for this element. There are several reasons for this. The first reason is a visual one. In a chain set it is easier to distinguish between 0 and 1 entries than between -1 and 1 entries. The second reason is that a summation of the 0 and 1 entries of a chain in a chain set results in the cardinality (number of elements) of that chain. This does not hold when we replace 0 by (-1) (see sectionXXXX for the cardinality of a chain and a chain set). *XXXX* Finally the notation $(-1)\,A$ or $-A$ instead of $0A$ would tend to obscure the distinction between the negation of a sentence A versus the selection of the negated element of *yn-seq* (A). The negation of a sentence A is an operator on its yn-sequence. The negated element of a yn-sequence is the second one.

Fig. 9.3 is an attempt to summarize and compare the notation of traditional logic with that of the probability logic. Note that according to the probability logic, it is the supplied information, i.e. the assertion of A, which we wish to represent and store in a knowledge base system. We are not interested in those cases in which the supplied information is false. This is in contrast to the truth table representation which, in general, also contains rows for which the represented statement is false (f entry in last column).

The last column of fig. 9.3 shows the probability logic in the chain set notation. The chain set notation is a standardized way for the representation in a knowledge base of the probability distribution over the yn-sequence of the ground universe.

The relation between the notation of natural language versus that of the probability logic is shown in figure 9.4. We see natural language is highly ambiguous, seen from the point of view of a mathematician. E.g., 'A=Drawer#1 contains knives' can stand both for the sentence itself, and for the outcome corresponding to the affirmation of the sentence, and for the assertion of this affirmed outcome. However, in most cases natural language resolves this ambiguity without difficulty by making use of the context in which the sentence appears.

It turns out that there are cases in which it seems natural to use a notation which imitates, in part, the context dependent notation of natural language. This is indicated in the last column of figure 9.4. Thus, we will sometimes use A instead of $1A$. E.g. '*IF A THEN NOT B*' or '$A \rightarrow \neg B$' instead of the more precise '*IF 1A THEN 0B*' or '$1A \rightarrow 0B$' respectively. On other occasions we will simply denote the yn-set of A by the letter A. In axiomatic theories of logic one often uses different fonts to resolve such ambiguities. However, we found that the use of different fonts for each of the four above-mentioned meanings of A would unnecessarily complicate the reading of this book. The slightly context dependent notation of the last column of figure 9.4 can always be retranslated to the precise notation of the previous column.

A = a phrase, (e.g. 'knives') or a whole declarative sentence (e.g. 'Drawer #1 contains knives').
$\neg A$=$NOT\ A$ = The negation of the phrase or sentence A.

	Natural Language Notation	Probability Logic	
		Precise Notation	Semi-Natural Notation
Outcome corresponding to the affirmation of A	A	$1A$	A
Assertion of the affirmed outcome	A	$P(1A) = 1$	$P(A) = 1$
Outcome corresponding to the negation of A	$NOT\ A$	$0A = 1\neg A$	$NOT\ A$
Denial of the affirmed outcome	$NOT\ A$	$P(0A) = 1$ $P(1A) = 0$	$P(NOT\ A) = 1$ $P(A) = 0$
yn-sequence of A	A	$yn\text{-}seq(A) =$ $\langle 1A, 0A \rangle$	
yn-sequence of $NOT\ A$	$NOTA$	$yn\text{-}seq(\neg A) =$ $\langle 0A, 1A \rangle$	
yn-set or yn-universe of A	A	$yn\text{-}set(A)=$ $\{1A, 0A\}$	A

Figure 9.4: *YES-NO notation in natural language versus probability logic. Comparing the context-dependent natural language notation for affirmation and negation with the context-independent, unambiguous notation of the probability logic. The last column shows a semi-natural notation of the probability logic sometimes used in this book.* ***figupdate2natprob***

Operator (in traditional logic) Prob Inducer (in prob logic)	Symbol	Equivalent 'English' Symbol
Negation	$\neg$ or (-1) or $[\sim]$	*NOT*
Conjunction	$\wedge$ or , or [&]	*AND*, *BUT*
Disjunction (inclusive)	$\vee$	*ORA*
Disjunction (exclusive)	$\vee_{ex}$	*ORE*
Implication	$\rightarrow$ or $[\supset]$	*IF THEN*
Equivalence	$\leftrightarrow$ or $\equiv$	*IFF* or *IS EQUIVALENT TO*

Figure 9.5: *Notation for the negation and connectives. The notation varies for different authors. Different symbols (separated by 'or') are equivalent. The symbols in square brackets are not used in our book. The third column shows equivalent symbols in 'English notation' which are often used in this book. 'B IFF A' stands for* (A$\rightarrow$B)$\wedge$(B$\rightarrow$A). *Alternative names for 'IFF' are 'IF AND ONLY IF' or 'IS EQUIVALENT TO'. The symbols stand for operators (truth functions) in traditional logic. In the probability logic the same symbols stand for specifiers or inducers of probability distributions.* ***figop***

9.3 The Connectives and the Negation in Traditional versus Probability Logic

This section sets up the probability distributions induced by the *AND*, *OR*, *IF THEN* and *IFF* connectives. The meaning and updating of the probabilities is discussed in sect. 9.4.

Let A and B be two statements, i.e. two declarative sentences. These can be combined by a connective *conn* to give a *compound* or *composite* statement λ,

$$\lambda = A \; conn \; B \; , \tag{9.16}$$

where *conn* is one of the connectives listed in figures 9.5, 9.6 and 9.7.

Statements without connectives are called *atomic*, in contrast to *composite* statements with connectives.

Fig. 9.5 lists the different notations for the negation and the connectives. The same symbols are used in the probability logic but with a somewhat different meaning, especially for the implication.

Note that in the probabilistic notation a comma has the meaning of *AND*. It is used to denote an outcome consisting of the affirmed or negated element of *yn(A)* and the affirmed or negated element of *yn(B)*; e.g. the outcome $(1A, 0B)$.

In traditional logic, the truth value of λ=*A conn B* is a function of the truth values of A and B respectively. The function does not depend on the meanings of the particular A and B. It depends only on the connective, and on the particular combination of truth values of A and B. The connectives are therefore called truth functional. This is also expressed by saying that the connectives are 'operators' which

operate on the truth values of A and B, and result in the truth value of the composite statement.

The negation $\neg = NOT$ is, in traditional logic, also a function which operates on its argument. It is therefore also truth functional, but it operates on one truth variable only (see fig. 9.3, first entry of bottom row). Because the negation is truth functional just like the connectives, many textbooks on logic consider the expression $NOT\ A$ to be also a composite one and call the $\neg$ operator a connective (see, e.g., [37, p. 5]). Because NOT does not connect two expressions, and because of the complete symmetry between negation and affirmation (see equations (9.10)-(9.15)), we do not follow this terminology here. When A is atomic, then we say that $NOT\ A$ is also atomic.

In summary, the truth functionality of the connectives and the negation in traditional logic expresses the fact that this logic considers the connectives and the negation to be functions from the domain of all possible combinations of truth values of A and B (four points in all for the connectives, two for the negation) to the range of truth values of the composite expression $\lambda{=}A\ conn\ B$. This range consists of the two values t and f.

xxx in fig. 9.6

3 xxx in fig. 9.7

In the probability logic we do not look upon the connectives as being functions whose value is the truth value of $\lambda = A\ conn\ B$. What we are interested in is the information supplied by the assertion of $\lambda = A\ conn\ B$. $A\ conn\ B$ is thus looked upon as *specifying* or *inducing* m-valued probabilities in the universe $A \times B$,

$$U = A \times B = \{(1A,\ 1B),\ (1A,\ 0B),\ (0A,\ 1B),\ (0A,\ 0B)\}\ . \tag{9.17}$$

We have here abbreviated the notation for the product universe $yn\text{-}set(A) \times yn\text{-}set(B)$ to $A \times B$.

For affirmation and negation we already saw in sect. 9.2 and fig. 9.3 that the assertion of A, and of $\neg A{=}NOT\ A$ induces probability distributions in the universe

$$U = A = \{1A, 0A\}\ . \tag{9.18}$$

In analogy to the yn-sequence of A for the negation, eq. (9.7), we define the yn-sequence of A, B,

$$yn\text{-}seq(A, B) = \langle (1A,\ 1B),\ (1A,\ 0B),\ (0A,\ 1B),\ (0A,\ 0B) \rangle\ . \tag{9.19}$$

Any probability distribution induced by a connective can now be represented by a sequence or list of four m-valued probability values, the first value being the probability of the first element of (9.19), the second the probability of the second element etc. .

Each connective has its own, specific probability distribution in $A \times B$. This distribution is independent of the particular meanings of A and B, just as the truth table of a connective in traditional logic is independent of these meanings.

The probabilities induced by each of the five connectives are shown in the second columns of figures 9.6, 9.7 (in the second column of fig. 9.3 for the negation).

Traditional Logic, Idea of A *conn* B	Probability Logic, Assertion of A *conn* B	Probability Logic in Chain Set Representation
	A ***AND*** B	
A B | $A \wedge B$ t t | t t f | f f t | f f f | f	$Prob(1A, 1B) = 1$ $Prob(1A, 0B) = 0$ $Prob(0A, 1B) = 0$ $Prob(0A, 0B) = 0$ $\sum = 1$	A | 1 B | 1 likelihood | 1 probability | 1
	A ***ORE*** B	
A B | $A \vee_{ex} B$ t t | f t f | t f t | t f f | f	$Prob(1A, 1B) = 0$ $Prob(1A, 0B) = m$ $Prob(0A, 1B) = m$ $Prob(0A, 0B) = 0$ $\sum = 1$	A | 1 0 B | 0 1 likelihood | 1 1 probability | m m
	A ***ORA*** B	
A B | $A \vee B$ t t | t t f | t f t | t f f | f	$Prob(1A, 1B) = m$ $Prob(1A, 0B) = m$ $Prob(0A, 1B) = m$ $Prob(0A, 0B) = 0$ $\sum = 1$	A | 1 1 0 B | 1 0 1 likelihood | 1 1 1 probability | m m m

Figure 9.6: *The AND, ORE (exclusive OR) and ORA (inclusive OR) connectives in traditional logic (column I) and in probability logic (columns II and III). Each chain set column refers to one of the outcomes of eq. (9.17). For the meaning of the likelihoods, see section XXXX. Only the probability row is needed for the storage of information supplied by one or more declarative sentences. The likelihood row is needed for the representation and answering of questions.* ***figupdate2notationandor***

Traditional Logic, Idea of *A conn B*	Probability Logic, Assertion of *A conn B*	Probability Logic in Chain Set Representation

IF A **THEN** B

A	B	$A \rightarrow B$
t	t	t
t	f	f
f	t	t
f	f	t

$Prob(1B|1A) = 1$
$Prob(1A,\ 1B) = m$
$Prob(1A,\ 0B) = 0$
$Prob(0A,\ 1B) = 0m$
$Prob(0A,\ 0B) = m$
$\sum = 1$

A	1	0	0
B	1	1	0
likelihood			
probability	m	0m	m

A **IS EQUIVALENT TO** B

A	B	$A \leftrightarrow B$
t	t	t
t	f	f
f	t	f
f	f	t

$Prob(1B|1A) = 1$ AND
$Prob(1A|1B) = 1$
$Prob(1A,\ 1B) = m$
$Prob(1A,\ 0B) = 0$
$Prob(0A,\ 1B) = 0$
$Prob(0A,\ 0B) = m$
$\sum = 1$

A	1	0
B	1	0
likelihood		
probability	m	m

Figure 9.7: *The IF THEN and IFF connectives in traditional logic (column I) and in probability logic (columns II and III). Each chain set column refers to one of the outcomes of eq. (9.17). See chapterXXXX for the derivation of the (joint) probability row of an IF THEN chain set, starting from the basic definition of 'IF 1A THEN 1B' as the specification 'Prob(1B|1A) = 1', and the assumption that none of the four marginal outcomes 1A, 0A, 1B, 0B has probability 0. The chain set of the IF THEN connective has no likelihood row for the following reasons: 1) The likelihood row of a chain set is never needed for the storage of information supplied by the declarative IF THEN statement. 2) The* question *'λ? =IF 1A THEN 1B?', to be answered on the basis of the information info, is treated in the chain set logic by multiplying the info chain set by 1A, and then asking the question '1B?' see section XXXX. 3) The $P(\lambda|ch)$ definition of a likelihood value (see XXX) is meaningless when λ is an IF THEN question.* ***figupdate2notationit***

Columns III of figures 9.6, 9.7 show the same probabilities in the chain set notation. Each chain (column) of a chain set corresponds to one outcome in $A \times B$. Outcomes with probability 0 are not listed.

As an illustration, the statement $\lambda{=}A\ ORA\ B$ induces the probability distribution $\langle m,\ m,\ m,\ 0\rangle$ over the yn-sequence of eq. (9.19),

$$Prob(1A,\ 1B) = m \quad Prob(1A,\ 0B) = m \quad Prob(0A,\ 1B) = m \quad Prob(0A,\ 0B) = 0 \tag{9.20}$$

This distribution is in the universe or space $A \times B$ of eq. (9.17). Just as for any other probability distribution, the sum of the values of the (joint) *probability row* of a chain set must be equal to 1.

Note that each each element of eq. (9.19) corresponds to a 0-1 chain (column) of the chain set representation. Each such 0-1 chain can be interpreted as a binary number. In the yn-sequence of (9.19) these numbers are set up in decreasing order. The same criterion for the order of the elements in a yn-sequence will be used also when the ground universe GU consists of more than two elements, i.e. when we have more than one connective; e.g. for $GU = \{A, B, C\}$.

The meaning of the values of the *likelihood row* of a chain set is discussed in sectionXXXX. Only the probability row is needed to represent the supplied information. The likelihood row is needed for the chain set representation of questions, i.e. of possible inferences whose probability we wish to find on the basis of the supplied information. In sectionXXXX we show that *IF THEN questions* or *inferences* are treated slightly differently from other questions in the probability logic. This is signalled by the lack of a likelihood row in the *IF THEN* chain set of fig. 9.7. *XXXX* *XXXX*

There exist cases of *IF THEN* inferences drawn on the basis of the probability logic which differ essentially from the corresponding inferences in propositional calculus. This happens just in the cases in which the inferences of traditional logic disagree completely with the inferences to be expected according to everyday logic. The expression 'everyday logic' is meant to include also the logic used in accepted mathematical derivations and textbooks.

As an illustration, consider the two propositions

$$C = (A \rightarrow B), \qquad D = (A \rightarrow \neg B)\ . \tag{9.21}$$

The inference of D from C has probability 0 in the chain set logic, I.e., the answer to the question $D?$, based on the information C is, as expected, 'no'. (See example 9.5.2.)

The analogous inference in traditional logic would require that the truth table of $(A \rightarrow B) \rightarrow (A \rightarrow \neg B)$ contains only f values in its last column. This is, however, not the case, The last column of this truth table actually contains three t, and only a single f value. The inference of D from C is thus not a contradiction in traditional logic.

9.4 Meaning and Updating of the Probabilities Induced by *AND* and *OR*

In figures 9.6, 9.7 we set up the probability values induced by '*A conn B*' over the four elements of the yn-set of eq. (9.17) for each of five different connectives '*conn*'. The meaning of the probability values 0 and 1 needs no further explanations because such a certainty value will never be changed by any updating. Futhermore, it is equal to the probability of occurrence of $u_i = (jA,\, kB), \quad j, k \in \{1, 0\}$ in an instance of an experiment without any additional information supply concerning the instance. In particular we see that no further discussion is needed concerning the meaning of the probabilities specified by an *AND* connective. One of these probabilities is one, the other being 0. By the statement '*A AND B*' an informant thus wishes to convey that both $1A$ and $1B$ are certain to occur. However, for the *OR* and *IF THEN* connectives the situation needs further clarification.

In Sect. 7.3 we defined probabilities in connection with the outcomes of a random experiment. The following thought experiment[1] defines the random experiment to which the probability values induced by the *OR* connectives refer. Any statement with an *OR* connective is considered to be an instance of an experiment whose purpose it is to find the probability distribution induced by that connective. We use the *ORA* connective (inclusive *OR* in the sense of '*A or B or both*') for our illustrations. The situation is similar for *ORE*, the exclusive *OR*.

Let L be a hypothetical person (L stand for 'Learner') who knows that *OR* is a connective. However, for some reason L does not know what probability distribution this connective induces. During a predetermined period of time, L now collects from conversations, from the literature, and from radio programs *all* statements containing an OR connective. This collecting act is carried out until L is in possession of a prespecified number N of disjunctive statements, e.g., $N = 1000$,

$$\begin{aligned} X_1 \quad &= A_1 \quad OR \quad B_1 \ , \\ &\vdots \\ X_{1000} &= A_{1000} \ OR \ B_{1000} \ . \end{aligned} \tag{9.22}$$

There need not be any connection between the meaning of the N declarative sentences $A_1, \ldots, A_N$, and similarly for the B sentences.

Let us assume that later on L has the means of ascertaining the true outcome $(jA_n, kB_n), \quad j, k \in \{1, 0\}$ for the real world situation to which X_n referred. For example, for 'X_n=The drawer contains knives OR forks', we assume that L opens the drawer and ascertains whether it contains knives and whether it contains forks.

For each of the N statements, L now asks an expert in English whether the true outcome could be expected on the basis of the statement X_n. If the expert says 'no',

[1] 'Thought experiments' (Gedankenexperimente) have been in use in theoretical physics since the famous discussions between Bohr and Einstein concerning quantum theory [6, p. 222 et seq.]. There is no reason why they should not be used in connection with the meaning of the logical connectives in natural language.

then the statement X_n is considered to be false, and is deleted from the set $\{X_n\}$ of objects to which the experiment refers.

Let $\{X_{n'}\}$ be the set of remaining statements for which the expert answered 'yes'. These are the statements which the expert considered to be a correct description of the situation pertaining to the true outcome. L finds that for the different elements of $\{X_{n'}\}$ there are cases for which the outcome $(1A,\, 1B)$ occurred, others for which $(1A,\, 0B)$ occurred, and still others for which $(0A,\, 1B)$ occurred. Furthermore, L notices that those outcomes for which the expert considered the OR sentence to be an incorrect description of the true situation were always of the type $(0A,\, 0B)$. On the basis of these results, L infers the probability distribution $\langle m\, m\, m\, 0\rangle$ of the last row of fig. 9.6 for OR. The three m-valued probabilities thus refer to the underlying distribution over the object set of all disjunctive statements.

The $\langle m\, m\, m\, 0\rangle$ distribution shows that the ORA connective expresses partial uncertainty. The informant who uses a disjunctive statement X_n wishes to convey that she is uncertain which of three outcomes is the correct one for the instance, in the world to which she refers. She is, however, certain that the outcome with probability 0 does not apply to the instance. Or that a priori none of the three outcomes $(1A_n,\, 1B_n)$ $(1A_n,\, 0B_n)$, $(0A_n,\, 1B_n)$ is certain to occur, and none of them is certain not to occur. This could not have been indicated by the assignment of the probability value $0m1$ to each of these outcomes. However, by updating of type 2, i.e. by narrowing down the object set through additional information supply concerning the given situation, an m-valued probability *can* always be updated to 0 or 1.

The additional information can be given given in the form of an additional statement *info* 2 concerning A_n and B_n. The original composite statement $X_n = A_n OR\, B_n$ will now be denoted by *info* 1.

When the knowledge base is supplied with both *info* 1 and *info* 2 (which are assumed to be true), then the resulting information is the conjunction of '*info* 1' and 'info 2',

$$info \;=\; info\,1 \;\; AND \;\; info\,2\;. \tag{9.23}$$

We say that '*info* 1 is updated by *info* 2', or that '*info* 2 is updated by *info* 1'.

Since we are referring to the outcome of a particular instance, our updating is of type 2, according to the updating rules of fig. 9.2. This means that a probability value of 0 or 1 in *info* 2 overrides a probability value m for the same outcome specified by *info* 1 and vice versa. The following example illustrates such updating.

Example 9.4.1 *A knowledge base is supplied with the following two items of information,*

$$\begin{aligned} info\,1 &= Drawer\ \#1\ contains\ knives\ ORA\ forks,\\ info\,2 &= Drawer\ \#1\ contains\ knives\ ORE\ forks. \end{aligned} \tag{9.24}$$

These give rise to a new state of the knowledge base corresponding to the single item of information,

$$kb\text{-}state\,2 = \; info\,1 \;\; AND \;\; info\,2\;. \tag{9.25}$$

Let us set

$$\begin{aligned} A &= \textit{Drawer \#1 contains knives,} \\ B &= \textit{Drawer \#1 contains forks.} \end{aligned} \tag{9.26}$$

According to fig. 9.6, info 1 and info 2 assign the same probabilities to the elements of $A \times B$, eq. (9.17), except that the outcome (1A, 1B) is assigned the probability m by info 1, and the probability 0 by info 2. According to rule 9.1.1, info 2 updates the probability of this outcome to 0, and the final probability distribution stored in the knowledge base is the same as info 2 = $A\,ORE\,B$ in the second row of fig. 9.6.

A new item of information,

$$\textit{info 3} = \textit{Drawer \#1 does not contain knives}\ , \tag{9.27}$$

will now update the probability of (1A, 0B) from m to 0 (because $1A$ cannot occur according to info 3) and we are left with the final and certain info $Prob(0A,\ 1B) = 1$, the other three elements of $A \times B$ having probability 0.

XXXX In sectionXXXX examples XXX, XXX we already treated the case of example 9.4.1 with the aid of the more automatic chain set procedures, using multiplication and prolongation of chain sets. In these examples, as well as in many other cases, the use of Bayes' postulate for the assignment of the probability values works satisfactorily. However, there exist cases in which Bayes' postulate is insufficient to
XXXX represent the available information. ExampleXXXX below illustrates such a case. The m-notation works satisfactorily in all cases. However, when a prolongation of the ground universe is needed, then there exist cases in which the correct multiplication and question answering procedures in the m-notation are more complicated
XXXX than the procedures which use Bayes postulate. (see XXXX).

The reason why the multiplication and prolongation operations for the chain sets
XXXX of section XXXX are equivalent to type 2 updating is the following.

The multiplication operation for chain sets, which represents the operation of conjunction between labels, excludes all chains which are not present in both chain sets. Thereby it automatically assigns the probability value 0 to a chain which is not present in both chain sets. This is equivalent to type 2 updating of a probability value m of a given chain in one chain set by the default probability value 0 of that chain in the other chain set in which this chain does not occur. As an illustration, let us use again example 9.4.1.

kb-state 2 of example 9.4.1, is represented by a chain set consisting of the two chains $\lceil 1\,0 \rceil$ and $\lceil 0\,1 \rceil$, each of these having a probability m. The single-chained chain set of *info* 3, eq. (9.27), must first be prolonged into the ground universe GU = {*A=knives, B=forks*}. The prolonged chain set consists of the single chain $\lceil 0\,\mathrm{b} \rceil$ which, consequently, must have the probability 1.

Multiplying these two chain sets we find that the chain $\lceil 1\,0 \rceil$ has the probability m in the first one and 0 in the second. Its probability is therefore updated to 0. The remaining chain, $\lceil 0\,1 \rceil$, has thus the probability 1.

Example 9.4.1 illustrates clearly the 'narrowing down of the object set effect' which characterizes type2 updating. *Info*2 eliminates all objects with the outcome $(1A, 1B)$, thereby leaving us only with objects having one of the outcomes $(1A, 0B)$ or $(0A, 1B)$. *Info*3 then eliminates objects with the outcome $(1A, 0B)$, and leaves us only with objects having the outcome $(0A, 1B)$. Consequently we are left with a certainty distribution for the instance to which the three items of information refer.

We note that an instance, or 'an instance of an experiment' refers to information concerning the description of a particular situation or fact in a real or imaginary world. For example, both *info*1 and *info*2 may be information concerning an instance of a throw of two particular dies at a particualr point of time. Or they may be, as in our example 9.4.1, information about whether a given drawer contains knives, and whether it contains forks, at a particular time. Or they may be information as to whether a particular instance of an animal, called 'Bobby' is a dog, or whether Bobby is a cat.

9.5 Meaning and Updating of the Probabilities Induced by *IF THEN*

The basic meaning of the statement

$$IF\ jA\ \ THEN\ kB, \qquad j,k \in \{0,1\}, \tag{9.28}$$

is

$$P(kB \mid jA) = 1\ , \quad \text{and consequently} \quad P(\ (1-k)B \mid jA) = 0\ . \tag{9.29}$$

A conjunction of two or more *IF THEN* statements gives rise to updating of type 1 of a single underlying probability distribution. This is discussed in sect. 9.6, and in more detail in sectionXXXX. *XXXX*

In sectionXXXX we show that the joint distribution of the first row of fig. 9.7 follows from eq. (9.29) under the assumption that each of the marginal probabilities in $A \times B$ is bigger than 0, *XXXX*

$$P(1A) = P(0A) = P(1B) = P(0B) = m\ . \tag{9.30}$$

These assumed marginal probabilities can, however, be type2 updated by a non-(*IF THEN*) statement. The well known modus ponens of traditional logic is a typical such case.

Thus, consider the conjunction

$$(IF\ 1A\ \ THEN\ 1B)\ AND\ 1A\ . \tag{9.31}$$

The first outer component of this composite statement induces the joint probability distribution of the first row of fig. 9.7. The second component induces the distribution Prob(1A)=1, thereby updating $Prob(1A, 1B)$ to 1, and the other joint probabilities

to 0. The end result is equivalent to '*1A AND 1B*'. The question '*1B?*' is now answered by 'yes'. This result corresponds to the modus ponens inference of traditional logic.

The next two examples are a foretaste of the treatment of the *IF THEN* connective in the probability logic. ChapterXXXX discusses this connective in much more detail.

XXXX

I have already said this: The basic meaning of *IF A THEN B* in the probability logic is $Prob(B|A) = 1$. How and under what conditions the joint probabilities of the first row of fig. 9.7 follow from the 1-value of $Prob(B|A)$ is discussed in sectionXXXX.

XXXX

Also this example has already been said

Example 9.5.1 ***The type 2 updating of IF THEN information.*** *A knowledge base is supplied with*

$$info1 = IF\ A\ THEN\ B \tag{9.32}$$
e.g. IF drawer #1 contains knives THEN drawer #1 contains forks.

This information is represented in the knowledge base by the chain set or by the four joint probabilities shown in the first row of fig. 9.7.

We now get the additional information supply

$$info2\ =\ A\ =\ Drawer\ \#1\ contains\ knives. \tag{9.33}$$

This effects a type 2 updating to 0 of both Prob(0A, 1B) and Prob(0A, 0B). The only nonzero joint probability which is left in the first row of fig. 9.7 is now Prob(1A, 1B) which is thus updated to 1. Since Prob(1A, 1B)=1, we have also that the marginal probability Prob(1B)=1. We can thus infer 'B=Drawer #1 contains forks'. Note that this is the modus ponens inference *of traditional logic. We remark again that the chain set procedures result in the same inference in a more automatic way.*

Transfer this example to first chapter on chain sets:

Example 9.5.2 ***The probabilistic treatment of the strange IF THEN inference of traditional logic.*** *This example is a case in which the probability logic gives an inference which is different from the corresponding inference in propositional calculus; and in which the inference of propositional calculus contradicts the expected one according to the meaning of the implication and the negation in natural language.*

Consider the information

$$info = IF\ A\ THEN\ B = (A\ \rightarrow\ B)\ . \tag{9.34}$$

It is stored in the data base either in the form of the probabilities of the first row and second column of fig. 9.7; or in the equivalent form of the chain set of that row.

On the basis of this information, we now wish to answer the question

$$qu = IF\ A\ THEN\ NOT\ B? = Does\ A\ imply\ NOT\ B? = (Prob(0B|1A)=1?\ ) \tag{9.35}$$

When the chain set procedures for answering questions are applied to this problem we find that Prob(0B|1A)=0. An even easier way to obtain this result is to use the basic interpretation in the probability logic of the assertion of info = (IF A THEN B) = (A → B), namely Prob(1B|1A)=1. The summing-up-to-one law for probabilities holds also for conditional probabilities having identical values of the conditioning variable. We have therefore Prob(0B|1A)=1–1=0, and conclude that,

We can never infer 'IF A THEN NOT B' from 'IF A THEN B'. (9.36)

The result of eq. (9.36) is just what we expect according to the meaning of 'IF A THEN B' in natural language. Strangely enough the inference of eq. (9.36) is not confirmed in propositional calculus.

To confirm the result of eq. (9.36) in propositional calculus we must show that the truth table of

$$(A \rightarrow B) \quad \rightarrow \quad (A \rightarrow \neg B)\,. \tag{9.37}$$

is a contradiction; in other words that the negation of (9.37) is a tautology, or that the last column of the truth table of (9.37) contains only f values. This is, however, not the case. Indeed, this column contains three t and only one f value. Consequently (9.37) is not a contradiction according to traditional logic.

9.6 Type 1 versus Type 2 Updating

Chapter 10

Compound Probabilities

10.1 Conditional Probabilities, Dependence, Cause and Effect

10.2 Cause and Effect versus Probabilistic Dependence.

Another, not infrequent, source of error in statistical inference is a confusion between statistical dependence and cause-and-effect.

For example, suppose that one finds that persons with lung cancer have a bigger probability of smoking than persons from a random sample of the total population. This does not justify the inference that lung cancer is the cause of smoking, and that smoking is the effect of lung cancer. A statistical dependence of variable 2 on variable 1 *can*, but need not indicate that 1 is the cause of 2. It may be that 2 is the cause of 1. Which of the two is correct can only be found by non-statistical methods. Namely by finding the temporal order of phenomenon 1 and phenomenon 2 in each object. In the smoking example we must find whether lung cancer appears after the start of the smoking habit or vice versa.

Even then the cause-to-effect reasoning may not be correct because both phenomenon 1 and phenomenon 2 may be caused by a third phenomenon. We may then find a statistical dependence between 1 and 2 although 1 is not the cause of 2, and 2 is not the cause of 1.

As an illustration, suppose that school children in a given country are classified according to whether they are natives of that country or immigrants (phenomenon 1).

An intelligence test set up by native psychologists, and performed in the language of the country, shows lower average scores for the immigrant children than for the native ones (phenomenon 2). This does not allow us to conclude that the difference in intelligence rating is caused by differences in racial characteristics. The lower scores are probably due to a third phenomenon, namely that the immigrant children do not know the language and the habits of the country well enough, and thus misunderstand some of the questions.

$$xxx \tag{10.1}$$

10.3 Forward and Backward Probabilities

10.4 Likelihood Reasoning

Likelihood reasoning starts out with all possible probability distributions and finds out which of these could have given rise to the particular sequence of observed data, and with what probability. It thus operates with

$$P(\text{observed sequence of data} \mid \text{assumed probability distribution } P(u)\ ) = P(\text{observation} \mid \text{assumed } P(u)\ )\ . \tag{10.2}$$

The last expression in eq. (10.2) is simply an abbreviation for the first one.

The conditional probability function of eq. (10.2), when considered as a function of the conditioning variable $P(u)$, is called the likelihood of $P(u)$.

The *maximum likelihood estimate* of the underlying probability distribution $P(u)$ is that (or one of those) assumed probability distributions $P(u)$ which makes the value of the conditional probability of eq. (10.2) a maximum,

$$P^{maximum\ likelihood}(u) = \text{that } P(u) \text{ which maximizes } [P(\text{observation} \mid P(u)\ )]\ . \tag{10.3}$$

To illustrate the maximum likelihood method, suppose again that we have a die with an unknown degree of loading. An experiment consisting of six throws of the die results in the sequence (4 4 4 4 4 4); i.e. the number 4 turns up each time. From this observed sequence we wish to find the maximum likelihood estimate of $P(u)$, $u \in \{1,\ 2,\ 3,\ 4.\ 5,\ 6,\ \}$, i.e. of the probability distribution for the outcomes 1, 2, 3, 4, 5, 6 of single throws of the die.

Now the distribution $P(4) = 1,\ \ P(1) = P(2) = P(3) = P(5) = P(6) = 0$ results in the likelihood value 1 for the observed outcome of six 4's. Since a likelihood is a probability, and since a probability value can never exceed 1, the above probability distribution $P(u)$ is the maximum likelihood estimate.

However, consider now the assumed probability distribution $P(4) = 0.99, P(3) = 0.01,\ \ P(1) = P(2) = P(5) = P(6) = 0$. For this distribution the probability of the observed sequence (4 4 4 4 4 4) is 0.99^6=0.94, assuming that the probabilities of the outcomes of successive throws are independent.

Thus the likelihood values of the two $P(u)$ distributions differ by only 6%. However, the distributions themselves differ markedly in quality. The maximum likelihood $P(u)$ distribution predicts the outcome 4 as a certainty; in contrast to the second distribution which has almost the same likelihood value. The last and inductive step of choosing that $P(u)$ whose likelihood is biggest can, therefore, give a qualitatively wrong result. If we had had a sequence of 128 observations instead of only 6, then a sequence of 128 4's, and therefore a qualitative error of this kind, would have the much smaller probability value of 0.28 in the case of the second $P(u)$. The maximum likelihood estimate $P(4) = 1,\ \ P(1) = P(2) = P(3) = P(5) = P(6) = 0$ of P(u) has therefore a much bigger probability of giving the sequence 4 4 . . . 4 (128 times) than

the estimate $P(4) = 0.99, P(3) = 0.01, \quad P(1) = P(2) = P(5) = P(6) = 0$. However the $P(u)$ estimate $P(4) = 0.999, P(3) = 0.001, \quad P(1) = P(2) = P(5) = P(6) = 0$. has a probability 0.999^{128}=0.88 of resulting in a sequence of 128 4's. This probability differs by only 12 % from the $P(4) = 1$ maximum likelihood estimate of the probability of an outcome of 128 4's. We see that no matter how long our sequence of observations is, an inductive inference of a certainty, in our case of $P(4) = 1, \quad P(1) = P(2) = P(3) = P(5) = P(6) = 0$, is never guaranteed to be true. We must always leave open the possibility of slight deviations from the certainty distribution.

This concludes our short overview of the likelihood method. More computational details can be found in the book by Duda and Hart [14, pp. 45-49, 198].

10.5 Frequencies Approach Probabilities

10.6 Bayesian or A Posteriori Reasoning

10.7 Likelihood versus Bayes

10.8 Ignorance

10.8.1 Bayes Postulate

10.8.2 Assumption of Prior Probs?

Chapter 11

Building up the Knowledge Base

11.1 The State of the Knowledge Base

In chapter 3, figs. 3.3-3.6, we demonstrated simplified versions of a possible representational form of the entries in the knowledge base or lexicon. Here we shall discuss, in a general way, what entries are being put into the lexicon. To be more specific, we do not wish to put into the lexicon descriptions of all semantically correct situations of the external world, only of those situations which do occur in this world. We will, however, mention the case of imagined *possible external worlds* in section. *xxx*

The expressions 'knowledge base', 'data base' and 'lexicon' will be assumed to be synonymous. We shall, however, distinguish between a knowlege or data base on the one hand, and a knowledge or data base *system* on the other. The latter consists of the lexicon plus all the procedures which go with it. These include procedures for the man-machine dialog, for the retrieval of information from the lexicon, for checking the consistency of new information with the old one, and for the insertion of newly supplied information which is found to be consistent with the existing one. Above all, the knowledge base system includes the important meaning-connected procedures, discussed in sect. 3.4, for establishing the cross references or pointers between different words in the lexicon.

We will assume that the procedures are all present in the database system when it is started up, and that they are not changed during the use of the system. However, the system is initially started up with an empty lexicon. The lexicon is gradually filled while the system is run in the 'information supply mode' in which the user 'Alex' supplies information to the machine 'Max'. At a subsequent dialog one can, of course, start up with a lexicon which has already been partially filled.

At each point of time one can talk about the 'state of the lexicon'. This state is simply the collection of lexicon entries as they exist at the given time. As the dialog proceeds, the lexicon may be modified, and it then enters a new state. The modification can consist of the addition of new entries, or of the addition of information to existing entries.

The case of removal of existing information from the lexicon is a complicated one because of the pointers between different lexicon entries. Removal of an entry, or of a

part of an entry, may thus necessitate removal of parts of other entries to which it is connected; which may again give rise to the necessity of removing still more information etc. . This problem is well known in all database design (see e.g. [**?**, p. 76]). The pointers are automatically set up during the insertion of information into the lexicon. The procedures for the deletion of information are therefore completely dependent on the corresponding insertion procedures. In principle the deletion of information problem can always be solved when the final form of the insertion procedures has been established. We will not treat it here.

11.1.1 Truth

The information supplied by Alex is assumed to be a true description of the external world. Said in another way, the system is assumed to accept information only from information suppliers Alex whom it considers to be reliable.

Philosophers consider the definition of truth to be an extremely difficult and many-faceted problem. (See, e.g., [20, pp. 86-134]). Here we will simply assume that the information supplier Alex is truthful in the following two ways.

1) Alex supplies the correct semantic description or meaning of words according to the usage in her language community. This correct meaning includes also possible relationships between words, such as the subset relationship between the entries of
figxxx fig. 3.4, subsequently expressed by the 'is a' and 'may be a' pointers of, e.g., figxxx. As an example, the sentence 'Every mammal is a vertebrate' expresses such a meaning-related truth.

2) The second way in which Alex is assumed to be truthful concerns the requirement that the information which she supplies to the system must be a true description of the state of the external world. We will assume for the moment that this world is the real one which we observe with the aid of our senses or of measurement apparatuses. The sentence 'The address of President George Bush is 'The White House, Washington D.C.' in February 1992, is an example of factual truth about the state
xxx of the real external world. Concerning other possible external worlds, see xxxxx.

In Kant's terminology, the above two types of truth are called *analytic* and *synthetic* truth respectively. These are supplied by analytic versus synthetic sentences [32, p. 91]. In this book we will also use the names *meaning-related* versus *factual* instead of *analytic* and *synthetic* respectively.

11.1.2 Time Dependence

The two types of truth that we mentioned in sect. 11.1.1 behave very differently as concerns their time dependence. We will discuss here shortly how possible time dependence may influence the contents and structure of the lexicon entries.

Very roughly we can say that the analytic or meaning-dependent truth of sect. 11.1.1 is, in the main, time independent while the synthetic of factual truth is often time dependent. The designer of a knowledge representation system must decide in advance which lexicon entries, or which parts of such entries are time independent, and

which are time dependent. And she must see to it that the procedures of the system take the time dependence problem into account.

As a first step, truth concerning the meaning of words will be assumed to be permanent, i.e. not to change with time. Consequently, the lexicon entries of common nouns, verbs, adjectives and adverbs should all be time independent. Examples of such words are: dog, house, idea; walk, eat, write; big, light, heavy; very, unusually. The reader may protest that the meanings of adjectives are context dependent. However, context dependence on the noun to which an adjective is attached has no connection with time dependence and vice versa.

In a given natural language, the meaning of a word may change slowly with time. If we wish to ensure our system against errors due to such changes, we can use two solutions. The first one is to require that the lexicon entry of a meaning-related word which denotes a universal concept (e.g. 'animal') must refer to a single point of time which is common to all such words. We can then set up a new lexicon in which the meanings are valid for a different point of time in the development of the language. A second possibility is to make use of the device, already mentioned in connection with eq. (1.1), for the storage and processing of more than one meaning of a word. An implementation of such a device is discussed in xxxxxsect *xxx*

Since the state of the external world changes with time, factual truth will often be time dependent. Information about the external world can concern the existence or non-existence of instances of a given class, the attribute *values* of instances, and the description of situations. All of these may be time-dependent. E.g., instances of living dinosaurs existed around 100 million years ago, but do not exist now. Attribute values of John which may change with time are his address, height (if he is a boy), profession etc. . Finally we have time dependent situations such as 'John is playing tennis' or 'John is taking a course in mathematics this semester'. The English language has the useful 'ing' ending of verbs to indicate such time limited validity of a description.

To solve the problem of time dependence of the descriptions of the external world, we have roughly the same two possibilities that we mentioned in connection with a possible time dependence of the meaning of a word. If we use different lexica for each point of time, and if we assume that the *meaning* of common nouns, verbs, adjectives and adverbs do not change within the total time interval covered by all the different lexica, then the lexicon entries for the above four syntactic categories will be the same in all the lexica. (An exception to this statement can occur in connection with information concerning the *existence* of instances of a common noun concept. The entries which may vary from one lexicon to the next are those of proper nouns and other descriptions of instances.

Should we, on the other hand, decide upon the solution of including in one lexion descriptions of different states of the external world, we will probably not insert into the lexicon the equivalent of separate entries for each state of, e.g. John, but will use a device to record the time dependent part of an entry in some tabular or other form which is equivalent to the representation of a function of time.

In this book we start out with the case in which all information in the lexicon is permanently true; or with the case in which the lexicon refers solely to one state, i.e.

to one point of time of the external world. A possible device for recording possible
xxx time dependence is described in sectionxxx.

Chapter 12

The Implication of Traditional Logic

12.1 Introduction

Part ?? of this book will introduce the chain set logic. The present part ?? provides a more specific background than part I for the use of chain sets.

We shall see that the construction of classification and quantification structures can always be formulated in terms of IF THEN statements. The treatment of such statements is therefore an essential part of our knowledge representation system.

In traditional logic IF THEN statements are treated basically with the aid of the truth table of the so-called *material implication*. The material implication or modifications of it are still an essential part of practically all logics in spite of all the criticisms to which this implication has been subjected.

Practically every elementary textbook on mathematical logic starts out with the truth tables of propositional calculus. This calculus is then followed by a completely new subject, namely predicate calculus, which is used for the treatment of quantification and classification problems.

The chain set structures of this book are, in a way, modified truth tables. However, chain sets can by themselves, or in combination with the structures of the Alex system, be used for classification and quantification purposes. The usual predicate calculus is therefore not part of our system. In contrast, it is important to recognize the similarities and the differences between chain sets and the traditional truth tables of propositional calculus, especially since the differences between these two systems are most marked for the implication. These differences concern both the truth table for *information* supplied in the form of IF THEN *statements*, and the procedure for drawing *inferences* or answering *questions* having an IF THEN form.

The present chapter chapter 12 is therefore devoted to a summary of traditional propositional calculus. Sect. 12.2 presents the well-known truth tables, including that of the material implication. The traditional method of drawing inferences through the use of the tautological implication is discussed in sect. 12.3, and criticisms of the material implication in sect. 12.4.

Classification and quantification information gives rise to pure or modified tree structures. Chapter ?? discusses these trees, and why they are inadequate for the representation of classification information in the general case.

Our representation of knowledge system will usually combine chain set structures with structures of the Alex knowledge representation system. The final chapter of part ??, chapter ??, gives a short introduction to the Alex system.

12.2 Truth Tables

Mathematical Logic deals with *sentences* or *statements* or *formulas* or *propositions* [37, chapter 1]. All these four concepts have approximately, but not quite, the same meaning. The difference in nuance is discussed in detail in Haack's book [20, chapter 6].

The main difference is between propositions versus the other three. Propositions are expressed by statements, i.e. by declarative sentences. According to Haack, a proposition is what is common to a set of synonymous declarative sentences. Two sentences express the same proposition if they have the same meaning. Thus the two sentences

$$\text{John is Jane's husband.}, \tag{12.1}$$

$$\text{Jane is John's wife.} \tag{12.2}$$

express the same proposition according to this definition. When there is no possibility of confusion, we will here often use the words 'sentence', 'statement', 'proposition' synonymously as a means for stylistic variations.

Traditional 2-valued logic deals solely with propositions that are either true or false. This can be expressed by saying that the sentences can assume *truth values* in the range $\{T, F\}$. Sentences that express the same proposition, such as (12.1), (12.2) have always the same truth value.

Propositional calculus does not distinguish between a truth value F on the one hand, and a logical inconsistency on the other, i.e. between synthetic and analytic lack of truth (see chapter 4). Thus the truth value of the sentences

$$\text{Napoleon Bonaparte was an Englishman.}, \tag{12.3}$$

$$\text{Napoleon Bonaparte was a Corsican AND NOT a Corsican.} \tag{12.4}$$

is the same in traditional logic, namely F.

In contrast, in the chain set logic every declarative sentence supplied to the knowledge base is assumed to be true. The falsity of sentence (12.3) would thus be expressed by supplying the information

$$\text{It is NOT TRUE that Napoleon Bonaparte was an Englishman.}, \tag{12.5}$$

or,

$$\text{Napoleon Bonaparte was NOT an Englishman.} \tag{12.6}$$

to the knowledge base.

If the sentence (12.4) were supplied to the knowlege base, then the chain set construction for 'Corsican AND NOT Corsican' would result in 'a chain set without chains' (see xxx), signifying that the supplied information is logically inconsistent. *xxx*

In propositional calculus, the truth values of composite sentences with negations and connectives (see figs xxx) are defined as functions of-, or operators on-, the truth *xxx* values of the component sentences. Consequently the operators for the negation and the connectives are said to be *truth functional*.

The main operators are denoted by $\neg$, $\wedge$, $\vee$, $\rightarrow$, corresponding to *not*, *and*, *or* (inclusive), and *if then* (or *implies*) in natural language. To these we also add the operators $\vee_{ex}$ for *either . . . or* (exclusive *or*) and $\leftrightarrow$ for *is equivalent to* or *if and only if*. Fig. 9.5 lists the operators. Instead of the symbols used in mathematical logic, we will often use those in the right hand column of fig. 9.5 because of their correspondence with natural language English.

Mathematical logic includes *not* among the *propositional connectives* [37, p. 5]. Since the negation does not connect two things, we prefer Quine's *logical particles* (see eq. (4.6) here) as a collective name for the words *not*, *and*, . . . in the last column of fig. 9.5. All of these are *connectives* except *not*. *NOT* is a unary function or operator, it acts on one argument only. All the other operators are binary. E.g., the truth value of '*A AND B*' is a function of the truth value of A and of the truth value of B.

A proposition with no logical particles is called an *atomic* propostion in traditional logic [37, p. 5]. All other propositions are *compound* or *composite*. Thus a sentence with a negation but no connectives is a composite one according to this terminology. Because of the symmetry between affirmation and negation, the chain set logic considers the negation of an atomic statement to be atomic also.

The truth tables of propositional calculus for the six operators are shown in fig. 12.1. In order to facilitate the comparison of the truth tables with the chain sets of part **??**, we have here used a somewhat different notation from the one that is usual in mathematical logic. In the first place we have replaced the symbols T and F for the two truth values by 1 and 0 respectively. In the second place we have transposed the truth tables. Rows have become columns and vice versa. In the third place we have used a horizontal ruled line to separate the entries for the truth values of the components from the bottom row, containing the truth values of the composite statement for different combinations of the truth values of the components. Thus, e.g., the second column of fig. 12.1 (c) says that when A is true and B is false, then $A\ ORA\ B$ is true.

12.3 Tautological Implication for Inferences

Inferences are drawn in propositional calculus by making use of the concept of tautology. A tautology is a composite proposition that is true for all possible combinations of truth values of its components. In other words, a proposition is a tautology iff

(a) λ=*NOT A*

A	1	0
λ	0	1

(b) λ=*A AND B*

A	1	1	0	0
B	1	0	1	0
λ	1	0	0	0

(c) λ=*A ORA B*

A	1	1	0	0
B	1	0	1	0
λ	1	1	1	0

(d) λ=*A ORE B*

A	1	1	0	0
B	1	0	1	0
λ	0	1	1	0

(e) λ=*IF A THEN B*

A	1	1	0	0
B	1	0	1	0
λ	1	0	1	1

(f) λ=*B IFF A*

A	1	1	0	0
B	1	0	1	0
λ	1	0	0	1

Figure 12.1: *The traditional truth tables of 2-valued logic for the six logical particles of fig. 9.5. For subsequent comparison with chain sets, the symbols T and F have been replaced by 1 and 0 respectively, and rows and columns have been transposed. The row below the horizontal ruled line shows the truth value of λ for the given combination of truth values of the components in that column.* ***figtruthtables***

the truth table of the proposition has only 1 entries in its bottom line. E.g., the proposition λ=*A ORA NOT A*, fig. 12.2, is a tautology; so is λ=*A ORE NOT A*.

Suppose now that we have two noncomposite or composite propositions called *info* and *qu*. These two symbols are mnemonics for 'information' and 'question' respectively. In propositional calculus they are often called 'premise' and 'conclusion'. The premise can be thought of as the conjunction of all the statements in a knowledge base.

According to propositional calculus, the conclusion *qu* can be inferred or deduced from the premise *info* if and only if the implication *info*$\rightarrow$*qu* is a tautology. One also says that *qu* follows from *info*, or that the theorem *qu* can be proved from the assumptions *info* (see, e.g., Kleene [37, p. 33]).

As an example, let us prove the transitive law in propositional calculus. This law says that the conclusion

$$qu = (A \rightarrow C) \tag{12.7}$$

follows from the premise

$$info = (A \rightarrow B) \wedge (B \rightarrow C)\ . \tag{12.8}$$

Fig. 12.3 shows the truth table of λ=(*info*$\rightarrow$*qu*) for all combinations of truth values of A, B, and C. The table has only 1-entries in the bottom row. Consequently *qu* follows from *info* according to propositional calculus. We will see that the transitive law holds also in the chain set logic, and that its derivation is accomplished in fewer steps there.

12.4 Criticisms of the Material Implication

λ=*A ORA NOT A*

A	1	1	0	0
$\neg A$	0	0	1	1
λ	1	1	1	1

Figure 12.2: *Truth table of propositional calculus for a tautological label λ. The truth values of the bottom row are all 1.* ***figaoranota***

λ=*IF* $((A \rightarrow B) \wedge (B \rightarrow C))$ *THEN* $(A \rightarrow C)$

A	1	1	1	1	0	0	0	0
B	1	1	0	0	1	1	0	0
C	1	0	1	0	1	0	1	0
$A \rightarrow B$	1	1	0	0	1	1	1	1
$B \rightarrow C$	1	0	1	1	1	0	1	1
$(A \rightarrow B) \wedge (B \rightarrow C)$	1	0	0	0	1	0	1	1
$A \rightarrow C$	1	0	1	0	1	1	1	1
λ	1	1	1	1	1	1	1	1

Figure 12.3: *Truth table for the derivation of the transitive law in propositional calculus. The implication 'premise$\rightarrow$conclusion' must be a tautology, where 'premise=$(A \rightarrow B) \wedge (B \rightarrow C)$' and 'conclusion=$A \rightarrow C$'.* ***figtrans1***

Bibliography

[1] Aristotle. *Prior and Posterior Analytics.* Everyman's Library, Dutton:New York, 1964; original ≈344 B.C. Edited and translated by John Warrington.

[2] Aristotle. *Aristotle's Categories and Propositions.* The Peripatetic Press, Grinnell, Iowa, 1980; original ≈344 B.C. Translated with Commentaries by Hippocrates G. Apostle.

[3] Bandler, W and Kohout, L. Fuzzy power sets and fuzzy implication operators. *Fuzzy Sets and Systems*, 4:13–30, 1980.

[4] Bandler, W. and Kohout, L.J. The interrelations of the principal fuzzy logical operators. In Gupta, M.M., Kandel, A., Bandler, W., and Kiszka, J.B., editors, *Approximate Reasoning in Expert Systems*, pages 767–780. Elsevier Science Publishers B.V.(North Holland), 1985.

[5] Bayes, Thomas. An essay towards solving a problem in the doctrine of chances. *Biometrika, Cambridge*, 45:293–315, 1958. Original published posthumously in *The Philosophical Transactions of the Council of the Royal Society of London* (1763), **53**, 370-418.

[6] Bohr, Niels. Discussions with einstein on epistemological problems in atomic physics. In Schilpp, Arthur, Paul, editor, *ALBERT EINSTEIN: Philosopher-Scientist*, pages 199–241. Tudor Publishing Company, 1951.

[7] Boole, George. *An Investigation of the Laws of Thought on which are Founded the Mathematical Theories of Logic and Probabilities.* Dover, New York; Original Edition Macmillan, 1854.

[8] Borkowsky, L. *Jan Lukasiewicz Selected Works.* North Holland, 1970.

[9] Charniak, Eugene and McDermott, Drew. *Introduction to Artificial Intelligence.* Addison-Wesley, 1985.

[10] Chomsky, Noam. *Syntactic Structures.* Mouton Publishers, 1957.

[11] Davidson, Donald. Truth and meaning. *Synthese*, 17:304–323, 1967.

[12] DeFinetti, B. *Theory of Probability: A Critical Introductory Treatment*, volume 1. Wiley, New York, 1974.

[13] Duda, R.O., Gaschnig, J.G., and Hart, P.E. Model design in the prospector consultant system for mineral exploration. In Michie, D., editor, *Expert systems in the micro-electronic age*, pages 153–167. Edinburgh University Press, 1979.

[14] Duda, R.O. and Hart, P.E. *Pattern Classification and Scene Analysis*. Wiley, 1973.

[15] Einstein, Albert. Autobiographical notes. In Schilpp, Arthur, Paul, editor, *ALBERT EINSTEIN: Philosopher-Scientist*, pages 1–95. Tudor Publishing Company, 1951.

[16] Feller, William. *An Introduction to Probability Theory and its Applications*, volume I. John Wiley, third edition, 1967.

[17] Findler, V., Nicholas. *Associative Networks, Representation and Use of Knowledge by Computers*. Academic Press, 1979.

[18] Fisher Box, Joan. *R. A. Fisher, The Life of a Scientist*. John Wiley & Sons, New york, 1978.

[19] Giles, R. Introduction to a logic of assertions. Mathematical preprint 1989-2, Department of Mathematics and Statistics, Queen's University, Kingston, Ontario, Canada K7L3N6., 1989.

[20] Haack, S. *Philosophy of Logics*. Cambridge University Press, 1978.

[21] Hartmann, R.R.K. and Stork, F.C. *Dictionary of Language and Linguistics*. Applied Science Publishers Ltd., London, 1972.

[22] Hatcher, William S. *The Logical Foundations of Mathematics*. PWS-Kent Publishing Company, Boston, 1989.

[23] Hisdal, E. The IF THEN ELSE statement and interval-valued fuzzy sets of higher type. *Int. J. Man-Machine Studies*, 15:385–455, 1981.

[24] Hisdal, E. Reconciliation of the yes-no versus grade of membership dualism in human thinking. In Gupta, M.M., Kandel, A., Bandler, W., and Kiszka, J.B., editors, *Approximate Reasoning in Expert Systems*, pages 33–46. North Holland, 1985.

[25] Hisdal, E. Infinite-valued logic based on two-valued logic and probability, part 1.1. Difficulties with present-day fuzzy set theory and their resolution in the TEE model. *Int. J. Man-Machine Studies*, 25:89–111, 1986.

[26] Hisdal, E. Infinite-valued logic based on two-valued logic and probability, part 1.2. Different sources of fuzziness and uncertainty. *Int. J. Man-Machine Studies*, 25:113–138, 1986.

[27] Hisdal, E. Are grades of membership probabilities? *Fuzzy Sets and Systems*, 25:325–348, 1988.

[28] Hisdal, E. Infinite-valued logic based on two-valued logic and probability, part 1.3. Reference experiments and label sets. Research Report 147, Institute of Informatics, University of Oslo, Box 1080 Blindern, 0316 Oslo 3, Norway, 1988,1990. ISBN 82-7368-053-3. Can also be found on http://www.ifi.uio.no/~ftp/publications/research-reports/Hisdal-3.ps.

[29] Hisdal, E. Infinite-valued logic based on two-valued logic and probability, part 1.4. The TEE model. Research Report 148, Institute of Informatics, University of Oslo, Box 1080 Blindern, 0316 Oslo 3, Norway, 1988,1990. ISBN 82-7368-054-1. Can also be found on http://www.ifi.uio.no/~ftp/publications/research-reports/Hisdal-4.ps.

[30] Hisdal, E. Explanatory versus postulate fuzzy set theory. In Eklund, Patrik, editor, *MEPP'92: Proceedings of the International Seminar on Fuzzy Control*, pages 42–52, Åbo Akademi University, Department of Computer Science, DataCity, Lemminkäinengatan 14-18, SF-20520, Åbo, Finland, 1992.

[31] Hisdal, E. Interpretative versus prescriptive fuzzy set theory. *IEEE Transactions on Fuzzy Systems*, 2:22–26, 1994.

[32] Hurford, James R. and Heasley, Brendan. *Semantics: a coursebook.* Cambridge Univesity Press, 1983.

[33] Jaynes, E.T. Confidence intervals vs bayesian intervals. In Harper, W.L. and Hooker, C.A., editors, *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, pages 175–213. D. Reidel Publishing Company, 1976.

[34] Jeffreys, Harold. *Theory of Probability.* Clarendon Press, Oxford, third edition, 1961.

[35] Kempson, Ruth. M. *Semantic Theory.* Cambridge University Press, 1977.

[36] Kingman, J.F.C. and Taylor, S.J. *Introduction to Measure and Probability.* Cambridge University Press, 1966.

[37] Kleene, S.C. *Mathematical Logic.* John Wiley, New York, London, 1968.

[38] Kolmogoroff, A. *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Julius Springer, Berlin, 1933.

[39] Laviolette, M. and Seaman, W., John. The efficacy of fuzzy representations of uncertainty. *To appear in International Journal of Aproximate Reasoning*, ≈1993.

[40] Lyons, John. *Language and Linguistics.* Cambridge University Press, 1981.

[41] Nilsson, Nils J. *Principles of Artificial Intelligence.* Springer Verlag, 1982.

[42] Ogden, C. K. and Richards, I.A. *The Meaning of Meaning.* Harcourt, Brace and Company, 1946.

[43] Peterson, W. Wesley. *Error Correcting Codes.* The M.I.T. Press, 1961. This is not the big book of the same name by Peterson and Weldon.

[44] Plato. *Phaedo.* Clarendon press, Oxford, 1975. Editor and Translator: Gallop, David.

[45] Quillian, M. Ross. Semantic memory. In Minsky, Marvin, editor, *Semantic Information Processing*, pages 227–270. MIT Press, 1968.

[46] Quine, Willard Van Orman. Two dogmas of empiricism. *Philosophical Review*, 60:20–43, 1951.

[47] Ramsay, A. *Formal Methods in Artificial Intelligence.* Cambridge University Press, 1988.

[48] Renyi, Alfred. *Foundations of Probability.* Holden-Day, Inc., 1970.

[49] Sagan, Carl. *Cosmos.* Macdonald Futura Publishers, London, 1980.

[50] Savage, L.J. *The Foundations of Statistics.* Dover, Mineola, NY, second revised edition, 1972.

[51] Schank, Roger C. *Conceptual Information Processing.* North Holland, 1975.

[52] Schank, Roger C. and Carbonell, Jaime G., Jr. Re: The gettysburg address. In Findler, Nicholas V., editor, *Associative Networks*, pages 327–362. Academic Press, 1979.

[53] Shafer, Glenn. *A Mathematical Theory of Evidence.* Princeton University Press, 1976.

[54] Shortliffe, E.H. *Computer-Based Medical Consultations: MYCIN.* Elsevier, New York, 1976.

[55] Sowa, John. F. *Conceptual Structures.* Addison-Wesley, 1984.

[56] Sowa, John. F. Editor. *Principles of Semantic Networks: Explorations in the Representation of Knowledge.* Morgan Kaufman, 1991.

[57] Stigen, Anfinn. *Tenkningens Historie*. Gyldendal Norsk Forlag, 1983.

[58] Sverdrup, Erling. *Lov og Tilfeldighet (Norwegian for 'Law and Chance')*. Universitetsforlaget, Munksgaard, Copenhagen, 1964.

[59] van Heijenoort, Jean, editor. *From Frege to Gödel*. Harvard University Press, 1967.

[60] Vogt, Hans. *Forelesninger i Almen Språk Vitenskap*. Lingvistisk institutt, University of Oslo, 2-nd edition, 1978.

[61] von Mises, Richard. *Probability, Statistics and Truth*. Macmillan, New York, 1939.

[62] von Mises, Richard. *Mathematical Theory of Probability and Statistics*. Academic Press, 1964. Edited and Complemented by Hilda Geiringer.

[63] Weber, S. A general concept of fuzzy connectives, negations and implications based on t-norms and t-conorms. *Fuzzy Sets and Systems*, 11:115–134, 1983.

[64] Winograd, Terry. Computer software for working with language. *Scientific American*, 251:91–101, September 1984.

[65] Wonnacott, Ronald J. and Wonnacott, Thomas H. *STATISTICS Discovering Its Power*. John Wiley & Sons, 1982.

[66] Zadeh, L.A. Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-3:28–44, 1973.

[67] Zadeh, L.A. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978.